Dipl.-Ing. Philipp Koncar, BSc.

Quantitative Text Analysis for Digital Humanities

Empirical Case Studies of Historic Sentiment, Multilingual Controversy and Employee Satisfaction

Doctoral Thesis

to achieve the university degree of

Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisor

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Institute of Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Graz, May 2021

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date

Signature

Abstract

Quantitative and computational text analysis methods hold great potential to benefit a plethora of diverse research areas, including but not limited to philosophy, history, economics, political studies, communication studies, museum studies, translation studies or ethnic studies. The application of such methods in the respective fields of humanities and social sciences even spawned an overarching discipline known as digital humanities and, as of today, we already see promising works in this area, such as big data analyses of historic plays related to literary studies or the automatic prediction of fake news linked to media studies. However, we still recognize plenty of opportunities and open questions for which the application of quantitative text analysis methods may find answers to. A steep learning curve for researchers, who often have little background in statistics and computer science, is perhaps preventing a broader application of such methods in the respective fields. In this thesis, I set out to explore how one can apply quantitative and computational methods to gain new insights for the digital humanities. The presented methods shall serve as guidelines and demonstrations of how to conduct quantitative text analysis in different contexts. In particular, I outline the application of such methods in three distinct case studies. First, I conduct an in-depth sentiment analysis on a collection of historic periodicals. The analysis highlights how to use existing approaches, that have originally been tailored for today's media, in a historic context and adds new input to literary and media studies. Second, I analyze the controversy of user comments posted in online social media and specifically bring attention to linguistic differences thereof. In doing so, I reveal that controversy is universal across languages and set an example for the application of quantitative methods in communication studies. Third, I study employee satisfaction expressed in online employer reviews to showcase how to use quantitative text analysis methods in management sciences. I consider such reviews through the lens of Herzberg's Two-Factor Theory,

a well-established and well-studied theory that explains the factors that influence employee satisfaction. As such, I demonstrate how one can link quantitative and computational methods to theories established in traditional research. Since I consider multilingual data in all case studies, the presented analyses are further related to linguistics as well as cultural studies. For the second and third case study, I exploit the quantified features in prediction experiments that further exemplify real-world applications in the respective contexts. Overall, my thesis provides guidelines for how to conduct quantitative analyses in diverse disciplines that are part of humanities and social sciences and, thus, adds fruitful input to the increasingly popular digital humanities. Based on the presented approaches, other researchers can conduct and adapt similar analyses to drive their research forward.

Kurzfassung

Die quantitative und computerbasierte Analyse von Texten birgt viele Vorteile für die Forschungsarbeit in zahlreichen Fachgebieten, wie zum Beispiel in der Philosophie, Geschichte, Museologie, Wirtschafts-, Politik-, oder Kommunikationswissenschaft. Die Anwendung computergestützter Methoden in den jeweiligen Bereichen der Geistes- und Sozialwissenschaften hat sogar zur Entstehung einer neuen übergreifenden Disziplin geführt - den digitalen Geisteswissenschaften. Aktuell gibt es schon vielversprechende Arbeiten auf diesem Gebiet, wie zum Beispiel die Analyse von historischen Theaterstücken in der Literaturwissenschaft oder die automatische Erkennung von Falschinformationen in der Medienwissenschaft. Dennoch existieren reichlich Potenzial und zahlreiche offene Fragen, für die quantitative und computerbasierte Methoden Antworten liefen können. Die mit diesen Methoden einhergehende Lernkurve für Forscher, die eventuell nur wenig Kenntnisse über Statistik und Informatik haben, könnte eine breitere Anwendung verhindern. Das Ziel dieser Arbeit ist es, diesem Problem entgegenzuwirken und zu untersuchen, wie quantitative und computerbasierte Methoden in unterschiedlichen Bereichen der digitalen Geisteswissenschaften verwendet werden können. Die gezeigten Ansätze sollen einerseits als Leitfaden dienen und andererseits demonstrieren, wie ähnliche Analysen und Methoden in unterschiedlichen Themengebieten und für verschiedene Sprachen angewendet werden können. Um dieses Ziel zu erreichen, werden drei Fallstudien vorgestellt. Die erste dieser Fallstudien beschäftigt sich mit einer detaillierten Sentiment-Analyse für mehrsprachige und historische Zeitschriften und liefert so neue Einblicke für Literatur- und Medienwissenschaften. Die zweiten Fallstudie analysiert und beschreibt Kontroversen in sozialen Medien und fokussiert sich dabei speziell auf sprachliche Unterschiede diesbezüglich. Die Ergebnisse dieser Studie zeigen, dass sich Kontroversen in sozialen Medien ähnlich und unabhängig von der Sprache verhalten und dienen als ein Beispiel für quantitative Analysen im Bereich

der Kommunikationswissenschaft. Abschließend beschäftigt sich diese Arbeit mit der Zufriedenheit und Unzufriedenheit von Arbeitnehmern, die sie in online Bewertungen und Rezensionen zum Ausdruck bringen. Diese Rezensionen werden unter anderem in Kombination mit der Zwei-Faktoren-Theorie von Frederick Herzberg, eine lang existierende und sehr bekannte Theorie um die Einflussfaktoren für Mitarbeiterzufriedenheit zu bestimmen, untersucht. Dadurch wird demonstriert, wie quantitative Methoden mit traditionellen Theorien aus der Wirtschaftswissenschaft in Verbindung gebracht werden können. Da in allen drei Fallstudien mehrsprachige Daten analysiert werden, bringen die gezeigten Ansätze und Ergebnisse auch neue Erkenntnisse für Sprach- und Kulturwissenschaften. Weiters werden die in der zweiten und dritten Fallstudie berechneten Charakteristiken für Vorhersagemodelle verwendet um mögliche praktische Anwendungen aufzuzeigen. Zusammengefasst bietet diese Dissertation eine Orientierungshilfe für die Anwendung von quantitativen und computergestützten Methoden in den unterschiedlichen Bereichen der digitalen Geisteswissenschaften. Die vorgestellten Verfahren sind klar und transparent und können so von Wissenschaftlern übernommen und angepasst werden um ihre eigene Forschung voranzubringen.

Acknowledgments

First and foremost, I want to express my thanks to my supervisor Denis Helic for all his good advice which made this work possible. Thank you, Denis, for all the encouraging words, the animating discussions and the ever respectful and trouble-free collaboration. Further, I want to thank the Graz University of Technology for the quality education and the Austrian Academy of Sciences for funding my PhD.

I also owe gratitude to the excellent researchers working at the Institute Centre for Information Modelling and the Know-Center GmbH both located in Graz, Austria. Working with them had nothing but a positive impact on the course of my doctoral studies. In particular, I want to thank my former project team members Alexandra Fuchs, Bernhard C. Geiger, Christina Glatz, Elisabeth Hobisch, Sanja Saric, Martina Scholger and Yvonne Völkl.

Fortunately, I was not alone during this endeavor and had fellow PhD candidates in similar situations and with whom I could share the various experiences. Specifically, I want to thank Lukas Eberhard, Aleksandar Karakas, Patrick Kasper, Niklas Hopfgartner, Lisa Posch, Thorsten Ruprechter, Tiago Santos as well as Matthias Wölbitsch for the fun times and interesting talks we had together. Special thanks to my former colleague Simon Walk, with whom I not only had the pleasure to work on exciting research papers but who also significantly influenced my decision to pursue a PhD.

Finally, I want to thank my family and friends who unconditionally supported me on this road. Mom and Dad, without your parental care and education, I would not have become the person I am today. Lea, no words can ever express my gratitude to you, so the least I can do is to dedicate this work to you. None of this would have been possible without you ♡.

Contents

1 Introduction

1.1 Motivation

Ever since knowledge was transferred in a written form, scholars tried to extract insightful information of texts to learn more about a variety of matters. Unsurprisingly, religion was the first decisive reason to conduct text analysis, mainly to better understand and visualize the bible [e.g., Herbermann, 1913] or to silence non-religious agendas emerging in traditional media in the Modern Age [e.g., Groth, 1948]. In the 20th century, social scientists began to utilize text analysis methods to improve the life of individuals and to better societies as a whole. For example, we saw studies focusing on how African Americans were represented in the press [Simpson, 1936] or how children's books reflect nationalism in different countries [Martin, 1936]. Only more recently, researchers started to use computational methods on traditional as well as Web data to gain even more insights into various social phenomena, such as political matters [e.g., Adamic and Glance, 2005; Gruzd and Roy, 2014] or health care issues [e.g., Daraz et al., 2018; Ayanouz, Abdelhakim, and Benhmed, 2020]. The usefulness of quantitative text analysis to learn more about human behavior and society is undisputed in our research community and even led to a distinct research discipline called *digital humanities*, which combines methods from computer science with traditional humanities and social sciences.

The advantages of quantitative methods are manifold and benefit a plethora of disciplines, such as education [e.g., Wegerif and Mercer, 1997], economics [e.g., Reinstein and Snyder, 2005], communication studies [e.g., Ziegele, Breiner, and Quiring, 2014] or anthropology [e.g., Acerbi et al., 2013]. For example, these methods allow for big data analyses which provide more generalizable results as well as bigger pictures of certain phenomena. The same quantitative methods can lead to faster problem solving as analyses that would take years to conduct by humans can be executed in a fraction of the time depending on a computers capacities. This further leads to better cost efficiency, which is a significant factor for many research projects, as computational methods can save time-consuming and tedious analyses of individual researchers. The latter is especially important for projects that focus on the analysis of text (e.g., reading through large text corpora to learn about sentiment expressed in the past is very intricate). Another advantage of quantitative methods is that they circumvent perception biases of individual researchers because such methods—if conducted properly—objectively reflect the investigated data. But perhaps the biggest quality of quantitative methods is that they can transfer and adapt easily to different contexts. Whether it are historic texts to learn more about the philosophy of the ancient greek or user comments posted in online social media to predict results of political elections, the approach to analyzing the considered sources often follows the same principles.

Since the digital humanities are still rather new, we see plenty of opportunities and open questions to which computational and quantitative text analysis can find answers. Providing guidelines as well as case studies for humanists and social scientists of how to apply such text analysis methods, which often stem from computer sciences or statistics, is important and could help to accelerate research projects in the respective fields. For example, think of the potential insights literary studies can provide into societies preceding the ones we know today. There are many sources of historic texts, such as pamphlets, periodicals or books, that captured the particularities and idiosyncrasies of distant times. Thanks to ambitious digitalization projects, many of these sources are available in digital forms and call for just as ambitious computational and quantitative analyses. In doing so, we can achieve new perspectives on how we see and interpret our history, precisely identify what moved people in the past as well as what brought them together or sparked controversies among them.

Controversies are thoroughly investigated in the field of communication studies for which computational methods bring advantages too. From a scholarly point of view, controversies are of a particular interest because they have been known to attract increased interest [Ziegele, Breiner, and Quiring, 2014], to increase critical thinking [Schommer-Aikins and Hutter, 2002;

Vydiswaran et al., 2012] but also to slow down decision making [Tjosvold, 1985]. Thus, controversies affected and affect the lives of many individuals in the past and present. With the rise of the Internet and today's dominating online social media, we perhaps see more controversies evolving around numerous topics, such as *climate change* or *gun control laws*, than ever before. Naturally, there already were extensive efforts to investigate the characteristics of this phenomenon on the Web [e.g., Rad and D. Barbosa, 2012; Tan et al., 2016; Zielinski et al., 2018; Hessel and Lee, 2019; Jasser et al., 2020] and especially on the social media platform Twitter [e.g., Guerra et al., 2013; Gruzd and Roy, 2014; Morales et al., 2015; Jang and J. Allan, 2018]. There are, however, plenty of other opportunities to further extend our knowledge about controversies, for example, by extending the analyses to other social media platforms or by making comparisons across platforms. In either way, the application of quantitative text analysis methods in the context of controversies could foster the public discourse.

Other fields, in which quantitative methods have already shown to bring benefits for a plethora of individuals, are management and organizational sciences. During the 20th century, employee satisfaction became more relevant to scholars, on the one hand to help employers because employee satisfaction has shown to be closely linked to business performance [e.g., P. C. Smith et al., 1969; Podsakoff and Williams, 1986; Ramayah and Nasurdin, 2006], and on the other hand to better the lives of employees themselves [e.g., Darling, Arn, and Gatlin, 1997; Rathi and Rastogi, 2008]. There have been numerous attempts in describing and defining employee satisfaction in traditional research [e.g., Hoppock, 1935; Blood, 1969; Schneider and Schmitt, 1976; Spector, 1985] and previous studies further identified a great potential of data from the Web, specifically online employer reviews, to benefit the research of employee satisfaction [Miles and Mangold, 2014; Dabirian, Kietzmann, and Diba, 2017; Green et al., 2019]. This calls for the application of text analysis methods on related Web data to learn more about employee satisfaction.

Note that the above examples only highlight a small selection of research areas for which quantitative text analysis methods may prove useful. As mentioned before, these methods can virtually be applied to any field that is related to digital humanities. The aim of my thesis is to demonstrate exactly this and to paradigmatically illustrate how the same quantitative and computational text analysis methods can be applied in diverse contexts. In the following sections, I outline how I achieve this goal. To be specific, I describe the problem statements, objectives and general approaches of my thesis in Section 1.2, the respective case studies in Section 1.3, my related publications in Section 1.4, the overall contributions and implications of this thesis in Section 1.5, as well as its structure in Section 1.6.

1.2 Problem Statements, Objectives and General Approaches

Problem Statements. Despite the usefulness of computational and quantitative text analysis methods, their application in the context of humanities and social sciences is still rather hesitant. Due to this, our research community misses plenty of opportunities as the respective research areas typically have a significant impact on the lives of individuals and societies as a whole. During my work for projects related to digital humanities, I noticed two problems that are potentially responsible for the slow approach between computer sciences and the respective fields in humanities and social sciences. Perhaps the biggest obstacle are steep learning curves of computational methods and the lack of transparent and easily interpretable guidelines for the application of them in the respective fields. In my experience, the humanists had excellent knowledge about the considered data (they have worked with it for several years at the time I joined the research project), spend lots of efforts to create well-curated and well-annotated digital editions, which they knew to be perfectly suitable for computational and quantitative analysis. While they already knew about certain approaches, such as topic modeling, sentiment analysis or named entity recognition, they lacked a quick start guide to actually conduct such analyses. As such, we require thorough guidelines that exemplify the application of quantitative text analysis methods in a straightforward and transparent way. Ideally, these guidelines illustrate the whole process, starting with necessary text preprocessing and ending with making actual interpretations of the considered data through visualizations and statistical hypothesis testing. Additionally, the demonstrated methods should be applicable to various different contexts in order to benefit a wide spectrum of disciplines, such as literary studies, philosophy, history, education, cultural studies or linguistics.

The second obstacle that I observed and may keep humanists and social scientist from applying computational methods is a certain skepticism of them towards such methods. On the one hand, the underlying methodologies involved with such approaches may be hard to interpret and untransparent for them. The latter holds especially true for some of the more recent text analysis methods that majorly rely on sophisticated machine learning approaches, such as neural networks. Hence, it is important to initially introduce humanists to methods that are transparent and leave no questions on how results were created. On the other hand, humanists have manually studied their data for many years and all of a sudden there are computational approaches that can generate results in no time, which may probably be worrying to some of them. Of course, quantitative approaches can never replace the qualitative work of these experts. They should rather be understood as complementary research methods to broaden our perspectives of certain matters. Therefore, it is of utmost importance to demonstrate how humanists can work in symbiosis with computer scientists as well as how long-existing insights and believes can be aligned with results gained from such novel methods.

To counteract these two obstacles, in this thesis, I showcase how one can apply transparent and easily interpretable text analysis methods to benefit research in diverse disciplines. More precisely, I present three case studies respectively related to literary studies, communication studies as well as management sciences. These case studies comprise separate articles on quantitative text analyses I conducted for different research projects I was part of during my doctoral research. In the first case study, I deal with the application of sentiment analysis (a specific kind of text analysis that tires to assess the emotions, opinions and attitudes conveyed by text) on a multilingual historic text corpus that has previously only been studied through manual close-reading experiences, a time-consuming and tedious process that allows for a broad understanding on a microscopic level but fails to provide a bigger picture of the data on a macroscopic level. A particular challenge of sentiment analysis in a historic context lies in the language forms of earlier times, including different spellings and meanings of words as compared to today's languages. In the second case study, I investigate controversy on online social media platforms. As mentioned before, controversies attract increased attention [Ziegele, Breiner, and Quiring, 2014] which is a challenge for providers of social media platforms. The increased user involvement often leads to heated debates among users that often result in personal attacks and insults [Sumi, Yasseri, et al., 2011]. Hence, the automatic prediction of controversy based on textual features was investigated extensively in existing communication studies [e.g., Rad and D. Barbosa, 2012; Tan et al., 2016; Zielinski et al., 2018; Hessel and Lee, 2019; Jasser et al., 2020]. A commonality of these existing works is that they focused solely on the English language, while other languages and especially a comparison between languages are lacking in previous research. To this day, we do not have any knowledge of whether or not cultural traits or the respective languages have an influence on the characteristics that constitute controversies in online social media and, as such, we cannot make any assumptions about the generalizability of previous findings. Finally, in the third case study, I focus on employee satisfaction expressed in online reviews. Traditional research of employee satisfaction is based on survey data manually collected in specific industries as well as on distinct methodologies. Hence, we lack comparable results across different countries as well as industries. One way to counteract this problem is to rely on user-generated data from the Web and employ quantitative text analysis methods to gain new insights from this data. The problem here is that, as of today, we do not know what to to expect from such reviews with regards to the well-established theories of employee satisfaction and, thus, comparisons to traditional studies may be hard to conduct.

Objectives. My thesis aims at demonstrating how one can apply quantitative text analysis methods to gain new insights in diverse disciplines related to digital humanities. For that, I present three case studies with a focus on literary and media studies, communication studies and management science, respectively. I specifically focus on these diverse fields to highlight how similar methods can be used to address completely different research questions. This showcases how easy these quantitative methods can be adapted and transferred to other contexts. The presented methods are all easily interpretable by scholars working in any field and are applicable to different languages, which further makes the presented guidelines more accessible to researchers around the globe. In particular, I aim to outline necessary data preprocessing steps, introduce suitable quantitative text analysis methods for the separate case studies (the majority of those are the same across case studies), suggest visualization techniques to illustrate computed results and employ statistical hypothesis testing that allows for assessing the significance of made observations.

Regarding the objectives of the individual case studies, I first demonstrate how to apply sentiment analysis on a historic, multilingual and computationally unexplored text corpus. The quantitative analysis shall, for example, reveal how sentiment developed over time, how topics were perceived in earlier times, how different entities (e.g., persons or countries) referred to each other with regards to sentiment, or how words conveying a sentiment were used in historic texts. The presented analysis is related to literary and media studies, but the the proposed approach to analyze sentiment may be of interest to a plethora of scholars who focus on the analysis of historic texts in other research areas. For example, this work may be relevant for scientists working in history, philosophy, sociology, cultural studies, linguistics, performing arts or religious studies. Second, I employ quantitative text analysis methods to investigate controversy in online social media. Contrary to existing studies focusing on this phenomenon, I aim to consider a multitude of the data, to combine previously separately investigated features and to compare the results across six different languages. In particular, this study investigates English, French, German, Italian, Portuguese and Spanish user comments posted on such a social media platform and, thus, I can compare the different characteristics of controversy across cultural boundaries. The proposed methods are related to communication studies and exemplify how to apply quantitative text analysis in a way so it is easily comparable across the respective languages. This is of use for all researchers who base their analysis on multilingual text and want to assess cultural as well as linguistic differences in the investigated data. Third, I provide new insights on the influential factors of employee satisfaction. For that, I rely on a novel and unexplored dataset collected from the Web. The objective of this case study is two-fold: (i) I assess how the findings of previous studies on influential factors for employee satisfaction are reflected by this novel data and (ii) I analyze the same data through the lens of Herzberg's Two-Factor theory, allowing me to assess what to expect from online employer reviews in terms of this traditional and well-established theory. In doing so, I highlight how to connect results gained through novel computational methods with findings gained from traditional research, which benefits the approach of humanities as well as social sciences towards computer science. Further, for the second and third case study, I exploit the quantified features to facilitate prediction experiments that demonstrate potential real-world use cases as a result of the conducted analyses.

General Approaches. I always follow the same procedure to empirically analyze the diverse phenomena of respective case studies. This comprises a description of necessary preprocessing steps, a detailed description of the employed text analysis methods, followed by the presentation of results and a discussion thereof. A prominent part in each case study is sentiment analysis, for which I apply the exact same method throughout the individual articles. Other than that, I consider text readability, part-of-speech tags, word embeddings as well as sub group discovery among other basic linguistic features. These features are then used to assess differences in the respective observations. In any case, these methods are easy to understand as well as to interpret and are transparent without any prior knowledge of such methods. In the second and third case study, I conduct prediction experiments that demonstrate the applicability of the quantified features. Similar to the text analysis methods, the employed prediction methods are straightforward and allow for an easy interpretation of results. As such, I can draw conclusions about the importance of different features and highlight the ones that are most predictive in the respective case studies.

1.3 Case Studies

My thesis comprises applications of quantitative text analysis in three distinct case studies and aims to demonstrate how one can conduct such analyses in different research contexts. In the first case study, I apply sentiment analysis to a historic dataset containing popular periodicals, which provides new insights into how texts published during the Age of Enlightenment conveyed sentiment. In the second case study, I analyze and predict multilingual controversy on a popular social media platform. I highlight linguistic and cultural commonalities as well as differences regarding controversy and provide information on most predictive features for it. Finally, in the third case study, I focus on employee satisfaction expressed in a novel and unexplored dataset comprising online employer reviews. In doing so, I uncover influential aspects for employee satisfaction across different countries and industries and contribute to the discussion of the well-known Two-Factor Theory introduced by Frederick Herzberg. I describe each case study as well as the corresponding problem, approach, contributions and findings in more details in the following sections.

Case Study I (Historic Sentiment): Existing Sentiment Methods Applied on Historic Texts

Problem. While sentiment analysis is popularly applied on Web data, little is known about the sentiment in historic data. However, historic data holds great potential in revealing the underlying processes of how the modern societies we know today have formed. The absence of sentiment analysis on historic texts is probably due to a lack of dedicated methods which were specifically designed for the languages of earlier times. Such dedicated methods are necessary to take into account the peculiarities of older languages, such as different word spellings or word meanings compared to today's languages. This especially holds true for sentiment analysis, where word meanings have a significant impact to successfully determine polarities. An obstacle for the implementation of dedicated methods is perhaps the underlying time-consuming process involved, as they typically require substantial manual annotations from experts. To circumvent this problem, one can only fall back on existing methods that were originally created for modern texts, typically stemming from the Web. Whether and how these existing methods are applicable in historic contexts is yet to be explored.

Approach. In Koncar, Fuchs, et al. [2020], my colleagues and I rely on a manually curated and annotated digital edition which contains more than 3 700 issues of Spectator periodicals published in French, German, Italian, Portuguese and Spanish between 1711 and 1822. The Spectator periodicals were an important tool to spread the ideas of the Enlightenment. Thus, they captured a very crucial period in human history and since they questioned customs and traditions as well as included public criticism of religion and regimes, they were emotionally charged and contained various opinions and attitudes that comprise a polarizing sentiment. Each of these issues covered one or more topics, such as politics, religion, marriage or the image of women, and adhered to the Spectator-specific communicative structure defined by multiple narrative levels as well as narrative forms, such as letters to the editors, allegories or dialogues. Based on existing sentiment dictionaries for the respective languages [Yanqing Chen and Skiena, 2014], we conduct a three-fold sentiment analysis. First, we compute the sentiment of individual issues and study whether or not sentiment changed over the years, how narrative levels and forms were reflected by sentiment as well as how the specific topics had been perceived by the authors back in the Age of Enlightenment. Second, we create sentiment networks that represent different entities (i.e., periodicals, places and persons) as nodes and sentiment between entities as edges. Thus, we can, for example, infer the polarity relation between different periodicals as well as between countries of the 18$^{\text{th}}$ century Europe. Third, we construct sentiment word networks in which nodes represent sentiment words and edges represent semantic relation between the respective words. Using these sentiment word networks, we explore how words conveying a sentiment diffused across the texts and investigate most important sentiment words.

Findings and Contributions. In Koncar, Fuchs, et al. [2020], colleagues and I find that results stemming from existing sentiment methods for the most part align with close reading experiences, suggesting that such methods are also applicable in historic contexts. However, we also saw room for improvements to further increase the quality of results. In detail, we report that sentiment conveyed in the different periodicals is depending on languages as well as cultures. For example, Spanish Spectator periodicals have a consistently more negative mean sentiment compared to Italian Spectator periodicals. The mean sentiment across all languages is stable throughout the years for all languages, however. This observation is contrary to our expectations, because we anticipated that certain events, such as wars, impacted the sentiment in periodicals. Regarding the various narrative levels and forms, we find that our results are inconclusive across languages, suggesting that sentiment may be more depending on stylistic characteristics of authors than on the given forms. In combination with insights from close-reading experiences, we made some noteworthy observations. For example, our analysis of topics revealed that Italian and Spanish authors of Spectator periodicals disguised their criticism of religion through rhetorical and stylistic devices to avoid censorship by clerics. The analysis of sentiment networks to investigate relations between persons revealed that so-called "untouchables" (e.g., *Dante Alighieri*) served as resolutions for negative examples provided in the Spectator periodicals. Finally, our sentiment word networks suggest a general mild to positive attitude towards the relevant topics discussed during the Age of Enlightenment, whereas negative words were used to distinctively discuss more critical issues.

Case Study II (Multilingual Controversy): Characteristics of Multilingual Controversy in Social Media

Problem. While controversies foster overall activity of users in online social media, they are also catalysts for heated debates among users and often otherwise objective discussions are at risk to derail. Due to this, many social media platforms rely on moderators that intervene if necessary. The implementation of methods or tools that can support moderators in carrying out this task has attracted much research in the past [e.g., Choi, Yuchul Jung, and Myaeng, 2010; Pennacchiotti and Popescu, 2010; Garimella et al., 2016]. However, the majority of these studies relies on textual features and focus on the English language, which is of course not the only language used in online social media. As such, we still lack a broader understanding of controversy on social media platforms as linguistic as well as cultural discrepancies may prevent adaptions of existing methods to other languages. Such language differences may, for example, be reflected in the sentiment expressed by users, a common feature utilized in the analysis and prediction of controversy [e.g., Dori-Hacohen and J. Allan, 2013; Mejova et al., 2014]. Further, the aforementioned studies typically focus on specific social media platforms, such as Twitter or Wikipedia, and we do not know whether and how common controversy features transfer to other, still unexplored platforms.

Approach. To tackle these problems, in Koncar, Walk, and Helic [2021] colleagues and I investigate controversy of user comments contained in a multilingual and previously unexplored dataset. In particular, we study 50 linguistically and thematically different Subreddits of the popular social news aggregation website Reddit. On the various Subreddits, users can express their agreement (through up-votes) or disagreement (through down-votes) by reacting to the comments of others. Practically, Reddit has its own algorithm to label controversial comments based on these user reactions. We utilize this mechanism to identify distinct characteristics between controversial and non-controversial comments, respectively for the different languages including English, French, German, Italian, Portuguese and Spanish. The characteristics we set our focus on stem from previous research of controversies on various platforms. Hence, we combine previously separate sets of features and transfer them to other languages as well as to a new dataset. We utilize the computed features to predict controversy in those Subreddits.

Findings and Contributions. Our findings presented in Koncar, Walk, and Helic [2021] suggest that, except for the topics discussed, controversy is reflected similarly by our features across languages. On the one hand, we show that controversial comments are significantly harder to read and convey significantly more negative sentiment as compared to non-controversial comments for all languages. Similar to previous findings by Mishne, Glance, et al. [2006] and Ziegele, Breiner, and Quiring [2014], we find that the discussion of controversial issues provokes increased involvement and participation of users independent of language. Our findings suggest that existing methods as well as features can be transferred to other languages and datasets, as long as the particularities of respective languages are taken care of (e.g., specifically tuned methods to compute sentiment and readability for each language). This observation is further corroborated by our prediction experiment, for which we observe that the language of comments is not predictive of controversy in any way. Instead, features which incorporate user involvement and participation and which are completely independent of language are most predictive of controversy on Reddit. Overall, we find that we can achieve moderate prediction performance based on the extracted features.

Case Study III (Employee Satisfaction): Online Employer Reviews and Their Benefit to Employee Satisfaction

Problem. Almost all of us spend most of their lifetime at work. Thus, our livability is significantly impacted by the satisfaction or dissatisfaction we experience at work. Existing works focused extensively on measuring and studying the factors that improve employee satisfaction and more recently we have seen researchers exploiting data from the Web, specifically online employer reviews, to learn even more about it. So far, little is known about how these findings gained from Web data correlate with the findings gained from traditional survey data. Another problem is that we lack knowledge about what insights are to expect from online employer reviews in terms of employee satisfaction. While there are many theoretical frameworks to study employee satisfaction, traditional research specifically tailors their surveys to be compliant with such frameworks. This is not possible when relying on existing user-generated data found on the Web. Hence, considering online employer reviews through the lens of existing theories and frameworks would reveal how such data can be used to learn about employee satisfaction.

Approach. Colleagues and I tackle these two problems by studying the influential factors for employee satisfaction based on a novel and unexplored dataset in Koncar and Helic [2020] and in Koncar, T. Santos, et al. [2021]. In the former publication, we consider the influence of employee benefits and employee position on employee satisfaction as well as the implication of employee satisfaction or dissatisfaction on the employment status. We specifically focus on these characteristics, as they have been studied extensively in past research [S. P. Brown and R. A. Peterson, 1993; Dienhart and Gregoire, 1993; Griffeth, Hom, and Gaertner, 2000; De Cremer, 2003; Cornelißen, 2009; De Cremer, Dijk, and Folmer, 2009; Hausknecht, Rodda, and M. J. Howard, 2009; Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016]. In the latter article, we use quantitative text analysis to investigate online employer reviews through the lens of Herzberg's Two-Factor theory, a well-known and well-studied theory to explain the interaction between influential factors for employee satisfaction and employee motivation. More precisely, in both works we analyze multi-aspect online employer reviews found on kununu, a so far unexplored reviewing platform on which employees anonymously rate and review their employers. The crawled dataset comprises more than two million of such reviews created for more than 380 000 employers either located in the United Stated (U.S.) or three European countries and operating in 43 different industries. Each review comprises an overall rating, which we utilize as an expression of employee satisfaction, additional information provided by employees, such as their position or the benefits they received, as well as free-from review text. In Koncar and Helic [2020], we solely rely on the additional information provided by employers and neglect the textual content of reviews. We conduct a straightforward logistic regression to predict employee satisfaction based on this additional information. In Koncar, Walk, and Helic [2021], we take it a step further and consider the textual content of reviews as well as the different review aspects. We link the latter to factors defined by the Two-Factor theory and, thus, can assess the factors that are most important to reviewers as well as evaluate the theory on this novel data.

Findings and Contributions. The findings in Koncar and Helic [2020] support most of existing studies in previous research conducted on manually acquired survey data. For example, we find a positive correlation between the number of benefits employees received and the employee satisfaction expressed by these employees in their reviews. Additionally, we report that former employees rate significantly more negative as compared to satisfied employees, which is also corroborating previous results. Unsurprisingly, these factors are also most predictive for employee satisfaction according to the results of our prediction experiment. In the case of employee position, for which existing studies showed no significant impact on employee satisfaction [Dienhart and Gregoire, 1993; Cornelißen, 2009], we find contradicting results as, for example, managers express significantly higher levels of satisfaction compared to other positions. The position of employees does not add significantly to the prediction performance, however. All of these results are similar across countries, with a few exceptions, such as the higher number of benefits received by managers in Germany compared to managers in the U.S. Either way, our analysis and findings demonstrate a way of how to exploit such reviews to gain new insights into the influential factors for employee satisfaction. The extension of this study in Koncar, T. Santos, et al. [2021] reveals that the factors, which according to Herzberg's Two-Factor theory, prevent dissatisfaction are more important than factors that foster satisfaction. As such, when studying such reviews, we should expect to learn more about factors that prevent dissatisfaction. The textual content, specifically words related to the different factors of the theory, has high predictive power for employee satisfaction. Our results add valuable input to the discussion of the theory as our novel dataset provides an opportunity to revisit it in today's context.

1.4 Publications

My thesis cumulates the following four publications:

Article 1: [Koncar, Fuchs, et al., 2020] Koncar, P., Fuchs, A., Hobisch, E., Geiger, B. C., Scholger, M. and Helic, D. (2020). Text sentiment in the Age of Enlightenment: an analysis of Spectator periodicals. *Applied Network Science*

Article 2: [Koncar, Walk, and Helic, 2021] Koncar, P., Walk, S. and Helic, D. (2021). Analysis and Prediction of Multilingual Controversy on Reddit. 13*th* *ACM Web Science Conference*

Article 3: [Koncar and Helic, 2020] Koncar, P. and Helic, D. (2020). Employee Satisfaction in Online Reviews. 12*th* *International Conference on Social Informatics* **Best Paper Nominee**

Article 4: [Koncar, T. Santos, et al., 2021] Koncar, P., Santos, T., Strohmaier, M. and Helic, D. (2021) What Herzberg's Two-Factor Theory Reveals About Employee Satisfaction in Online Employer Reviews. **Under review at** *Business & Information Systems Engineering*

Note that I address author contributions in the case study chapters the respective publications are related to.

1.5 Contributions and Implications

My thesis provides guidelines for the application of quantitative text analysis methods in humanities and social sciences and, thus, is a significant contribution to the digital humanities. I employed the same methods in fundamentally different case studies, each with a separate research question and focusing on a multilingual dataset. Hence, I facilitate the access of quantitative text analysis methods for a plethora of researchers working in diverse research areas around the globe. Specifically, the contributions of the individual case studies cumulated in this thesis are as follows:

Historic Sentiment: With the empirical analysis of sentiment conveyed by the Spectator periodicals, my thesis demonstrates how to apply existing sentiment methods to historic data. Further, this work contributes to a better knowledge of Spectator periodicals, complementing the close-reading experiences of experts working in the fields of humanities. In particular, the detailed analysis of how topics had been perceived as well as of the sentiment relations between different entities extends our knowledge about the Age of Enlightenment, a crucial period in human history as it shaped the society as we know it today.

Multilingual Controversy: The empirical study of multilingual controversy on the social media platform Reddit reveled that controversy is universal across languages and, thus, it suggests that existing methods to analyze and predict controversy can be transferred to other languages as long as linguistic characteristics are considered. With the conducted prediction experiment, the thesis highlights the predictive strengths of a combination of previously separately considered features and, as such, the proposed method can help to inform the implementation of more sophisticated approaches to automatically support moderators and administrators of these platforms by, for example, steering their attentions towards controversial discussions, allowing them to intervene timely if necessary.

Employee Satisfaction: The analysis of online employer reviews demonstrates how to exploit such reviews to learn about employee satisfaction. The findings stem from a novel and unexplored dataset and can inform employers on how to improve working conditions to foster the satisfaction and, thus, the motivation of their employees. For example, the conducted prediction experiment revealed that employee satisfaction can be accurately predicted by solely considering the number of benefits as well as the position of employees. This may suggest that employers need to introduce more benefits and provide career opportunities to satisfy their employees. With the analysis of such reviews through the lens of Herzberg's Two-Factor theory, my thesis reveals what to expect of such reviews in terms of the theory, indicating that reviewers devote more attention towards factors, that according to the theory, solely prevent dissatisfaction. Further, it provides an opportunity to evaluate and reconsider the factors of the theory based on its evaluation on a large, up-to-date and previously unexplored dataset.

Overall, the results and findings of this thesis broaden our knowledge in three different research fields and, thus, demonstrate the general usefulness and applicability of quantitative text analysis. Scholars in literary studies gain new insights into how sentiment was conveyed in periodicals published during the Age of Enlightenment, adding fruitful input to its discussion and potentially bringing clarity to assumptions made based on close-reading experiences. The findings on controversy in online social media increases our knowledge in the fields of communication studies and can inform the development of novel and improved tools to automatically predict controversy. This can help moderators and administrators to keep their platforms free of heated and unobjective discussions, which may promote the finding of conflict resolutions and, thus, may improve the social discourse. Regarding management sciences, this thesis is of particular interest to employers as the provided results can support them in their management decisions to improve the satisfaction of their employees and, thus, their motivation. Thus, these findings can better the lives of a plethora of individuals who spend substantial amounts of their time at work, which potentially benefits societies as a whole.

1.6 Structure of This Thesis

The remainder of this thesis is structured as follows. First, I provide an overview of related work in Chapter 2. In particular, I review existing research on text analysis in Section 2.1, including its history, descriptions of commonly applied preprocessing techniques and extracted text features as well as more detailed sections on text readability and sentiment analysis. Subsequent to that, I review the existing literature related to Spectator periodicals in Section 2.2, to controversy on the Web and in social media in Section 2.3 and to employee satisfaction in Section 2.4.

I present the publication related to Case Study I (Historic Sentiment) in Chapter 3, the publication related to Case Study II (Multilingual Controversy) in Chapter 4 and the two publications related to Case Study III (Employee Satisfaction) in Chapter 5. Please note that I disclose my personal contributions to each of these publications respectively in Section 3.1, Section 4.1 and in Section 5.1. In Table 1.1, I provide an overview of the separate publications, including the case study they are related to, their related research field, their main focus as well as their main contributions.

Finally, in Chapter 6, I conclude my thesis with a digest of the main findings and contributions of respective case studies in Sections 6.1, the potential implications and limitations of this work respectively in Section 6.2 and Section 6.3 as well as related future work in Section 6.4.

Table 1.1: **Overview of Publications.** The table lists the individual publications cumulated in this thesis as well as their related research field, case study (CS), focus and main contribution.

Article	Research Field	CS	Focus	Main Contribution
Article 1 [Koncar, Fuchs, et al., 2020]	Literary studies	Case study I	Sentiment analysis in Spectator periodicals	Sentiment analysis on a historic and computationally unexplored dataset to gain insights of how sentiment conveyed during the Age of Enlightenment.
Article 2 [Koncar, Walk, and Helic, 2021]	Communication studies	Case study II	Analysis and prediction of controversy on Reddit	Analysis and prediction of controversy considering linguistic and cultural differences and conducted on an unexplored dataset.
Article 3 [Koncar and Helic, 2020]	Management sciences	Case study III	Analysis and prediction of employee satisfaction in online employer reviews	Analysis and prediction of employee satisfaction based on its interaction with employee benefits, employee position as well as employment status expressed in online employer reviews comprised in a novel and unexplored dataset.
Article 4 [Koncar, T. Santos, et al., 2021]	Management sciences	Case study III	Analysis of online employer reviews through Herzberg's Two-Factor Theory	Study of online employer reviews through the lens of Herzberg's Two-Factor theory, revealing what to expect from such reviews in the context of the theory and, thus, bridging the gap between studies relying on traditional survey data and those relying on Web data as well as evaluating the theory on unexplored and current data.

2 Related Work

In this chapter, I address various topics related to the individual publications I present in this thesis. To begin with, I describe general concepts of text analysis as well as the history behind it in Section 2.1. Further, I describe existing research related to the methods used to conduct my analyses, including text preprocessing, readability and sentiment analysis. In the subsequent sections, I focus on the respective backgrounds of the three case studies presented in this thesis. First, I outline research related to the Spectator periodicals, a popular journalistic genre which emerged during the 18^{th} century and formed societies during the Age of Enlightenment in Section 2.2. I use a manually curated corpus of Spectator periodicals to to study sentiment conveyed in texts published during that time. Second, I review studies related to controversy on the Web and social media platforms, such as Twitter or Reddit, in Section 2.3. Third, in Section 2.4, I refer to existing research in the context of employee satisfaction. I cover different definitions of employee satisfaction, including the well-known Two-Factor Theory, and research from the past as well as more recent works exploiting data from the Web.

2.1 Text Analysis

Text analysis, also referred to as text mining, is a subfield of content analysis and nowadays focuses on computational approaches to automatically interpret texts and to combine formal statistical methods with humanistic interpretive techniques [Ignatow and Mihalcea, 2017]. A broad spectrum of research fields is relying on text analysis, for example, to study political events [e.g., Adamic and Glance, 2005; Conover et al., 2011; Grevet, Terveen, and Gilbert, 2014; Gruzd and Roy, 2014; Morales et al., 2015], to enhance marketing strategies [e.g., Yubo Chen, Fay, and Q. Wang, 2003; Godes and Mayzlin, 2004; Dellarocas, X. M. Zhang, and Awad, 2007] or to understand the helpfulness of product reviews [e.g., S.-M. Kim, Pantel, et al., 2006; Diaz and V. Ng, 2018; Eberhard et al., 2018; MSI Malik and Hussain, 2018]. Commonly, text analysis covers two parts: (i) information retrieval including approaches and methods, such as web crawling, to collect and build a text corpus and (ii) natural language processing (NLP) which comprises a set of advanced statistical methods, such as sentiment analysis [e.g., Pang and Lee, 2008; Yanqing Chen and Skiena, 2014; Hutto and Gilbert, 2014] or topic modeling [e.g., Zuo, J. Zhao, and K. Xu, 2016; Gerlach, Peixoto, and Altmann, 2018], to gain new information about your text corpus.

While text analysis is relatively new in the fields of computer science, it has much history in social sciences [e.g., Schlesinger and Walworth, 1938; Berelson, 1952; Wegerif and Mercer, 1997; Reinstein and Snyder, 2005]. Before social scientist began to adapt computational methods, they had to rely on time-consuming and tedious processes, for example, by manually studying transcriptions of political speeches or newspaper articles. More recently, social scientists use different computational text analysis methods to study various matters, for example, in education [e.g., Wegerif and Mercer, 1997], economics [e.g., Reinstein and Snyder, 2005], communications [e.g., Ziegele, Breiner, and Quiring, 2014] or even anthropology [e.g., Acerbi et al., 2013].

In the following sections, I first describe the history of text analysis and then continue with the background of text preprocessing techniques (which are required to produce accurate results), text readability as well as sentiment analysis, as I heavily rely on these methods throughout the individual publications.

2.1.1 History of Text Analysis

Early Beginnings Initiated by Religion

The analysis of text has a long history that started within a religious context. Arguably, the first systematic text analysis occurred in the 13th century, in which the Dominican friar Hugh of Saint-Cher together with a plethora of other friars created the first concordance (i.e., an index of words to refer to where they occur in a text) of the bible (i.e., the *Latin Vulgate*) [Herbermann, 1913]. As religions have always been captivated by the written word, unsurprisingly, other traces of early text analysis date back to the 17^{th} century and inquisitorial pursuits by the Church. Back then, the first known dissertations about newspapers were defended by individuals to achieve degrees in theology [Krippendorff, 2018]. With the rise of the printing press, the Church became worried about the spread of nonreligious topics and tried to suppress it in moralizing ways [Groth, 1948]. A first well-documented quantitative text analysis was related to the Church as well and was conducted in Sweden during the 18^{th} century [Dovring, 1954]. The Swedish state Church analyzed the *Songs of Zion*, a set of popular hymns originating from an anonymous author that were considered as a threat to the orthodox clergy. The case became controversial throughout the analysis and many prestigious scholars debated whether or not the religious symbols in the hymns were a danger to the Church or not [Krippendorff, 2018].

Analysis of Newspapers

Around the beginning of the 20^{th} century, text analyses moved to a nonreligious context and focused on the emerging market of newspapers. Text analysis provided opportunities to study the possibilities in forming public opinions as well as to find answers for many ethical questions related to the publication of newspapers. One of the first quantitative study of newspapers was conducted by Speed [1893], who asked “Do newspapers now give the news?”. He found that during 1881 and 1893, newspapers in New York reduced their coverage of scientific, literary and religious topics in favor of gossip, scandals and sports. Similarly, Mathews [1910] uncovered that a New York newspaper devoted more space to demoralizing and trivial topics as opposed to noteworthy matters. Wilcox [1900] believed that the profit motive was the cause for the cheap coverage of newspapers, while Fenton [1910] even reported that this type of reporting is responsible for a growth in crime rates. However, there was also a study conducted by White [1924], indicating a general demand for actual facts in newspapers.

The textual analysis of newspapers helped in laying the foundation for journalism and the discussion thereof [Krippendorff, 2018]. For example, a well-known text analysis of newspapers was conducted by the Viennese Löbl [1903], where he introduced a classification scheme to analyze the structure of newspapers. His work contributed significantly to the idea of "newspaper science" (i.e., journalism or "Publizistik" in German).

Transition to Other Contexts

During the 1930s, text analysis methods used on newspapers were transferred by sociologists to study various social phenomena such as stereotypes or nationalism. For example, Simpson [1936] studied how African Americans were represented in Philadelphian newspapers. Martin [1936] analyzed how nationalism was expressed in children's books published in the United States, Great Britain as well as other European countries. Another example is the work of Schlesinger and Walworth [1938], in which they studied how textbooks from the United States described wars they participated in and how these compared to descriptions in textbooks published in countries of former war enemies.

In the 1940s, balance, fairness and bias became more important to journalists as well as researchers. For example, Janis and Fadner [1943] introduced the *coefficient of imbalance*, trying to explain how biased newspapers reported about a specific entity. Further, Allport and Faden [1940] analyzed newspapers in context of rumor transition and showed how information in an issue of a newspaper changes within an institution as it wanders from desk to desk and finally ends on the printed page.

Starting with the 1930s, text analysis was also applied in a broader scholarly context, for example in sociology [Becker, 1932], physics [Rainoff, 1929], communication [Berelson, 1952] and again journalism [Tannenbaum and Greenberg, 1961]. During the two world wars, text analysis was increasingly applied for the investigation of propaganda [George, 1959]. Ever since the extensive use of propaganda in World War I [Lasswell, 1971] and the effective impact of propaganda in Europe before World War II, the military used text analysis methods to uncover individuals that wanted to influence others for malicious intents [Krippendorff, 2018].

The Beginnings of Quantitative and Computational Text Analysis

Around the 1960s, researchers saw a great potential in using computers to complement existing manual text analysis methods. For example, Sebeok and Zeps [1958] used information retrieval techniques to analyze about 4 000 Mari (i.e., a Finno-Ugric ethnic group) folktales. Hays [1960] started with conceptualizing a computer system to automatically analyze documents related to politics. Similarly, Stone, Dunphy, and M. S. Smith [1966] introduced the *General Inquirer*, a system build on computational analysis methods to study text. This system found many applications, ranging from advertising and marketing to psychotherapy and literary studies [Krippendorff, 2018]. It was further suggested by Sedelow [1989] to replace the dictionary, that serves as the basic core of the General Inquirer, with a thesaurus (i.e., a dictionary containing more information of respective words, such as synonyms and antonyms) to better reflect word meanings. In 1993, Miller et al. [1990] introduced the well-known *WordNet*, a network-based approach to describe relations between words which is still applied and extended for the purpose of text analysis as of today [e.g., AlMousa, Benlamri, and Khoury, 2021; Arıcan et al., 2021; Özçelik et al., 2021].

With the increasing popularity of computational methods to analyze text, question arose about how they compare to manual and human-based approaches. Schnurr, Rosenberg, and Oxman [1992] evaluated the *Thematic Apperception Test* (a personality test in which a person responds to ambiguous stimuli) [Murray, 1943] to results based on computational methods and found disagreement between the two. However, their results were later rebutted by Zeldow and McAdams [1993]. In another work, Nacos et al. [1991] compared human-labeled data of political news to computer-labeled instances of the same data created by Fan [1988], who studied the influence of public information on forming opinions of individuals. Here, significant correlations between the two datasets were found, suggesting that computational methods are a promising way to analyze text.

2.1.2 Natural Language Processing

The popularity of computational text analysis methods lead to a new research area known as Natural Language Processing (NLP). It combines the fields of computer science, probability theory and statistics with linguistics, aiming to make human language accessible to computers. As of today, NLP is omnipresent in our daily lives, including, for example, text classification to keep our inboxes free from spam [e.g., Gharge and Chavan, 2017], automatic machine translation of posts submitted to social media websites [e.g., Honnet et al., 2017], writing with chat bots for health care assistance [e.g., Ayanouz, Abdelhakim, and Benhmed, 2020] as well as spelling correction in word processors [e.g., Altarawneh, 2017].

Typically, NLP comprises diverse subfields, such as sentiment analysis [e.g., Solangi et al., 2018], text summarization [e.g., Merchant and Pande, 2018], topic modeling [e.g., Jelodar et al., 2019] or authorship attribution [e.g., Stamatatos, 2017]. Concrete examples of applications include, for example, hate speech detection [e.g., A. Schmidt and Wiegand, 2017] or fake news detection [e.g., Oshikawa, Qian, and W. Y. Wang, 2018] in online social media.

In the following sections, I provide an overview of selected NLP techniques and describe readability and sentiment analysis in more detail as the publications cumulated in my thesis rely heavily on these two subfields. Note that most of the related works I present here are addressing NLP techniques for the English language. However, advances for other languages (especially for those using anything other than the Latin alphabet) have been made [e.g., K.-F. Wong et al., 2009; Habash, 2010].

Text Preprocessing

Besides the removal of noise (e.g., links or HTML tags in data originating from the Web which are irrelevant for one's research question), one may consider to apply further preprocessing steps, for example, to reduce size of data or to improve quality of results. Typically, these steps include but are not limited to tokenization, stop words removal as well as stemming or lemmatization. I will briefly discuss the concepts and ideas behind those approaches in the following paragraphs. Note that preprocessing techniques are depending on the language of the texts to analyze. Here, I focus on methods suited for English, but I point the interested reader to the works of, for example, Le et al. [2020] for French, Blombach et al. [2020] for German, Artese and Gagliardi [2019] for Italian, Gomes, Adán-Coello, and Kintschner [2018] for Portuguese and Orellana, Trujillo, and Cedillo [2020] for Spanish preprocessing techniques to provide concrete guidance for the languages occurring in the datasets used in my thesis.

Tokenization. Tokenization describes the process of separating a string of text into units or pieces often referred to as *tokens* [Hickman et al., 2020]. Typically, these units represent words, phrases or symbols in a text and are used in the subsequent tasks of NLP. Depending on the aim of the text analysis, it is suggested to always remove punctuation and other non-alphabetic characters at this step [Banks et al., 2018]. However, since punctuation also conveys stylistic characteristics, the removal of it strongly depends on the research question. [Pennebaker, Boyd, et al., 2015]. For example, punctuation is necessary for the correct computation of readability metrics, such as the Flesch Reading Ease [Flesch, 1948], and is used to amplify polarity of words in sentiment analysis [e.g., Hutto and Gilbert, 2014]. Further, special tokenizers have been introduced to capture emoticons [Kern, Park, et al., 2016], which are often represented by a combination of punctuation and other symbols.

Stop Words Removal. The removal of stop words aims at excluding words that are so common in a text, such as "the", "a", "my", that they add no useful information to the text analysis. Removing stop words also reduces the size of the data and, thus, benefits the execution time of NLP tasks [Hickman et al., 2020]. Existing research suggests to always remove stop words [Banks et al., 2018], independent of text length and corpus size [Kobayashi et al., 2018] unless one's research question focuses on stylometry (i.e., stop words are also distinctive of author style) [Kern, Park, et al., 2016].

Stemming & Lemmatization. In the process of stemming and lemmatization, different variants of words are transformed and conflated into their root (canonical) form (i.e., their *stem* or *lemma*, respectively). For example, the words "reflection", "reflecting" and "reflected" would be reduced and combined to the single word "reflect". Both approaches were originally conducted to reduce execution time of computational text analysis methods [Hickman et al., 2020]. Stemming and Lemmatization differ in their methods to find root forms. The former uses rule-based heuristics to truncate suffixes from words (e.g., "organs" → "organ" and "organic" → "organ") with the risk of combining words with different meanings [Vijayarani, Ilamathi, Nithya, et al., 2015]. Common methods and rule sets for stemming include the *Porter Stemmer* [Porter, 1980; Porter, 2001] (which is also available for languages other than English) or the *Lovins Stemmer* [Lovins, 1968].

Contrary, more sophisticated lemmatization methods rely on dictionaries of words and their respective variants and meanings. This allows lemmatization to keep words and their distinct meanings when reducing them to their root form (e.g., "organs" → "organ" while "organic" stays the same) [Balakrishnan and Lloyd-Yemoh, 2014]. As such, lemmatization provides more interpretable results compared to stemming [Schütze, Manning, and Raghavan, 2008]. Existing lemmatization algorithms include rule-based [Juršic et al., 2010] and classification-based [Gesmundo and Samardzic, 2012] approaches as well as methods relying on neural networks introduced in more recent research [Kondratyuk et al., 2018].

Text Features

To conduct the actual computational text analysis, one has to extract features, which are interpretable by computers, from the texts. In this section, I will briefly describe selected text features used throughout the individual publications of my thesis.

Part-of-Speech Tagging. Besides basic linguistic features, such as the number of characters, syllables and words in a text, a variety of other approaches exist to extract features and information from text. A very commonly applied technique, which is also a precursor for many other NLP tasks (e.g., for sentiment analysis [Dell'Orletta, 2009] or named entity recognition [Budi et al., 2005]), is *part-of-speech* (POS) tagging, in which each word in a sentence is assigned to its part-of-speech, such as noun, verb or adjective [Voutilainen, 2003]. Many stochastic methods (e.g., Markov models) to conduct POS tagging have proven to be accurate in past approaches [e.g., DeRose, 1988; Church, 1989; Mcteer, Schwartz, and Weischedel, 1991; Thede and Harper, 1999], almost always outperforming simple rule-based approaches [e.g., Klein and Simmons, 1963; Brill, 1992]. More recently, POS taggers based on deep learning [Deshmukh and Kiwelekar, 2020] as well as domain specific approaches for social networks, such as Twitter [Fanoon and Uwanthika, 2019] or Reddit [Behzad and Zeldes, 2020], have been explored.

Network Representations of Text. Networks have been used to represent and analyze text in past research. Typically, the nodes in such networks represent a word and edges between those nodes represent a semantic or syntactical relation between the words. For example, *word adjacency networks*, which capture syntactical relations between words, have been used for stylometry [Diego Raphael Amancio, 2015] and authorship attribution [Segarra, Eisen, and Ribeiro, 2015; Segarra, Eisen, Egan, et al., 2016]. Roxas and Tapang [2010] used such networks to uncover and classify differences in the structure of prose and poetries. In a different context, word co-occurrence networks have been used to automatically determine word usages [Véronis, 2004] or for lexical acquisition [Widdows and Dorow, 2002]. Diego R Amancio, Oliveira Jr, and Costa [2012] as well as Silva and Diego R Amancio [2012] used networks for the automatic disambiguation of word senses (i.e., finding the meaning of ambiguous words based on the context). Another work by Antiqueira et al. [2009] demonstrated how word networks can be used for text summarization. Additionally, word networks have been utilized for topic modeling (i.e., finding abstract topics that occur in a text corpus) [Z. Liu et al., 2010; Zuo, J. Zhao, and K. Xu, 2016; Gerlach, Peixoto, and Altmann, 2018] and sentiment analysis [Yanqing Chen and Skiena, 2014].

Word Embeddings. A useful way to represent text through numbers are word embeddings, which are fixed length vector representations of words. Baroni, Dinu, and Kruszewski [2014] describe two different kinds of word embeddings: (i) count-based models and (ii) prediction-based models. The former base on corpus-wide statistics whereas the latter build on machine learning techniques [Almeida and Xexéo, 2019]. One of the first approaches for count-based word embeddings was proposed by Salton, A. Wong, and C.-S. Yang [1975], who introduced the *vector space model* and suggested that each document in a collection is represented by a t-dimensional vector, where t is the number of distinct terms in all documents. The elements in the vectors can be binary or any real number (eventually weighted as used in TF-IDF) and provide information about whether or not words are occurring in the respective documents. Using these vector space models, one can then apply computational methods, for example, to compute similarity between documents. More sophisticated count-based approaches to create word embeddings include *Latent Semantic Analysis* [Deerwester et al., 1990], *Hyperspace Analogue to Language* [Lund and Burgess, 1996] and its improved version *COALS* [Rohde, Gonnerman, and Plaut, 2006]. Perhaps the most popular among these approaches is *GloVe* introduced by Pennington, Socher, and Manning [2014], which has been found by the authors to significantly outperform other count-based methods.

Prediction-based methods to create word embeddings are strongly linked to *neural language models*, which served as a foundation for some of the most popular methods we know today [Almeida and Xexéo, 2019]. The first approach was documented by Bengio, Ducharme, et al. [2003] and early results suggested that such methods can model language better than more traditional count-based techniques. Soon, advances towards better efficiency and other improvements followed [Bengio, Senécal, et al., 2003; Morin and Bengio, 2005; Mnih and Hinton, 2007]. Arguably one of the most influential studies regarding prediction-based word embeddings were conducted by Tomas Mikolov, Kopecky, et al. [2009], Tomáš Mikolov, Karafiát, et al. [2010], and Tomáš Mikolov, Yih, and Zweig [2013]. These intermediary works ended in the well known *word2vec* model based on two different architectures: (i) the continuous bag-of-words (CBOW) model as well as the (ii) skip-gram model [Tomas Mikolov, K. Chen, et al., 2013; Tomas Mikolov, Sutskever, et al., 2013]. The main difference between the two is that the former predicts the center word based on its context, whereas the latter predicts the context based on a center word. Word2vec not only reflects syntactic but also semantic relations, which led to the tremendous success of the model. Many relations, such as female-male, and even analogies, such as Vienna is to Austria what Berlin is to Germany, can be recreated by arithmetical operations on word vectors created through word2vec. In 2016, Bojanowski et al. [2017] as well as Joulin et al. [2016] introduced *FastText*, which further increases the performance of prediction-based word embeddings by not training word embeddings but instead training n-gram embeddings.

Text Readability

Assessing the readability (i.e., the difficulty to read) of a text is a popular subfield in NLP [e.g., Rello et al., 2012; Lenzner, 2014; Temnikova, Vieweg, and Castillo, 2015; Daraz et al., 2018; Pancer et al., 2019]. Ever since knowledge started to transfer in a written form, the comprehensibility of text has been a concern for scholars. For example, Zakaluk and Samuels [1988] reported that savants of the ancient Greek thought and argued about the clarity of texts used in rhetorical training for law students. In the 20^{th} century, more scientific and systematic approaches to study readability emerged [Collins-Thompson, 2014]. Dale and Jeanne S Chall [1949] provided a first formal definition of text readability, defining it as the essence of text characteristics that influence the understanding, level of interest as well as the reading speed of a reader. Such characteristics include, for example, the overall text length or the mean number of words in a sentence.

One of the first formulas to predict readability was introduced by Vogel and Washburne [1928]. The authors based their work on a sample of 700 books, which had been read and evaluated by pupils for the paragraph meaning section of the Stanford Achievement Test. More precisely, authors extracted textual characteristics (i.e., the number of unique words in a 1 000 words sample, the number of prepositions in the sample, the number of clauses in a 75 sentence sample as well as the number of words not in the Thorndike list of the 10 000 most frequent words [Thorndike, 1921]) and defined a formula to provide grade-level rankings for each book. Some of the mentioned textual characteristics served as a model for other, more modern formulas [R. C. Anderson and Davison, 1986].

In general, most traditional formulas to estimate text readability rely on some combination of word difficulty (semantic features) and sentence complexity (syntactic features) [see Jeanne Sternlicht Chall, 1958; Zakaluk and Samuels, 1988]. For example, the *Flesch Kincaid Grade Level* (FKGL) [Kincaid et al., 1975] provides a grade-level ranking based on the mean number of words per sentence as well as the mean number of syllables per word:

$$FKGL = 0.39 * \left(\frac{\text{number of words}}{\text{number of sentences}}\right) + 11.8 * \left(\frac{\text{number of syllables}}{\text{number of words}}\right) - 15.59 \tag{2.1}$$

In this case, higher scores indicate a harder to read text and that they require a higher grade level to understand it.

Other formulas, such as the *Flesch Reading Ease* [Flesch, 1948], the *Dale-Chall readability formula* [Dale and Jeanne S Chall, 1948], the *SMOG Index* [Mc Laughlin, 1969] or the *Gunning Fog Index* [Gunning, 1952] work with similar principles and slightly adjusted formulas, often resulting in different scales (e.g., the Flesch Reading Ease ranges from 0 to 100, where higher values indicate an easier to read text). While all of these formulas were intended for the English language, researchers have adapted the formulas for other languages as well. For example, Franchina and Vacca [1986] introduced an adaption of the Flesch Reading Ease for Italian and Amstad [1978] for German.

Readability formulas were applied numerously in existing research based on Web data [Milne, Culnan, and Greene, 2006; Korfiatis, Rodríguez, and Sicilia, 2008; Hu, Bose, Koh, et al., 2012; Temnikova, Vieweg, and Castillo, 2015; Fang et al., 2016; Fabian, Ermakova, and Lentz, 2017; Daraz et al., 2018; Pancer et al., 2019]. For example, Pancer et al. [2019] analyzed 4 000 Facebook posts from *Humans of New York* (a popular blog that publishes portraits and interviews of individuals collected on the streets of New York City) over the course of three years. They found that posts that are easier to read are liked, commented on and shared more frequently as compared to harder to read posts, suggesting that readability has an impact on the engagement of users participating on social media websites. In another context, Fang et al. [2016] focused on factors that influence the value of online tourism reviews and reported that better readability increases the helpfulness of such reviews. Korfiatis, Rodríguez, and Sicilia [2008] found similar results for online book reviews. Hu, Bose, Koh, et al. [2012] studied how readability and sentiment analysis can be used to detect the manipulation of online reviews. Additionally, Temnikova, Vieweg, and Castillo [2015] studied readability of crisis communication on Twitter in English-speaking countries. Based on their findings, the authors describe the factors that negatively influence reading comprehension and state several recommendations to write clearer and easier to understand crisis messages. Daraz et al. [2018] investigated the readability of online health information in the United States and Canada by considering 13 different readability formulas. They concluded that online health information is inappropriate for general public use as the texts are too hard to read, suggesting that this may lead to misinformation as well as negatively impact health in these countries. Similar to that, Fabian, Ermakova, and Lentz [2017] analyzed 50 000 privacy policies of popular English websites and found that such policies are hard to read. As another contribution of their work, authors mentioned a redundancy in applying multiple readability formulas as most of them provide similar results. Their results on readability of privacy policies are supporting earlier findings of Milne, Culnan, and Greene [2006], who uncovered that privacy policies are becoming harder to read already in the year 2006.

The aforementioned works demonstrate the varying applications of readability formulas in the context of Web data. Besides data originating from the Web, text readability was also extensively applied, for example, to study the quality of textbooks [Flory, Phillips Jr, and Tassin, 1992; Chiang-Soong and Yager, 1993] or the clarity of questions in surveys [Harmon, 2001; Lenzner, 2014], to support people with learning disabilities and dyslexia [Abedi et al., 2011; Rello et al., 2012] and even by the military to increase comprehension of various technical instructions [Kincaid et al., 1975].

Sentiment Analysis

Another popular subfield of NLP is sentiment analysis which is the computational study of people's emotions, attitudes and opinions towards a specific entity (e.g., another person or a topic) conveyed by textual data [Pang and Lee, 2008]. Sentiment analysis is one of the most actively studied aspects of NLP and attracts much attention since the beginning of this millennium [B. Liu, 2020]. Arguably the first mention of the term *sentiment analysis* occurred in the works of Nasukawa and Yi [2003], who manually created a set of terms (or opinion words) to automatically compute the sentiment expressed in camera reviews extracted from websites. However, earlier works also studied sentiment and opinions while not specifically referring to the term sentiment analysis. For example, Hearst [1992] introduced a *direction-based text interpretation* model which assesses whether the author of a text is in favor, neutral or opposed of a specific event. Another example is the work of Hatzivassiloglou and McKeown [1997], in which authors proposed an algorithm to create a set of words based on their positive or negative semantic orientation.

In their survey paper, B. Liu and L. Zhang [2012] reported three main approaches to sentiment analysis: (i) dictionary based, (ii) classification based with supervised learning and (iii) classification based with unsupervised learning. In my thesis, I rely on the first category (dictionary based) which is why I focus on these approaches for the rest of this section. However, I point the interested reader to the works of Pang, Lee, and Vaithyanathan [2002], Ye, Z. Zhang, and Law [2009], Maas et al. [2011], and L. Zhang, S. Wang, and B. Liu [2018] for examples on how to conduct sentiment analysis through the other two classification-based approaches.

Dictionary based approaches rely on lists of words (or other lexical units, such as word bi-grams or phrases) for which the sentiment or polarity is known. Typically, a word in a dictionary may be represented as a tuple of the form *(word, sentiment)*, where the word can be any word of a given language and the sentiment can be binary (i.e., positive or negative), multi-class (e.g., extremely positive, slightly positive, neutral, slightly negative, extremely negative), or any continuous value (e.g., a score ranging from -1 to $+1$, where values close to -1 represent a negative sentiment, values close to $+1$ represent a positive sentiment and values close to 0 represent a neutral sentiment) [Ahire, 2014]. The creation of such dictionaries can be conducted in two ways: (i) a time-consuming and tedious manual creation of dictionaries and (ii) the automatic creation of dictionaries by extending a small set of seed words through various computational approaches [B. Liu and L. Zhang, 2012]. In case of the former, multiple annotators label a predefined set of words for which some sort of agreement is computed (e.g., the mean of numerical labels) [e.g., Hutto and Gilbert, 2014]. Regarding the latter, a small set of words (usually also manually annotated) is used in conjunction with another form of lexical dictionary or resource (e.g., WordNet [Miller et al., 1990]) to automatically extend the set of seed words [e.g., S.-M. Kim and Hovy, 2004]. Once the dictionary is created, one can compute the sentiment on different levels (e.g., sentiment or document level) by, for example, using a naive approach which sums the sentiment values of words with a labeled sentiment. If the resulting sum is below zero, the text would be considered to convey a negative sentiment, otherwise it would be considered to convey a positive sentiment.

Among the well-known sentiment dictionaries is VADER, an approach specifically design for shorter, English texts originating from social media sites and introduced by Hutto and Gilbert [2014]. This dictionary is based on more than 7 000 words that were manually annotated by 10 independent individuals with a score ranging between -4 (negative) and $+4$ (positive). The authors then defined the sentiment of each word by averaging over the ten annotations. Their proposed algorithm to compute sentiment scores is based on a set of rules that, for example, accounts for negations and considers punctuation to weaken or amplify resulting sentiment scores. Similar to that, Thelwall et al. [2010] introduced *SentiStrength*, a program based on sentiment dictionaries stemming from sentiment terms included in the General Inquirer [Stone, Dunphy, and M. S. Smith, 1966], the Linguistic Inquiry and Word Count (LIWC) [Pennebaker, Francis, and Booth, 2001] program as well as ad-hoc additions of words made by the authors themselves during their evaluation process. This approach is also available for languages other than English [Thelwall, 2017]. The aforementioned LIWC [Pennebaker, Francis, and Booth, 2001] is another famous dictionary based approach to analyze sentiment. In contrast to previously described works, LIWC extends the mere sentiment analysis and identifies multiple characteristics of authors, such as social relations, honesty and thinking style.

The applications of sentiment analysis (using either dictionary based or classification based approaches) on Web data are manifold. For example, V. K. Singh et al. [2013] as well as Thet, Na, and Khoo [2010] conducted a sentiment analysis to study movie reviews. Similarly, Srujan et al. [2018] classified Amazon book reviews through sentiment analysis. In a different context, Abel et al. [2017] performed such an analysis on German employer reviews. Alaei, Becken, and Stantic [2019] provided an overview of sentiment analysis applications in tourism and hospitality. They compiled many different works exploiting reviews written on different platforms, such as TripAdvisor.com or Booking.com, to study how to best increase revenue. More recently, Barkur and Vibha [2020] studied Tweets posted in India during the first lockdown due to the COVID-19 outbreak. The authors concluded that the majority of India's population had a positive attitude towards the lockdown and were supportive of the actions taken by the government.

Sentiment analysis found an increased interest in digital humanities and literary studies in recent years. In a historic contexts, applications range from the prediction of happy endings in German novels of the 19$^{\text{th}}$ century [Jannidis et al., 2016], over the analysis of plays written by Gotthold Ephraim Lessing in the 18$^{\text{th}}$ century [T. Schmidt, Burghardt, and Wolff, 2018], to the understanding of sentiment in subgenres comprised in a corpus of 19$^{\text{th}}$ century Spanish American novels [Henny-Krahmer, 2018]. In more modern contexts, sentiment analysis was applied to study the emotional arcs of stories [Reagan, Mitchell, et al., 2016] or, in a similar fashion, to categorize literary genres of stories through the development of sentiment within those stories [E. Kim, Padó, and Klinger, 2017]. These examples highlight the diverse applicability of sentiment analysis in the different research fields.

2.1.3 Extension to Previous Research

The aforementioned literature demonstrates the value of computational text analysis to expand our knowledge in different fields. In my thesis, I apply text analysis on unexplored datasets to shed light on three different case studies. First, I apply sentiment analysis to historic Spectator periodicals (see Section 2.2) by relying on existing sentiment dictionaries [see Yanqing Chen and Skiena, 2014] and various network representations of entity relations as well as sentiment words. The Spectator periodicals were previously only analyzed through close-reading approaches and my work is the first to apply computational methods to this data. Since such periodicals had their heyday in the 18$^{\text{th}}$ century, they hold great potential in gaining new insights on how the Age of Enlightenment shaped our society as we know it today. Second, I analyze and predict controversy (see Section 2.3) of user comments posted on the popular social media platform Reddit. In doing so, I extend previous studies thereof by considering a multitude of data as well as by combining different features, such as word usage, writing style or user involvement, that have previously only been considered in separate works. Third, I apply textual analysis to investigate online employer reviews and assess what they reveal about employee satisfaction (see Section 2.4). While text analysis has been applied to study employee satisfaction in the past, I consider several different textual features that have not been considered together in previous research.

2.2 The Spectator Periodicals

In the first case study presented in my thesis, I focus on Spectator periodicals published during the Age of Enlightenment, a groundbreaking period in human history as it formed the society that we know today. The Spectator periodicals significantly influenced their readers during that time regarding, for example, the image of women [Messbarger, 1999; Carr, 2014] or religious beliefs [D. Allan and Virtue, 1993]. In the following sections, I briefly describe the history of Spectator periodicals, specifically focusing on why they inspired numerous translations and imitations all across Europe.

2.2.1 The Path to the Spectator Periodicals

Before the news in the printed press became "periodical", it was published rather occasional. For example, between 1450 and 1600, news spread through pamphlets or proclamations which covered religious matters, such as the Protestant Reformation, battles of wars or natural disasters [Seguin, 1963]. With the beginning of the 17th century, we saw the introduction of actual newspapers (and even "newsbooks") that were published periodically (daily, weekly, monthly or quarterly) and aimed to provide new information (thus it is called "news") to its readers. The issues of these periodicals were numbered so that the readers would know whether they skipped an issue and missed information [Schröder, 1995]. By the 1660s, there were about 50 newspapers in Europe, published in different languages, including German, French, Italian, English as well as Latin [Schröder, 1995]. Further, Ries [1977] showed that newspapers borrowed from each other (e.g., a news paper published in Copenhagen regularly translated and adapted German news published in Hamburg), suggesting a cultural translation of the news. With an increased demand of newspapers, we saw that specializations of them emerged in France and which can be categorized into three types: (i) political newspapers, (ii) social newspapers and (iii) scholarly newspapers. For example, in the 17th century France, the *Gazette* covered political information (always in accordance with the requirements of the government) [Feyel, 2000], while the *Mercure galant* focused on the world of fashion and interior decoration [Dotoli, 1983]. The *Journal des savants* planed to cover scholarly information, such as short summaries of new books or legal decisions (stemming from the fact that many French scholars were lawyers back in the time). Eventually, it ended up discussing literature, church history and natural philosophy [Morgan, 1929]. All three types of periodicals were imitated outside of France, again suggesting a process of cultural exchange [M. L. Pallares-Burke, 2007].

Starting with the 18th century, a new type of periodicals was introduced with the publication of *The Tatler* (1709 – 1711), *The Spectator* (1711 – 1714) and *The Guardian* (1713), all published by *Richard Steele* and *Joseph Addison* in England. These three periodicals had enormous success in a market where periodical publications were commercially unsuccessful and could only be sustained through partisan political support [Downie, 1993]. The determining factor for their success was the combination of the three previously separate fields of the press (political, social, scholarly). As such, they could reach a broader public and introduced a more secular discourse. In addition, due to its popularity, *The Spectator* created and defined a new genre of periodicals and journalism, today known as *Spectator periodicals* [M. L. Pallares-Burke, 2007; Krefting, Nøding, and Ringvej, 2015].

2.2.2 Translations and Imitations

Due to their success, the Spectator periodicals have been translated and imitated all across Europe. The first imitation was Justus van Effen's *Le Misantrope*, a French periodical published in the Netherlands in 1711 [Rau, 1980]. According to Stephen [1920], imitations in England were countless. In France, more than 100 imitations of the *The Spectator* have been published [M. L. Pallares-Burke, 2007]. Similarly, we know of multiple imitations in Germany, such as *Der Patriot* (published in Hamburg in 1724) or *Der Freymaurer* (published in Leipzig in 1738) [Maria Lúcia G Pallares-Burke, 1994]. Further, there were multiple (partial) translations of *The Spectator* to German and French in 1714 [Gilot and Sgard, 1981; Martens, 2017], followed by Danish, Dutch, Italian, Portuguese, Spanish and Swedish [Gustafson, 1932; M. L. Pallares-Burke, 2007; Krefting, 2018]. These translations were often perceived better than respective imitations as ,for example, stated in a letter by *Johann Wolfgang von Goethe* to his sister, in which he mentioned that the German imitations copy the outward appearance but not the actual essence of *The Spectator* [M. L. Pallares-Burke, 2007]. The sheer popularity of these translations was further corroborated by the study of Mornet [1910], who analyzed 500 catalogues of French private libraries and found French translations of *The Spectator* to appear the most.

The countless imitations and translations also rose questions about cultural translations. Maria Lâucia G Pallares-Burke [2002] mentioned that a translator of a text as well as a cultural translator faces the dilemma between intelligibility and fidelity. Controversies over whether or not a translation or imitation of *The Spectator* was truthful occurred within the so-called *Republic of Letters* (an intellectual community of the Age of Enlightenment) [M. L. Pallares-Burke, 2007]. Questions about what it meant to be a *Spectator* and to write as a *Spectator* were asked. The evolving debates focusing on these issues were mainly discussed in the European cities Copenhagen, Paris and Zürich [M. L. Pallares-Burke, 1996]. In Copenhagen, Holberg [1748] mentioned an internal war between *Spectators* and referred to other Danish periodicals trying to convert readers to morals too harshly and quickly, where instead a good *Spectator* should be a "gentle teacher", who eradicates vices and faults among people. Similarly, in Paris comparisons between new Spectator periodicals and the traditional model introduced by Addison and Steele were made. The authors of these comparisons noted that to be a true *Spectateur*, one needs to obey to a certain way of writing and needs to be denounced in the Republic of Letters otherwise [M. L. Pallares-Burke, 2007]. In Zürich, Bodmer [1728], who were the authors of the first German imitation, led a campaign against *Der Patriot* and *Die vernünftigen Tadlerinnen*, which allegedly did not adhere to authenticity and impartiality.

2.2.3 Extension to Previous Research

The numerous existing studies focusing on the Spectator periodicals highlight the interest of scholars. As the periodicals helped in shaping a more modern society during the Age of Enlightenment, they contain valuable information about this crucial process in human history. In contrast to existing studies, which heavily relied on time-consuming and tedious close-reading experiences (i.e., manually reading through each issue of a periodical), I apply the first large-scale sentiment analysis to the corpus of Spectator periodicals. This multilingual analysis covers different parts, including but not limited to how sentiment developed over the time, how different topics, such as politics and religion, had been perceived, how different stylistic elements affected sentiment as well as how periodicals referred to other periodicals with regards to sentiment. As such, I provide new insights into how Spectator periodical expressed sentiment, which allow for comparisons across five different languages and add fruitful input, for example, to the discussion of unethical translations and imitations through the analysis of sentiment conveyed between periodicals.

2.3 Controversy on the Web

The analysis and impact of controversial contributions on the Web has been studied extensively in various contexts, including weblogs [Adamic and Glance, 2005; Mishne, Glance, et al., 2006], news articles [Choi, Yuchul Jung, and Myaeng, 2010; Mejova et al., 2014; Siersdorfer et al., 2014; Ziegele, Breiner, and Quiring, 2014], Twitter [Guerra et al., 2013; Gruzd and Roy, 2014; Morales et al., 2015; Jang and J. Allan, 2018], Reddit [Tan et al., 2016; Hessel and Lee, 2019; Jasser et al., 2020], search engines [Yom-Tov, Dumais, and Q. Guo, 2014; Dori-Hacohen, Yom-Tov, and J. Allan, 2015; Koutra, Bennett, and Horvitz, 2015] or Wikipedia [Rad and D. Barbosa, 2012; Dori-Hacohen and J. Allan, 2013; Dori-Hacohen and J. Allan, 2015; Zielinski et al., 2018]. Most of these studies rely on aforementioned text analysis methods, such as POS tagging [e.g., Jang and J. Allan, 2018], readability [e.g., Siersdorfer et al., 2014] or sentiment analysis [e.g., Mejova et al., 2014; Zielinski et al., 2018]. Additionally, networks have been utilized to investigate controversy [e.g., Yom-Tov, Dumais, and Q. Guo, 2014; Dori-Hacohen, Yom-Tov, and J. Allan, 2015; Koutra, Bennett, and Horvitz, 2015]. I will describe these works in detail in the following sections.

Tjosvold [1985] described controversy as a special kind of conflict (which exists whenever incompatible activities occur [Deutsch, 1973]), in which the ideas, opinions, information and theories of one person are incompatible to those of another person. Further, these differences lead to at least a temporarily delay of reaching consensus. Similarly, according to Merriam-Webster Dictionary, controversy is defined as "a discussion marked especially by the expression of opposing views"[1]. Zielinski et al. [2018] even introduced a formal computational model to define controversy based on a group of people (or a set of communities), their opinion, and any given object they discuss. In either way, controversy comprises the discussion of a certain topic that sparks agreement and disagreement among individuals and communities. Popular examples include *gun control laws* [e.g., Spitzer, 2020] or *abortion laws* [e.g., Sarvis and Rodman, 1973].

2.3.1 Controversy in News Articles, Wikipedia and Search Engines

Choi, Yuchul Jung, and Myaeng [2010] introduced a method relying on text and sentiment analysis to automatically detect controversial issues and their subtopics in online news articles. Similar to this approach, Mejova et al. [2014] introduced a new and crowd-sourced lexicon consisting of controversial and non-controversial terms that serves as a basis for sentiment analysis in context of controversy in online news articles. Their results indicate that controversial news articles include biased language without conveying strong emotions. Siersdorfer et al. [2014] analyzed controversy of comments posted to articles published on Yahoo! News as well as to YouTube Videos based on the number of up- and down-votes and the words that appeared in them. Ziegele, Breiner, and Quiring [2014] conducted qualitative surveys with users who comment on online news articles. They used their gained insights to quantitatively analyze controversy of user comments posted on two major German news outlets (Spiegel.de and Bild.de) and found that controversy attracts much attention and fosters increased engagement of users. Akin to that, Mishne, Glance, et al. [2006] reported similar behavior for users commenting on various weblogs.

Controversy of articles on Wikipedia has been investigated by Rad and

[1]Link to website: `https://merriam-webster.com/dictionary/controversy` (accessed on 20.03.2021)

D. Barbosa [2012]. The authors used the underlying revision history to detect controversial articles by identifying mutual reverts or bipolarity in the revision history. This method has been further adapted by Dori-Hacohen and J. Allan [2013] to predict the controversy of arbitrary Web pages by mapping them to Wikipedia articles that were already labeled as controversial or non-controversial. Here, the authors used sentiment analysis as a baseline, for which they report a high recall. Further, Dori-Hacohen and J. Allan [2015] fully automated this approach by using meta data of articles, ultimately allowing it to extend to a larger scale and to improve the overall classification performance. In a similar context, Zielinski et al. [2018] computed sentiment of talk pages to predict the controversy of the respective Wikipedia articles.

In the context of search engines, controversy has been studied to analyze how aggregated results impact the opinions of individuals. For example, Dori-Hacohen, Yom-Tov, and J. Allan [2015] indicated the political power of search engines and addressed the problems of defining whether a topic is controversial or not. Yom-Tov, Dumais, and Q. Guo [2014] analyzed search engine data and showed that users are more likely to read websites promoting opinions that are similar to their own. Koutra, Bennett, and Horvitz [2015] focused on how tragic events interfere with the user search behavior on controversial topics such as gun control laws. Their key findings indicated that people usually tend to search for information they agree with. However, when tragic events are directly affecting them, they are more open to consider content outside their own opinion.

2.3.2 Controversy in Social Media

The majority of work related to controversy in social media focused on political debates. For example, Adamic and Glance [2005] analyzed the political blogosphere during the 2004 U.S. Election. They investigated interactions between liberal and conservative blogs and found that both conservative as well as liberal blogs are linking more frequently among themselves. Their results have been confirmed by Conover et al. [2011], conducting a similar analysis on Twitter. Another study based on Twitter data conducted by Guerra et al. [2013] investigated new ways to quantify polarization in social networks extracted from the platform. The authors discovered that non-controversial networks can also be divided into modular communities and, hence, the traditional modularity score is not a measure of antagonism between communities. Morales et al. [2015] examined the propagation of controversial topics through social networks and evaluated their model on Twitter conversations regarding the Venezuelan president Hugo Chávez. They concluded that powerful and well-known parties managed to draw all attention to them, ultimately weakening the effect of tweets from minorities. Similar to this, Gruzd and Roy [2014] evaluated the 2011 Canadian Federal Election. They found that communities on Twitter were forming around similar political interests. However, they also found evidence of cross-ideological links between the rivalling communities, helping to start wider debates. Jang and J. Allan [2018] proposed a method based on textual features, such as readability and POS tags, to summarize stances in controversial discussions found on Twitter.

Reddit is another online social media platform that has been studied with regards to controversy. For example, Hessel and Lee [2019] extracted basic linguistic features (e.g., the number of words), text readability, various word embeddings as well as other discussion features (e.g., structure in the discussion tree) from comments posted in six different Subreddits (i.e., separate communities on Reddit dedicated to a specific topic, such as politics or sports) to predict and analyze controversy. The authors found that structural features are most predictive for controversy and that they better generalize to other communities than domain specific conversational (i.e., word embeddings) features. Tan et al. [2016] studied the Subreddit *r/ChangeMyView*, a community specifically dedicated to the discussion of controversial topics. Among other features, authors considered text readability and found that, when users try to persuade others, comments in such discussions are often harder to read and more complex. Jasser et al. [2020] studied 17 million comments posted on Reddit through a network-based analysis and concluded that discussions including many controversial comments result in higher degrees of activity, similar to the works form Mishne, Glance, et al. [2006] in the context of weblogs and Ziegele, Breiner, and Quiring [2014] in the context of news article comments.

Similar to the study of Jasser et al. [2020] conducted on Reddit, social media websites have been used to investigate controversy based on the underlying networks that form between users on such platforms. The general assumption of such network-based approaches is that interactions about controversial topics result in networks with a clustered structure, where each cluster represents a community with a fixed opinion. Users within one community reinforce the opinions of each other. This so-called *echo chambers* behavior has been investigated in numerous studies [An, Quercia, and Crowcroft, 2014; Grevet, Terveen, and Gilbert, 2014; Flaxman, Goel, and Rao, 2016; Jacobson, Myung, and Johnson, 2016; Shin and Thorson, 2017]. Akoglu [2014] introduced a polarization metric calculated on signed bipartite opinion networks and labeled controversial nodes by exploiting network effects and political domain knowledge. Garimella et al. [2016] extended this idea to identify controversial topics apart from the political domain and introduced the random-walk controversy (RWC) score, a measure to quantify controversy in social networks. Further, Garimella et al. [2017] tried to minimize controversy in a network extracted from Twitter. They proposed an edge-recommendation algorithm that efficiently reduces the networks's RWC score.

2.3.3 Extension to Previous Research

The existing literature on controversy often focused on one specific topic, such as the discussion of political elections in social media or the knowledge transfer in online news articles. Further, these studies typically only considered texts written in English. In my thesis, I conduct a quantitative analysis on a multilingual dataset that comprises a plethora of different topics to study controversy. More precisely, my extension to previous studies is three-fold: First, I demonstrate how to analyze and predict controversy by considering a combination of numerous features introduced in separate works. These features include word usage patterns, writing styles, sentiment as well as various user involvement features. Second, while all the aforementioned studies investigated controversy based on English texts, I consider English, French, German, Italian, Portuguese and Spanish texts in my analysis and, thus, I can infer cultural differences regarding controversy on the Web. Third, I analyze a multitude of texts originating from 50 Subreddits that deal with varying topics and, thus, my work exceeds previous studies focusing on Reddit [Tan et al., 2016; Hessel and Lee, 2019] in terms of generalization.

2.4 Employee Satisfaction and Motivation

Employee satisfaction (also referred to as job satisfaction; i.e., how employees feel about their work [Janssen, 2001]) and its relation to employee motivation and engagement has been studied extensively since the second half of the 20th century [e.g., Harter, F. L. Schmidt, and Hayes, 2002; Lundberg, Gudmundson, and Andersson, 2009; Bahadori et al., 2015; Dabirian, Kietzmann, and Diba, 2017; Holmberg, Caro, and Sobis, 2018; Tichaawa and Idahosa, 2020]. In the following sections, I provide an overview of the various definitions of employee satisfaction and describe some works focusing on employee satisfaction conducted in an offline and pre-Internet context. Among those, I specifically discuss Herzberg's Two-Factor Theory, a well-known and well-analyzed theory to describe the drivers behind employee satisfaction and motivation, which serves as a framework for one of the publications presented in this thesis. Finally, I discuss more recent studies related to employee satisfaction conducted in an online context through utilizing data originating from the Web.

Researchers focusing on employee satisfaction had different definitions of it [e.g., Hoppock, 1935; Vroom, 1964; Blood, 1969; Locke, 1976; Schneider and Schmitt, 1976; Choo and Bowley, 2007]. An early definition by Hoppock [1935] described it as a combination of psychological, physiological and environmental conditions that causes individuals to be truthfully satisfied with their jobs. While the author mentioned the influence of external factors, he was convinced that employee satisfaction is something that mainly depends on internal factors of employees (i.e., the way an employee feels). Blood [1969] saw it in a similar fashion and defined employee satisfaction to be solely depending on the values one brings to work. Contrary to these assumptions, Vroom [1964] thought of it to be dependent on the roles of employees at the workplace, which is similar to the definition by Schneider and Schmitt [1976], who saw it to be completely depending on organizational conditions and not on predispositions of employees. Locke [1976] defined employee satisfaction to be a mixture of both the conditions at work as well as the characteristics of employees. Further, he stated that employee satisfaction is closely related to the engagement of employees. A more recent definition by Choo and Bowley [2007] described it to be the result of job performance (e.g., by achieving goals). Perhaps the best conception of employee satisfaction with regards to my thesis is the more general definition by Ellickson and Logsdon [2002] or Spector [1985] and Spector [1997], who stated that employees develop either positive (satisfaction) or negative (dissatisfaction) attitudes towards their work based on their perceptions. The more a work environment fulfills their needs, values and personal traits, the more likely it is for employees to be satisfied.

2.4.1 Herzberg's Two-Factor Theory

Another well-known theory to explain employee satisfaction, which is also related to the definitions of Ellickson and Logsdon [2002] and Spector [1985] and Spector [1997], is the Two-Factor Theory introduced by Frederick Herzberg in 1959. Herzberg created a two-dimensional model of factors that influence employee satisfaction based on his opinion that it cannot be measured on the same continuum [F. Herzberg, Mausner, and Snyderman, 1959; Stello, 2011]. Initially, two studies—one with 13 employees of different professions and one with 39 managers—were conducted to decide on the influencing factors, which was further extended by Herzberg with another survey conducted with 203 accountants [F. Herzberg, Mausner, and Snyderman, 1959; F. I. Herzberg, 1966]. The distilled hypothesis was that certain factors lead to satisfaction whereas the absence of others lead to dissatisfaction. The former set of factors were related to the need for growth and self-actualization and became known as *motivation factors*. Typical motivation factors include, for example, responsibility at work or promotion prospects. On the other hand, the factors that contribute to dissatisfaction were related to the need of avoiding unpleasantness and became known as *hygiene factors*. Examples for hygiene factors include the compensation for one's work or the communication between team members. Important to note is the difference between the two types of factors. According to Herzberg, motivation factors are intrinsic and work to increase employee satisfaction, whereas hygiene factors are extrinsic and reduce employee dissatisfaction. This means that as soon as hygiene factors deteriorate below a threshold that is acceptable to employees, their dissatisfaction rises. This is not true for the inverse and, thus, when hygiene factors are optimally fulfilled, they prevent dissatisfaction but do not foster any satisfaction. The latter can only be achieved when fulfilling motivation factors. Analogue to hygiene factors, the absence of motivation factors does not increase dissatisfaction [F. Herzberg, Mausner, and Snyderman, 1959]. Therefore, Herzberg specifically emphasized that the opposite of satisfaction is no satisfaction and that the opposite of dissatisfaction is no dissatisfaction [F. I. Herzberg, 1966; F. Herzberg et al., 1968].

The Two-Factor Theory is one of the most influential theories in the context of employee satisfaction [Alshmemri, Shahwan-Akl, and Maude, 2017]. As such, the theory has been studied extensively in existing research and put to the test in various industries, for example, in the fields of healthcare. Holmberg, Caro, and Sobis [2018] evaluated the reasons for personnel shortages in Swedish mental health care through Herzberg's Two-Factor Theory. The authors interviewed 25 nurses and found the lack of career advancements as a partial reason for the shortages. Bahadori et al. [2015] applied the theory to study the motivation of 600 workers in the military health organizations of Iran. They reported that salary was the most important hygiene factor and responsibility the most important motivation factor for these workers. Similarly, Alshmemri, Shahwan-Akl, and Maude [2016] collected data of 272 Saudi Arabian nurses to find which factors contribute to employee satisfaction. For that, authors used the the Two-Factor Theory as a theoretical framework. Their results revealed that nurses in Saudi Arabia were in general dissatisfied at the time of the study, with female nurses being more dissatisfied than male nurses, suggesting possible implications for the recruitment process of female nurses.

In the context of hospitality and tourism, the Two-Factor Theory was applied, for example, to investigate the motivation of seasonal workers. For that, Lundberg, Gudmundson, and Andersson [2009] questioned 613 seasonal workers and qualitatively analyzed the thereby collected data. The authors found support for the theory, but also uncovered discrepancies regarding the needs of such workers in some cases (e.g., the study revealed that migrant workers had higher needs for internal communication). Civre, Lovec, Fabjan, et al. [2013] studied the motivation of Slovenian employees working in tourism. They found that, as according to the theory, motivation factors need to be fulfilled to satisfy employees, whereas hygiene factors were not statistically significant to increase satisfaction. Balmer and Baum [1993] demonstrated the general applicability of the theory to study the hotel choices of guests in hospitality. More recently, Tichaawa and Idahosa [2020] conducted a similar analysis through the Two-Factor theory to investigate satisfaction levels of festival attendees in Cameroon. Based on survey data of 324 individuals, they also found that the theory holds true in this context.

Education and academia are other prominent sectors which saw numerous applications of the Two-Factor Theory. For example, Ghazi, Shahzada, and Khan [2013] interviewed 300 university teachers and found that such employees were satisfied with both hygiene and motivation factors. Contrary to their expectations, the motivation of teachers was depending more on the fulfillment of hygiene factors. Similarly, Khanna [2017] surveyed 478 university teachers from North India and found support for the theory. Notably, the authors found differences in the significance of hygiene and motivation factors based on the home town of academics. Chu and Kuo [2015] applied Herzberg's theory to evaluate the motivation and evolvement of Taiwanese elementary school teachers. Their results suggested that both hygiene and motivation factors have a positive and significant affect on employee engagement when tested separately, with the exceptions of monetary rewards (hygiene factor) and recognition (motivation factor). Interestingly, when the authors studied both factors together as independent variables, hygiene factors had no significant influence on motivation. The authors stated that this reflects the initial definition of the Two-Factor Theory and indicates that the theory still holds true in today's society. Another example in the context of education was conducted by DeShields Jr, Kara, and Kaynak [2005], who used the Two-Factor Theory to study the motivation and satisfaction of 160 undergraduate business students at a state university in Pennsylvania. Here, the authors utilized hygiene factors to capture performance of advising staff and motivation factors to capture performance of classes and faculties and again found support for Herzberg's Two-Factor Theory.

Herzberg's Two-Factor Theory has been applied in other industries as well, further evincing its popularity. For example, Alfayad and Arif [2017] investigated the influence of employee voice (i.e., employees communicate their views and thoughts to employers) on employee satisfaction by applying the Two-Factor Theory on feedback from 300 Jordanian non-managerial employees. The authors found that the acknowledgement of employee voice increases employee motivation and satisfaction and, thus, suggested that organizations need to support employees' expressions of ideas as it fosters organizational effectiveness. Hur [2018] investigated how hygiene factors and motivation factors differently affect employee satisfaction between the private and public sectors. The authors found no difference between the two sectors and, thus, support for the Two-Factor Theory. Kotni and Karumuri [2018] applied Herzberg's Two-Factor Theory on data from 150 salesmen of the retail sector and found that they were more satisfied with the fulfillment of hygiene factors as compared to motivation factors, suggesting discrepancies in Herzberg's theory (similar to the findings of Ghazi, Shahzada, and Khan [2013]).

2.4.2 Other Studies of Employee Satisfaction

Employee satisfaction is known to contribute to better employee engagement [Steinhaus and Perry, 1996; Weiss, 2002], is linked to the overall performance of businesses [Kumar and Pansari, 2015] and influences the lives of numerous individuals who go to work during the majority of their lives [Greenhaus and Beutell, 1985; Ernst Kossek and Ozeki, 1998]. Naturally, employee satisfaction has been of interest to researchers for a long period in time [e.g., J. L. Howard and Frink, 1996; Harter, F. L. Schmidt, and Hayes, 2002; M. G. Brown, 2006; DiMicco et al., 2008; Luo, Zhou, and Shon, 2016]. Two approaches emerged: (i) studying employee satisfaction as an independent variable to benefit organizations as well as (ii) investigating employee satisfaction as a dependent variable to benefit employees.

For example, in the case of the former, Harter, F. L. Schmidt, and Hayes [2002] analyzed 7 939 business units in 36 different companies to measure the impact of employee satisfaction on customer satisfaction, productivity, profit, employee turnover and accidents. There study revealed significant relations between them and implied that changes in management processes to increase employee satisfaction may also increase business-unit outcomes.

Similarly, M. G. Brown [2006] found a strong link between employee satisfaction and customer satisfaction as well as revenue of organizations. Chi and Gursoy [2009] investigated the direct and indirect (via the influence of employee satisfaction on customer satisfaction) relationship between employee satisfaction and financial performance of hospitality companies. The authors found no significant direct but significant indirect impact on financial performance. More recently, Luo, Zhou, and Shon [2016] analyzed multi-aspect employer reviews on the platform *Glassdoor*[2] and reported a positive correlation between overall employee satisfaction and business performance. Notably, authors also discovered a negative correlation for some review aspects, including *safety*, *communication* and *integrity*.

Contrary, J. L. Howard and Frink [1996] studied employee satisfaction as and dependent variable and how the restructuring of organizations can benefit employees. The Author indicated the usefulness of restructuring organizations to increase satisfaction and stated a significant impact of employee satisfaction on the overall quality of life. More recent studies take advantage of Web data to benefit employee satisfaction, as previous empirical research [Miles and Mangold, 2014; Dabirian, Kietzmann, and Diba, 2017; Green et al., 2019] already highlighted a great potential in online employer reviews to complement existing management measurement methods, such as annual employee surveys. For example, DiMicco et al. [2008] launched a dedicated online social network (similar to Facebook) for employees working at IBM. They found that the implementation of the network increases satisfaction due to the fact that employees get to know each other more easily, which made them more eager to work together with their colleagues. Saha et al. [2019] used data from *Linkedin*[3] to study role ambiguity (i.e., unclear responsibilities and degree of authority of employees) and its effects on employee wellbeing. Their proposed method can help to identify role ambiguity in organizations and demonstrates the potential of analyzing data from the Web and using gained insights to improve life at work. Dabirian, Kietzmann, and Diba [2017] extracted 38 000 reviews of the highest and lowest ranked employers on Glassdoor in order to identify what employees care about and made suggestions to employers on how to become a great place to work. Stamolampros et al. [2019] used

[2] Link to website: `https://glassdoor.com`

[3] Link to website: `https://linkedin.com`

Glassdoor to gather about 300 000 reviews of employers operating in tourism and hospitality to study influencing factors for employee satisfaction. They found that leadership and cultural values are predictors for high employee satisfaction. Similarly, Yeonjae Jung and Suh [2019] collected more than 30 000 Korean employer reviews from *Jobplanet*[4] and used Latent Dirichlet Allocation [Blei, A. Y. Ng, and Jordan, 2003] to infer factors leading to employee satisfaction. Symitsi et al. [2021] extracted about 350 000 reviews from Glassdoor to find the drivers behind employee satisfaction and found *culture and values* as well as *senior leadership* to matter the most.

Online employer reviews have also been studied in contexts other than employee satisfaction. For example, Marinescu et al. [2018] described a selection bias in online employer reviews sourced from Glassdoor. Authors showed that individuals with extreme opinions are more motivated to share their experiences as compared to individuals with moderate opinions. They also stated that to provide incentives for reviewing may be a solution to this problem. Chandra [2012] used Glassdoor reviews to uncover different perspectives of work-life-balance in eastern and western cultures. More recently, Green et al. [2019] analyzed reviews from Glassdoor and their influence on stock returns. Their results indicate that companies for which reviews become more positive over time significantly outperform companies for which reviews become more negative over time. In another work, Könsgen et al. [2018] studied how review discrepancy effects job seekers for which they relied on 25 827 reviews collected from the German version of *Kununu*[5]. The authors found that high levels of discrepancies lead to increased intentions to avoid submitting applications to respective employers.

2.4.3 Extension to Previous Research

I extend the previous research on employee satisfaction by analyzing a novel and unexplored dataset, comprising more than two million English and German reviews of employers operating in 43 different industries. Thus, I provide comparable insights across different countries and sectors, allowing

[4]Link to website: `https://jobplanet.co.kr/`
[5]Link to website: `https://kununu.de/`

to derive cultural and as well as industrial characteristics of employee satisfaction expressed in such reviews. More precisely, I present two analyses that investigate employee satisfaction in different approaches. First, I take up existing influential factors for employee satisfaction already explored in existing research, comprising the impact of employee benefits [Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016], the influence of employee positions [Dienhart and Gregoire, 1993; De Cremer, 2003; Cornelißen, 2009; De Cremer, Dijk, and Folmer, 2009] as well as the effect of employee satisfaction on the employment status [S. P. Brown and R. A. Peterson, 1993; Griffeth, Hom, and Gaertner, 2000; Hausknecht, Rodda, and M. J. Howard, 2009]. Contrary to these studies, which relied on survey data manually and separately collected by researchers in distinct countries and industries, I conjunctly study these aspects through a large collection of online employer reviews originating from different cultural and industrial backgrounds. Thereby, I analyze otherwise detachedly considered aspects of employee satisfaction on one large dataset, delivering new and more generalized insights into the influence of such factors. In the second analysis, I consider the same dataset through Herzberg's Two-Factor Theory and reveal how to interpret online employer reviews in terms of this well-established theory. In doing so, I uncover whether hygiene or motivation factors are more relevant to reviewers of different countries and industries. My results add fruitful input to the discussion of the theory as well. For example, the potential findings on the generalizability of the theory across different countries and industries may be of special interest to the research community, as the theory has been criticized for that in the past [Furnham, Forde, and Ferrari, 1999; Parsons and Broadbridge, 2006; Y. Li, 2018].

3 Case Study I: Historic Sentiment

3.1 Related Publication and Author Contributions

Article 1: [Koncar, Fuchs, et al., 2020] Koncar, P., Fuchs, A., Hobisch, E., Geiger, B. C., Scholger, M. and Helic, D. (2020). Text sentiment in the Age of Enlightenment: an analysis of Spectator periodicals. *Applied Network Science*

For this work, as a first author, I was responsible for cleaning and preprocessing the (publicly available) data, conducting the sentiment analysis and all experiments associated therewith, preparing the results as well as interpreting the results together with Alexandra Fuchs and Elisabeth Hobisch, both experts working in the fields of humanities and contributing the necessary background knowledge. My supervisor, Denis Helic, contributed to the selection of analysis methods and further assisted in the discussion of the presented findings. Note that all the experiments were conducted on data which is publicly available through a manually curated digital edition provided by the *Centre for Information Modeling* in Graz and which was handed over by Martina Scholger in the *TEI* format.

The main idea for this work was proposed by the initial team of the *DiSpecs* project (Distant Spectators: Distant Reading for Periodicals of the Enlightenment), comprising Alexandra Fuchs, Bernhard C. Geiger, Elisabeth Hobisch, Martina Scholger and myself, to answer one of four research questions related to the quantitative, computational and distant-reading analysis of the Spectator periodicals. This work was based on a previous and preliminary work written by Koncar and Helic [2019]. All authors were involved in the writing of this article.

3.2 Text Sentiment in the Age of Enlightenment: An Analysis of Spectator Periodicals

3.2.1 Abstract

Spectator periodicals contributed to spreading the ideas of the Age of Enlightenment, a turning point in human history and the foundation of our modern societies. In this work, we study the spirit and atmosphere captured in the spectator periodicals about important social issues from the 18th century by analyzing text sentiment of those periodicals. Specifically, based on a manually annotated corpus of over 3 700 issues published in five different languages and over a period of more than one hundred years, we conduct a three-fold sentiment analysis: First, we analyze the development of sentiment over time as well as the influence of topics and narrative forms on sentiment. Second, we construct sentiment networks to assess the polarity of perceptions between different entities, including periodicals, places and people. Third, we construct and analyze sentiment word networks to determine topological differences between words with positive and negative polarity allowing us to make conclusions on how sentiment was expressed in spectator periodicals.

Our results depict a mildly positive tone in spectator periodicals underlining the positive attitude towards important topics of the Age of Enlightenment, but also signaling stylistic devices to disguise critique in order to avoid censorship. We also observe strong regional variation in sentiment, indicating cultural and historic differences between countries. For example, while Italy perceived other European countries as positive role models, French periodicals were frequently more critical towards other European countries. Finally, our topological analysis depicts a weak overrepresentation of positive sentiment words corroborating our findings about a general mildly positive tone in spectator periodicals.

We believe that our work based on the combination of the sentiment analysis of spectator periodicals and the extensive knowledge available from literary studies sheds interesting new light on these publications. Furthermore, we

demonstrate the inclusion of sentiment analysis as another useful method in the digital humanist's distant reading toolbox.

3.2.2 Introduction

During the Age of Enlightenment (starting in the 18^{th} century), so-called *spectator periodicals* were a popular way of distributing information to a non-academic audience and providing a platform for debating a plethora of topics, such as politics, religion and literature. Originating from London, the idea of these periodicals quickly spread all across Europe through translations and imitations adapting the characteristic narrative system and micro-narrations to local context in respective countries and languages [Fuchs, Ertler, and Holzer, 2019]. As some periodicals questioned customs and traditions or even included public criticism, these periodicals were emotionally charged and contained opinions with polarizing sentiment.

In this paper, we leverage the sentiment conveyed in spectator periodicals to learn more about the characteristics of the Age of Enlightenment, including, for example, important topics for the people living back then, cultural idiosyncrasies as well as the textual peculiarities regarding sentiment found in the periodicals. For that, we conduct a three-fold sentiment analysis on a large dataset comprising spectator periodicals written in five major European languages to shed light on: (i) the factors influencing sentiment, such as temporal aspects or topics discussed, (ii) the relations between periodicals, places as well as people, and (iii) the usage of sentiment conveying words in the literature of the 18^{th} century.

While text sentiment on the Web (especially in social media) and in literature has been studied extensively in recent years, respectively in computer sciences [B. Liu and L. Zhang, 2012] and digital humanities [Sprugnoli et al., 2016; Moreno-Ortiz, 2017; T. Schmidt and Burghardt, 2018], we still lack a broader understanding of sentiment in texts originating from earlier, pre-digital times and conventional media. Hence, our work not only contributes to the comprehension of the Age of Enlightenment, but also demonstrates a method to analyze the literature of the 18^{th} century through automatic text processing.

Approach. Building upon our previous work [Koncar and Helic, 2019], we analyze a publicly available and manually annotated dataset of more than 3 700 issues of spectator periodicals published in French, German, Italian, Portuguese and Spanish between 1711 and 1822. Each issue deals with one or more topics (e.g., politics or marriage) and follows the genre-specific communicative structure, comprising multiple entangled narrative levels and several narrative forms (e.g., letters to the editors, dream sequences, allegories, and dialogues).

To study text sentiment, we first utilize existing sentiment dictionaries to compute the sentiment of individual issues, allowing us to investigate the development of sentiment over time, the impact of narrative forms on sentiment and how different topics had been perceived back in the Age of Enlightenment.

Second, we construct and analyze sentiment networks, in which nodes represent entities inferred from the manual annotation of the periodicals (such as the name of the periodical or its author, place names or person names) and edges signal the sentiment between these entities. These sentiment networks allow us to study, for example, the polarity relation between countries of Europe in the 18^{th} century.

Third, we construct sentiment word networks in which nodes represent sentiment conveying words and edges represent semantic relation between these words. We then use these networks to explore how positive and negative words are distributed over the texts and whether they appear in chunks that signal strongly polarizing views on given topics or whether they diffuse more evenly across larger portions of text. Additionally, we investigate the most important words expressing sentiment, allowing us to infer if periodicals focused more on positive or negative sides of discussed topics. For that, we compute basic network metrics, including (i) degree distributions and (ii) clustering coefficients to assess the distributions of words, (iii) assortativity to measure biases in semantic relations between words of similar sentiment, and (iv) centralities to find most central and important words in networks.

Findings. We observe differences in the mean sentiment conveyed in spectator periodicals between the five languages. For example, while Spanish periodicals have a consistently more negative sentiment, Italian and French periodicals have a more positive sentiment, indicating cultural dependencies on how to address important issues. The influence of narrative forms on sentiment is inconclusive, as results vary significantly across languages. Notable commonalities are positive *selfportraits* and rather negative *utopias* for all five languages. Our results on topics suggest similar dependence on language and places.

However, using the insights from our close reading experience reveals some noteworthy observations. For example, the criticism of religion in Italian and Spanish periodicals was disguised by authors through rhetorical and stylistic devices in order to avoid censorship. Further, our sentiment networks depict how so-called "untouchables", such as *Dante Aligheri* and *Francesco Petrarca*, were used to resolve negative examples in periodicals. We find distinguishable characteristics between positive and negative word usage through our sentiment word networks. For example, we observe low transitivity and a tendency towards low degrees for negative words as well as higher transitivity and a tendency for mid-range degrees for positive words. Combining these observations with the results from our centrality analysis, which depict a weak overrepresentation of positive words among top central words, we find that spectator periodicals had, in general, a mild to positive attitude towards topics of the Age of Enlightenment and used the majority of negative words distinctively to discuss critical issues.

Contributions. To the best of our knowledge, our work is the first to investigate text sentiment in spectator periodicals published during the Age of Enlightenment in such a broad scope. By combining distant reading results with insights acquired during close reading, we extend the knowledge of spectator periodicals and of this decisive period in human history. Further, we publish the code of our analysis[1], opening up possibilities to learn more about the characteristics of texts originating from the 18th century and to compare them to today's texts and media regarding, for example, the attitude towards important societal topics then and now.

[1]Code available at `https://github.com/philkon/sentiment-spectator-extended`

3.2.3 Related Work

The Spectator Press. The foundation for the journalistic genre of spectator periodicals was laid by *The Tatler* (1709 – 1711), *The Spectator* (1711 – 1714) and *The Guardian* (1713) which were published by *Richard Steele* and *Joseph Addison* in England and combined previously separate fields of journalism, such as political, social and scholarly information, into one genre. Therefore the spectator press was the first print-medium able to reach a broader public with a secular discourse [Melton, 2001].

Issues of *The Spectator* enjoyed great popularity (circulation of around 1 600 exemplars [King, 2018]), were translated and imitated quickly and numerously all across Europe and, thus, created a new genre of periodicals and journalism [M. L. Pallares-Burke, 2007; Krefting, Nøding, and Ringvej, 2015]. The first imitation by Justus van Effen, *Le Misantrope* (1711 – 1712), was soon followed by (partial) translations to French and German in 1714 [Gilot and Sgard, 1981; Martens, 2017], followed by Danish, Dutch, Italian, Portuguese, Spanish and Swedish [Gustafson, 1932; M. L. Pallares-Burke, 2007; Krefting, 2018]. Spectator periodicals significantly influenced the society regarding, for example, the image of women [Messbarger, 1999; Carr, 2014] or religious beliefs [D. Allan and Virtue, 1993].

In this paper, we work with a manually annotated dataset including French, German, Italian, Portuguese and Spanish periodicals published between 1711 and 1822.

Sentiment Analysis. A common task in Natural Language Processing (NLP) is sentiment analysis, aiming to investigate emotions, attitudes and opinions expressed in textual data [Pang and Lee, 2008]. Basic methods rely on dictionaries comprising lists of words for which the sentiment or polarity (either positive or negative) is known. Popular dictionary-based methods include SentiStrength [Thelwall et al., 2010] or VADER [Hutto and Gilbert, 2014], both specifically introduced for short texts originating from social media platforms, such as Twitter or Facebook. Another widely used dictionary-based method is LIWC [Pennebaker, Francis, and Booth, 2001], aiming to identify characteristics of authors by automatically analyzing their texts. For that, authors introduced additional dedicated dictionaries capturing, for example, social relations, honesty and thinking style. Further, machine learning approaches for sentiment classification have been studied [Pang, Lee, and Vaithyanathan, 2002; Ye, Z. Zhang, and Law, 2009; Maas et al., 2011], including neural networks [L. Zhang, S. Wang, and B. Liu, 2018]. As most of studies in the field of sentiment analysis (or, in general, all NLP areas) focus on the English language, we encounter a significant scarcity of dictionaries for non-English languages. Addressing this issue opened a whole new research area known as *cross-lingual sentiment classification* [Q. Chen, C. Li, and W. Li, 2017], trying to transfer existing English models to other languages. Commonly, this transfer happens by using machine translation techniques to project English models to the target language or vice versa [Prettenhofer and Stein, 2010; Xiao and Y. Guo, 2013; Yanqing Chen and Skiena, 2014].

In recent years, there has been an increased interest in applying and exploring sentiment analysis in digital humanities projects, especially in literary studies. Jannidis et al. [Jannidis et al., 2016] utilize sentiment analysis for predicting happy endings in German novels from the 19^{th} century using a lexicon-based approach. Schmidt et al. [T. Schmidt, Burghardt, and Wolff, 2018] investigate 18^{th} century plays of Gotthold Ephraim Lessing evaluating their results against a manually annotated text corpus. Henny-Krahmer [Henny-Krahmer, 2018] explored the relation between sentiments and subgenres in a corpus of 19^{th} century Spanish American novels, differentiating direct speech and narrated text.

In our work, we rely on a basic dictionary approach introduced by Chen and Skiena [Yanqing Chen and Skiena, 2014] as it allows for easy interpretation of results and, as opposed to other works, is available for the five languages contained in our dataset.

Networks to Represent Texts. Network representations of texts have been studied extensively [Antiqueira et al., 2009; Diego R Amancio, Oliveira Jr, and Costa, 2012; Silva and Diego R Amancio, 2012; Cong and H. Liu, 2014; Diego Raphael Amancio, 2015; Kulig et al., 2015]. Depending on the application, there are multiple ways of how to model texts as networks. In cases where semantics are important, models connect words with semantic relations or words that co-occur in the same context, for example, a sentence or paragraph [Widdows and Dorow, 2002; Véronis, 2004; Diego R Amancio, Oliveira Jr, and Costa, 2012]. If structure or style is important, words are connected based on syntactical relations, with word adjacency networks [Roxas and Tapang, 2010; Diego Raphael Amancio et al., 2011] being a well-known approach. Basically, this model links adjacent words in texts with each other, captures stylistic characteristics of texts and is language independent.

Network representations of texts have also been used for topic modeling. For example, Gerlach et al. [Gerlach, Peixoto, and Altmann, 2018] introduced an approach based on bipartite networks of documents and words and used existing community detection methods to identify topics. Zuo et al. [Zuo, J. Zhao, and K. Xu, 2016] introduced WNTM, a topic model based on co-occurrence networks and specifically designed for sparse and short texts. Further, Liu et al. [Z. Liu et al., 2010] built the Topical PageRank for co-occurrence networks to extract keyphrases that summarize documents.

In this paper, we combine the network representations of spectator periodicals with sentiment analyses through two different approaches: (i) we create sentiment networks to study the relation between different entities and (ii) we construct sentiment word networks to analyze word usage patterns. While most of the presented studies focused on the English language, we investigate metrics of networks based on French, German, Italian, Portuguese and Spanish.

3.2.4 Dataset, Preprocessing and Sentiment Dictionaries

We now describe the dataset we use for our analysis as well as the necessary preprocessing steps to prepare data. Further, we explain how we leverage existing sentiment dictionaries to assess the sentiment conveyed in spectator periodicals.

Dataset. We conduct our analysis on a collection of spectator periodicals that is manually annotated and curated by experts working in the fields of humanities. For that, we leverage *The Spectators in the international context*, a digital scholarly edition project which aims on building a central repository[2] for spectator periodicals [Scholger, 2018; Ertler et al., 2011-2020]. The annotated periodicals follow the XML-based Text Encoding Initiative (TEI) standard [Consortium, 2020], which provides a vocabulary on how to represent texts in digital form, and are publicly available through the digital edition[3].

Overall, our dataset contains 3 718 issues of 67 distinct periodicals written in French (1 672 issues in 25 periodicals), German (35 issues in 4 periodicals), Italian (1 320 issues in 14 periodicals), Portuguese (44 issues in 1 periodicals) and Spanish (647 issues in 23 periodicals). Periodicals were published during different time periods allowing us to analyse a time span of 111 years. Each issue follows the same literary style and includes annotations of the *levels of representation* and *narrative forms*. The *levels of representation* reflect the fact that spectator periodical texts are organized according to the model of interlaced forms of discourse [Ertler, 2004], which reminds of the framed construction of an Italian novel or—metaphorically—of the principle of Russian dolls. *Narrative forms*, such as dream sequences or metapoetical frames, were used by authors to create a more complex atmosphere of entertainment, allowing them to address the messages of virtue in a playful context and through multiple points of view.

Additionally, our dataset includes the following information for each issue: the author, the date and country of publication, one or multiple manually

Table 3.1: **Dataset Statistics.** This table lists the number of periodicals, issues, known authors, anonymous publications, topics, text passages and the time spans for which our dataset contains publications, respectively for each of the five languages.

	French	**German**	**Italian**	**Portuguese**	**Spanish**
# Periodicals	25	4	14	1	23
# Issues	1 672	35	1 320	44	647
# Known Authors	15	3	11	1	20
# Anon. Publications	456	27	1	0	108
# Topics	37	27	34	27	32
# Text Passages	8 855	151	7 040	236	4 008
Time Span (Years)	1711 - 1795	1723 - 1765	1716 - 1822	1752 - 1754	1735 - 1804

[2]The central repository is hosted, preserved and presented as part of an OAIS-compliant, certified institutional repository for Humanities' research data, the *Humanities' Asset Management System (GAMS)*; Link: `https://gams.uni-graz.at`.

[3]`https://gams.uni-graz.at/spectators`

annotated topics (out of a list comprising 38 distinct topics[4]), as well as mentioned people (including real or fictional), places (i.e., countries, cities or other geographic features) and works (e.g., other periodicals, novels or plays). The list of topics is based on existing literature focusing on the spectator periodicals [Rosa, 1966; Rau, 1980; Boulard, 2000; Ertler, 2003; Lévrier, 2007] and was further adapted during the close reading process. In Table 3.1, we list a detailed comparison between languages and summarize statistics of the dataset. Note that due to the smaller number of German and Portuguese periodicals, we refrain from making any conclusive statements about those two languages.

Preprocessing. We start our analysis by parsing and processing the TEI encoded XML files. For each issue included in our dataset, we extract and aggregate text according to levels of representation and narrative forms into *text passages* (cf. Figure 3.1 for an illustration). We further extract authorship information, dates of publication, manually assigned topics and mentioned people, places and works. Next, we normalize author names (e.g., *Eliza Haywood* and *Eliza Fowler Haywood* $\Rightarrow$ *Eliza Fowler Haywood*). If the publication date of individual issues is not known, we set it to the date of the first known publication of the corresponding periodical. E.g., if a periodical was published between $1711 - 1712$, then we assume that every issue was published in 1711.

Sentiment Dictionaries. We determine words expressing a positive or negative sentiment by using dictionaries that have already been evaluated for French, German, Italian, Portuguese and Spanish in a previous work [Yanqing Chen and Skiena, 2014]. There, authors extracted most frequent words of Wikipedia articles and created a knowledge graph to combine similar words of different languages through using Wiktionary, machine translation (via Google translate), transliteration links and WordNet. Starting with sentiments of English vertices based on a dictionary with 1 422 (32%)

[4]**List of topics in alphabetical order:** America (West India), Apologetic of Spain, Austria, Autopoetical Reflection, Charity, Critics on Nobility, Culture of Conversation, Economy, Education and Formation, England, Family, Fashion, Foreign Societies, France, Friendship, Germany, Happiness, Idea of Man, Image of Men, Image of Women, Italy, Law, Love, Manners and Customs, Morale, Nature, Other Countries, Passion, Philosophy, Politics, Reason, Religion, Science, Spain, Structure of Society, Superstition, Switzerland, Theatre Literature Arts.

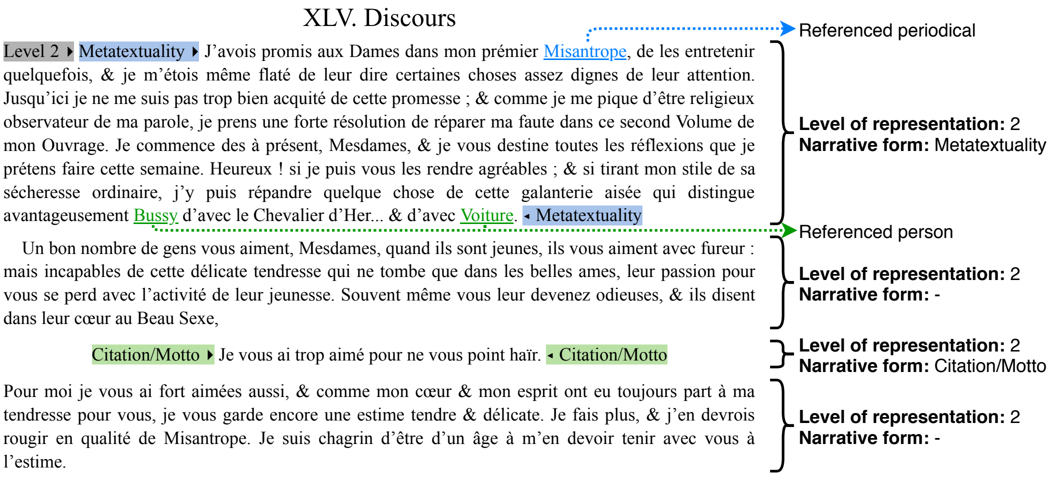


Figure 3.1: **Preprocessing Illustration.** This figure depicts an excerpt from *Le Misantrope* (Vol.2\004 (1711 – 1712) by Justus van Effen; as seen in the digital edition) and illustrates how we aggregate text during preprocessing. In particular, we aggregate all text with the same level of representation and narrative form (indicated by colored labels) into text passages, respectively for each issue. In this example, we arrive at three text passages: One with level of representation 2 and narrative form *Metatextuality* (i.e., the first paragraph), one with level of representation 2 and narrative form *Citation/Motto* (i.e., the third paragraph), and one with level of representation 2 and no distinct narrative form (i.e., the second and fourth paragraph). Further, we extract mentioned periodicals (indicated with blue color), people (real or fictional; indicated with green color) and places (none in this excerpt) for our analysis.

positive and 2 956 (68%) negative words, authors propagated sentiments to vertices of other languages and created dictionaries for 136 languages, each including a list of words with both a negative and a positive sentiment. For our work, we used the resulting French (4 653 words; 35% positive; 65% negative), German (3 974 words; 38% positive; 62% negative), Italian (4 491 words; 36% positive; 64% negative), Portuguese (3 953 words; 35% positive; 65% negative) and Spanish (4 275 words; 36% positive; 64% negative) dictionaries.

3.2.5 Text Sentiment in the Spectator Press

In the first part of our analysis, we investigate the sentiment expressed in the texts of periodicals. Specifically, we ask the following questions:

(i) **Sentiment Over Time:** How did sentiment develop over time? Did historic events influence the emotions of authors and, if so, to what extent?
(ii) **Sentiment of Narrative Forms:** Did sentiment depend on the various narrative forms? For example, was a selfportrait more positive than a dream sequence?
(iii) **Sentiment of Topics:** How have the different topics been perceived in the 18$^{\text{th}}$ century? Are important social issues for the Age of Enlightenment, such as religion, more emotionally discussed than other matters?

We conduct this analysis for each of the five languages contained in our dataset and compare individual results.

Computing Sentiment. Using the respective sentiment dictionaries described earlier, we compute the sentiment score s of each text passage in our dataset with $s = (W_p - W_n)/(W_p + W_n)$, where W_p is the number of positive words in a text passage and W_n is the number of negative words in a text passage. Hence, the sentiment score is a value ranging between -1 and $+1$, where values close to -1 are considered as negative, values close to $+1$ as positive, and where values close to zero indicate a neutral sentiment.

To assess the applicability of sentiment dictionaries, we compute the coverage of sentiment words (i.e., the fraction of words in the dictionaries that are actually contained in issues), respectively for each language. For French, Italian and Spanish, we report a coverage ranging between 71% and 76%, suggesting that the dictionaries extracted on modern texts are also suitable for languages of the 18$^{\text{th}}$ century. In case of German and Portuguese, the coverage is lower with 36% and 25% respectively. However, we argue that this observation is due to the limited number of issues for these languages. Specifically, we cover 72% of negative and 79% of positive words for French, 33% of negative and 41% of positive words for German, 74% of negative and 81% of positive words for Italian, 23% of negative and 31% of positive words for Portuguese as well as 69% of negative and 75% of positive words for Spanish.

Sentiment Over Time. We first investigate the development of sentiment over time, following the intuition that temporal proximity to certain events, such as political unrest, impacts the emotions of authors. We report mean sentiment of text passages per year in which issues were published in the respective languages in Figure 3.2. Overall, sentiment varies over time for all of the five languages. For German, Portuguese and Spanish, the mean sentiment is slightly negative, indicating that German, Portuguese and Spanish periodicals express more negative emotions. The overall mean sentiment is -0.18 for German, -0.19 for Portuguese and -0.13 for Spanish. In contrast, for French and Italian periodicals, mean sentiment is slightly positive throughout the years with an overall mean sentiment of 0.03 and 0.19 respectively.

In contrast to our initial hypothesis, we observed that no general statements can be given regarding the evolution of sentiment over time. Rather, sentiment is tightly connected to the individual periodicals, each of which has typically been published over short periods of time. We now discuss the particularities of each language in the following paragraphs.

For French periodicals (cf. Figure 3.2a), we observe three peaks where sentiment is, on average, more positive compared to the remaining years. The first one of these peaks in 1728 is related to the publication of *La Spectatrice*, the first French spectator periodical presumably written by a woman. Therefore, the more positive sentiment in this year may be caused by the discursively created “female voice”, not daring to express criticism as directly as her male colleagues. The second peak in 1753 is related to the publication of *Le Spectateur moderne*. Of this periodical, only one issue has been preserved and, hence, the peak for 1753 might not be representative. The positive peak in 1786 is caused by the publication of the first issue of *Les Chiffons*, an entertaining and satiric social critic. For the former two periodicals, it was the first issue of the respective periodical and we hypothesize that the authors may have used a milder tone to attract new readership.

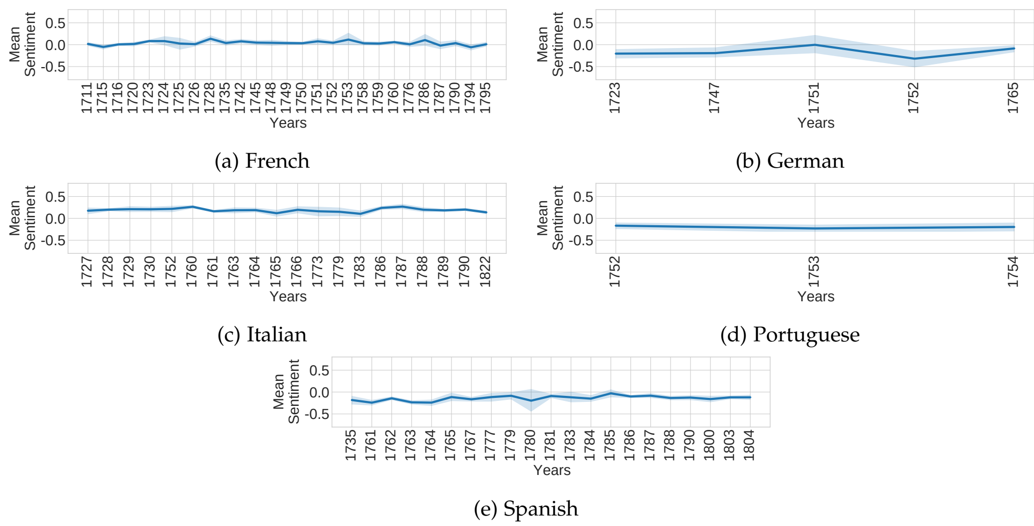


Figure 3.2: **Mean Sentiment Over Years.** This figure illustrates the mean sentiment of text passages (with 95% confidence intervals) over the years in which issues were published in respective languages. The mean sentiment varies over languages with French and Italian periodicals being mostly positive, whereas German, Portuguese and Spanish periodicals have mostly negative sentiment.

Due to the very small quantity of text material, the sentiment over time for German periodicals (cf. Figure 3.2b) is strongly depending on the individual periodicals. Each year represents only one periodical or even only one or two issues of one periodical. In this case, we note the negative sentiment in the year 1752, in which the third issue of *Die Zuschauerin* was published. This periodical focused on the role of women in family and society and presumably used more negative portraits and examples in the issue published in 1752.

In case of Italian periodicals (cf. Figure 3.2c), we observe a particularly positive sentiment in the years 1786, 1787 and 1788 in which the periodical *Donna galante ed erudita* was published. This periodical mainly addressed the positive aspects of fashion, healthcare, theatre, and women. On the other hand, we speculate that the slightly more negative sentiment in the year 1765 is due to the *Il Caffè* authors' critical attitude towards politics and economy.

The Portugal of the 18th century faced economic, social and political problems due to the reign of terror, fear as well as political and religious persecution. Further, censorship was widespread. The Portuguese periodical *O Anonymo* was published from 1752 to 1754, describing this situation with a lot of criticism, suggesting the negative mean sentiment of Portuguese periodicals (cf. Figure 3.2d). The hopeless situation began to change with statesman *Marquis of Pombal*, who introduced many reforms based on the ideas of the Enlightenment during the second half of the 18^{th} century [Dill, 2015].

For the Spanish periodicals (cf. Figure 3.2e), we report a negative sentiment for 1735, the first year in which a Spanish periodical was published. We argue that this is due to *El Duende crítico*, a critical periodical similar to a pamphlet [López, 2002] and published in the course of this year. The temporary and slightly more positive sentiment for the year 1762 may be caused by the publication of *El Pensador*, a moderate critic of the Spanish society and protégé of the Spanish King. Similarly, the slight increase in sentiment in 1765 seems to be related to the first publication of *El Belianís literario* a periodical that focused on literary critics. However, as this periodical is a satiric critic, making use of exaggerated praise or criticism, our sentiment analysis might not correctly infer the expressed sentiment and results may be inaccurate. We further address this issue in the limitations of our work. Another notable year for Spanish periodicals is 1785, for which the mean sentiment is most positive. This could be due to *El Censor*, the most famous Spanish spectator and one of the most durable publications with extraordinary literary quality [Guinard, 1973]. It was one of the most critical spectator periodicals and therefore had problems with censorship. During its life in press ($1781 - 1787$), *El Censor* was prohibited twice through censorship, which led to an interruption of the publication each time. After the second break, it was again published in 1785 [Guinard, 1973]. As *El Censor* was the only Spanish spectator published in this year, we argue that the positive sentiment may result from strategies of the authors to avoid further problems with censorship by means of a lighter tone or better concealing of criticism, such as satire or indirect criticism. One could also argue that the negative sentiment in Spanish periodicals stemmed from an increased criticism of Spanish authors due to the slow progress of the Enlightenment in Spain. However, we argue that this is a common prejudice [Astigarraga, 2015] and perhaps the more negative tone of the Spanish spectator press may be related to an inner cultural conflict between conservatives and progressives in Spain during that time [Von Tschilschke, 2009]. In this conflict, the spectator periodicals were a medium chosen mainly by the progressives in order to spread their reformist ideas.

Sentiment of Narrative Forms. As spectator periodicals deliberately avoided news, they had only limited subjects to address. In order to add variety to their texts, they embedded the same moral messages into different narrative forms, allowing authors to address the same subjects from different points of view. For example, to illustrate the importance of female virtues, the authors can use a *General Account* on a vicious women punished for her misconduct just as well as an *Allegory* representing female virtues. In other words, the message would be the same, but the sentiment could vary. In consequence, the different narrative forms are likely to vary in sentiment. To determine whether there are any trends in polarity for the specific narrative forms, we investigated them separately in our analysis.

To assess if the sentiment of a narrative form is more positive or more negative relative to the language mean (cf. Figure 3.2), we standardize sentiment scores: From each sentiment value we subtract the language mean and divide by the standard deviation of the respective language.

In Figure 3.3 we illustrate the mean standardized sentiment of each narrative form for each of the five languages contained in our dataset. Note that not every text passage is annotated with a narrative form. Thus, it is easily possible that all narrative forms have a sentiment that is more positive or more negative than the language mean. We observe such a case for German, for which all narrative forms have negative relative sentiment. The apparent discrepancy is resolved by the fact that text passages without annotation have positive relative sentiment.

Overall, results vary across languages. In case of French, Italian and Portuguese periodicals, the majority of narrative forms convey a more positive sentiment as compared to the language mean, while the majority of narrative forms in Spanish periodicals expressed a more negative sentiment.

Focusing on the differences across languages, we note that *Utopia* is rather negative compared to the language mean, for both French (mean standardized sentiment equals -0.22) and Spanish periodicals (mean standardized

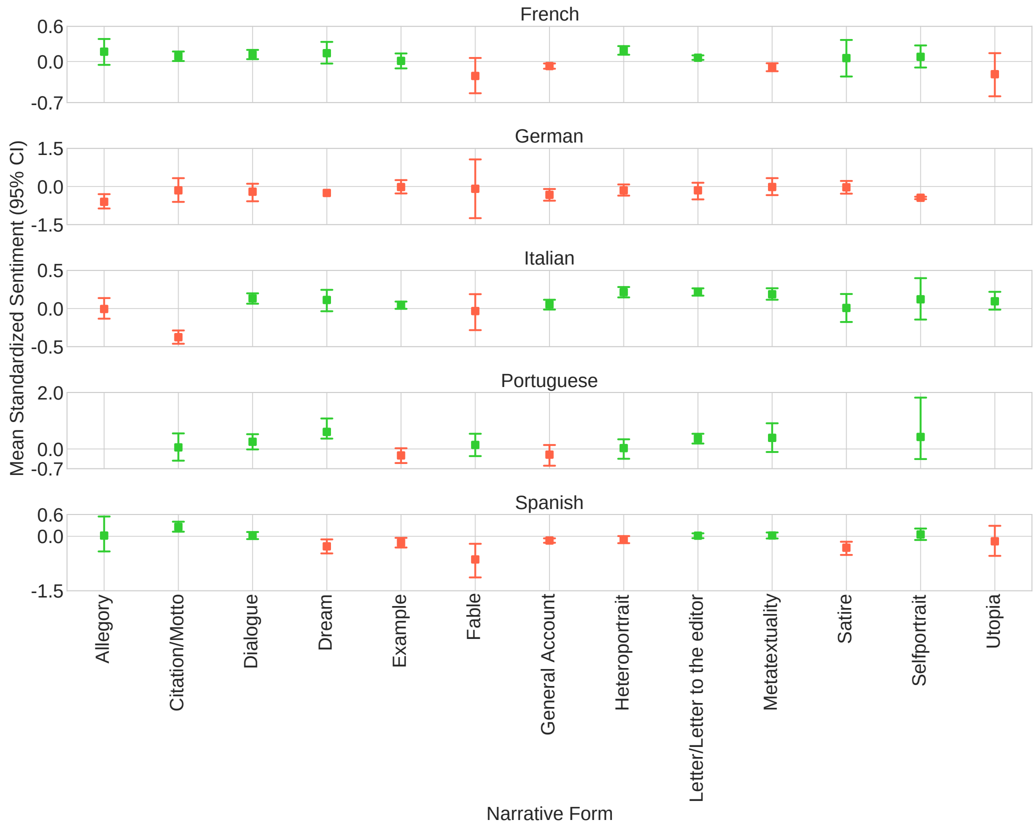


Figure 3.3: **Mean Standardized Sentiment For Narrative Forms.** This figure illustrates the mean standardized sentiment (with 95% bootstrap confidence intervals) for each narrative form, respectively for each language contained in our dataset. Red squares indicate a more negative sentiment and green squares indicate a more positive sentiment compared to the language mean. Note that not all text passages follow a specific narrative form and that German and Portuguese do not include all narrative forms. We observe significant differences across languages.

sentiment equals -0.14), whereas it is positive for Italian periodicals (mean standardized sentiment equals 0.10). Investigating on that, we conclude that this observation might have multiple reasons. On the one hand, *Utopia* was not only used to annotate the strict literary definition, which describes nearly perfect and high quality societies (i.e., we would expect a positive sentiment for this narrative form). Instead it was also used to annotate fanciful narrations. Such narrative forms are frequent in Spanish periodicals, including *El Duende de Madrid* comprising a complete fanciful narration in which authors of published discourses were represented as goblins in a darker setting, potentially explaining the negative mean standardized sentiment for *Utopia* in Spanish periodicals. On the other hand, for French periodicals the *Utopia* annotation is more accurate to the literary definition, but French periodicals also include dystopic narrations, explaining the negative sentiment for French. The authors of Spanish periodicals seem to have used *dream sequences* for such dystopic narrations, reflecting the rich literary heritage of Spanish satiric dream narrations of the baroque period [Gómez Trueba, 1999] and explaining the more negative sentiment of *Dream Sequences* for Spanish. Regarding the more positive sentiment for *Utopia* in Italian periodicals, we conclude that these utopias mainly comprise dialogue series between Odysseus and various animals in the *Osservatore veneto* [Fuchs and Ertler, 2014]. Here, similar to utopias of French and Spanish, negative examples are shown, however, due to the transformation of the moral instruction to Circe's Island[5], it is integrated into a harmonious scene. The moral instruction then takes place based on a striking, pleasing and nonviolent language. Some of the utopias integrated in other Italian periodicals include positive visions of the future. Overall, it turns out that utopias in the Italian periodicals are not used to scare, but to instruct in a pleasant way. Thus, the positive sentiment captures the overall positive atmosphere which is apparently a characteristic of the Italian periodicals.

The narrative form *Letters* or *Letters to the Editor* have a more positive sentiment across languages, with the exception of German periodicals. Here the mean differences are 0.06 for French, 0.22 for Italian, 0.36 for Portuguese and 0.02 for Spanish. These results are surprising to us since readers, though generally polite, also expressed their honest opinions and did not always agree with the spectators in their letters. In fact, the positive sentiment for this narrative form might be an artifact from the sentiment dictionaries we used. In the 18th century, correspondence was still a very formal act, often related to rhetoric conventions [Vellusig, 2000] and the dictionaries might not be able to assess correct sentiment due to their creation on modern texts.

[5]"Circe" refers to the goddess of magic in Greek mythology.

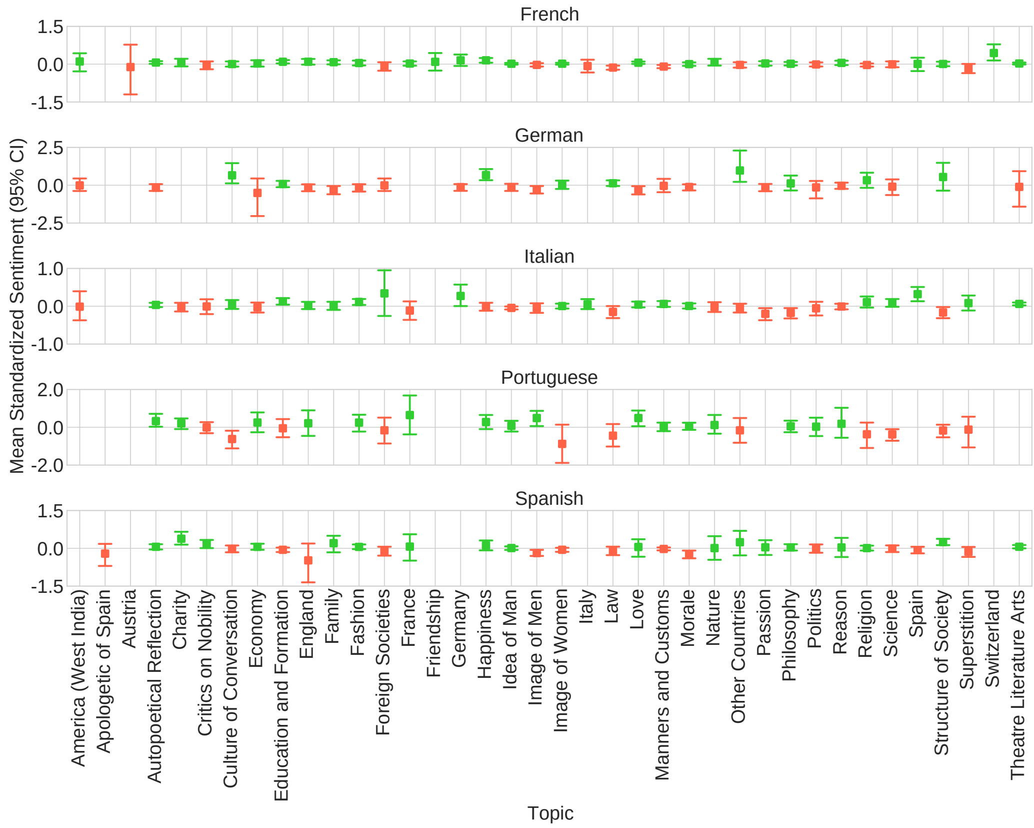


Figure 3.4: **Mean Standardized Sentiment For Topics.** This figure illustrates the mean standardized sentiment (with 95% bootstrap confidence intervals) for each topic, respectively for each language contained in our dataset. Red squares indicate a more negative sentiment and green squares indicate a more positive sentiment compared to the language mean. Note that not all languages include each topic. Overall, we observe significant differences across languages.

We also note that the *Selfportrait* is more positive than *Heteroportrait* for Spanish periodicals, suggesting that authors of the periodicals as well as their correspondents tend to present themselves in a positive way, whereas they use others as examples for bad behavior. This might be related to a general human tendency of blaming others and not oneself for your own mistakes [Weiner, 1995; Shaver, 2012].

Sentiment of Topics. We now investigate how different topics had been perceived in periodicals. For that, we standardized sentiment in a similar fashion as in the previous study on narrative forms in order to overcome the general sentiment biases of languages (cf. Figure 3.2). This allows us to infer whether or not the sentiment of topics is more positive or negative compared to the language mean.

In Figure 3.4 we depict the mean standardized sentiment of each topic contained in our dataset for each of the five languages. Again, sentiment varies significantly per topic over all five languages.

In the case of French periodicals, *Germany* and *Switzerland* are among the most positive topics, while, for example, *Superstition* and *Law* had been perceived more negatively compared to the French mean. For German, we find that *Other Countries*, followed by *Happiness* and *Structure of Society* are among the more positive topics, whereas topics related to interpersonal relationships including *Love*, *Family* as well as *Image of Men* are more negative compared to the language mean. For these two languages, we observe that French periodicals wrote more positively about Germany but German periodicals wrote more negatively about France. This may be due to the fact that French culture was dominant in arts and literature during the 17^{th} and 18^{th} century which, hence, was discussed polemically by other countries in Europe. Regarding Italian periodicals, we report more positive sentiment for *Spain* and *Foreign Societies* [Fuchs and Ertler, 2018]. The topics *Passion* and *Philosophy* comprise a more negative text for Italian periodicals when compared to their language mean. Among positive topics for Portuguese periodicals we find *France*, *Image of Men* and *Love*, while negative topics include *Image of Women*, *Law* and *Culture of Conversation*. Finally, *Charity*, *Structure of Society* and *Other Countries* are more positive topics for Spanish periodicals compared to its sentiment mean. Notable negative topics for Spanish include *England*, the *Apologetic of Spain* and *Morale*.

Analyzing the results of French in more detail, we find it surprising that the topic *Image of Women* has a positive mean standardized sentiment, whereas the topic *Image of Men* has a negative mean standardized sentiment. The discussion of gender roles is one main topic of many enlightened authors and women were often considered a deviation of the male norm [Honegger, 1991; Steinbrügge, 2016]. Therefore, and keeping in mind the close reading experience, it would have seemed more logical to us to find a negative sentiment value for *Image of Women*. However, considering the plethora of different French periodicals, this observation turns out to be potentially caused by the quantitative dominance of *Jean-François de Bastide*'s voluminous periodicals *Le Nouveau Spectateur*, *Le Monde comme il est* and *Le monde*. One main topic of these periodicals is the *Love* discourse [Fischer-Pernkopf, Mussner, and Ertler, 2018], which of course accompanies a more positive sentiment on the opinion of women.

For Italian, we observe that the topics *Politics* and *Structure of Society* were generally shown in a gloomy light. For example, northern Italy is governed by foreign rulers who were not perceived in a positive way, although the Italians knew how to negotiate with them. The fact that *Charity* and *Fortune* convey a negative sentiment may be explained by the description of charitable actions, which often included the example of tragic destinies. These charitable actions are often connected to the criticized structures of society and the differences between city and countryside, which becomes particularly visible in the Venetian periodicals comprising the *Gazzetta veneta* and *L'Osservatore veneto* by *Gasparo Gozzi* or the *Gazzetta veneta* by *Pietro Chiari*. The Italy of the 18^{th} Century often looked hopefully at other foreign countries due to its own backwardness. After all, it was important to imitate the foreign, supposedly more progressive societies, explaining the more positive sentiment for the topic *Foreign Societies*. Also notable is the positive sentiment of the topic *Religion*, since a certain criticism of religion is present in Italian periodicals. Especially the church representatives were sometimes criticized quite harshly. Nevertheless, the need of caution had to be stressed, in order to prevent becoming a victim to censorship. Subsequently, rhetorical methods can only be perceived at a closer look, rather than on the surface as our results suggest.

The fact that among Spanish periodicals topics such as *Image of Men* and *Image of Women*, *Culture of Conversation* and *Manners and Costumes* have a negative sentiment could be due to the fact that those were the main areas of criticism for the Spanish authors, as direct criticism of Church or the Monarchy were not allowed [Von Tschilschke, 2009]. Hence, social interaction and in general the society were in the center of the enlightened attention in Spain. The comparatively negative sentiment of the topic *Apologetic of Spain* seems to be due to the fact that many spectator authors were among the adversaries of the Spanish apologists. As a reaction to criticism on Spanish literature from inside and outside the country, the apologists started defending the Spanish culture no matter what. In opposition to that, the authors of *El Censor*, *El Apologista Universal* and *El Corresponsal del Censor* started ridiculing the apologists [Guinard, 1973]. Notably, as the inflexible stratification of society [Ertler, 2004] was one reason of the discontent of enlightened authors in Spain, we would have expected a more negative sentiment for *Structure of Society* and *Critics on Nobility*.

3.2.6 Sentiment Networks

Our dataset contains annotated persons, places and works. In the second part of our analysis, we leverage these annotations to find answers to the following questions:

(i) **Sentiment Between Periodicals:** How did periodicals write about other periodicals? Was the tone among periodicals supportive and appreciative, or did they also criticize each other?
(ii) **Sentiment Between Languages and Places:** How had countries, cities and other places been perceived by different language communities? Was the emerging nationalism of the 18^{th} century reflected in periodicals?
(iii) **Sentiment Between People:** How did authors of spectator periodicals write about other (real or fictional) people? Were authors emotional or did they write about others on a factual basis?

To answer these questions, we construct three directed sentiment networks in which nodes represent the different entities (i.e., either periodicals, places or people) and edges represent sentiment (i.e., either positive or negative). There is a directed edge from one entity to another if the former has referenced the latter in one of the related text passages. Since the manual annotation of the dataset includes a language-independent normalization (e.g., "Dieu" and "Dios" are normalized to "God"), the constructed sentiment networks summarize data from all available languages. The representation of interacting entities in the form of graphs allows for simultaneously visualizing (i) the sentiment of how entities were perceived (edge color), (ii) how often an entity referenced another entity (edge width), and (iii) how

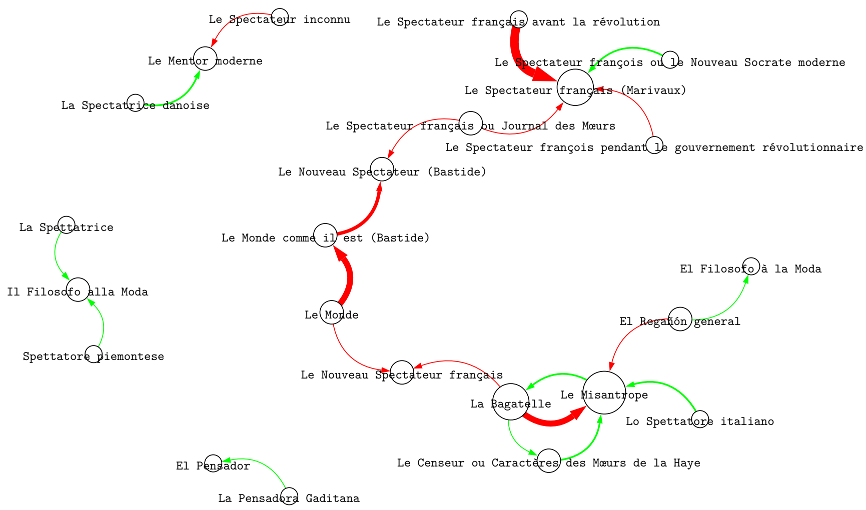


Figure 3.5: **Sentiment Between Periodicals.** This figure depicts the network of periodicals (represented by nodes) referencing other periodicals contained in our dataset. Directed edges indicate if a periodical referenced other periodicals and the color of edges illustrates the mean standardized sentiment of the text passages in which a respective periodical was referenced. Node sizes represent the degree of periodicals and edge widths represent the multiplicity of references.

often an entity was referenced (node size). As such, we find this form of representation useful for the interpretation of the results.

Sentiment Between Periodicals. We start by investigating the sentiment with which periodicals refer to other periodicals. Hence, nodes in the network represent periodicals published in the different languages and all across Europe. The edges signal the mean sentiment, normalized over the mean sentiment of the referring periodical, that was expressed in the text passages where a periodical was mentioned. We depict the resulting network in Figure 3.5. The two Italian periodicals *Spettatore piemontese* and the *Spettatrice* positively refer to the *Filosofo alla moda*, which is a translation of the French translation *Le Spectateur ou le Socrate modern* of the English prototype *The Spectator*. Further, the *Filosofo alla moda* has been the first periodical published in Italy and due to its popularity it was published once again in 1749 with the title *Lo Spettatore italiano o sia il Socrate moderno. Opera tradotta dal francese e publicata già col titolo di Filosofo alla moda ovvero maestro universale*. Therefore this text can be considered as a role model for other Italian periodicals and, as such, the text expresses positive sentiment as expected.

For Spanish periodicals, we observe a connection between *La Pensadora Gaditana* and *El Pensador*, because the publication of the presumably female *La Pensadora Gaditana* is a reaction to the disparaging representation of women in one of the most read Spanish periodicals, *El Pensador*. Despite that, we observe a positive sentiment for this references and argue, that this is potentially due to the appreciation of *El Pensador*, the first and a highly successful Spanish spectator periodical.

The subnetwork of French periodicals illustrates the starting point of the genre in French: *Le Misantrope*, the first French imitation of *The Spectator* published by *Justus van Effen* in the Netherlands, and *Le Spectateur François*, the first genuinely French spectator periodical published by *Pierre Carlet de Marivaux*. Being the first periodicals of their kind, they became the main points of reference for further French publications. Our network also shows the chain of references between the three publications of *Jean-François de Bastide* which were published consecutively: *Le Nouveau Spectateur* (1758 − 1760), *Le Monde comme il est* (1760) and *Le Monde* (1760 − 1761). The negative sentiment with which the successors refer to their predecessors is surprising, as one would expect that an author would use its previous successful work as a positive point of reference. However, we see a similar negative connection between *Justus van Effen*'s periodicals *La Bagatelle* (1718 − 1719) and its predecessor *Le Misantrope* (1711 − 1712). Note that the reference from the earlier *Le Misantrope* to the later *La Bagatelle* stems from the fact that both periodicals were annotated on the basis of a book edition published in 1742. These book editions often included modified texts (e.g., footnotes or prefaces), introducing the possibility of references to "future" periodicals.

Sentiment Between Languages and Places. We now analyze how the different language communities referenced to places, such as countries or cities. In case of this bipartite network, nodes with outgoing edges represent the languages and nodes with incoming edges represent either a country, city or fictional place. Again, we normalize sentiment over the mean of a language and connect languages to places if they have been referenced to in the respective text passages. For the purpose of better visualization, we only keep edges with at least 10 occurring references, and preserve only those

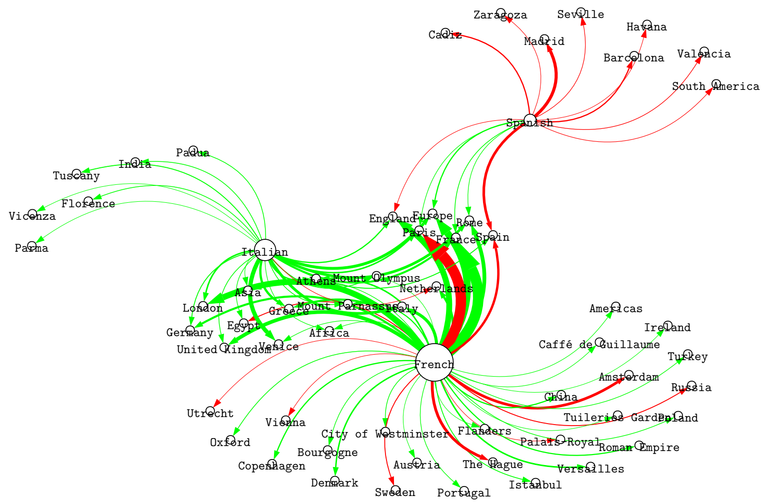


Figure 3.6: **Sentiment Between Languages and Places.** This figure depicts the respective languages in which periodicals were published and the places referenced by them (both represented by nodes). Directed edges indicate that a periodical referenced a place and the color of edges illustrates the mean standardized sentiment of the text passages in which a respective place was referenced. Node sizes represent the degree of languages and edge widths represent the multiplicity of references. Note that we only show edges with a minimum multiplicity of 10.

nodes that are connected by the remaining edges. We depict the resulting reference network in Figure 3.6.

Starting with Italian, we first notice the references to mythological Greek places, such as the Mount Parnassus or Mount Olympus, indicating that the Antiquity and the ancient mythology is still important for story-telling in the 18th century. Very interesting are also the positive references to other countries, such as England, France, Germany, as well as cities, such as Paris or London. During the 18th century, Italy suffered from a weak self-image, which motivated it to imitate other European countries and cities. Furthermore, since other European cities caught up economically, Italian cities, such as Venice, lost their economic preeminence, without actually becoming economically inferior [Vaussard, 2001]. Thus, these other European countries and cities were generally referenced positively in spectator periodicals. The references to Asia and India are indicators for the *Chinoiserie* which was popular especially in the late 17^{th} and the 18^{th} century. Not only in art, but also in literature the Chinese style and the style of other countries in East Asia was imitated [Beevers, 2008]. The interest in distant foreign countries correlates also with the trading interests. The only country which appears in a very negative sentiment context are the Netherlands. Potentially the strong protestant context of the Netherlands was a provocation. The positive sentiment concerning England in Italian periodicals could be based on England's strong influence on the Enlightenment, which led to its perception as a role model. Another possible explanation is that the sentiment may arise from those Italian periodicals being translations of the English periodicals.

The Spanish texts generally refer to Europe, Paris, France and Rome with a positive sentiment, which reflects the cultural orientation of the Spanish Enlightenment towards other parts of Europe [Von Tschilschke, 2009]. The Spanish territory (which at that moment included the colonies in South America) was, however, presented with a more negative sentiment. We argue that this is related with the goal of spectator periodicals to examine and criticise the current state of their countries and societies, which often resulted in authors presenting their surroundings in a more negative way. This same explanation may be applied to the negative sentiment values we find in the French periodicals for Paris, Utrecht, The Hague and Amsterdam. These cities were the main publication places for French spectator periodicals, which is why they are in the center of criticism. Note that this observation is contrary for Italian periodicals, which refer to Italian cities positively. The clearly negative sentiment in references to Spain in French periodicals seem to be due to a cultural debate, in which French authors presented French culture as generally superior to Spanish culture [Von Tschilschke, 2009] and used Spain as a negative example of anti-enlightened politics, society and literature. London, England and in general the United Kingdom have a strong influence on the Enlightenment and therefore were perceived in a more positive way in French periodicals.

Sentiment Between People. Finally, we investigate references to other authors and people both real and fictional. In this network, we represent individuals as nodes and directed edges between nodes indicate that an author referenced a person. We standardize sentiment over the mean of an author and, again, keep only edges with multiplicity greater than 10. We illustrate the resulting reference network in Figure 3.7.

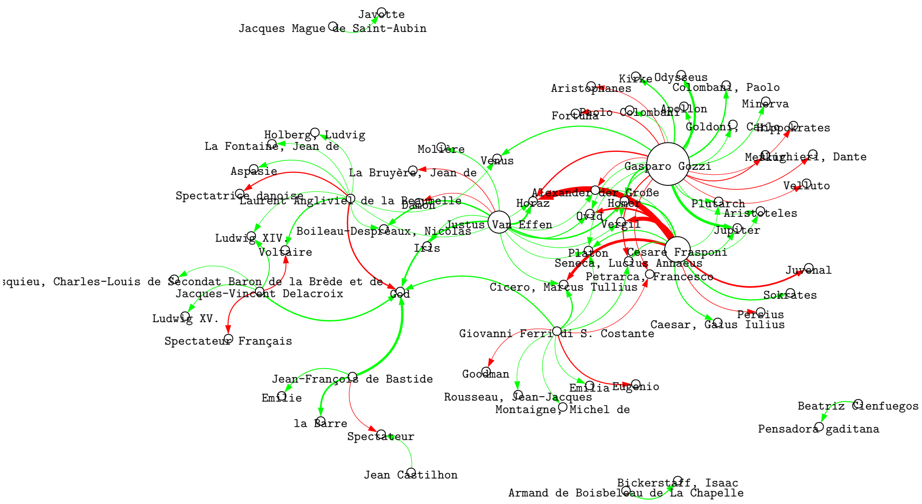


Figure 3.7: **Sentiment Between People.** This figure depicts the authors of periodicals and the people (real and fictional) referenced by them. Directed edges indicate if an author referenced a person and the color of edges illustrates the mean standardized sentiment of the text passages in which a respective person was referenced. Node sizes represent the degree of authors and edge widths represent the multiplicity of references. Note that we only show edges with a minimum multiplicity of 10 and do not consider anonymous publications for this analysis.

High degree nodes in the network include authors of three different Italian periodicals, each showing different references. *Cesare Frasponi* is the author of the *Filosofo alla moda*, which is a translation of the English *The Spectator*. *Gasparo Gozzi* is the author of the *Gazzetta veneta* and the *L'Osservatore veneto* and *Giovanni Ferri* is the author of the *Spettatore italiano*. The former periodical, *Filosofo alla moda*, includes mostly references to the Antiquity, especially to ancient philosophers and authors. However, it barely includes references to mythology. Contrary to our expectations, we observe a negative sentiment in references to *Horaz*, a leading Roman lyric poet in the beginnings of the Roman empire. His principle *prodesse et delectare* was very important for moral periodicals, which is why we would have expected a positive sentiment. Further, Italian authors barely referenced any other Italian authors due to the fact that the *Filosofo alla moda* is a translation from the English *The Spectator*.

We observe a wider range of references in the periodicals published by

Gasparo Gozzi. His references to the Antiquity comprise both philosophers and mythological figures, such as *Horaz*, *Plutarch*, *Platon* and *Apollo*. The sentiment of these references is either positive or negative, depending on the message and the example signaled in the text passage the person or figure was mentioned in. A noteworthy result are the references to *Fortuna*, which stemmed from the superstitious fears in peoples life before the Enlightenment. People started to overcome this fear during the 18th century, however, readers needed to be reminded once in a while by negatively reflecting this superstition. Further, Italian authors include numerous references to the own historical Italian tradition of literature, such as *Petrarca*, *Dante* and *Goldoni*. The reference to the latter reflect the discussion between *Goldoni* and *Chiari* about the modern Italian theatre, which captivated the whole world of theatre and literature in Italy at the time [Hösle, 1993]. The fact that *Goldoni* appears in positive contexts may provide evidence that *Gozzi* took sides with *Goldoni* during this discussion.

The periodical *Spettatore italiano* of *Giovanni Ferri*, which is the latest publication contained in our dataset (1822), contains far less references. The *Querelle des Anciens et des Modernes*, a literary and artistic debate at the turn of the 17th and 18th century, is still present in this text, but the micro-narrations with examples from the everyday life of the present are much more important and numerous. In addition, the Antiquity in the *Spettatore italiano* was overcome and not seen as a timeless valid model anymore. Italian authors are barely found, due to the fact that the periodical takes care of everyday life examples and is less concerned with literature. Noteworthy here is the appearance of *Eugenio* and *Goodman*, two fictional figures, which are the leading voices in a plethora of micro-narrations. We observe that these two characters seem to be connected with a negative sentiment, but, in fact, they often co-occur when there are other characters with negative sentiment. Indeed, the function of *Eugenio* and *Goodman* in these cases is to correct the wrong opinions of other figures.

Finally, we discuss why *Horaz*, *Petrarca* and *Dante* appear in a negative sentiment even though interpretation of results leaves us with no reliable statement. *Horaz* is the universal model of the spectator genre and the Italian writers *Dante* and *Petrarca* are also models in the Italian periodicals who were "untouchables". Similar to *Eugenio* and *Goodman*, their appearance in a negative context can be explained by the fact that the spectator periodicals generally created a lot of negative examples which were then solved by *Horaz*, or with reference to the *Canzoniere* published by *Petrarca* and the *Divina Commedia* published by *Dante*.

Regarding French periodicals, we observe an overlap with Italian periodicals focusing on antique names, such as *Horaz*, *Cicero*, *Seneca* or *Homer*. As these were also frequently referenced in the original English *The Spectator*, we argue that these references (at least partially) exist due to the literary fashion introduced by *Addison* and *Steele*, the publishers of *The Spectator*. In the French periodicals, these references are mainly positive, which suggests that they served as moral or literary models for the authors of spectator periodicals. In contrast, only the authors of French periodicals referred to more recently influential French authors from the 17^{th} century, such as *Jean de La Fontaine*, *Jean de La Bruyère*, *Nicolas Boileau* and *Molière*. Their literary works generally served as a model of orientation for authors of the 18^{th} century. Therefore, the negative sentiment in the references to *Jean de La Bruyère* by *Justus van Effen* probably results from the content of *La Bruyères' Charactères*, which incorporates mainly negative character-portraits of men to illustrate their vices. *Nicolas Boileau* was one of the most influential literature theorists of this era and is mainly referred to with positive sentiment. French results further clearly expose the intellectual framework of the periodicals by *Jacques-Vincent Delacroix* and *Laurent Angliviel de La Beaumelle*. *La Spectatrice danoise* published in Denmark by the latter represents an intermediate position between the French cultural reference framework (*La Fontaine*, *Boileau*, *Voltaire*) and the Danish context (*Ludvig Holberg*). The sentiment in both cases seems to be equally positive. *Voltaire* constitutes an interesting point in this network. Whereas *La Beaumelle* refers to him in a positive way, *Delacroix'* periodicals seem to have a more negative view on this polemic author.

In comparison, the Spanish periodicals seem to include rather little reference to other authors or thinkers. Only in *La Pensadora Gaditana* there are numerous references to the fictional (female) author of the periodical, *Beatriz Cienfuegos*, which is apparently due to her own discourse, but also to numerous letters of readers published in the periodical. The fact that the sentiment value for these references are mainly positive is probably due to the fact that the authors present the fictional authors in a positive way, but also many readers address them with certain respect in their letters.

From the results of *Les Chiffons* by *Jacques Mague de Saint-Aubin* and *Le Philosophe Nouvelliste*, the French translation of *The Tatler*, we can draw the same conclusions.

3.2.7 Sentiment Word Networks

In the third part of our analysis, we set our focus on the textual characteristics of periodicals and investigate how sentiment was conveyed by them. Specifically, we are interested in how positive and negative words are used in the text and if there are any significant differences in their usage. We structure our analysis around the following questions:

(i) **Commonly Used Sentiment Words:** What sentiment words are frequently used in periodicals, and are positive or negative words more common? Which sentiment words have stronger connectivity and which words act as mediators between other sentiment words?
(ii) **Sentiment Motifs:** Are there common local sentiment motifs such as closed triads of sentiment words? Are there small local groups of sentiment words that frequently co-occur in close vicinity?
(iii) **Sentiment Connectivity:** Globally, do sentiment words tend to connect to other words of the same polarity or is there strong mixing of words with opposite sentiment? In other words, are there global clusters with a predominant sentiment?

Network Extraction. We construct sentiment word networks in a similar fashion to the method introduced by Montemurro and Zanette [Montemurro and Zanette, 2013]. The authors proposed to create networks in which nodes represent the most informative words in a corpus of texts and in which edges indicate that the connected words share semantic similarity.

We construct sentiment word networks as follows: Rather than extracting the most informative words, we simply restrict our attention to the words in the sentiment dictionaries that we used in our previous analyses. Then, two words are connected if they are regularly used in the same context.

Specifically, as a first step we use Spacy[6] and its respective language models to lemmatize texts of periodicals in order to combine inflected word forms into single representations. Further, we remove words that occur in more than 80% of all text passages and filter stop words[7] introduced by the automatic translation of English sentiment dictionaries to respective languages. We then represent each word by a vector counting the number of occurrences of this word in respective text passages. Two words are finally connected if the cosine similarity between these words is significant above a 99% confidence level, where significance is computed using 1000-fold permutation. Specifically, we repeatedly evaluate similarities on randomized frequency vectors, allowing us to compute p-values, defined as the fraction of times the random similarities were equal to or greater than those from the original texts, respectively for each pair of words. In Table 3.2 we list the top ten significant edges according to their p-value for each of the five networks.

We create networks respectively for each of the five languages contained in our dataset and analyze (i) degree distributions as well as degree, betweenness and closeness centralities to find commonly used sentiment words and their connectivity patterns and roles, (ii) local clustering coefficient distributions to find sentiment motifs, and (iii) degree and sentiment assortativity to analyze the sentiment connectivity.

We present results separately for positive and negative word nodes, except for assortativity, for which we consider the networks as a whole. Note that

Table 3.2: **Most Significant Edges.** This table lists the top ten significant edges in networks according to their p-value, respectively for each language.

French	German	Italian	Portuguese	Spanish
propre - naturel	vorgeben - ansprechen	primo - celebrare	empenhar - chato	primero - dudar
repoussant - détresse	unruhig - versprechen	meglio - dimostrare	perturbar - impor	estrafalario - crítica
amour - cher	eilig - ohnmächtig	natura - dimostrare	hipocrisia - energia	errores - crítica
miséricordieux - détresse	unvermeidlich - bedauern	picche - bandiera	impedimento - malicioso	equitativo - crítica
charme - amour	genießen - falls	poetico - chiamare	energia - impor	ángel - lógico
charmant - amour	vollenden - untergraben	dimostrare - onore	respeitar - hum	emprendedor - crítica
toujours - indifférence	gewaltig - lindern	dimostrare - opera	lentamente - homicida	mejoría - ansia
chagrin - amour	erträglich - gewähren	imperterrito - banderuola	infringir - homicida	críticos - crítica
tromper - indifférence	veranlassen - opfern	spade - bandiera	hipocrisia - perturbar	lúcido - menguante
peine - oublier	dreist - zweifeln	dimostrare - passione	convencido - agitar	tajante - cruento

[6] `https://spacy.io` (version used: 2.2.3)
[7] `https://github.com/Alir3z4/python-stop-words`

we only consider the largest connected component of each network for the remainder of our analysis (the minor components in networks have less than three nodes each; only the German network has one minor component with 29 nodes).

In Table 3.3 we list the number of nodes and edges of our sentiment word networks including the percentage of positive and negative nodes in each network. We observe that positive nodes represent the minority for each network, albeit with a less pronounced imbalance than in the original dictionaries. Specifically, we observe a ratio of 37% positive nodes for French, 46% positive nodes for German, 38% positive nodes for Italian, 42% positive nodes for Portuguese and 41% positive nodes for Spanish. We depict the resulting sentiment word networks for German, our smallest network, in Figure 3.8.

Commonly Used Sentiment Words. We start with investigating the commonly used words in periodicals by reporting cumulative distribution functions (CDF) of the degree for all nodes, positive nodes only and negative nodes only in Figure 3.9, respectively for each of the five languages contained in our dataset. Our results depend on the sizes of networks. For the three larger networks of French (cf. Figure 3.9a), Italian (cf. Figure 3.9c) and Spanish (cf. Figure 3.9e), we observe that negative nodes have higher probabilities for lower degrees in networks, whereas positive nodes have higher probabilities for mid-range degrees. In other words, negative words have a more narrow degree distribution than positive words, and the median degree of negative words is smaller than of positive words.

To test for significance of the difference between the distributions for positive and negative nodes in the networks, we conduct two-sample Kolmogorov-

Table 3.3: **Network Sizes.** This table lists the number of nodes (with ratios of positive and negative nodes) and edges of the semantic sentiment networks, respectively for each language.

	French	**German**	**Italian**	**Portuguese**	**Spanish**
# Nodes	2 624	207	2 711	685	2 474
... thereof positive	982 (37%)	95 (46%)	1 039 (38%)	286 (42%)	1 007 (41%)
... thereof negative	1 642 (63%)	112 (54%)	1 672 (62%)	399 (58%)	1 467 (59%)
# Edges	45 172	1 589	44 512	2 784	31 576

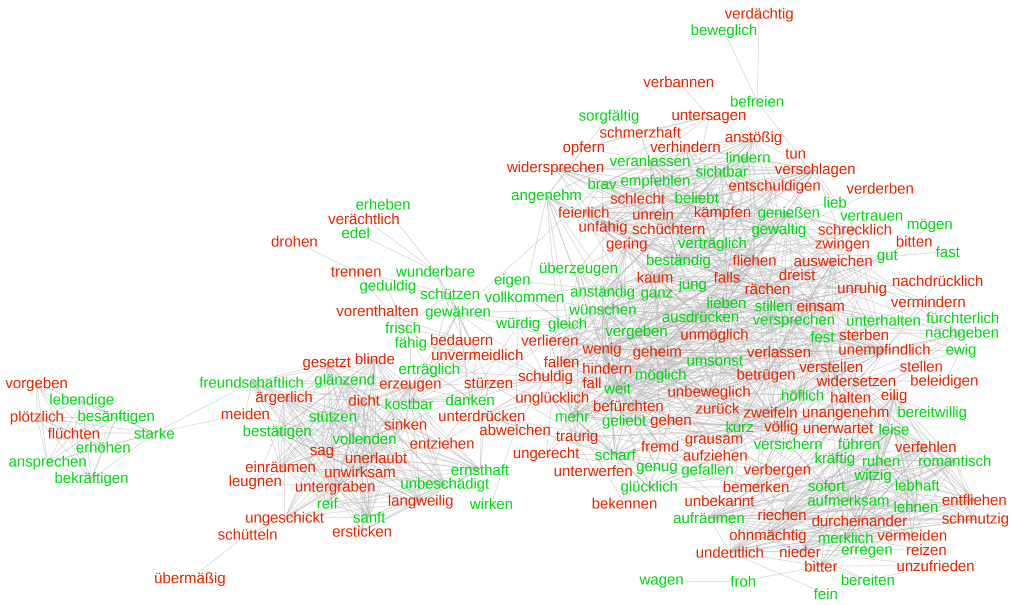
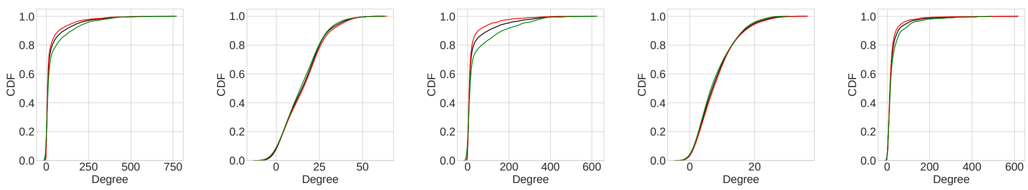


Figure 3.8: **German Sentiment Word Network.** This figure illustrates the sentiment word network for German where red word nodes represent words conveying a negative sentiment and green word nodes represent words conveying a positive sentiment.

Smirnov tests with the commonly used significance level $\alpha = 0.05$. The null hypothesis of this nonparametric test is that both samples are randomly drawn from the same distribution. Note that our samples and distributions fulfill all other assumptions for this test (independent samples, at least an ordinal level of measurement as well as continuous variables). Rejecting the null hypothesis when comparing distributions of positive and negative nodes indicates a significant difference between these distributions. To account for multiple hypothesis tests we perform Bonferroni correction [Bonferroni, 1935] and divide our significance level by 5. For French, Italian and Spanish networks, the p-value is smaller than 0.0005, indicating a significant difference between degree distributions of positive and negative nodes. These results suggest that a large number of negative words appear



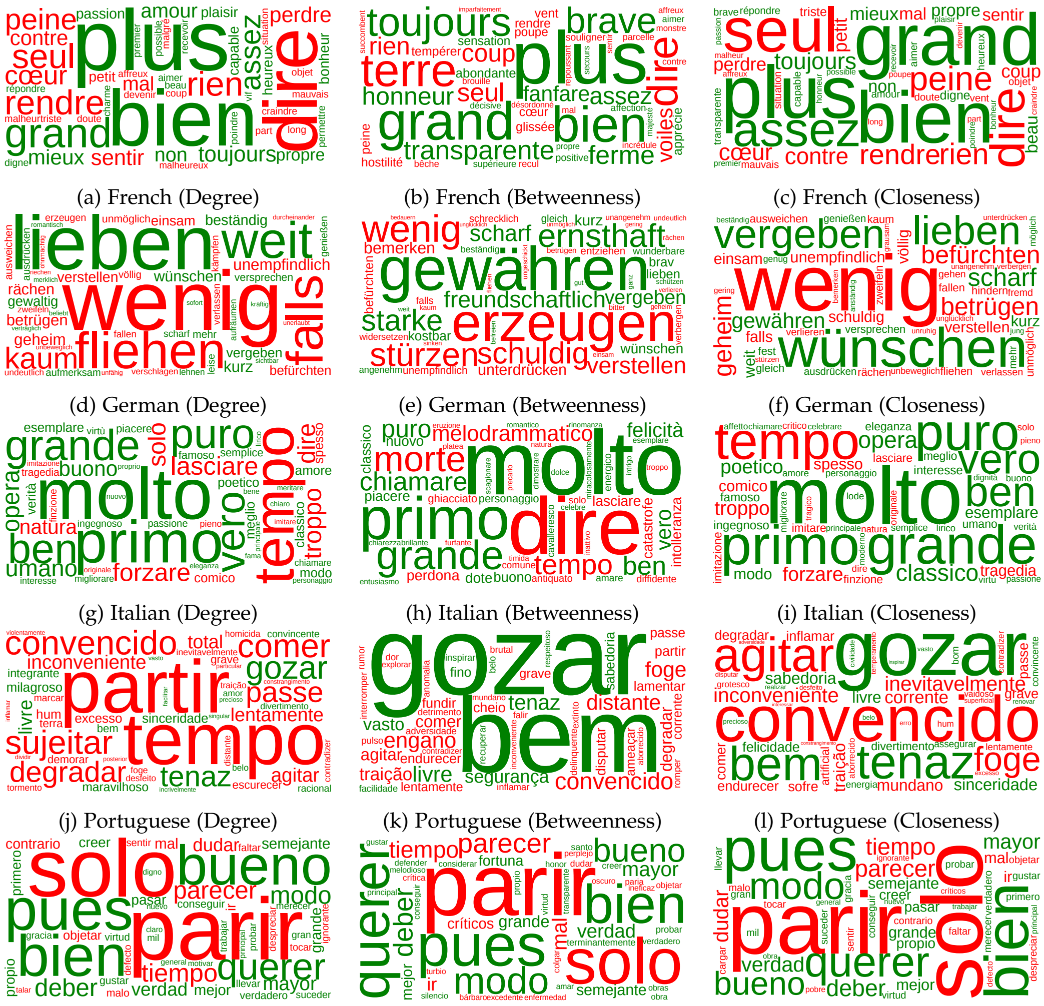
(a) French (b) German (c) Italian (d) Portuguese (e) Spanish

Figure 3.9: **Degree Distributions.** This figure depicts the CDF of degrees of all nodes (black lines), positive nodes (green lines) and negative nodes (red lines) respectively for each of the five languages contained in our dataset. In case of the three larger networks, including French (cf. Fig. 3.9a), Italian (cf. Fig. 3.9c) and Spanish (cf. Fig. 3.9e), negative nodes have higher probability for low degrees whereas positive nodes have higher probability for mid-range degrees. The difference between degree distributions of positive and negative words is significant (p-values < 0.0005) for each of the three networks according to two-sample Kolmogorov-Smirnov tests. Results are different for the two smaller German (cf. Fig. 3.9b) and Portuguese networks (cf. Fig. 3.9d) for which there is no significant difference between positive and negative distributions according to a two-sample Kolmogorov-Smirnov test (p-values > 0.05).

with a small number of other sentiment-conveying words and that positive words, on average, appear with a substantial number of other sentiment-conveying words. A plausible explanation for this observation in larger networks would be that the basic attitude of periodicals is positive. Also, more negative parts of the text include a larger number of various negative words specifically tailored to the given context.

In case of the two smaller networks of German (cf. Figure 3.9b) and Portuguese (cf. Figure 3.9d), the two-sample Kolmogorov-Smirnov test (again with $\alpha = 0.05$ and Bonferroni correction) resulted in a p-value of 0.83 and 0.39, respectively. We conclude that there is no significant difference between the degree distributions of positive and negative words for those two networks. We believe that this is a consequence of the small corpus size for these languages and, hence, refrain from drawing any conclusions based on these observations.

We extract commonly used sentiment words in periodicals by reporting the top 50 central words (according to the degree) for French in Figure 3.10a, for German in Figure 3.10d, for Italian in Figure 3.10g, for Portuguese in



(a) French (Degree) (b) French (Betweenness) (c) French (Closeness)

(d) German (Degree) (e) German (Betweenness) (f) German (Closeness)

(g) Italian (Degree) (h) Italian (Betweenness) (i) Italian (Closeness)

(j) Portuguese (Degree) (k) Portuguese (Betweenness) (l) Portuguese (Closeness)

(m) Spanish (Degree) (n) Spanish (Betweenness) (o) Spanish (Closeness)

Figure 3.10: **Top Central Words.** This figure depicts the top 50 most central words according to degree (left column), betweenness (center column) and closeness (right column) centrality, respectively for French, German, Italian, Portuguese and Spanish (top to bottom rows). The majority of top central words convey a positive sentiment for Italian and Spanish and a negative sentiment for French, German and Portuguese across all centrality metrics. Note that words stem from the languages of the 18th century and respective European countries and, thus, do not relate to potentially ambiguous meanings of modern day languages.

Figure 3.10j and for Spanish in Figure 3.10m. Top central words for French include “bien” (“well”), “plus” (“more”) and “dire” (“to say”). For German, the top central words according to degree centrality include “lieben” (“to love’), “wenig” (“little”) and “fliehen” (“to flee”). In case of Italian, the top words include “molto” (“very” or “much”), “primo” (“first”) and “tempo” (“time” or “period”). The most commonly used words in Portuguese periodicals include “partir” (“to leave”), “tempo” (‘time”) and “convencido” (“convinced”). Finally, Spanish top words comprise “solo” (“alone” pr “simple”), “parir” (“to calve” or “to lamb”) and “pues” (“because” or “since”).

Regarding the ratios of positive and negative words among the top 50 central words, we find mixed results for the five languages. For French, we find an equal proportion of positive and negative words. The majority of top words is negative for the smaller German (54% negative words) and Portuguese (64% negative words) networks, while for Italian and Spanish we report a majority of positive words with 70% and 66%, respectively.

To find the sentiment words that bridge meaning and connect otherwise distant groups of sentiment words appearing together, we compute betweenness centrality for each of the five networks. We depict the top 50 central words according to betweenness centrality for French in Figure 3.10b, for German in Figure 3.10e, for Italian in Figure 3.10h, for Portuguese in Figure 3.10k and for Spanish in Figure 3.10n. Top words according to betweenness centrality are similar to those according to degree centrality. We observe notable differences for the smaller German and Portuguese networks. Top words for the former include “gewähren” (“to grant” or “to award”) or “erzeugen” (“to create”, “to produce” or “to make”), while top words for the latter include “bem” (“well” or “right”) and “gozar” (“to enjoy”).

Similar to degree centrality, the majority of top words according to betweenness centrality is positive for Italian (58% positive words) and Spanish (62% positive words), while we observe higher proportions of negative words for German (60% negative words) and Portuguese (62% negative words). For the French network, which has equal ratios according to degree centrality, we report a higher ratio of negative words with 58%. These results are contrary to our previous work [Koncar and Helic, 2019] with simple co-occurrence networks for which we find higher ratios of negative words among top central words for all of the three centrality metrics.

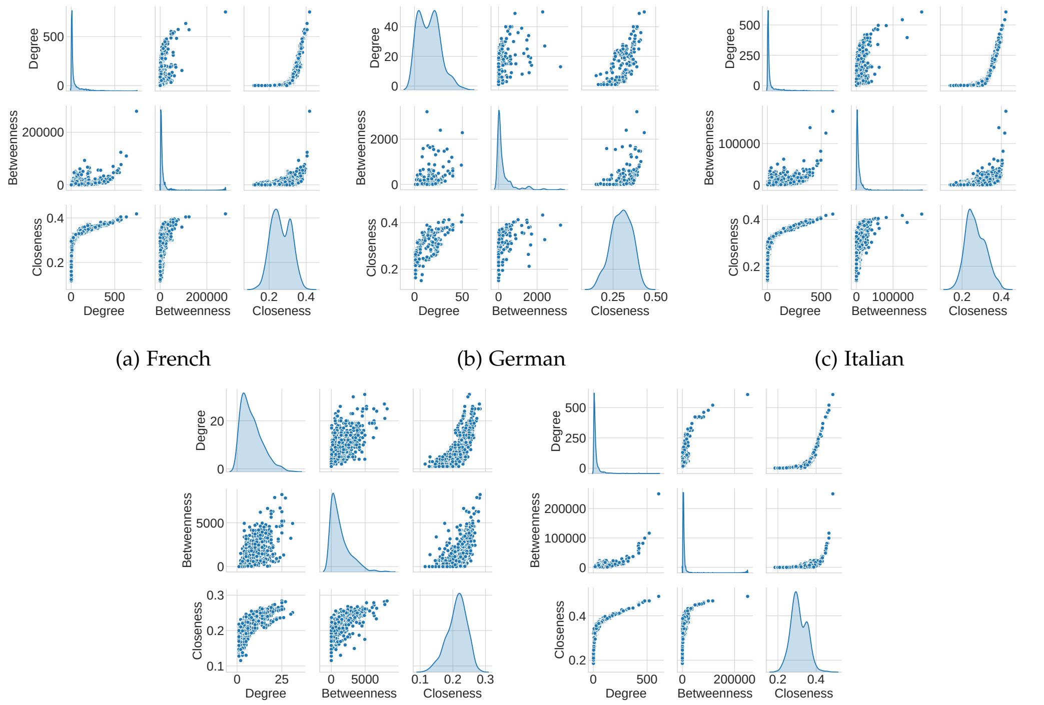
Completing our analysis of sentiment words connecting other sentiment words, we also investigate closeness centralities. Again, we consider the top 50 central words, respectively for each of the five languages contained in our dataset. We illustrate results of closeness centrality for French in Figure 3.10c, for German in Figure 3.10f, for Italian in Figure 3.10i, for Portuguese in Figure 3.10l and for Spanish in Figure 3.10o.

Top words for French, Italian and Spanish are again similar to those according to degree centrality. We observe notable differences for German, such as "vergeben" ("to forgive" or "to award") or "wünschen" ("to wish"), and Portuguese, such as "agitar" ("to shake" or "to stir up") or "tenaz" ("tenacious").

Again, we observe similar results to degree centrality, with a majority of positive words for Italian (66% positive words) and Spanish (62% positive words), a majority of negative words for German (64% negative words) and Portuguese (62% negative words), as well as equal ratios of positive and negative words for French.

Finally, we depict correlations between degree, betweenness and closeness centrality in Figure 3.11, respectively for each of the five languages. We compute Spearman's rank correlation coefficients among each pair of metrics and for each language and find that metrics positively correlate with each other (in all cases p-values are smaller than 0.0005), suggesting that words are equally central among different metrics. Our computed Spearman's correlation coefficients are ranging from 0.56 to 0.76 between degree and betweenness centrality, from 0.65 to 0.87 between degree and closeness centrality and from 0.35 to 0.69 between betweenness and closeness centrality across the three languages.

Summarizing, the results for degree, betweenness and closeness centrality in combination with ratios of negative words in networks (cf. Table 3.3) suggest an overrepresentation of the minority class (positive nodes), corroborating similar results for heterophilic networks found in previous work [Karimi et al., 2018]. To further investigate our observations, we compute proportion tests based on one-sample z-tests for each of the centrality metrics and

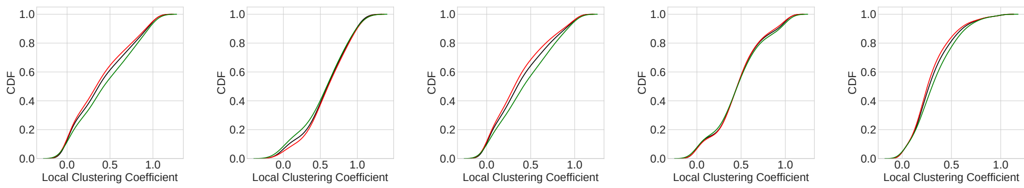


(a) French (b) German (c) Italian

(d) Portuguese (e) Spanish

Figure 3.11: **Centrality Metric Correlations.** This figure illustrates the correlations between the three centrality metrics, respectively for each language. For all metrics and languages, we observe a strong positive correlation, suggesting that words are similarly central across metrics.

languages. This allows us to compare the proportion of negative words among top central words to the proportion of negative words in sentiment word networks. Note that this test assumes continuous data following a normal distribution. Since we can not assume normality, we have to first test whether we can use the normal approximation method by checking if $np \geq 10$ and $n(1-p) \geq 10$, where n is the sample size (in our case 50) and p is the sample mean, which we confirm for all of our cases. In 5 out of 15 cases, the ratio of negative words among top central words is significantly smaller than ratios in networks (p-values < 0.0005). The remaining 5 tests are insignificant as p-values are greater than 0.05 (all cases with greater ratios among top words as compared to networks), all comprising results for



(a) French (b) German (c) Italian (d) Portuguese (e) Spanish

Figure 3.12: **Local Clustering Coefficient Distributions.** This figure depicts the CDF of local clustering coefficients of all nodes (black lines), positive nodes (green lines) and negative nodes (red lines) respectively for each of the five languages contained in our dataset. In case of the three larger networks, including French (cf. Fig. 3.12a), Italian (cf. Fig. 3.12c) and Spanish (cf. Fig. 3.12e), negative nodes have higher probability for lower local clustering coefficients whereas positive nodes have higher probability for mid-range local clustering coefficients. The difference between local clustering coefficient distributions of positive and negative words is significant (p-values < 0.0005) for each of the three networks according to two-sample Kolmogorov-Smirnov tests. Results are different for the two smaller German (cf. Fig. 3.12b) and Portuguese networks (cf. Fig. 3.12d) for which there is no significant difference between positive and negative distributions according to a two-sample Kolmogorov-Smirnov test (p-values > 0.05).

our two smaller German and Portuguese networks. These results suggest that there is, indeed, a slight overrepresentation of positive words among top central words.

Sentiment Motifs. We look for groups of sentiment words that frequently co-occur in close vicinity by computing local clustering coefficients for each of the five languages. We depict CDF of local clustering coefficients for all nodes, positive nodes only and negative nodes only in Figure 3.12, respectively for each of the five language contained in our dataset. Again, our results vary depending on the size of networks. Contrary to our expectations (following previous observations about a negative correlation between the degree and the local clustering coefficient [Ravasz and Barabási, 2003]), for French (cf. Figure 3.12a), Italian (cf. Figure 3.12c) and Spanish (cf. Figure 3.12e) networks, we observe that negative nodes have higher probabilities for lower local clustering coefficients, whereas positive nodes have higher probabilities for mid-range local clustering coefficients. This suggests that negative words have less chance to co-occur frequently with other sentiment conveying words and that positive words are more often used together in groups of similar context.

We again test for the significance of the difference between the distributions for positive and negative nodes by conducting Kolmogorov-Smirnov tests (again with $\alpha = 0.05$ and Bonferroni correction). For the three larger networks, comprising French, Italian and Spanish, we observe p-values smaller than 0.005, indicating a significant difference between distributions of positive and negative words. In case of our smaller German (cf. Figure 3.12b) and Portuguese (cf. Figure 3.12d) networks, the p-values are 0.92 and 0.97, respectively, depicting an insignificant difference between positive and negative words. Similar to degree distributions of smaller networks, we argue that this is a consequence of the small corpus size and refrain from further conclusions.

Overall, we observe a wide range of local clustering coefficients in all networks, suggesting that sentiment words form pairs as well as triangles when considering their co-occurrences.

Sentiment Connectivity. We now study if words of similar frequencies and similar sentiment (i.e., positive or negative) form global clusters with a predominant sentiment in respective sentiment word networks. For that, we compute degree and sentiment assortativity.

Our results for degree assortativity of French, Italian and Spanish range between -0.10 and 0.04, depicting that high and low degree nodes connect almost randomly with each other. In case of German and Portuguese, networks are slightly more assortative, with values of 0.22 and 0.32, respectively. This observation is contrary to previous findings, where Cantwell and Newman [Cantwell and M. Newman, 2019] depicted slightly disassortative (based on degree) word networks constructed through part-of-speech tags.

Regarding sentiment assortativity, we observe similar result to our previous work [Koncar and Helic, 2019] where we constructed networks based on simple co-occurrences and found that networks are non-assortative. Here, sentiment assortativity lies between -0.02 and 0.07 for all networks. In any case, the results for sentiment assortativity suggest that positive and negative words are used randomly together and do not form any clusters based on their polarity. One plausible hypothesis for this observation could be that authors of spectator periodicals aimed at being impartial, which is possibly reflected in their explanations depicting both positive as well as negative aspects of the subject.

3.2.8 Discussion

Our analysis sheds light on the characteristics of sentiment conveyed by spectator periodicals published during the 18^{th} century. Assessing how sentiment evolves over time, we find that periodicals are, in general, rather stable over time and differences in sentiment are language-dependent and the individual periodicals themselves. It is, however, interesting that the polarity of mean sentiment across all languages is rather low, suggesting that authors succeeded in their aim to be neutral and impartial. This would be different to results of sentiment analyses on today's media, for which polarity is shown to be more extreme [Godbole, Srinivasaiah, and Skiena, 2007; Dos Rieis et al., 2015], especially in cases of social conflicts in divided societies [Makrehchi, 2014]. However, note that this more extreme polarity may be an artifact of shorter texts analyzed in studies focusing on today's media. Typically, the ratio of sentiment words compared to non-sentiment words may be higher in shorter texts (e.g., a tweet) which, thus, results in a more polarizing sentiment as compared to longer texts (e.g., an issue of a periodical). To test whether or not text length influences the comparison between sentiment of spectator periodicals and sentiment of today's media, we compute the mean sentiment of all text passages with a maximum length of 280 characters (i.e., comparable to the length of a tweet) across all five languages. Here, we find a mean sentiment of 0.027 (95% confidence intervals: 0.015 and 0.039), indicating that spectator periodicals may have been more neutral as compared to today's media. Nevertheless, we suggest an in-depth analysis of sentiment differences between spectator periodicals and today's media for future work.

Regarding the relation between sentiment and narrative forms, we find that our results are inconclusive. A potential explanation for this observation could be that narrative forms have been used by authors through different means, each with a different interpretation and incorporation of distinct styles. A further analysis of the different styles of authors is a promising subject for future work, and can be accomplished, for example, by authorship identification.

Our analysis on topics reveals both differences, such as disparate views on religion, and commonalities, such as a general more critical stance towards politics, among countries in Europe of the 18^{th} century. We observe that topics dealing with interpersonal relationships were perceived positively across countries, which may be due to the general tenor of treating other people with respect and the particular aim of spectator periodicals to increase the morality of their readers. In our analysis, we also report a critical discussion of foreign societies (especially in case of French and Spanish periodicals), which could indicate the rising nationalism during the 18^{th} century [Mann, 1996].

Overall, our analysis sheds light on how important issues of the Age of Enlightenment were perceived by the people living during that time and contributes to the knowledge of the starting point of modern societies.

We further investigate how periodicals wrote about other periodicals and find mixed results. While Italian and Spanish periodicals were positive about other periodicals in these languages, French periodicals were rather negative about other French periodicals. Hence, drawing conclusions from these observations is vague and we suggest a more detailed analysis for future work, allowing us to learn more about how sentiment and information spread across periodicals.

The results from our sentiment word networks reveal differences and commonalities to our previous results produced with simple co-occurrence networks [Koncar and Helic, 2019]. While we observe higher local clustering coefficients for negative words in simple co-occurrence networks, we report the opposite behavior with our new sentiment word networks, suggesting that when considering semantic affinities, negative words form less local motifs in networks. Typically, words in networks form local clusters when they co-occur within the same local domains (e.g., focusing around one distinct topic or text passage). To us it seems that this is not the case with words conveying negative sentiment, further corroborating our initial hypothesis that negative words are used distinctively to discuss critical issues. We find a commonality in the top central words, for which we observe a weak overrepresentation of positive words in both sentiment word networks and simple co-occurrence networks which, combined with our results from degree distributions, suggest a general mild to positive tone in spectator periodicals.

Limitations. Our sentiment analysis of spectator periodicals is a challenging task, especially due to the absence of dedicated and comparable methods for 18^{th} century texts. It is therefore inevitable to fall back on existing methods (lemmatization and sentiment dictionaries). At the first glance one may argue that these existing methods are originally intended for modern day texts, and that there is no reason to believe that they yield valid results for 18^{th} century texts. Our coverage analysis, however, reveals that the majority of words in the used sentiment dictionaries indeed appeared in the respective texts, thus putting our analyses on solid grounds.

Additionally, the selection of other sentiment dictionaries may alter our results as suggested by previous research [Reagan, Tivnan, et al., 2015]. To assess if our results are robust with respect to the selection of sentiment dictionaries, we repeated our calculation of the development of the sentiment over time with a selection of alternative publicly available sentiment dictionaries[8]. In the case of French, we observe similar trends over the years but the mean polarity is shifted towards a more positive sentiment as compared to results from original dictionaries. For Italian and Spanish, we observe similar mean polarity and trends. Thus, we obtain qualitatively comparable results with the alternative selection of sentiment dictionaries. However, a more in-depth analysis on the influence of sentiment dictionaries on the quantitative results represents an interesting avenue for future research, potentially further strengthening our results.

Regarding potential issues with lemmatization for sentiment word networks conducted with models for modern day languages, we redid all related experiments without lemmatization. We find similar results for degree distributions, local clustering coefficients as well as degree and sentiment assortativity, albeit noticing a larger degree assortativity for the smaller German (0.43 compared to 0.22) and Portuguese (0.59 compared to 0.33)

[8] In particular, we use FEEL [Abdaoui et al., 2017] for French, an OpeNER sentiment lexicon for Italian [Russo, Frontini, and Quochi, 2016] and a Spanish dictionary found on GitHub (`https://github.com/aylliote/senti-py`).

networks. In the case of centralities, we report similar ratios of positive and negative words among top central words of all centrality measures (with some variances in top central words). As such, we obtain qualitatively comparable results without lemmatization and suggest a detailed analysis on the impact of lemmatization for future work.

We nevertheless believe that the implementation of adjusted models could further improve the quality of analysis, and thus leave this task for future work. Further, we address the fact that our approach is data hungry, as our sentiment word networks only connect words if similarity is above a certain significance level, which can only be achieved if sufficient data is available. For example, the small corpus size of German and Portuguese texts resulted in these networks having multiple components, effectively leaving them useless for the analysis. The significance level also impacts the resulting sizes of networks and using other levels could alter the quantitative and qualitative results of the analysis of sentiment word networks. We tried an additional significance ($\alpha = 0.05$), but results varied only minimally for the three larger networks.

3.2.9 Conclusions

In this work, we conducted a three-fold analysis of sentiment conveyed by spectator periodicals published during the Age of Enlightenment. We first set our focus on the relation between sentiment and annotations contained in our multilingual dataset and found discrepancies across languages depending on narrative forms and the discussed topics. In the second part of our analysis we constructed sentiment networks to assess the relation between different entities. Here, we found a rather critical interaction between periodicals, positive references across countries but negative references in domestic cases, indicating a tendency of authors to be more critical with fellow countrymen, as well as a general positive tone between authors. Finally, we constructed sentiment word networks to study the differences between positive and negative word usages. Specifically, we observed low transitivity and a tendency towards low degrees for negative words as compared to positive ones which have higher probabilities for mid-range degrees as well as for higher transitivity. The weak overrepresentation of positive words among the most central words in networks combined with our results of degree distributions indicate a rather mild to positive tone across all periodicals in which critiques comprise a plethora of different negative words.

Our work serves as a basis for promising future work, for which we plan to implement dedicated sentiment dictionaries for the language of the 18th century as well as the consideration of additional languages, such as Danish or Dutch spectator periodicals. Alternatively, other views on sentiment could be considered, for example, an existing machine learning approach to infer the emotional arcs of spectator periodicals [Reagan, Mitchell, et al., 2016]. Further, we are interested in detailed comparison of sentiment characteristics between today's media and the spectator periodicals from the Age of Enlightenment. In particular, we are interested in applying existing methods [Ferraz de Arruda et al., 2018] to study and compare the evolution of topics now and then.

Overall, this paper broadens our knowledge about spectator periodicals and demonstrates a distant reading method to analyze sentiment conveyed by them. The public availability of data and code for reproducibility lays the foundation for extended studies of spectator periodicals as well as other publications originating from the 18th century.

4 Case Study II: Multilingual Controversy

4.1 Related Publication and Author Contributions

Article 2: [Koncar, Walk, and Helic, 2021] Koncar, P., Walk, S. and Helic, D. (2021). Analysis and Prediction of Multilingual Controversy on Reddit. 13^{th} *ACM Web Science Conference*

My contributions to this article were the collection, cleaning and preprocessing of the (publicly available) data, the design and conduction of all experiments needed for the analysis and the prediction of controversy on Reddit as well as the visual preparation of the produced results. The interpretation of these results and the writing of the paper were done collaboratively by all three authors.

The main idea for this work, which is the multilingual analysis of controversy on Reddit, was proposed by me in order to have a relation to the first article presented in this thesis. Thus, the languages considered for this analysis coincide with the languages of the Spectator periodicals. This work is an adaption of a previously rejected paper in which the same authors already analyzed and predicted controversy on Reddit, but on a smaller scale focusing only on the English language. Similarly to this earlier version, all three authors contributed to the selection of methods, to the framing and writing of, as well as the interpretation and discussion of results in the paper presented here.

4.2 Analysis and Prediction of Multilingual Controversy on Reddit

4.2.1 Abstract

Social media users express their opinions about arbitrary subjects, including controversial matters such as the 2020 U.S. presidential election or climate change. Controversial topics typically attract user attention, which often lead to fruitful, but sometimes also heated discussions potentially segregating the community. Understanding features that are predictive of controversy in social media can improve moderation of communities and therefore the public discourse. In this paper, we analyze and predict controversy on the multilingual social platform Reddit. In particular, we compare a large set of textual and user activity features in controversial and non-controversial comments posted in six different languages. Using these features we perform a prediction task and study their predictive strengths for controversy. Our results indicate that, regardless of the language, controversial comments are harder to read, more negative and users follow up faster and more frequently to such comments. Moreover, with our prediction experiment (ROC AUC = 0.79) we find that across all languages user activity is the most predictive of controversy on Reddit. Our results contribute to an improved understanding of controversy in social media and can serve as a foundation for tools and models to automatically detect controversial content posted on such platforms.

4.2.2 Introduction

Discussions in social media frequently evolve around controversial topics, such as gun laws or abortion, that separate users into agreeing and disagreeing communities [Misra and Walker, n.d.]. While the discussion of controversial topics may lead to new insights [Vydiswaran et al., 2012], break down user stereotypes [Schommer-Aikins and Hutter, 2002], raise quality of collaborative efforts due to a higher diversity in debates [Shi et al., 2019], or increase (anonymous) user attention [Ziegele, Breiner, and Quiring, 2014], it is also a catalyst for disputes among opposing communities, eventually resulting in destructive discussions [Sumi, Yasseri, et al., 2011].

Due to the sheer amount of discussions taking place on social media platforms, researchers and practitioners already recognized the automatic identification of controversy as an indispensable tool for monitoring of such discussions. In practice, this allows moderators and mediators to intervene timely and resolve conflicts or to advise users to include references backing up their claims in debates on social media. Hence, the prediction of controversy has been studied extensively in existing research [Mishne, Glance, et al., 2006; Dori-Hacohen and J. Allan, 2013; Mejova et al., 2014; Siersdorfer et al., 2014; Ziegele, Breiner, and Quiring, 2014; Jang and J. Allan, 2018; Zielinski et al., 2018; Hessel and Lee, 2019]. However, most of these studies focused on English content and platforms such as Twitter [Guerra et al., 2013; Gruzd and Roy, 2014; Morales et al., 2015; Jang and J. Allan, 2018] or Wikipedia [Rad and D. Barbosa, 2012; Dori-Hacohen and J. Allan, 2013; Dori-Hacohen and J. Allan, 2015; Zielinski et al., 2018].

Research Question. In this work, we study controversy on the multilingual social news aggregation website Reddit, which our community has not yet widely analyzed with respect to controversy. In particular, we ask how commonly studied discussion features, such as *word usage* [Mejova et al., 2014; Siersdorfer et al., 2014; Hessel and Lee, 2019], *writing style* [Siersdorfer et al., 2014; Jang and J. Allan, 2018; Hessel and Lee, 2019], *sentiment* [Dori-Hacohen and J. Allan, 2013; Mejova et al., 2014; Zielinski et al., 2018], as well as *user involvement* [Mishne, Glance, et al., 2006; Ziegele, Breiner, and Quiring, 2014] are predictive of controversy on Reddit and whether their predictive properties carry over to languages other than English.

Approach. We define controversy as a discussion topic that separates users into agreeing and disagreeing groups, which captures a general controversy definition (e.g., from Wiktionary[1]) as a debate of contrary and opposing views. More precisely, on Reddit users can post all types of digital content (e.g., text, videos, pictures and links to other websites) that other users can comment on as well as up-vote or down-vote to indicate agreement or disagreement, respectively. Reddit automatically labels controversial submissions or comments based on these up- and down-votes. Here, we

[1] https://en.wiktionary.org/wiki/controversy

operationalize these controversy labels and analyze over 123 million English, German, French, Italian, Portuguese and Spanish comments posted on 50 different discussion boards (Subreddits).

We base our analysis on well-established features of social media users and postings previously studied in settings different to ours [Mishne, Glance, et al., 2006; Dori-Hacohen and J. Allan, 2013; Mejova et al., 2014; Siersdorfer et al., 2014; Ziegele, Breiner, and Quiring, 2014; Jang and J. Allan, 2018; Zielinski et al., 2018; Hessel and Lee, 2019]. In particular, we study word usage in comments and perform subgroup discovery to find words that have been distinctively used in controversial and non-controversial comments. We then compute various textual features (e.g., readability, POS tags) to study differences in writing styles and sentiment as well as structural and temporal features (e.g., number of replies, time to first reply) to investigate differences in user involvement. For comparison, we use statistical hypothesis testing to identify significant differences in feature distributions between controversial and non-controversial comments. Next, to answer our research question regarding the predictive strengths of individual features for controversy on Reddit, we perform a range of prediction experiments. Finally, we repeat the same analysis for six languages to learn more about cultural and linguistic differences regarding controversy in online social platforms. Note that all data and code used for our work is publicly available[2].

Findings & Contributions. Overall, our results indicate that, except for word usage, our features reflect controversy similarly across languages and cultural differences. For example, we find that controversial comments are harder to read and contain significantly higher fractions of negative sentiment as compared to non-controversial comments. We confirm results from previous research [Mishne, Glance, et al., 2006; Ziegele, Breiner, and Quiring, 2014] and find increased user participation in discussions of controversial topics also on Reddit. Our prediction experiments reveal that user involvement features are most predictive of controversy on Reddit, regardless of language.

Our findings contribute to a better understanding of controversy in social media by uncovering its universal properties across languages and by identifying features highly predictive of controversy. As such, our work

[2]`https://github.com/philkon/reddit-controversy`

can inform the development of novel and existing models to advance the automatic detection of controversy, hence improving the civil discourse by supporting moderators to timely intervene in user separating debates and disputes.

4.2.3 Related Work

Existing research studied controversy, for example, in the context of weblogs [Adamic and Glance, 2005; Mishne, Glance, et al., 2006], news articles [Choi, Yuchul Jung, and Myaeng, 2010; Mejova et al., 2014; Siersdorfer et al., 2014; Ziegele, Breiner, and Quiring, 2014], Twitter [Guerra et al., 2013; Gruzd and Roy, 2014; Morales et al., 2015; Jang and J. Allan, 2018], search engines [Yom-Tov, Dumais, and Q. Guo, 2014; Dori-Hacohen, Yom-Tov, and J. Allan, 2015; Koutra, Bennett, and Horvitz, 2015] or Wikipedia [Rad and D. Barbosa, 2012; Dori-Hacohen and J. Allan, 2013; Dori-Hacohen and J. Allan, 2015; Zielinski et al., 2018]. We now briefly recap some of the studies that are most relevant for our work.

Word Usage. Hessel and Lee [2019] predicted controversy of comments in six Subreddits by capturing textual content through TFIDF and word2vec models. Mejova et al. [2014] used crowdsourcing to manually identify controversial and non-controversial words, which they used to label news articles of 15 major U.S. news outlets. Siersdorfer et al. [2014] defined controversy of comments from YouTube and Yahoo! News based on comment ratings (i.e., up- and down-votes) and investigated words occurring in them. Similar to that, we analyze words that are distinctively used in both controversial and non-controversial comments but extend our analysis to six different languages and 50 different Subreddits in total.

Writing Style. Hessel and Lee [2019] and Siersdorfer et al. [2014] used basic textual features, such as the number of words or sentences as well as readability, to predict and analyze controversy in comments. Jang and J. Allan [2018] further incorporated POS tags to summarize stances in controversial Twitter discussions. We take up and extend such features to study their influence on controversy on Reddit.

Sentiment. Dori-Hacohen and J. Allan [2013] introduced a method to detect the controversy of arbitrary web pages. Authors used sentiment analysis as a baseline, which bares a high recall but performs worse than their proposed method. Mejova et al. [2014] found that sentiment and emotions are tempered in controversial news articles. Zielinski et al. [2018] detected controversy of Wikipedia articles by considering the sentiment of their respective talk pages. We also investigate and compare sentiment of comments for all six languages.

User Involvement. Ziegele, Breiner, and Quiring [2014] first conducted qualitative interviews with users who comment on news articles and then performed a quantitative analysis of user comments taken from Spiegel.de and Bild.de (both German news outlets) and their respective Facebook pages. Authors reported that controversy attracts much attention and provokes increased user engagement. Mishne, Glance, et al. [2006] support these findings and uncovered similar behavior for comments posted in various weblogs. In our work, we also study the impact of controversial comments on user involvement on Reddit. Further, we go beyond the listed works and combine user involvement with word usage, writing style and sentiment characteristics to automatically detect controversy in such comments.

4.2.4 Dataset and Descriptive Analysis

Dataset. On Reddit, users can express opinions about submissions (i.e., new threads) and comments (i.e., replies to existing threads) from other users by commenting as well as by up- or down-voting them. Each contribution has a score, which aggregates up- (counting as $+1$) and down-votes (counting as -1). Reddit automatically labels contributions as controversial if the number of up- and down-votes is both high (indicating that many users read the comment) and close to each other, resulting in a score close or equal to 0. The detailed procedure of how Reddit labels controversial contributions is not publicly available. However, the general idea behind this largely reflects the definition of controversy (e.g., from Wiktionary), specifying it as a debate or discussion that involves contrary and opposing views. Additionally, controversial contributions are visible to all users and can be filtered at the top of each comment listing. Thus, Reddit users can specifically search for (or ignore) controversial comments and are aware of comments controversiality.

Reddit is structured into Subreddits, each addressing a different topic, such as politics or sports. While the majority of Subreddits addresses English speakers, there are also Subreddits in which users communicate in other languages. Usually, these non-English Subreddits are dedicated to residents of a given country and cover multiple topics at once (e.g., in the German Subreddit r/Austria, Austrians discuss news and sports but also post memes). We analyze 50 different Subreddits, which either address English, French, German, Italian, Portuguese or Spanish speaking users, are thematically different and have varying numbers of users and comments to cover a wide spectrum of possible Subreddits. We use a publicly available dataset[3] including Reddit's controversy labels and extract all comments posted in these Subreddits during the year 2019.

In preprocessing, we remove all hyperlinks, HTML tags, comments posted in submissions with less than ten comments as well as comments containing less than one word (empty and deleted comments). Further, we automatically detect the language of all comments through the Compact Language Detector 2[4] and remove all comments for which the detected language deviates from the expected Subreddit language. As such, we obtain a total of 123,026,308 comments for our analysis. In Table 4.1, we provide detailed statistics of our dataset, including the Subreddits we chose for our analysis. Note that, due to the smaller number of non-English users, the majority of comments is in English. We selected English Subreddits based on the variety of topics (e.g., news, sports, politics) and Subreddits in other languages based on their activity levels[5].

Preliminary Descriptive Analysis. In Figure 4.1, we depict selected characteristics of our dataset. Regarding the length of Submissions (i.e., the number of comments posted in a discussion thread), we observe that the majority of submissions on Reddit receive smaller numbers of comments

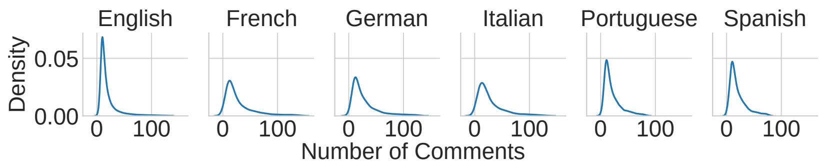
[3] https://files.pushshift.io/reddit

[4] `https://github.com/CLD2Owners/cld2`; We keep all comments for which the Compact Language Detector 2 only detected one language with at least 90% confidence.

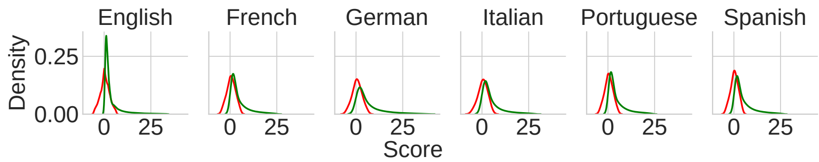
[5] Specifically, we manually checked available Subreddits in respective (non-English) languages and selected those with frequent user activity and at least 1,000 members.

Table 4.1: **Dataset Statistics.** The table lists an overview of our dataset, including the number of unique users, submissions, comments, controversial comments as well as the Subreddits (numbers in brackets show the ratio of controversial comments in respective Subreddits) we extracted comments from, respectively for each language.

Language	**# Users**	**# Submissions**	**# Comments**	**# Controversial**	**Subreddits**
English	5, 225, 561	2, 029, 507	115, 186, 784	3, 176, 622 (2.76%)	r/AskReddit (0.99%), r/facepalm (3.46%), r/funny (3.78%), r/me_irl (2.26%), r/nfl (3.47%), r/philosophy (5.18%), r/politics (4.20%), r/sports (4.50%), r/StarWars (5.42%), r/technology (5.98%), r/todayilearned (4.27%), r/worldnews (7.84%)
French	34, 092	33, 752	1, 431, 249	87, 638 (6.12%)	r/france (6.27%), r/FranceLibre (2.47%), r/jeuxvideo (0.22%), r/montreal (5.71%), r/Quebec (5.47%), r/rance (1.47%)
German	67, 546	49, 707	1, 867, 328	94, 256 (5.05%)	r/Austria (4.87%), r/Dachschaden (4.44%), r/de (5.34%), r/de_IAmA (1.61%), r/Finanzen (2.52%), r/FragReddit (1.58%), r/ich_iel (1.37%), r/rocketbeans (16.22%), r/wasletztepreis (1.67%)
Italian	16, 621	13, 145	669, 072	23, 008 (3.44%)	r/Italia (1.89%), r/italy (3.58%), r/ItalyInformatica (0.65%), r/Libri (0.07%), r/litigi (1.44%)
Portuguese	43, 995	65, 488	1, 765, 419	74, 820 (4.24%)	r/brasil (3.62%), r/BrasildoB (2.90%), r/brasilivre (3.99%), r/circojeca (0.38%), r/portugal (5.75%), r/PORTUGALCARALHO (2.45%), r/PrimeiraLiga (7.10%)
Spanish	57, 713	69, 140	2, 106, 456	76, 394 (3.63%)	r/argentina (3.07%), r/chile (4.13%), r/Colombia (3.60%), r/es (7.94%), r/espanol (0.19%), r/mexico (3.93%), r/podemos (1.83%), r/spain (10.48%), r/uruguay (5.80%), r/vzla (1.65%), r/yo_elvr (0.68%)



(a) KDE of Number of Comments by Submission



(b) KDE of Scores (green = non-controversial; red = controversial)

Figure 4.1: **Characteristics of Our Dataset.** The figure illustrates selected key characteristics of our dataset, including kernel density estimations (KDE) of submission length as well as comment scores, respectively for each language contained in our dataset. We observe that the majority of submissions receive only minimal attention, whereas a small number of submissions receive more comments (cf. Figure 4.1a; distribution truncated at **120** comments which is still above the **95**$^{\text{th}}$ percentile). As expected based on the definition of controversy by Reddit, we report higher probabilities for scores around zero for controversial comments (red color) compared to non-controversial comments (green color) for all six languages (cf. Figure 4.1b; distributions truncated at -5 and 30 which still is below the **5**$^{\text{th}}$ and above the **95**$^{\text{th}}$ percentile, respectively).

(mean submission length over all Subreddits = 54.42; median submission length over all Subreddits = 14.0), which is independent from language (cf. Figure 4.1a). Only a small number of submissions receive higher numbers of comments for all six languages, suggesting that not all contributions can attract substantial amounts of attention.

As expected, we observe that controversial comments have scores around zero (cf. Figure 4.1b; mean score over all Subreddits = 0.60; median score over all Subreddits = 0.00) independent from language. Non-controversial comments across all analyzed languages have a median score of 2.00 and a mean score of 18.72, indicating a rather positive attitude of users on Reddit.

In Table 4.1, we list the ratios of controversial comments for each of our 50 Subreddits. In general, we observe rather small ratios (ranging between 16.22% and 0.07%) of controversial comments, indicating that most discussions on Reddit do not lead to conflicts or disputes among its users, which again supports our assumption of a general positive mood on the platform.

For English, most controversy can be found in the Subreddit *r/worldnews* (7.84%). We argue that the world affairs discussed in this Subreddit across multiple continents and countries invite a plethora of different views, which increases the probability for conflicts among users. On the contrary, *r/AskReddit*, a Subreddit in which users can pose arbitrary questions to fellow users, has the smallest ratio (0.99%) of controversial comments among English Subreddits. This indicates an open-minded and welcoming community, willing to answer a wide range of user questions.

For other languages, Subreddits that are dedicated to languages, countries or cities are most controversial (e.g., *r/france*, *r/es*). Two exceptions are *r/rocketbeans* for German Subreddits and *r/PrimeriaLiga* for Portuguese Subreddits which are most controversial for respective languages. The former addresses viewers of the German live streaming channel *Rocket Beans TV*, which deals with topics related to the gaming industry or other issues focusing on a younger audience, potentially involving a more reckless user behavior and, thus, resulting in more controversial comments. In the latter, users specifically discuss the *Primeira Liga*, the highest division of the Portuguese football league system. Here, we argue that the rivalry among fans is clearly reflected in this Subreddit's controversial comments.

To infer the influence of user activity on resulting numbers of controversial comments, we compute Pearson correlation coefficients between the total number of comments and the number of controversial comments in submissions, respectively for each Subreddit. We find significantly positive correlations for all Subreddits (exceptions are *r/Libri*, *r/espanol* and *r/jeuxvideo* for which the correlation is non-significant), but with varying strengths. For example, $\rho = 0.89$ for *r/worldnews* and $\rho = 0.84$ for *r/chile*, but $\rho = 0.27$ for *r/ItalyInformatica* and $\rho = 0.31$ for *r/rance*. This suggests that the ratio of controversial comments is not only impacted by activity levels (e.g., the more attention the more controversial comments), but also depends on topics discussed and their individual communities.

4.2.5 Empirical Results

We first investigate word usage and then analyze a set of 23 writing style, sentiment and user involvement features for which we report median and mean differences between distributions of controversial and non-controversial comments and use statistical hypothesis testing to assess whether these differences are significant.

Word Usage

Motivated by existing research [Mejova et al., 2014; Siersdorfer et al., 2014; Hessel and Lee, 2019], which found that certain words (e.g., "abuse", "killing" or "race") are more related to controversy than others, we analyze word usage differences between controversial and non-controversial comments posted on Reddit. For that, we perform subgroup discovery and adopt the method from Hofland and Johansson [1982], which is based on contingency tables and chi-squared (χ^2) tests to assess which words have significantly different distributions in two text corpora. More precisely, for each language we look at the sets of the top 500 words (after removing stop words[6]) included in controversial and non-controversial comments and build the union of those two sets. The union of the top words contains 581 words for English, 565 words for French, 568 words for German, 586 words for Italian, 560 words for Portuguese and 580 words for Spanish. Next, for each language and each word from the respective union we build a 2×2 contingency table, which keeps the count of a given word, as well as the total count of all other words in both controversial and non-controversial comments. The null hypothesis of the χ^2 test (which we perform with Yates Correction [Yates, 1934]) states that the occurrence of a given word is independent of the controversy of the comment. Hence, words for which we can reject this null hypothesis are used distinctively in either controversial or non-controversial comments.

Results. In Figure 4.2, we depict the top 25 significant words with regard to their χ^2 values and to their relative frequencies in controversial and non-controversial comments (to decide where their usage is significantly

[6] Link to stop words lists: `https://github.com/Alir3z4/python-stop-words`

(a) English (b) French (c) German

(d) Italian (e) Portuguese (f) Spanish

Figure 4.2: **Word Usage Results.** The figure depicts the top **25** words according to χ^2 (represented through word sizes) respectively for controversial (red color) and non-controversial (green color) comments and each of our six languages. We observe differences in discussed topics not only between controversial and non-controversial topics but also between languages. For example, English (cf. Figure 4.2a) controversial comments focus more on politics and foreign countries, whereas French (cf. Figure 4.2b), German (cf. Figure 4.2c) and Italian (cf. Figure 4.2d) controversial comments address racism and immigration.

higher; all p-values < 0.0005), respectively for each of the six languages. We observe that, independent from language, top controversial words reflect topics that have been discussed in broader public during 2019, whereas non-controversial words cannot be related to topics that easily. These findings are similar to those from previous studies [Mejova et al., 2014; Siersdorfer et al., 2014; Hessel and Lee, 2019]. For example, English top controversial words include political persons, such as *Bernie Sanders* and *Donald Trump*, but also the foreign countries *China* and *Israel*. Top controversial words of other languages relate to racism, fascism, immigration, religion or gender equality.

We find that top words largely reflect local issues of respective countries. For example, English top words, to a large extent, address U.S. politics as well as foreign countries in the focus thereof. Note that *r/politics* and *r/worldnews* have the highest absolute numbers of English controversial comments (845, 089 and 682, 402 respectively), explaining the majority of top words being related to politics. In the case of French, controversial comments address *Islam* and *Moslems*, which may be caused by the lengthy series of terrorist attacks. For German top controversial words, we observe that an increasing rightward shift of society is reason for disputes. Similarly, Italians discuss rightwing politics as well as migration, which could be due to the ongoing refugee crisis. Portuguese top words focus on politics in Portugal as well Brazil, but also include *Benfica*, which stems from the Portuguese football club *Sport Lisboa e Benfica*, suggesting strong rivalry among sport fans in Portugal. Spanish controversial top words address gender equality as well as abortion, which has very restrictive laws in Latin America [Galli, 2020].

Writing Style

We now set our focus on how writing style of comments on Reddit reflects controversy. For that, we follow existing research [Siersdorfer et al., 2014; Jang and J. Allan, 2018; Hessel and Lee, 2019] and consider differences in text length, readability and POS tags between controversial and non-controversial comments.

Statistical Hypothesis Tests. To assert whether the difference between the two types of comments is significant, we use statistical hypothesis testing. In particular, our null hypothesis assumes equal distributions for controversial and non-controversial comments. Thus, we first perform the Brown-Forsythe test (at a significance level $\alpha = 0.05$) to assess the equality of variances between distributions. Based on these results, we then use the median test in cases of unequal variances, and in case of equal variances we use the Mann-Whitney-Wilcoxon test, which has a higher statistical power (cf. [Feltovich, 2003]). We select these tests as they make no assumptions about the underlying distributions (manual inspection of kernel density estimation plots revealed many different shapes). To counteract the problem of multiple comparisons, we perform the Bonferroni correction [Bonferroni, 1935] reducing the commonly used significance level $\alpha = 0.05$ for the entire set of n comparisons to α/n. In our work, we test 23 features (including features of subsequent sections), hence, $n = 23$ and $\alpha \approx 2.17 \times 10^{-3}$. Note that we provide a detailed overview of hypothesis test results as well as differences in distribution medians and means for all features and languages in Table 4.2.

Text Length. For each comment we extract the *number of characters*, the *number of syllables*, *the number of words* and the *number of sentences*. Further, we investigate the *characters to sentences ratio* and the *words to sentences ratio*.

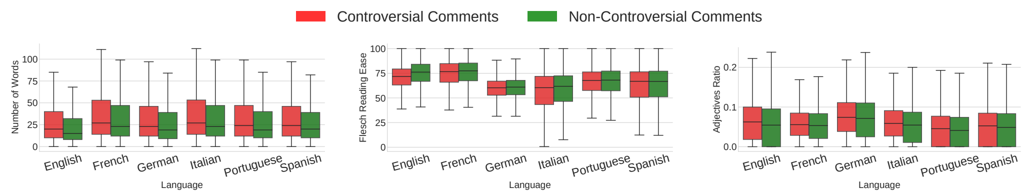
Readability. We determine the readability of comments by computing the *Flesch Reading Ease* [Flesch, 1948]. Since the original formula is intended for English only, we use its respective derivatives for French [Kandel and Moles, 1958], German [Amstad, 1978], Italian [Franchina and Vacca, 1986], Portuguese [Aguirre et al., 2017] and Spanish [Fernández Huerta, 1959] comments. These measurements are comparable and represent the reading difficulty of a text by a score ranging between 0 and 100, where higher values indicate easier to read texts and lower values indicate harder to read texts. Note that in this case we limit our dataset to comments with a minimum of 100 words as the readability formulas might return inaccurate values otherwise. Hence, for readability, we analyse $5,363,089$ English, $122,204$ French, $120,578$ German, $54,832$ Italian, $121,549$ Portuguese as well as $136,772$ Spanish comments.

POS Tags. Using Spacy's POS tagger[7], we extract the *ratio of nouns*, the *ratio of verbs*, the *ratio of adjectives*, the *ratio of adverbs* and the *ratio of pronouns* of each comment.

Results. We illustrate distributions of selected writing style features for controversial and non-controversial comments as well as each language in Figure 4.3. Starting with comment text length, we find that controversial comments are significantly longer across all languages (cf. Figure 4.3a). The number of characters, syllables and sentences positively correlate with the number of words (all Pearson $\rho > 0.754$ with p-values < 0.0005), further strengthening this observation. Similarly to comment length, controversial comments of all languages have significantly higher characters to sentences and words to sentences ratios.

In Figure 4.3b, we illustrate distributions of the Flesch Reading Ease for controversial and non-controversial comments, respectively for each lan-

[7] `https://spacy.io` (version used: 2.2.3)



(a) Number of Words (b) Flesch Reading Ease (c) Ratio of Adjectives

Figure 4.3: **Writing Style Results.** The figure depicts box plots for selected writing style features comparing non-controversial (green color) and controversial (red color) comments for all six languages. Horizontal black lines indicate the median and the first and third quartile. Whiskers indicate minimum and maximum values still within 1.5 interquartile ranges. Note that we do not depict outliers for better representation of data and that these characteristics apply to remaining box plots in this paper as well. We report that controversial comments have more words (a), are harder to read (b) and contain more adjectives (c) compared to non-controversial comments in any of the six languages.

guage. Overall, we find that comments on Reddit are between fairly difficult and fairly easy to read. According to median differences, controversial comments are harder to read as compared to non-controversial comments for all languages. The median difference is significant for English, French as well as German (all p-values $<$ our corrected α), but non-significant for Italian (p-value $= 0.007$), Portuguese (p-value $= 0.096$) and Spanish (p-value $= 0.168$). These results in combination with text length features suggest a more complex content for controversial comments.

Finally, we report significantly higher ratios of nouns (except for English with a p-value of 0.004) and adjectives (cf. Figure 4.3c) in controversial comments across all languages. Regarding the ratio of verbs, we observe significantly lower ratios for English, Portuguese and Spanish controversial comments as well as significantly higher ratios for German and Italian controversial comments. For French comments, there is no significant difference (p-value $= 0.208$) in the ratio of verbs. The ratio of adverbs is significantly higher for English, German and Spanish controversial comments, whereas it is not significantly higher for French (p-value $= 0.016$), Italian (p-value $= 0.346$) and Portuguese (p-value $= 0.211$) comments. The ratio of pronouns is significantly lower in English and French controversial comments, while it is significantly higher in German and Portuguese controversial comments. We observe no significant differences in the ratio of pronouns for Italian (p-value $= 0.004$) and Spanish (p-value $= 0.012$) comments. Overall, our POS tag features imply that controversial comments are written more impersonally than non-controversial comments.

Sentiment

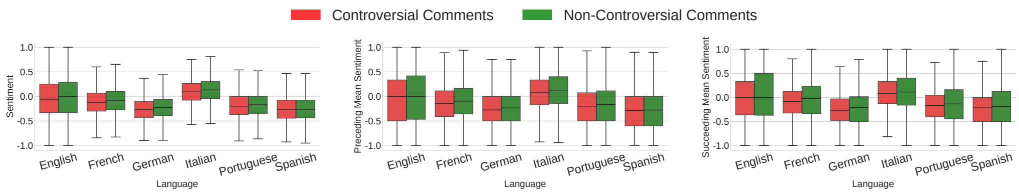
According to previous research [Dori-Hacohen and J. Allan, 2013; Mejova et al., 2014; Zielinski et al., 2018], sentiment is predictive of controversy. To compute the sentiment of comments posted on Reddit, we rely on existing sentiment dictionaries that have already been created[8] and evaluated for our six languages in existing work [Yanqing Chen and Skiena, 2014]. Using the respective sentiment dictionaries for our languages, we compute the sentiment s of each comment with $s = (W_p - W_n)/(W_p + W_n)$, where W_p is the number of positive words in a comment and W_n is the number of negative words in a comment. Hence, s ranges between -1 and $+1$, where values close to -1 are considered as negative, values close to $+1$ as positive, and values close to zero as neutral sentiment.

Besides computing the sentiment for each comment, we also investigate the *preceding mean sentiment* (i.e., the mean sentiment over preceding comments) and the *succeeding mean sentiment* (i.e., the mean sentiment over succeeding comments).

Similar to writing style features (cf. Section 4.2.5), we only consider comments with at least 100 words for the analysis of sentiment to prevent inaccurate values. Further, we use the same hypothesis testing approach to assess the significance of differences between controversial and non-controversial comments.

Results. Figure 4.4 depicts the distributions for controversial and non-controversial comments and each language. In general, we observe that the

[8] Authors extracted most frequent words of Wikipedia articles and created a knowledge graph to combine similar words of different languages through Wiktionary, machine translation (via Google translate), transliteration links and WordNet. Starting with sentiment of English vertices based on an existing dictionary, authors propagated sentiment to vertices of other languages and created dictionaries for 136 languages.



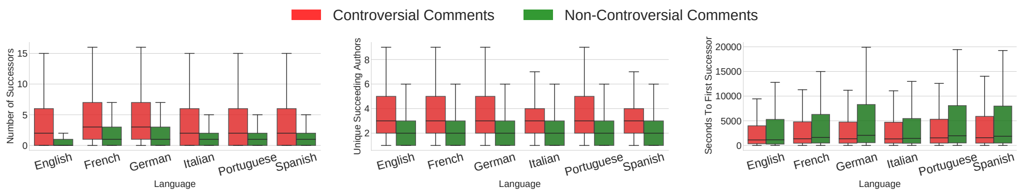
(a) Sentiment (b) Preceding Mean Sentiment (c) Succeeding Mean Sentiment

Figure 4.4: **Sentiment Results.** The figure depicts box plots for our sentiment features comparing non-controversial (green color) and controversial (red color) comments for all six languages. We find that, independent form language, sentiment is significantly more negative in discussion evolving around controversial comments than in those around non-controversial ones.

median sentiment of comments is rather negative for all languages, except for Italian, where the median sentiment is slightly positive. However, controversial comments have a significantly more negative sentiment compared to non-controversial comments for all languages, except Spanish (p-value $= 0.327$). The preceding mean sentiment (cf. Figure 4.4b) suggests that comments preceding controversial ones are also significantly more negative than those preceding non-controversial ones. This difference is not significant for Spanish comments (p-value $= 0.188$). The succeeding mean sentiment (cf. Figure 4.4c) is significantly more negative for controversial comments in all languages, indicating a general negative attitude for discussion threads with controversial comments.

User Involvement

Previous studies [Mishne, Glance, et al., 2006; Ziegele, Breiner, and Quiring, 2014] have shown that controversial topics attract a lot of attention in online discussions. Hence, we investigate whether users of our Subreddits exhibit a similar behavior and compute eight features capturing structural and temporal aspects of discussion threads. We inspect the *number of predecessors* (i.e., the number of preceding comments), the *number of successors* (i.e., the number of succeeding comments), the *number of unique preceding users* as well as the *number of unique succeeding users*. Analogously, we analyze the *time from predecessor*, the *mean time between predecessors* (i.e., the mean



(a) Number of Successors (b) Unique Succeeding Users (c) Time to First Successor

Figure 4.5: **User Involvement Results.** The figure depicts box plots for selected user involvement features comparing non-controversial (green color) and controversial (red color) comments for all six languages. We observe that controversial comments attract more attention (a and b) more quickly (c) compared to non-controversial comments, regardless of language.

time between all preceding comment), the *time to the first successor* and the *mean time between successors* (i.e., the mean time between all succeeding comments), providing insights into how fast comments attract attention. Note that we measure time in seconds for all temporal features.

We use the same hypothesis testing approach described in the Writing Style Section (cf. Section 4.2.5) to infer whether differences between controversial and non-controversial comments are significant.

Results. We depict distributions for selected user involvement features in Figure 4.5. The number of predecessors is significantly lower for controversial comments in any language (except English for which it is significantly higher), indicating that such comments are posted closer to the original submission of the discussion. Similarly, the number of unique preceding users is significantly lower for controversial comments in all languages except German (p-value $= 0.025$), Spanish (p-value $= 0.149$) and English (here the number is significantly higher). Controversial comments receive more attention as the number of successors and the number unique succeeding users is significantly higher for them across all languages.

The time from predecessor suggests that controversial comments in any language are posted significantly faster to preceding comments than non-controversial ones. Similar to that, the mean time between predecessors is significantly lower for controversial comments across languages. The time to the first successor (cf. Figure 4.5c) indicates that controversial comments are attracting attention significantly faster. Further, the mean time between successors is significantly lower for controversial comments, except for French (p-value = 0.128) and Italian (p-value = 0.023).

Summary of Empirical Results

For word usage, we find that controversial comments often include words related to topics frequently addressed in the public discourse of respective language, such as the refugee crisis in Italy or abortion in Latin America. On the other hand, non-controversial comments in all languages include more moderate words or words that cannot clearly be related to a topic, such as "time" or "friend".

For writing style, sentiment and user involvement features, we find significant differences for 122 out of 138 cases according to our hypothesis tests. We list median and mean differences for each feature and language in Table 4.2. Our results for writing style and sentiment differences suggest that controversial comments are significantly longer, harder to read, more impersonal and more negative than non-controversial comments. Our user involvement analysis reveals that controversial comments attract more user attention than non-controversial comments and that users are quicker to follow up on a controversial comment.

Overall, our results on writing style, sentiment and user involvement are similar across languages, indicating that discussions on controversial issues on Reddit follow common modalities shared between languages. However our word usage analysis suggests a different relevance of topics across languages and countries.

4.2.6 Predicting Controversy

Experimental Setup

Based on our empirical results, we now conduct a prediction task and investigate the predictive power of various features for controversy on Reddit. Note that the aim of this prediction task is not to achieve best performance, but rather to investigate what features are most predictive of

Table 4.2: **Distribution Differences and Hypothesis Test Results.** The table lists the median and mean (in brackets) differences between controversial and non-controversial distributions for each of our **23** features and each of the six languages. Depending on the differences in variances between non-controversial and controversial comment distributions, we either use the median test or the Mann-Whitney-Wilcoxon test (indicated with colored cells). Underlined values indicate significance (p-values of hypothesis tests $<$ our Bonferroni corrected $\alpha \approx 2.17 \times 10^{-3}$). We can reject our null hypothesis for **122** out of **138** tests.

Feature	English	French	German	Italian	Portuguese	Spanish
Number of Characters	29.0 (35.764)	25.0 (20.529)	29.0 (28.024)	26.0 (28.142)	26.0 (28.53)	24.0 (18.076)
Number of Syllables	7.0 (8.712)	5.0 (4.946)	8.0 (7.579)	8.0 (9.494)	8.0 (9.376)	8.0 (5.873)
Number of Words	5.0 (5.431)	4.0 (3.094)	4.0 (3.932)	4.0 (4.266)	5.0 (4.638)	4.0 (3.022)
Number of Sentences	0.0 (0.306)	0.0 (0.097)	0.0 (0.199)	0.0 (0.131)	0.0 (0.269)	0.0 (0.071)
Characters to Sentences Ratio	7.455 (7.699)	6.0 (8.304)	7.0 (6.994)	7.667 (10.337)	5.75 (5.746)	9.0 (10.37)
Words to Sentences Ratio	1.0 (1.068)	1.0 (1.223)	0.75 (0.936)	1.0 (1.504)	0.929 (0.864)	1.571 (1.578)
Flesch Reading Ease	−4.41 (−4.116)	−0.862 (−1.476)	−0.675 (−0.743)	−1.409 (−1.696)	−0.329 (0.085)	0.112 (−0.309)
Ratio of Nouns	0.0 (−0.005)	0.003 (0.002)	0.002 (0.0)	0.004 (0.003)	0.0 (0.002)	0.003 (0.003)
Ratio of Verbs	0.0 (−0.001)	0.001 (0.0)	0.002 (0.001)	0.002 (0.002)	−0.002 (−0.002)	−0.004 (−0.003)
Ratio of Adjectives	0.008 (0.005)	0.002 (0.002)	0.003 (0.003)	0.004 (0.003)	0.004 (0.003)	0.004 (0.003)
Ratio of Adverbs	0.002 (−0.0)	0.0 (0.0)	−0.001 (−0.001)	−0.0 (−0.001)	−0.001 (−0.001)	0.002 (0.0)
Ratio of Pronouns	−0.006 (−0.006)	−0.004 (−0.003)	0.002 (0.001)	0.001 (0.001)	0.001 (−0.001)	0.0 (−0.001)
Sentiment	−0.059 (−0.031)	−0.029 (−0.033)	−0.042 (−0.043)	−0.04 (−0.04)	−0.028 (−0.023)	0.0 (−0.002)
Preceding Mean Sentiment	0.0 (−0.03)	−0.046 (−0.039)	−0.041 (−0.041)	−0.038 (−0.041)	−0.033 (−0.026)	−0.006 (−0.002)
Succeeding Mean Sentiment	0.0 (−0.078)	−0.06 (−0.042)	−0.054 (−0.059)	−0.028 (−0.027)	−0.032 (−0.038)	−0.025 (−0.04)
Number of Predecessors	1.0 (0.398)	0.0 (−0.473)	0.0 (−0.274)	0.0 (−0.288)	0.0 (−0.447)	0.0 (−0.205)
Number of Successors	2.0 (3.413)	2.0 (2.978)	2.0 (3.018)	1.0 (2.507)	1.0 (2.912)	1.0 (2.429)
Unique Preceding Users	1.0 (0.485)	−1.0 (−0.194)	0.0 (−0.086)	0.0 (−0.106)	0.0 (−0.068)	0.0 (−0.048)
Unique Succeeding Users	1.0 (0.328)	1.0 (0.972)	1.0 (0.86)	1.0 (0.845)	1.0 (1.06)	1.0 (0.835)
Time From Predecessor	−2768.0 (−17558.083)	−1677.0 (−10135.157)	−2602.0 (−14428.74)	−887.0 (−10041.005)	−2116.5 (−13544.689)	−2084.0 (−14370.263)
Time To First Successor	−30.0 (−7102.537)	−260.0 (−3316.991)	−697.0 (−6166.192)	−97.5 (−2917.029)	−426.0 (−5341.833)	−303.0 (−6640.838)
Mean Time Bet. Predecessors	−454.625 (−3028.215)	−2010.875 (−4830.73)	−1465.917 (−4552.319)	−4522.5 (−8133.433)	−1430.9 (−6668.522)	−2483.5 (−8757.049)
Mean Time Bet. Successors	333.486 (−6697.816)	−37.4 (−3719.733)	−911.786 (−6424.904)	107.635 (−3577.459)	−323.429 (−5278.054)	−168.264 (−5918.054)

controversy. As such, our proposed method needs to be easily interpretable and may not reach the performance of more sophisticated approaches, such as neural networks.

Features. We use all features from the previous word usage (cf. Section 4.2.5), writing style (cf. Section 4.2.5), sentiment (cf. Section 4.2.5) and user involvement (cf. Section 4.2.5) analyses. In the case of word usage features, we count the number of the top 25 controversial and non-controversial words in comments, respectively for each language, and report these features as the *number of controversial words* and the *number of non-controversial words*. Additionally, we include the language of comments as well as the Subreddit they had been posted in. Note that we use one-hot-encoding to transform categorical features (language and Subreddit) and that we apply robust scaling (due to the presence of outliers) to numerical features.

Prediction Samples. We remove all comments that have less than 100 words in order to exclude unreliable textual features. This leaves us with a total of 5,919,024 comments (218,053 controversial and 5,700,971 non-controversial). To address the unbalance between classes, we perform random undersampling and finally obtain 436,106 comments (376,302 English; 16,796 French; 14,092 German; 4,540 Italian; 12,204 Portuguese; 12,172 Spanish).

Model and Evaluation. We employ *Gradient Boosted Decision Trees* (GBDTs) as implemented in scikit-learn[9]. We tune hyper parameters of the GBDTs through a grid search[10] and evaluate our model by using ten-fold cross-validation for which we report mean ROC AUC values over the ten cross-validation folds.

Prediction Results

We report a mean ROC AUC of 0.79, indicating that we can achieve moderate prediction performance and improve a random baseline of 0.50 by 0.29. In Figure 4.6, we depict the importance of selected features (we exclude all features with importance smaller than 0.01) to assess which of the

[9] `https://scikit-learn.org/0.23` (version used: 0.23.2)
[10] We include parameters for the best model in our GitHub repository.

Figure 4.6: **Feature Importances.** The figure illustrates mean feature importances over the ten-fold cross-validation. Our model achieves a mean ROC AUC of **0.79**. Note that, for visualization purposes, we only show features with an importance of at least **0.01**. The error bars indicate **95**% bootstrap confidence intervals. We observe that involvement features (pink color) are most predictive, while other features, such as writing style (orange color), contribute less to the prediction of controversy on Reddit.

previously analyzed features are most predictive of controversy. Here, we find involvement features to carry most predictive strengths. In particular, the number of unique succeeding users (0.32), the number of predecessors (0.18), the mean time between predecessors (0.09), the time from predecessor (0.06), the Subreddit *r/AskReddit* (0.06) as well as the number of controversial words (0.05) are most predictive. Other features have importance values equal or smaller than 0.03.

These results suggest that the more users participate in a discussion, the likelier it is to include a controversial comment. Other aspects of comments, such as the words it contains or the sentiment it conveys, are less important when predicting controversy on Reddit. Further, we see that languages have no or only minimal influence on prediction performance, suggesting that controversy on Reddit behaves similarly across languages. Interestingly, only three Subreddits are at least to some extent predictive of controversy. These include *r/AskReddit*, *r/politics* and *r/worldnews*. Combining these findings with ratios of controversial comments per Subreddit (cf. Table 4.1), we argue that it is either particularly unlikely (*r/AskReddit*) or particularly likely (*r/politics*, *worldnews*) for a comment to be controversial in these Subreddits.

4.2.7 Discussion

We now connect our results to our research question and compare our findings to those from existing controversy studies [Mishne, Glance, et al., 2006; Dori-Hacohen and J. Allan, 2013; Mejova et al., 2014; Siersdorfer et al., 2014; Ziegele, Breiner, and Quiring, 2014; Jang and J. Allan, 2018; Zielinski et al., 2018; Hessel and Lee, 2019] conducted in other contexts.

Word Usage. We found that controversial comments often include terms that are frequently discussed in the public discourse, such as "abortion", "society" or "politics". Contrary, non-controversial comments include terms, such as "friend" or "mother", that cannot be related to a specific topic that easily. In particular, we find that our extracted top words coincide with controversial and non-controversial words manually extracted by crowdworkers from news articles provided by NewsCred [Mejova et al., 2014]. We observe that our extracted controversial and non-controversial top words are respectively related to words frequently found in accepted and not accepted comments posted on YouTube or Yahoo! News [Siersdorfer et al., 2014]. This suggests that controversy on Reddit behaves similarly to controversy in other contexts. Further, we found that language communities on Reddit not only discuss topics related to their respective countries, but also global issues, implying a possible domain transfer across cultures. This is different when studying Subreddits in a micro-perspective, where text features, such as TFIDF or word embeddings, are very community specific [Hessel and Lee, 2019].

Writing Style. We observed that controversial comments are longer, harder to read and more impersonal compared to non-controversial comments. We argue that this is due to an overall more complex writing style used in controversial comments to persuade or deceive opinions of other users. Similar findings were obtained by Tan et al. [2016], where authors analyzed *r/ChangeMyView*, a Subreddit specifically dedicated to the understanding of contrasting views.

Sentiment. We saw that controversial comments convey a more negative sentiment than non-controversial comments. Further, comments preceding and succeeding a controversial comment also have more negative sentiment, suggesting that discussion threads with controversial comments have a general negative mood. Our findings are again similar to existing controversy studies in the context of edit wars on Wikipedia [Zielinski et al., 2018] or online news articles [Mejova et al., 2014].

In future work, we are interested in further analyzing the particular negative sentiment associated with controversial comments. For example, we would also expect controversial comments in which their authors write positive about a controversial topic (e.g., a user is for abortion and not against it) and, thus, convey a positive sentiment. Also, the investigation of a potential bias introduced by sentiment dictionaries, in which terms related to controversies are generally labelled as negative, can be promising for future work.

User Involvement. Similarly to the existing studies [Mishne, Glance, et al., 2006; Ziegele, Breiner, and Quiring, 2014], we found that users on Reddit feel particularly attracted to controversial comments. In our data, we observed correlations between the temporal and structural user involvement features. For example, there are weak negative correlations between the time to first successor and the number of successors (Spearman $\rho = -0.164$; p-value < 0.0005), suggesting that the faster a controversial comment gets its first reply, the more total replies it receives. However, this result may be somewhat obfuscated due to Reddit's way of displaying comments (i.e., the easy reachability of lower discussion tree levels), where deeper levels of a discussion tree do not attract as much attention as the lower levels due to the position bias in users perception of Web pages [Lamprecht et al., 2017]. Further, additional clicks are required to expand discussion branches, requiring additional effort from users to see comments further along the discussion trees.

Predicting Controversy. User involvement features discriminate the most between controversial and non-controversial comments. Especially the number of preceding comments and the number of unique succeeding authors indicate the influence of increased exposure (through commenting right from the beginning of a discussion) on comment controversy. This also indicates that comments posted at a later time of a discussion might get lost in the crowd. Overall, we suggest to consider word usage, writing style, sentiment and Subreddits next to user involvement features when predicting controversy, as they add useful information to controversy prediction, similar to what has been shown in existing studies [Dori-Hacohen and J. Allan, 2013; Zielinski et al., 2018; Hessel and Lee, 2019]. To assess the performance gain of these features in our context, we redo our prediction experiment (cf. Section 4.2.6), but this time only include user involvement, Subreddit as well as language features. We achieve a ROC AUC of 0.77 and, thus, observe a gain of 0.02 when including word usage, writing style and sentiment features.

Multilingual Controversy. The insignificant importance of languages in our prediction model further confirms that controversy on Reddit is language agnostic. However, we are aware that this might be due the unbalanced distribution of languages in our prediction samples (the majority of comments is in English). Thus, to control for the number of comments per language we perform an additional prediction experiment. In particular, we randomly draw 2, 270 (this is the number of Italian controversial samples and also the minimum across languages) controversial comments for each language for which we then repeat the experiment as described in Section 4.2.6. With this prediction experiment we confirm our previous findings as we again report that language is not important (all importance values < 0.01) and user involvement features are once more most predictive. Overall, this indicates that existing controversy detection models can be transferred to other languages, as long as they incorporate features similar to our user involvement features and are not based on words specific to a certain language.

Limitations. The scope of our work is limited to 50 Subreddits with comments posted during the year 2019. Thus, we observe a relatively short time period and also only a small fraction of all Subreddits hosted by Reddit. However, we argue that our results are representative for Reddit, as our analysis included some of the largest and most active Subreddits in their respective languages. Further, our work investigates controversy as defined by Reddit, which is different from other definitions based, for example, on Twitter [Pennacchiotti and Popescu, 2010; Guerra et al., 2013; Gruzd and Roy, 2014; Morales et al., 2015]. We leave a comparison of different controversy definitions to future work. Also, we want to note that we analyzed correlations and not causality. For example, a comment receiving much attention may not necessarily be controversial as it can also be popular due to other reasons. We leave the study of the causality to future work. Finally, we did not consider sarcasm and its influence on sentiments expressed in comments, a separate and non-trivial problem, which is out of scope for this paper.

4.2.8 Conclusions

In this paper we analyzed word usage, writing style, sentiment and user involvement in the context of controversial and non-controversial comments posted in six languages on Reddit. We performed subgroup discovery and computed a set of 23 features for comments posted in 50 different Subreddits and used statistical hypothesis testing to infer differences between controversial and non-controversial comments. Most notably, we found that users are more engaged and that they write more complex and negative comments when debating controversial issues. We observed this behavior for all languages, suggesting that controversy on Reddit is universal across languages, except for the fact that languages reflect local issues of respective countries. Further, we demonstrated that our analyzed and described features are predictive of controversy on Reddit, with user involvement features having the most predictive strengths independent from language. Our approach enables moderators to timely intervene if required or to simply automatically flag submissions that are in danger of derailing conversations.

For future work, we want to further extend our analysis to other datasets. We also plan to characterize users involved in controversial discussions. Moreover, we intend to experiment with early prediction of controversial comments using machine learning.

5 Case Study III: Employee Satisfaction

5.1 Related Publications and Author Contributions

Article 3: [Koncar and Helic, 2020] Koncar, P. and Helic, D. (2020). Employee Satisfaction in Online Reviews. 12^{th} *International Conference on Social Informatics* **Best Paper Nominee**

As the first author of this paper, I was responsible for the crawling, collecting, cleaning and preprocessing of the novel dataset comprising online employer reviews, the design and conduction of all experiments, including the analysis and the prediction of controversy, the visual preparation of results as well as the interpretation and discussion of results, the latter in collaboration with my supervisor Denis Helic.

The main idea for this paper was proposed by me in order to exploit additional results produced during the creation of the fourth article presented in this thesis. These additional results allowed for a comparison of existing findings from traditional studies with findings gained from Web data in the context of employee satisfaction. Based on the feedback provided by Denis Helic, it became an independent article. Both authors contributed to the framing and writing of this paper.

Article 4: [Koncar, T. Santos, et al., 2021] Koncar, P., Santos, T., Strohmaier, M. and Helic, D. (2021) What Herzberg's Two-Factor Theory Reveals About Employee Satisfaction in Online Employer Reviews. **Under review at** *Business & Information Systems Engineering*

As the primary author of the fourth article presented in this thesis, I contributed to the crawling, collecting, cleaning and preprocessing of the same dataset used in Article 3, the design and conduction of all experiments and the visual preparation of results. Denis Helic and Tiago Santos decisively assisted in the selection of methods to conduct the analysis and the prediction experiment. All four authors contributed to the interpretation and discussion of the presented results.

This work was originally supposed to be a straightforward empirical analysis of online employer reviews and I initially mentioned the Two-Factor Theory only in the Discussion section of the article. However, the main idea for this article, which is now the consideration of online employer reviews through the lens of Herzberg's Two-Factor theory, changed after Denis Helic suggested to shift the focus entirely to the theory. The framing of the paper was further elaborated by Markus Strohmaier. Again, all four authors were involved in the the writing of the article.

5.2 Employee Satisfaction in Online Reviews

5.2.1 Abstract

Employee satisfaction impacts the efficiency of businesses as well as the lives of employees spending substantial amounts of their time at work. As such, employee satisfaction attracts a lot of attention from researchers. In particular, a lot of effort has been previously devoted to the question of how to positively influence employee satisfaction, for example, through granting benefits. In this paper, we start by empirically exploring a novel dataset comprising two million online employer reviews. Notably, we focus on the analysis of the influencing factors for employee satisfaction. In addition, we leverage our empirical insights to predict employee satisfaction and to assess the predictive strengths of individual factors. We train multiple prediction models and achieve accurate prediction performance (ROC AUC of best model = 0.89). We find that the number of benefits received and employment status of reviewers are most predictive, while employee position has less predictive strengths for employee satisfaction. Our work complements

existing studies and sheds light on the influencing factors for employee satisfaction expressed in online employer reviews. Employers may use these insights, for example, to correct for biases when assessing their reviews.

5.2.2 Introduction

Employee satisfaction contributes to better employee engagement [Steinhaus and Perry, 1996; Weiss, 2002] and is strongly connected to the overall performance and productivity of businesses [Kumar and Pansari, 2015]. Further, employee satisfaction impacts the lives of many employees as they spend substantial amounts of their time at work [Greenhaus and Beutell, 1985; Ernst Kossek and Ozeki, 1998]. Hence, employee satisfaction has attracted much attention in existing research [Blood, 1969; Locke, 1976; Schneider and Schmitt, 1976; Corneließen, 2009; Artz, 2010; M. E. Malik, Nawab, et al., 2010; Aydogdu and Asikgil, 2011]. While some studies considered the advantage of high employee satisfaction for businesses (e.g., increased performance, reduced employee turnover or absences) [P. C. Smith et al., 1969; Podsakoff and Williams, 1986; Ramayah and Nasurdin, 2006], other studies focused on how to increase and foster employee satisfaction (e.g., through changing positions or granting benefits) [Darling, Arn, and Gatlin, 1997; Rathi and Rastogi, 2008]. Here, we set our focus on the latter and extend existing research by conducting a large-scale analysis of online employer reviews contained in an unexplored dataset. The potential of such reviews to complement traditional management measurements has been depicted in previous research [Miles and Mangold, 2014; Dabirian, Kietzmann, and Diba, 2017; Green et al., 2019], allowing to overcome problems with traditional assessment methods (e.g., annual employee surveys), such as employees that are reticent about feedback out of fear for consequences [Milliken, Morrison, and Hewlin, 2003] or managers not open for criticism [Holland, Cooper, and Hecker, 2016].

Research Question. Based on findings in previous research, in this paper we ask how employee benefits [Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016], employee positions [Dienhart and Gregoire, 1993; De Cremer, 2003; Corneließen, 2009; De Cremer, Dijk, and Folmer, 2009] as well as employment status [S. P. Brown and R. A. Peterson, 1993; Griffeth, Hom, and Gaertner, 2000; Hausknecht, Rodda, and M. J. Howard, 2009] interact with employee satisfaction expressed in online employer reviews.

Approach. To answer our research question, we empirically analyze online employer reviews found on kununu, a so far unexplored reviewing platform where employees can anonymously rate their employers. In particular, we adhere to existing research [S. P. Brown and R. A. Peterson, 1993; Dienhart and Gregoire, 1993; Griffeth, Hom, and Gaertner, 2000; De Cremer, 2003; Cornelißen, 2009; De Cremer, Dijk, and Folmer, 2009; Hausknecht, Rodda, and M. J. Howard, 2009; Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016] and set our focus on the influence of employee benefits, position, and employment status on employee satisfaction. Our dataset comprises more than 2 200 000 reviews of more than 380 000 employers operating in 43 different industries. On kununu, reviews comprise an overall rating ranging between one ("very bad") and five ("very good") stars as well as additional details, such as the position of reviewers and the benefits they had received. We interpret these overall ratings as an expression of employee satisfaction. Finally, we conduct a logistic regression to predict employee satisfaction, allowing us to assess the predictive strength of individual influencing factors.

Findings & Contributions. Overall, our results empirically confirm previous findings in existing research. For example, we observe that higher numbers of employee benefits positively influence employee satisfaction expressed in online employer reviews [Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016]. Further, we find that former employees review more negatively as compared to current employees, which reflects previous results suggesting that dissatisfaction causes employees to quit [S. P. Brown and R. A. Peterson, 1993; Griffeth, Hom, and Gaertner, 2000; Hausknecht, Rodda, and M. J. Howard, 2009]. However, we also find that employees of higher positions (e.g., managers) review more positively, suggesting that employee position may have an influence on employee satisfaction, contradicting previous findings stating the opposite [Dienhart and Gregoire, 1993; Cornelißen, 2009]. Lastly, with our prediction experiment, we find that the number of benefits granted to and the employment status of employees have the highest predictive strengths for employee satisfaction. When we combine all individual features, we achieve a mean ROC AUC of 0.89.

Our work contributes to the understanding of the influencing factors for employee satisfaction expressed in online employer reviews. Thus, we add fruitful input to the debate of employee satisfaction in social and management sciences.

5.2.3 Related Work

Employee Satisfaction. Employee satisfaction has been interpreted in a number of ways in previous research. For example, Blood [1969] stated that employee satisfaction is depending on the values one brings to the job. On the contrary, Schneider and Schmitt [1976] defined it as entirely depending on organizational conditions and not on predispositions of employees. Locke [1976] thought of it more as an interaction between work conditions and individual employees and stressed that satisfaction is relevant to the engagement of the latter. More precisely, he defined employee satisfaction as a positive or pleasant emotional state resulting from people's appreciations of their own job.

Previous studies focused on the positive influence of high employee satisfaction on commitment and engagement of employees. For example, researchers investigated the positive relationship of employee satisfaction and engagement of hotel managers [Gunlu, Aksarayli, and Perçin, 2010] or teachers [M. E. Malik, Nawab, et al., 2010]. Other works studied the positive influence of employee satisfaction on turnover [Aydogdu and Asikgil, 2011] or on organizational commitment [Lumley et al., 2011].

On the contrary to these studies, we analyze employee satisfaction as a dependent variable (i.e., aspects such as pay or rewards that benefit employee satisfaction). Existing studies demonstrated, for example, how to effectively reward employees for increased employee satisfaction [Darling, Arn, and Gatlin, 1997; Artz, 2010; Tessema, Ready, and Embaye, 2013] or how employee position influences employee satisfaction [Dienhart and Gregoire, 1993; Cornelißen, 2009]. We base our work on these previous studies to analyze a novel dataset of more than two million online employer reviews, allowing us to find new insights regarding the influencing factors of employee satisfaction expressed in online reviews.

Online Reviews. Numerous works have studied online employer reviews and most works focused on the reviewing platform *glassdoor*[1]. For example, Marinescu et al. [2018] described a selection bias in online reviews, meaning that people with extreme opinions are more motivated to share their experiences as compared to people with moderate opinions. Chandra [2012] depicted differences in work-life balance between eastern and western countries using reviews written on glassdoor. Luo, Zhou, and Shon [2016] inferred important characteristics impacting employee satisfaction and named innovation and quality as one of the most important. In a more recent work, Green et al. [2019] analyzed employer reviews on glassdoor and their influence on stock returns. Their results indicate that companies with improvements in reviews (i.e., reviews becoming more positive over time) significantly outperform companies with declines (i.e., reviews becoming more negative over time). Dabirian, Kietzmann, and Diba [2017] extracted $38,000$ reviews of highest and lowest ranked employers on glassdoor in order to identify what employees care about. Contrary to these works, we investigate online employer reviews found on kununu and also consider a larger quantity of reviews.

Online reviews have also been studied extensively in other contexts. As early as in 2001, Chatterjee [2001] studied how negative online reviews influence consumers. Yubo Chen, Fay, and Q. Wang [2003] investigated online reviews of automobiles and depicted that the intention to write reviews is depending on the price and quality of products. X. Li and Hitt [2008] as well as P.-Y. Chen, Wu, and Yoon [2004] studied online book reviews, where the former found that the average rating declines over time and the latter found that recommendations in reviews can have a positive impact on sales. Other works focused on the impact of online reviews on hotel business performance [Vermeulen and Seegers, 2009; Ye, Law, and Gu, 2009]. More recent and established works investigated the helpfulness of online reviews mostly written on Amazon [S.-M. Kim, Pantel, et al., 2006; Diaz and V. Ng, 2018; Malik and Iqbal, 2018; MSI Malik and Hussain, 2018],

[1] `https://glassdoor.com`

but also on the video games platform Steam [J. L. Barbosa, Moura, and R. L. d. S. Santos, 2016; Eberhard et al., 2018] or the review aggregation website Metacritic [Kasper et al., 2019; T. Santos et al., 2019]. Our work separates from these studies, which analyze reviews of material goods, as we focus on reviews of work experiences.

5.2.4 Dataset and Empirical Results

Dataset

We conduct our analysis on a novel dataset comprising reviews found on kununu, a platform that offers employees the possibility to anonymously review their employer. The platform is provided in English as well as in German for employers located in Austria, Germany, Switzerland (all since 2007) and the USA (since 2013). Reviews written on kununu consist of 13 individual aspect ratings, such as *company culture* or *teamwork*, each ranging between 1 (“very bad”) and 5 (“very good”). These aspect ratings are aggregated into an *overall rating* for each review. Additionally, reviews contain a headline (120 characters at a maximum) and information about the job status of a reviewer (i.e, whether they are a current or former employee). The position of employees (i.e., either “employee”, “management”, “temporary”, “freelancer/self-employed”, “co-op”, “apprentice” or “other”), the benefits and perks (e.g., life insurance or free parking) as well as free-form text are optionally disclosed by reviewers.

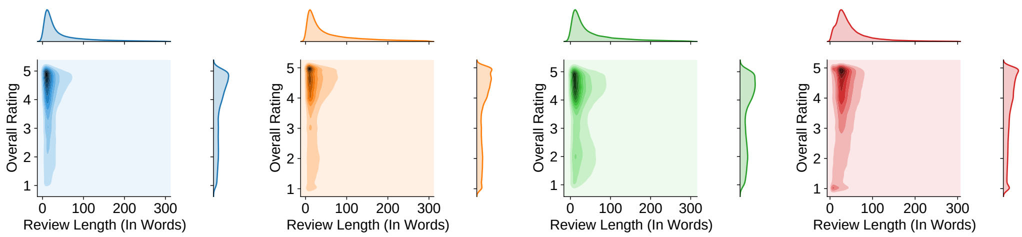
We automatically crawled[2] all reviews present on kununu up to the end of September 2019. Since kununu is bilingual, we had to normalize German names (e.g., of positions) to English ones in order to compare reviews across languages. We list descriptive statistics of our preprocessed dataset, comprising 2 240 276 reviews of 385 736 employers operating in 43 different industries, in Table 5.1.

[2]We used multiprocessing and multiple web proxies to shorten the execution time of the crawling and implemented other measures (e.g., testing for missing or duplicate data) as well as manually checked a selection of reviews to assure data integrity.

Preliminary Descriptive Analysis. We observe a larger number of reviews with positive overall ratings as compared to reviews with negative overall ratings (overall rating median = 3.85). Most notably, reviews of employers located in the USA seem to be more controversial with slightly higher probabilities for one star ratings as compared to European countries. We report a slight increase in the mean overall rating over time for the three European countries, whereas this mean for the USA decreases during the first years only to catch up with other countries in 2017. Regarding the length (in words) of optional review texts, we report long-tailed distributions for each of the four countries, indicating that the majority of reviews contains none or only a few words, whereas only a small number of reviews contain longer texts (review length median = 24). To see if reviews with optional texts are rather positive or negative, we compute Spearman's rank correlation coefficients between the number of characters in reviews with optional text and their respective overall ratings for each of the four countries. We depict bivariate kernel density estimations (KDE) between these two in Figure 5.1. Here, we find a weak negative correlation (all p-values < 0.0005) for each European country with ρ ranging between -0.20 and -0.17. In case of reviews of employers located in the USA, the negative correlation is much

Table 5.1: **Dataset Statistics.** The table lists descriptive statistics of our preprocessed dataset comprising employer reviews written on kununu up to the end of September 2019.

# Industries	43
# Employers	385 736
# Reviews	2 240 276
. . . # thereof for Austria (since 2007)	139 760
. . . # thereof for Germany (since 2007)	1 255 641
. . . # thereof for Switzerland (since 2007)	114 514
. . . # thereof for USA (since 2013)	730 361
. . . # thereof including positions	2 239 189
. . . # thereof including perks or benefits	1 855 491
Overall rating median	3.85
Review length (in words) median	24
# Benefits median	5



(a) Austria (b) Germany (c) Switzerland (d) USA

Figure 5.1: **Review Length and Overall Ratings.** The figure depicts bivariate KDEs of review length in words (distributions truncated at 300 words which is still above the 95th percentile for all languages) and overall ratings of reviews. We observe that most of the reviews contain none or only a few words as well as weak negative correlations between the two, suggesting that dissatisfied employees rather devote time and effort to write optional text.

weaker with $\rho = -0.08$ (p-value < 0.0005). This suggests that dissatisfied employees may rather invest time to address issues as compared to satisfied employees.

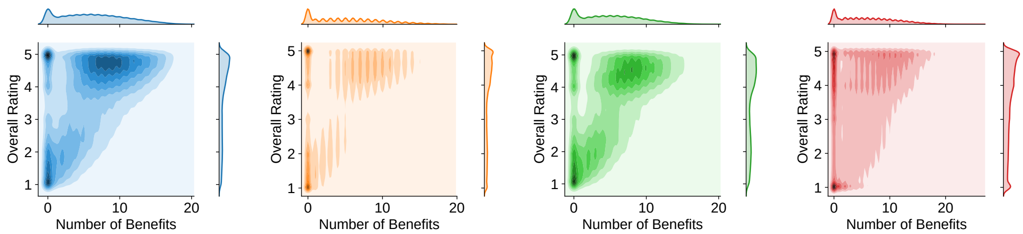
Overall, we report cultural, temporal and textual differences across reviews and an in-depth analysis of these differences might be promising for future work.

Empirical Results

Next we study the relationship between employee satisfaction and employee benefits, employee position as well as employment status.

Employee Benefits. To investigate the influence of benefits on employee satisfaction, we compute Spearman's rank correlation coefficients between the number of benefits received and the overall rating of reviews, respectively for each of the four countries in our dataset.

We depict bivariate KDEs of the number of benefits and overall ratings of reviews in Figure 5.2. Overall, we observe a positive correlation between the number of benefits received and overall ratings for each of the four countries. This suggests that the more benefits employees receive, the higher

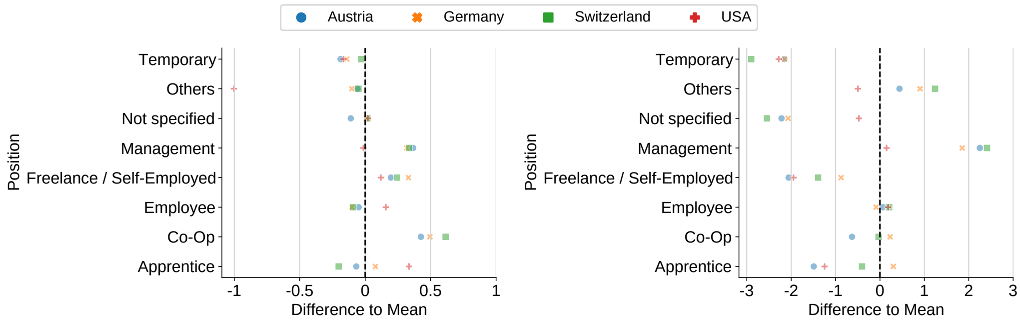


(a) Austria (b) Germany (c) Switzerland (d) USA

Figure 5.2: **Influence of Benefits on Employee Satisfaction.** We plot bivariate KDEs of the number of benefits and overall ratings of reviews, respectively for each country. Overall, we observe positive correlations between the two for all countries, with the exception of the USA for which we find weaker correlations. This observation supports previous findings [Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016] and indicates the positive influence of employee benefits on employee satisfaction.

is their overall rating in their reviews. However, strength of correlation varies across countries. While the mean Spearman's rank correlation coefficient for European countries is 0.46 (all p-values < 0.0005), this correlation is smaller for reviews of employers located in the USA with $\rho = 0.29$ (p-value < 0.0005), despite the fact that USA reviews have, on average, the second largest number of benefits per review (Austria: 6.32, Germany: 5.59, Switzerland: 5.38, USA: 6.30). Hence, we observe cultural differences, indicating that benefits for employees working for companies in the USA are not as influential as for employees of European companies.

Our findings on the positive influence of employee benefits for employee satisfaction are similar to those in existing research [Tessema, Ready, and Embaye, 2013]. Further, studies reported that benefits become more important as part of the compensation of employees [DeCenzo, Robbins, and Verhulst, 2016] and have grown in relevance to employees as well as in their variety [J. M. Newman, Gerhart, and Milkovich, 2016]. Also, not all benefits positively correlate with satisfaction to the same extent and some of them even negatively correlate with each other [Artz, 2010]. Thus, we analyze what benefits were granted most to satisfied employees (reviews with overall rating ≥ 4) and compare that to benefits granted to dissatisfied employees (reviews with overall rating ≤ 2) in order to infer if benefits are



(a) Difference to Overall Ratings (b) Difference to Number of Benefits

Figure 5.3: **Influence of Position on Employee Satisfaction.** In this figure we illustrate how employee position influences employee satisfaction in online employer reviews. In Figure 5.3a, we depict differences in overall ratings compared to country means, respectively for each position (note that we have no co-ops in the USA based version of kununu). Overall, we observe that, contrary to our expectations, employee positions have influence on overall ratings. For example, managers give a higher overall rating compared to the mean overall rating in all European countries. In Figure 5.3b, we depict differences in the number of benefits received compared to country means, respectively for each position. Again, we observe that benefits depend on position. For example, managers get, on average, two additional benefits compared to the country mean.

equally relevant.

Here, we observe that satisfied employees receive different benefits as compared to dissatisfied employees. In particular, satisfied employees receive benefits, such as *flexible working hours*, *401k* or *fitness programs*, that have a high impact on work and life quality. On the contrary, dissatisfied employees receive benefits, such as *parking* or *discounts*. Hence, online employer reviews reflect similar behavior as reported in previous research [Artz, 2010].

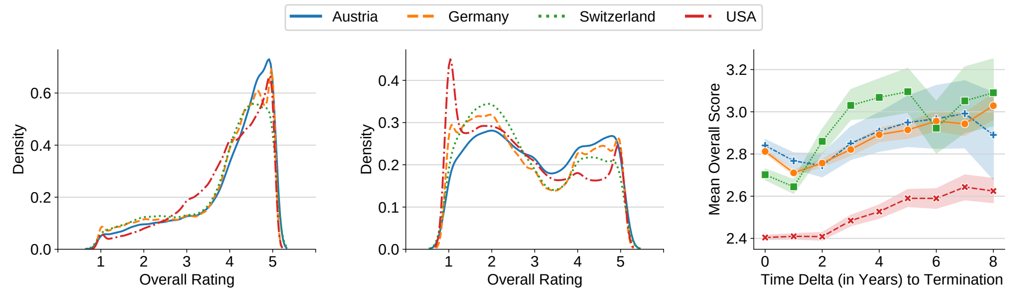
Employee Position. We study the influence of employee positions on employee satisfaction by computing for each position and country the difference to the respective country mean of overall ratings. We assess the significance of mean differences by checking for overlaps of 95% bootstrap confidence intervals.

In Figure 5.3, we depict the results of our analysis on employee positions. Mean differences are significant in all cases, with the exception of reviews without a position specified ("Not specified") for all four countries, other positions ("Others") for European countries, normal employees and apprentices in Austria, temporaries in Germany and Switzerland as well as managers in the USA. Specifically, we observe that co-ops (i.e., working students) have most positive reviews across all European countries (our dataset contains no reviews of co-ops for the USA), rating significantly more positive compared to the respective country mean (about half a star; 0.50 in numbers). Managers rate second most highest in Austria, Germany and Switzerland (all three significant), whereas managers in the USA rate slightly more negative compared to the country mean (non-significant). Apprentices rate significantly higher in the USA compared to European Countries, suggesting that interns and trainees are more satisfied with their education and instruction in the USA.

Our findings contradict existing research which suggests that employee position does not significantly influence employee satisfaction [Dienhart and Gregoire, 1993; Cornelißen, 2009]. Additionally, existing studies showed that higher positions (e.g., managers) often receive more compensation [De Cremer, 2003; De Cremer, Dijk, and Folmer, 2009]. To investigate whether this behavior is reflected in online employer reviews, we compare the number of benefits granted to employee positions to the respective country means. Again, we assess significance of mean differences by checking for overlaps of 95% bootstrap confidence intervals.

We depict results for the influence of position on the number of benefits received in Figure 5.3b. Differences are significant in all cases, except for normal employees in all four countries, co-ops in Germany and Switzerland as well as managers in the USA. We observe that managers in German speaking countries, on average, receive two benefits more compared to the respective country means. This behavior is different to managers in the USA, who, on average and according to our dataset, receive the same number of benefits as the country mean and perhaps suggesting higher levels of equality among different employee positions in the USA. Unsurprisingly, temporary and self-employed personnel receive fewer benefits in all four countries as they are not permanently employed.

Overall, our findings contradict existing results [Dienhart and Gregoire, 1993; Cornelißen, 2009] that suggest no influence of position on employee



(a) Current Employees (b) Former Employees (c) Time Since Termination

Figure 5.4: **Influence of Employment Satisfaction on Employment Status.** The figure illustrates how employee satisfaction potentially influences employment status. We plot KDEs of overall ratings for current (a) and former (b) employees, indicating that current employees have higher probabilities for positive overall ratings as compared to former employees who have higher probabilities to review more negatively. Further, we depict the time (in years) that lies between writing a review and termination of former employees (with 95% bootstrap confidence intervals), suggesting that overall ratings increase with time passed.

satisfaction. However, we confirm existing studies suggesting that higher positions receive more compensation [De Cremer, 2003; De Cremer, Dijk, and Folmer, 2009].

Employment Status. We analyze the potential impact of employee satisfaction on employment status by computing two-sample Kolmogorov–Smirnov tests between overall rating distributions of current and former employees, respectively for each country. Further, we investigate the impact of time lying between creations of reviews and terminations of former employees (up to 8 years as data becomes sparse with longer time periods).

We illustrate the results for this analysis in Figure 5.4. In general, current employees rate rather positively, having a mean overall rating of 3.86 across all countries, whereas possible frustration of former employees is clearly reflected in their mean overall rating of 2.81. We depict the distribution of overall ratings for current and for former employees in Figure 5.4a and in Figure 5.4b respectively. In any case, distributions are significantly different (all p-values < 0.0005) according to the two-sample Kolmogorov–Smirnov tests. For European countries, probabilities are similar for either negative or positive reviews, while probabilities for one star reviews are much higher for reviews of employers in the USA, suggesting higher frustration levels for their employees after termination as compared to employees of European companies.

Our findings reflect reports in existing research which suggest that employee satisfaction is the main reason for staying with or leaving an employer [Hausknecht, Rodda, and M. J. Howard, 2009]. Further, employee satisfaction is strongly, positively correlated with organization commitment [S. P. Brown and R. A. Peterson, 1993] and employee dissatisfaction is antecedent to forming an intention to quit [Griffeth, Hom, and Gaertner, 2000], providing a potential explanation for our findings.

In Figure 5.4c, we depict mean overall ratings according to the time that lies between the terminations of former employees and the creations of their reviews, respectively for each country. Our results suggest that the longer a termination lies in the past, the more positive are the reviews. While the mean overall rating of reviews written one month after termination is 2.59, this value is 3.91 after ten months, and mean values increase even more with longer time deltas.

Overall, our findings support our initial expectation that former employees review more negatively as compared to current employees. This effect, however, seems to weaken with more time lying between the termination of former employees and creation of their reviews.

5.2.5 Predicting Employee Satisfaction

We conduct a logistic regression to predict employee satisfaction based on our empirical results. We distinguish between reviews from satisfied and dissatisfied employees based on the overall rating contained in reviews. More precisely, we consider reviews with overall ratings less than or equal to the first quartile (overall rating ≤ 2.42) as expressions of dissatisfaction and reviews with overall ratings equal to or greater than the third quartile (overall rating ≥ 4.54) as expressions of satisfaction. We create classes this way to counteract the general bias towards more positive reviews (cf. Table 5.1 for the overall rating median). We implement our logistic regression with default parameters[3] and train multiple models respectively with the following feature spaces[4]:

(i) *Country:* The country the reviewed employer is located in.
(ii) *Year:* The year the review was written in.
(iii) *Review Length:* The length of reviews defined by the number of words.
(iv) *Benefits:* The number of benefits the reviewing employee received.
(v) *Position:* The position of the reviewing employee.
(vi) *Employment Status:* The employment status of the reviewing employee.
(vii) *All:* The combination of all the above feature spaces.

Note that we use one-hot encoding to transform categorical features (country, year, position and employment status) and that we standardize numerical features (review length and benefits). Further, we remove all reviews that are missing any of the described features. This leaves us with 430 998 positive and 419 309 negative reviews. To evaluate and compare performance of our models, we use ten-fold cross validation and report mean ROC AUC values over the folds.

Results. In Figure 5.5 we illustrate the mean ROC AUC respectively for each feature space. When we consider individual feature spaces, the number of benefits is most predictive for employee satisfaction with a mean ROC AUC of 0.78, followed by the employment status with a value of 0.74. Our remaining feature spaces perform notably worse. The model we train with the year the review was written results in a mean ROC AUC of 0.61, similar to the model using the length of reviews which results in a value of 0.60. We report that the position of employees has very low predictive power with a mean ROC AUC of 0.57, only followed by the country the reviewed employer is located in, which performed the worst with 0.52 and only minimally better than a random baseline (0.5). However, when we combine all the individual feature spaces, we can further increase performance by 0.11 (compared to the model trained on the number of benefits) to 0.89, indicating that the features provide complementary information on employee satisfaction.

[3] As implemented in scikit-learn 0.22.2 (`https://scikit-learn.org/0.22/`).
[4] We do not consider the time delta between termination of employees and creation of their reviews as this feature is only available for 14% of reviews in our dataset.

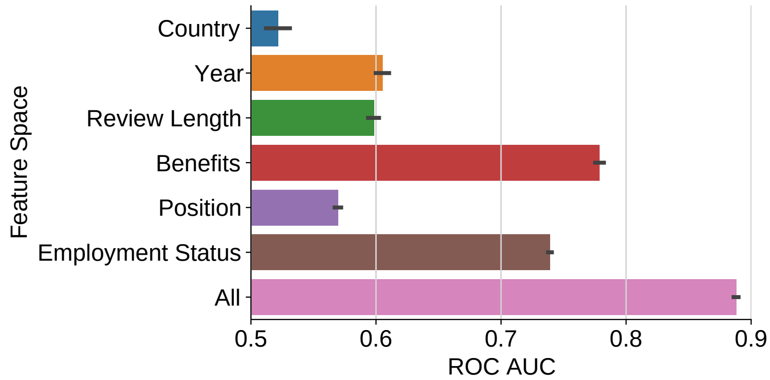


Figure 5.5: **Prediction Results.** In this figure we depict the results of our prediction experiment with which we assess the predictive strength of individual feature spaces and the combination of those for the prediction of employee satisfaction. The error bars indicate 95% bootstrap confidence intervals. We observe that the number of benefits received and the employment status of employees have the highest predictive strengths, while other feature spaces perform only minimally better than the random baseline (ROC AUC of 0.5). The combination of all features further improves the performance to a mean ROC AUC of 0.89, demonstrating that we can accurately predict employee satisfaction and that the consideration of all features is reasonable.

5.2.6 Discussion

With our empirical analysis of online employer reviews we provide new insights into the influencing factors for employee satisfaction expressed in such reviews. Our prediction experiment demonstrates the predictive strengths of individual features for predicting employee satisfaction. In the following, we discuss our findings and connect our results to our initial research question.

Support For Existing Studies in Online Employer Reviews. Overall, we find support for findings in previous studies [S. P. Brown and R. A. Peterson, 1993; Griffeth, Hom, and Gaertner, 2000; De Cremer, 2003; De Cremer, Dijk, and Folmer, 2009; Hausknecht, Rodda, and M. J. Howard, 2009; Artz, 2010; Tessema, Ready, and Embaye, 2013; DeCenzo, Robbins, and Verhulst, 2016; J. M. Newman, Gerhart, and Milkovich, 2016]. One exception are our findings on positions, suggesting that, contrary to previous works [Dienhart and Gregoire, 1993; Cornelißen, 2009], employee positions influence employee satisfaction. In particular, we observed higher employee satisfaction for managers. One possible explanation for this observation could be that managers review more positively in order to increase the reputation of their company (under the assumption that they are currently employed at respective companies). To test for this, we compute mean overall ratings for current and former managers. Here, we find a mean overall rating of 4.12 for current managers ($n = 193\,872$) and 2.72 for former managers ($n = 72\,258$), suggesting support for our assumption.

We observed that benefits have less impact on overall ratings for reviews of employers located in the USA as compared to European countries. After further investigating this result, we notice that USA based reviews include fewer benefits related to work-life balance (e.g., *flexible workhours* or *home office allowed*) as compared to German speaking countries. Thus, one possible explanation for the weaker correlation could be the lack of work-life benefits, which according to existing research are important for high employee satisfaction [Thompson, Beauvais, and Lyness, 1999; Muse et al., 2008].

In the case of the influence of employee positions on overall ratings, we found that managers of European companies are more satisfied as compared to managers of companies located in the USA. This observation can be a reflection of different leadership styles between the USA and European countries, where for the former pressure on managers might be higher because decision making is much more egalitarian in the USA compared to European countries [Meyer, 2018]. However, this could also be due to the fact that managers in German speaking countries enjoy more advantages as compared to those working for companies located in the USA (cf Figure 5.3b). Also notable is that freelancers rate more positive in all four countries compared to respective country means, supporting exiting work indicating that freelancers are more satisfied with their job because of higher levels of freedom [Ryan, 2009; Massey and Elmore, 2011; Deprez and Raeymaeckers, 2012].

We found that former employees rate more positive the more time lies between the creation of reviews and their termination. This suggests that frustration of reviewers dissipates over time ("time heals wounds"), which follows the intuition that forgiving becomes easier over time [McCullough, Fincham, and Tsang, 2003]. However, we suggest a more detailed analysis of this observation for future work.

Prediction. Finally, our prediction experiment revealed that the number of benefits and the position of employees are most predictive for employee satisfaction. However, only when considering the combination of all features, we achieved the best prediction performance. Note that more sophisticated approaches, such as deep learning, might further improve the prediction performance. Based on our observations, we suggest that employers should assign more responsibilities to their employees as well as grant them more freedom, especially related to work-life balance. Overall, these observations are strongly related to long existing theories of management sciences, such as Herzberg's Two-Factor Theory [F. Herzberg, Mausner, and Snyderman, 1959; F. I. Herzberg, 1966; F. Herzberg, 2017] which is based on similar suggestions.

Limitations. We based our work on employer reviews found on kununu, only one platform among many others providing similar reviewing possibilities. Despite the large amount and the variety of data, the quality of our analysis may be improved by considering additional platforms, such as *glassdoor.com*. However, note that our analysis requires adjustments to other platforms as they use different rating mechanics and consider other employer characteristics. Thus, our analysis is biased towards the particularities of kununu. Further, we acknowledge a potential bias introduced by reviewers, such as different interpretations of rating scales or herding behavior, as suggested by existing research [Lauw, Lim, and K. Wang, 2012]. We leave an in-depth analysis of this phenomenon on kununu for future work. For our prediction experiment, we defined employee satisfaction based on the quartiles of overall rating distributions. While small adjustments to this definition did not noticeably alter our results, other definition may result in different findings.

5.2.7 Conclusions

In this paper, we investigated online employer reviews comprised in an unexplored dataset to shed light on the influencing factors for employee satisfaction. We obtain comparable results from online employer reviews to results from existing research based on e.g., survey data. The only exception to this are our results regarding the influence of employee position on employee satisfaction where we find that they are more important in online reviews as compared to previous findings. Further, we observe cultural differences across employers, for example, benefits have less impact on employee satisfaction in the USA as compared to European countries. With our prediction experiment we depicted the predictive strengths of our individual findings, suggesting that the number of benefits and the employment status convey the most information for predicting employee satisfaction. When we combined all different features, we achieved a mean ROC AUC of 0.89, demonstrating the we can accurately predict satisfied and dissatisfied employees based on only a handful of features. Employers may use our findings to correct for biases when assessing their reviews or adapt management measures, such as shifting parts of compensation towards more benefits, as we demonstrated that they are most influential for employee satisfaction in reviews.

For future work, we plan to extend our analysis to learn more about the reviewing behavior of employees, for example, by considering the individual review aspects or by investigating industrial differences. Further, we want to adapt our analysis methods to other datasets comprising online employer reviews.

5.3 What Herzberg's Two-Factor Theory Reveals About Employee Satisfaction in Online Employer Reviews

5.3.1 Abstract

Online employer reviews hold great potential to learn about employee satisfaction and motivation, both key to economic success. However, little is known about how to interpret such reviews and how to derive actual management measurements from them. Through the lens of Herzberg's

Two-Factor theory, which identifies aspects influencing employee motivation and satisfaction, this paper investigates more than 2 million multi-aspect online employer reviews written in two different languages. This allows for a thorough understanding of what online employer reviews can reveal about employee satisfaction and motivation. For that, review aspects and ratings are leveraged to study factors that influence employee satisfaction with respect to the theory. Based on the gained insights, a prediction experiment is conducted to forecast employee satisfaction for individual employers and to measure the predictive strengths of factors. This analysis identifies relevant aspects for satisfied and dissatisfied employees working in 43 different industries and 4 different countries. Overall, the results illustrate that aspects, which according to Herzberg prevent dissatisfaction are most relevant for reviewers, while aspects that foster satisfaction are rather incidental. The prediction experiment achieves a mean balanced accuracy of 0.87, suggesting that these factors are predictive for employee satisfaction. This work is the first large-scale study of online employer reviews through the lens of Herzberg's Two-Factor Theory, providing comparative results across different employers, industries and countries. Moreover, the paper contributes to the knowledge of organizational and management theories as well as social sciences by adding useful input to the discussion of the Two-Factor Theory.

5.3.2 Introduction

Employee satisfaction is closely linked to employee motivation and their resulting performance and, as such, is crucial to economic success of a plethora of businesses [Dobre, 2013; Kumar and Pansari, 2015]. Therefore, businesses have an inherent interest in measuring and raising the levels of employee satisfaction [S.-H. Chen et al., 2006].

Traditionally, measuring employee satisfaction is based on both a solid theoretical framework identifying important aspects contributing to satisfaction as well as empirical analyses of data collected through interviews and surveys in businesses. A prominent example of such a framework is the *Two-Factor Theory* [F. Herzberg, Mausner, and Snyderman, 1959], introduced in 1959 by Frederick Herzberg. His theory defines two sets of factors: (i) *hygiene* *factors* capturing the surroundings and doings of a job, such as company culture, salary and working conditions and (ii) *motivation factors* supporting self-actuation of employees through, for example, responsibility, advancement and recognition. According to the theory employers should first fulfill hygiene factors to prevent dissatisfaction and then satisfy motivation factors to pave the way for high employee satisfaction and productivity. Due to a high potential of enhancing productivity of businesses, the theory has been applied extensively in empirical research by, for example, evaluating how businesses support hygiene and motivation factors in their organizations [Balmer and Baum, 1993; Tietjen and Myers, 1998; DeShields Jr, Kara, and Kaynak, 2005; Lundberg, Gudmundson, and Andersson, 2009; Alfayad and Arif, 2017; Holmberg, Caro, and Sobis, 2018; Hur, 2018; Kotni and Karumuri, 2018].

More recently, Web data and large-scale empirical analyses complemented organizational sciences [Miles and Mangold, 2014; Luo, Zhou, and Shon, 2016; Dabirian, Kietzmann, and Diba, 2017; Green et al., 2019; Das Swain et al., 2020]. For example, previous empirical research indicated a great potential in online employer reviews to complement existing management measurement methods for improving performance [Miles and Mangold, 2014; Dabirian, Kietzmann, and Diba, 2017; Green et al., 2019]. Such reviews have, for example, been studied in context of employer branding [Dabirian, Kietzmann, and Diba, 2017], organizational culture [Das Swain et al., 2020] and also employee satisfaction [Luo, Zhou, and Shon, 2016], for which authors studied the impact of satisfaction on corporate performance. However, studying employee satisfaction as a dependent variable based on online employer reviews is, to our knowledge, still new to our community. We lack a theoretical framework for measurement and interpretation of employee satisfaction in online reviews and believe that a combination of a theoretical framework and review data leads to (i) deeper and more general insights through context defined by the theory, and (ii) further development and refinement of the theory due to the new empirical results.

Research Questions. In this paper, we investigate online employer reviews through the lens of Herzberg's Two-Factor Theory and ask the following research questions. From theoretical perspective, we ask how we can interpret online employer reviews in the context of the Two-Factor Theory. Empirically, we ask whether both hygiene and motivation factors are equally important and relevant to reviewers. Finally, we ask whether the data grants evidence for refinement of the Two-Factor Theory in our digital age.

Approach. We answer our research questions with a novel dataset containing more than 2 200 000 reviews of more than 380 000 employers operating in 43 different industries extracted from the employer review platform *kununu*[5]. In particular, on kununu, employees of companies located in Austria, Germany, Switzerland and the United States of America (USA) can review their employers regarding different aspects (e.g., *support from management*) by providing ratings ranging between one and five stars and by (optionally) writing review text. We link these aspects to either hygiene and motivation factors and interpret review ratings as an expression of employee satisfaction (see Figure 5.6).

Thus, we analyze the characteristics of review aspects for which we define and test multiple hypotheses (see Table 5.2), each derived from the Two-Factor Theory and focusing on (i) the attention devoted to review aspects, (ii) the sentiment conveyed in review aspects, (iii) the readability of review aspects, (iv) the content of review aspects, and (v) how each of the previous generalizes across cultural, industrial and employment status differences. Finally, we illustrate the utility of our findings with a prediction experiment.

Findings. We observe high importance and relevance of hygiene factors in online employer reviews, indicating that their purpose exceeds solely preventing dissatisfaction and that they also foster satisfaction. Regarding our hypotheses, we find that dissatisfied employees rather write (optional) text and tend to focus more on aspects related to hygiene factors as compared to satisfied employees. Further, reviews from dissatisfied employees convey a more negative sentiment, while they are also more polished and readable (contradicting existing studies that suggest opposite behavior in other contexts [Korfiatis, Rodríguez, and Sicilia, 2008; Korfiatis, García-Bariocanal, and Sánchez-Alonso, 2012; Y. Zhao, X. Xu, and M. Wang, 2019]) than reviews from satisfied employees. We report that dissatisfied employees tend to focus more on terms related to hygiene factors and satisfied employees more on terms related to motivation factors. Finally, we uncover some evidence for the generalization of our results across industrial, cultural and

[5]Link to website: `https://kununu.com`

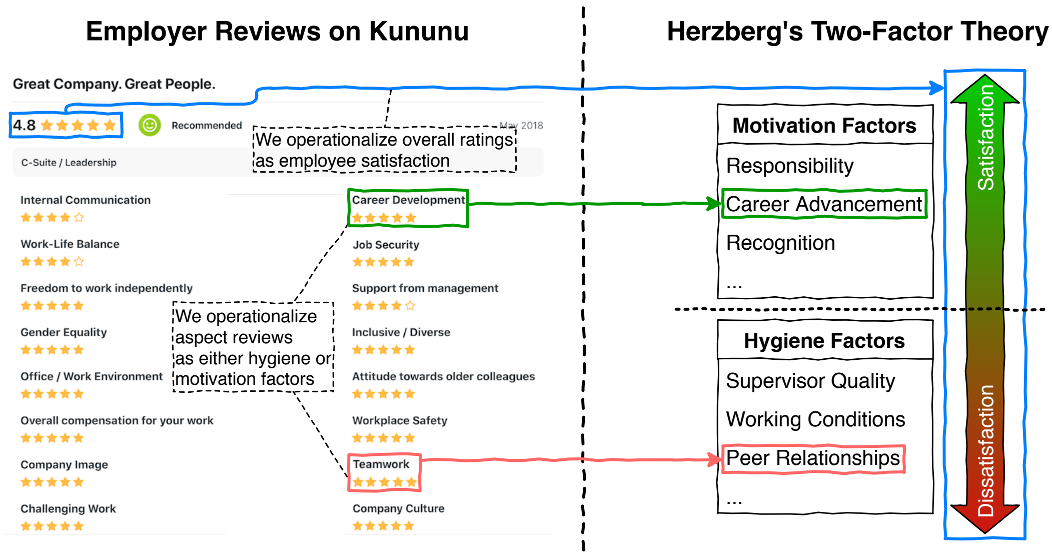


Figure 5.6: **Studying Online Employer Reviews Through the Lens of Herzberg's Two-Factor Theory.** The figure illustrates how we apply Herzberg's Two-Factor Theory to online employer reviews on kununu. On the left we depict an excerpt of a positive employer review found on kununu and on the right we depict the general principles of the theory with hygiene factors preventing employee dissatisfaction and motivation factors fostering employee satisfaction. In particular, we operationalize the overall ratings of reviews (displayed in the top left corner of the excerpt and highlighted in blue) as employee satisfaction and individual review aspects and their characteristics (e.g., aspect review length) as either hygiene factors (example highlighted in red) or motivation factors (example highlighted in green). Note that the exemplary terms for hygiene and motivation factors stem from words that are frequently related to them in existing research (see Table 5.6 in the Appendix section). This operationalization allows us to assess the influence of individual aspects on employee satisfaction through the lens of Herzberg's Two-Factor Theory. Our novel and large-scale dataset comprises employer reviews in German and English from four different countries and 43 different industries, enabling us to study how our findings generalize across different countries and industries.

employment status differences in the context online employer reviews. We leverage our empirical results to accurately predict employee satisfaction of individual companies, achieving a maximum balanced accuracy score of 0.87. Our different feature spaces indicate that textual content as well as textual style of reviews have high predictive power.

Contributions. We point out that hygiene factors are of high importance and relevance to online reviewers of employers. Thus, when analyzing online reviews to learn about employee satisfaction and motivation, we should identify and focus on hygiene factors, while the potential to learn about motivation factors from online reviews is limited. Further, we add fruitful input to the discussion of the Two-Factor Theory by applying it on a novel dataset. This bares potential implications for the theory, as we revisit hygiene and motivation factors for the digital age.

5.3.3 Related Work

Analyses of Online Reviews. A large body of previous work studied online reviews in different fields, such as their impact on hotel business perfor-

Table 5.2: **Hypotheses.** The table lists our individual hypotheses as well as their main focus and our respective findings (✓ = strong support; ∼ = weak support; × = rejection).

Focus	#	Hypothesis	Result
Attention	1.1	Satisfied employees write more reviews on motivation factors.	×
	1.2	Dissatisfied employees write more reviews on hygiene factors.	✓
	1.3	Satisfied employees write longer reviews for motivation factors.	✓
	1.4	Dissatisfied employees write longer reviews for hygiene factors.	✓
Sentiment	2.1	Satisfied employees write more positively about motivation factors.	×
	2.2	Dissatisfied employees write more negatively about hygiene factors.	✓
Readability	3.1	Satisfied employees write more readable reviews about motivation factors.	×
	3.2	Dissatisfied employees write less readable reviews about hygiene factors.	×
Content	4.1	Satisfied employees use more words related to motivation factors.	✓
	4.2	Dissatisfied employees use more words related to hygiene factors.	✓
Generalization	5.1	Results are independent of cultural context.	∼
	5.2	Results are independent of industry.	∼
	5.3	Results are independent of employment status (i.e., current or former).	∼

mance [Vermeulen and Seegers, 2009; Ye, Law, and Gu, 2009], analyzing the success of books [P.-Y. Chen, Wu, and Yoon, 2004; X. Li and Hitt, 2008] and movies [Reinstein and Snyder, 2005; Duan, Gu, and Whinston, 2008], understanding the characteristics of video game reviews [Eberhard et al., 2018; T. Santos et al., 2019], as well as recommending airlines [Siering, Deokar, and Janze, 2018] and predicting the satisfaction of their customers [Lacic, Kowald, and Lex, 2016].

Our work extends these studies with a novel dataset, which contains multi-aspect [McAuley, Leskovec, and Jurafsky, 2012] online reviews of employers and which we already analyzed in a preliminary work [Koncar and Helic, 2020]. We thus add to previously mentioned works, as those mainly focus on reviews of material and experience goods, while employer reviews of the kind we analyze capture further sensitive topics, such as gender biases or safety at the workplace.

Most works investigating online reviews of employers focused on the website *Glassdoor*[6]. For example, a previous study [Marinescu et al., 2018] described a selection bias in online employer reviews, where people with extreme opinions are more motivated to share their experiences as compared to people with moderate opinions. Other researchers studied workplaces [Luo, Zhou, and Shon, 2016; Dabirian, Kietzmann, and Diba, 2017], cultural drivers of work-life balance and employee satisfaction [Chandra, 2012] as well as organizational culture [Das Swain et al., 2020] by leveraging reviews on Glassdoor. Further research on such reviews links their valence to financial [Green et al., 2019] and workforce-related [Könsgen et al., 2018] impacts, thereby underscoring their relevance to employers.

Studying online platforms and social networks to benefit employees, employers and organizations as a whole has much history. For example, existing studies focused on why and how employees use social networking at work [DiMicco et al., 2008] or how employee engagement spreads in organizational social media [Mitra et al., 2017]. Similarly, Shami et al. [2014] analyzed texts of internal and external social media platforms to extract emotions and opinions of employee chatter. De Choudhury and Counts [2013] investigated emotional patterns during times of high and low productivity. In another work, Guy et al. [2016] studied how users use the "like

[6]Link to website: `https://glassdoor.com`

button" in an organizational context and how it may relate to organizational commitment. More recently, Saha et al. [2019] used data from *Linkedin*[7] to study role ambiguity (i.e., unclear responsibilities and degree of authority of employees) and its effects on employee wellbeing. Their proposed method can help to identify role ambiguity in organizations and demonstrates the potential of analyzing data from the Web and using gained insights to improve life at work.

In contrast to the mentioned works, we analyze a multiple of reviews and consider both English and German reviews allowing us to assess cultural influences on online employer reviews. Further, our predictive models for employer reviews are of independent interest as they may, for example, support employers in identifying problematic areas and grasping longitudinal and cross-sectional trends.

Herzberg's Two-Factor Theory. Herzberg defined the Two-Factor Theory by collecting feedback from 203 accountants and engineers, asking them in which situations they felt either good or bad about their work [F. Herzberg, Mausner, and Snyderman, 1959]. Leveraging the gathered feedback, he defined two different sets of needs that both contribute to employee satisfaction, namely (i) hygiene factors and (ii) motivation factors. Hygiene factors cover basic needs that are not directly related to the content of a job but rather represent the surroundings of it, such as the compensation for an employee's work, the company culture or the interpersonal communication. On the contrary, motivation factors relate to the self-actualization needs of employees, focusing, for example, on responsibilities, achievement and the actual content of the work itself. Motivation factors follow the idea that humans strive for always improving themselves [F. I. Herzberg, 1966; F. Herzberg, 2017], a fact that can only be satisfied by altering the content of work [Tietjen and Myers, 1998]. According to Herzberg's theory, the satisfaction of hygiene factors can prevent dissatisfaction and poor performance of employees but only the satisfaction of motivation factors encourage high employee satisfaction and, as such, high productivity. Note that the absence of motivation factors does not necessarily lead to dissatisfaction among employees.

[7] Link to website: `https://linkedin.com`

Ever since the introduction of the Two-Factor Theory, several empirical analyses tested for it in different industries or showcasing its general applicability. For example, Lundberg, Gudmundson, and Andersson [2009] investigated work motivation of seasonal workers in hospitality and tourism, finding support for the Two-Factor Theory, but also uncovering discrepancies in the needs of seasonal workers. In a related context, Balmer and Baum [1993] demonstrated the general applicability of the theory by using it to investigate guest motivation in hospitality. DeShields Jr, Kara, and Kaynak [2005] used the theory to study the motivation and satisfaction of business students, translating hygiene factors to capture performance of advising staff and motivation factors to capture performance of classes and faculties. Again, researchers find support for Herzberg's Two-Factor Theory.

More recent works studied, for example, the influence of employee voice (i.e., employees communicate their views to employers) on employee satisfaction by applying the Two-Factor Theory on feedback from 300 non-managerial employees [Alfayad and Arif, 2017]. Here, researchers found that acknowledgment of employee voice pushes motivation and therefore increases employee satisfaction. Holmberg, Caro, and Sobis [2018] investigated reasons for shortages of nursing personnel in Swedish mental health care using the Two-Factor Theory. They based their analysis on interviews with 25 nursing personnel demonstrating the usefulness of Herzberg's theory and identifying the lack of career advancements as a partial reason for these shortages. In another study, Hur [2018] reported differences between public and private sectors in how hygiene factors and motivation factors affect employee satisfaction. Kotni and Karumuri [2018] applied Herzberg's Two-Factor Theory on data from 150 salesmen of the retail sector and found that they are more satisfied with hygiene factors as compared to motivation factors, suggesting discrepancies from the Two-Factor Theory.

With our work we complement previous studies by applying Herzberg's Two-Factor Theory to online employer reviews, allowing us to investigate how hygiene and motivation factors influence employee satisfaction expressed in reviews and how to interpret them in the context of the theory.

5.3.4 Dataset and Methods

Dataset

Kununu is a platform allowing employees to anonymously review their employers and operates in Austria, Germany and Switzerland since 2007 and also in the USA since 2013. Hence, kununu is bilingual (German and English), but reviews can be composed in any language[8]. Each review on kununu consists of an *overall rating* ranging between 1 and 5, where 1 represents "very bad" and 5 represents "very good" experiences (as described by kununu). The overall score aggregates a variety of individual review aspects, each also ranging from 1 to 5 and grouped into four sections: (i) *company culture*, (ii) *diversity*, (iii) *work environment*, and (iv) *career*. We list the 13 individual aspects and their descriptions provided by kununu and stating what an aspect is about in Table 5.4. Note that there are five additional review aspects only available for the USA version of kununu (comprising *Inclusive / Diverse*, *Handicapped Accessibility*, *Workplace Safety*, *Job Security*, *Challenging Work*). For better comparability among countries, we exclude these five aspects from our analysis.

In addition to ratings, reviewers must specify a headline (maximum of 120 characters), whether they are a former or a current employee of the reviewed company, as well as whether or not they would recommend the employer to friends (answer not shown in reviews and only used internally by kununu). Reviewers can optionally state their position (i.e., either "employee", "management", "temporary", "freelancer", "co-op", "apprentice" or "other"), as well as include suggestions for improvements, what they like and dislike about the company, and comments on any of the aspect ratings (aspect reviews) in free-form text. Finally, kununu provides employer profile pages stating the country and industry they operate in.

Data Acquisition & Preprocessing. For our analysis, we automatically extracted (see the Appendix section for a detailed description of this process) all reviews present on kununu (comprising Austrian, German, Swiss and USA versions) up to the end of September 2019. Extracted reviews include

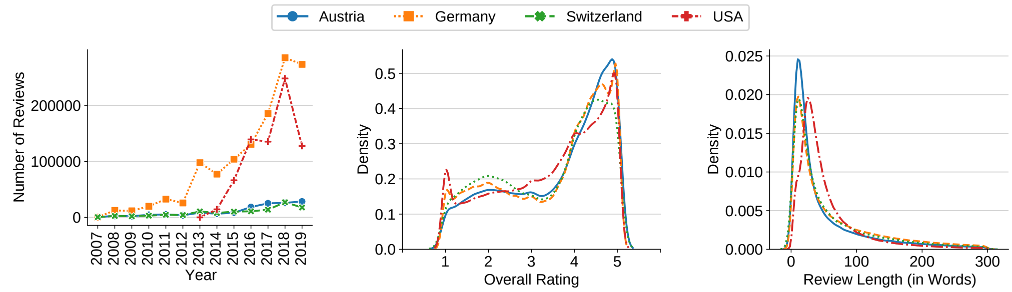
[8]Note that due to the primary focus on German and English, the number of reviews written in other languages is negligible.

the name and industry of the reviewed employer, overall and aspect ratings, all free-form texts, the review date and the employment state (either current or former employee) of the reviewer. As kununu is bilingual, we link and normalize German names of review aspects (e.g., *Kollegenzusammenhalt* ⇒ *Teamwork*) and industries (e.g., *Dienstleistung* ⇒ *Service and Support*) to English ones, allowing us to compare German and English reviews. In Table 5.3, we list descriptive statistics of our preprocessed dataset, comprising 2 240 276 reviews of 385 736 employers written over a time span of twelve years.

Preliminary Descriptive Analysis. In Figure 5.7, we depict selected characteristics of our dataset. In 2007, kununu had a comparatively small number of reviews where each of the three German-speaking countries had no more than 550 reviews. However, kununu grew rapidly over the years and aggregated more than 445 000 reviews in 2019 alone (up to the end of September; see Figure 5.7a). We report a slight increase in the mean and variance of overall ratings over time for each of the four countries contained in our dataset, indicating that reviews became more positive and reviewers more divided with time. We observe slightly different behavior for the reviews in the USA version of kununu, for which the mean overall rating decreases continuously in the first four years only to catch up with German versions in 2017, after which the variance of overall ratings starts to decrease. We interpret this as a first indicator for cultural differences reflected in our dataset.

Table 5.3: **Dataset Statistics.** The table lists descriptive statistics of our crawled dataset comprising employer reviews written on kununu up to the end of September 2019.

# Industries	43
# Employers	385 736
# Reviews	2 240 276
... # thereof for Austria (since 2007)	139 760
... # thereof for Germany (since 2007)	1 255 641
... # thereof for Switzerland (since 2007)	114 514
... # thereof for the USA (since 2013)	730 361
... # thereof including free-form text	1 662 250



(a) Number of Reviews over Time (b) KDE of Overall Rating (c) KDE of Review Length

Figure 5.7: **Characteristics of Our Dataset.** The figure depicts selected key characteristics of our dataset, including the number of reviews over time as well as kernel density estimations of overall ratings and review length, respectively for each of the four countries contained in our dataset. In Figure 5.7a, we present the number of reviews over time, depicting a steady increase throughout the years, especially after 2014, for the German and the USA based version of kununu. The two smaller countries, Austria and Switzerland, show similar behavior, but exhibit much smaller numbers in reviews. In Figure 5.7b, we illustrate the kernel density estimation (KDE) for the overall rating. We observe higher probabilities for positive ratings as compared to negative ones for all four countries. Reviewers on the USA based version of kununu seem to be slightly more controversial as indicated by the higher probability for one-star reviews in comparison to German speaking countries. Regarding length of reviews (in words) consisting of optional free-from text, we observe long-tailed distributions for all four countries, suggesting our dataset includes many reviews with no or only short review texts, while only a limited number of reviews have longer review texts (see Figure 5.7c; distribution truncated at 300 words, which lies still above the 95$^{\text{th}}$ percentile).

Regarding ratings, we observe a large number of reviews with positive overall ratings and a lower number of reviews with negative overall ratings in all four countries (see Figure 5.7b). Reviews in the USA are slightly different from those of remaining countries with overall ratings being more controversial which, again, depicts cultural differences among reviewers.

For the text length (measured in number of words) of reviews with optional review texts[9], we observe long-tailed distributions for each country (see Figure 5.7c), indicating that the majority of reviews contains only a few

[9]For this analysis, we combined texts of individual aspects and other review sections, respectively for each review.

words, whereas only a small number of reviews contain substantially longer free-form texts (the 95th percentile is 278 words). After manually inspecting reviews, we report that the majority of reviewers specifically address a selection of aspects, suggesting that they devote attention only to those aspects that are relevant to them.

5.3.5 Methodology

We investigate online employer reviews through the lens of Herzberg's Two-Factor Theory. For that, we first explain how review aspects relate to Herzberg's hygiene and motivation factors as well as how we leverage review ratings to measure employee satisfaction. We then address our overarching research question of how to interpret online employer reviews in the context of the Two-Factor Theory by defining a set of 13 hypotheses derived from it. Finally, we illustrate how to use our findings to accurately predict employee satisfaction on a company level.

Assigning Review Aspects with the Two-Factor Theory

Following the definition of Herzberg's Two-Factor Theory and providing the description of aspects as stated by kununu (see Table 5.4), we let three independent annotators assign the 13 review aspects to either hygiene or motivation factors. Note that we provide a detailed description of this annotation process in the Appendix section. We assess the inter-rater agreement between annotators by computing Fleiss' kappa [Fleiss, 1971] resulting in a value of 0.597, suggesting a moderate to substantial agreement among the three annotators. Annotators determined that *company culture*, *internal communication*, *teamwork*, *work-life balance*, *office and work environment*, as well as *environmental friendliness* relate to hygiene factors as they all address the surroundings of work done. On the other hand, all annotators consider *freedom to work independently* and *career development* as clear motivation factors as they are related to the content of the work done. Remaining factors comprising *support from management*, *gender equality* and *attitude towards older colleagues* are, according to our annotators, not clearly assignable to either motivation or hygiene factors. Note that our annotators are not the first to encounter such issues as the exact distinction between hygiene and motivation factors has been criticised in existing research [Parsons and Broadbridge, 2006; Y. Li, 2018]. After further discussing the respective assignments, the annotators agreed that the three aspects relate to both hygiene and motivation factors based on the following explanations: Regarding *support from management*, our annotators find that its description (see Table 5.4) includes the style of leadership, which can be interpreted as hygiene factor, but also addresses involvement in decision making, which can be interpreted as a motivation factor. Similarly, in case of *gender equality* and *attitude towards older colleagues* the description addresses equal treatment among colleagues (hygiene factor), as well as equal career opportunities (motivation factor). Hence, annotators linked these three factors to both hygiene and motivation factors. We list the resulting assignments, respectively for each of the 13 aspects, in Table 5.4.

Definition of Employee Satisfaction

We define *low* and *high* levels of employee satisfaction based on the overall rating of reviews. In particular, we consider reviews with an overall rating less than or equal to the first quartile (overall rating = 2.42) to represent low employee satisfaction (negative reviews) and reviews with an overall rating equal to or greater than the third quartile (overall rating = 4.54) to represent high employee satisfaction (positive reviews). This leaves us with 561 515 negative reviews and 594 409 positive reviews. The remaining 1 084 352 reviews are considered as reviews with neutral employee satisfaction and are therefore neglected for the remainder of the study. Slight changes to these thresholds do not qualitatively impact our results.

Hypotheses

We advance 13 different hypotheses, each focusing on distinct characteristics of employer reviews and derived from Herzberg's Two-Factor Theory.

Table 5.4: **Overview of Individual Review Aspects on kununu.** The table lists the 13 aspects of the four sections (*Company Culture*, *Diversity*, *Work Environment* and *Career*), which employees can review, along with their description taken from kununu (as of September 2019) as well as the factors independent experts assigned with them according to Herzberg's Two-Factor Theory (see Section 5.3.5).

	Kununu Review Aspect	**Kununu Description**	**Herzberg Factor**
Company Culture	Company culture	How is the overall company culture?	Hygiene
	Internal Communication	How is internal communication? To what extent are you informed about company results, successes, and challenges through regular communication?	Hygiene
	Teamwork	How are co-workers at working together and interacting in an honest, direct manner?	Hygiene
	Work-Life Balance	How does the company value work-life balance? Are families considered? Is there pressure to work long hours?	Hygiene
	Support from Management	Does leadership set realistic expectations, communicate clear goals, and involve employees in the decision making process?	Both
	Freedom to work independently	To what extent are you trusted to work independently?	Motivation
Diversity	Gender Equality	Are women treated equally and given the same career opportunities?	Both
	Attitude towards older colleagues	Does the company hire older workers? Are senior colleagues appreciated, supported, and given equal opportunities?	Both
Work Environment	Office / Work Environment	Is the work environment comfortable and suited to do the work you are doing? Is there proper ventilation, lightning, temperature control and technology available?	Hygiene
	Environmental Friendliness	To what extent does the company demonstrate concern or awareness for environment?	Hygiene
Career	Overall compensation for your work	Overall, do you feel that you are fairly compensated for your work?	Hygiene
	Company Image	Are you proud to work for your company?	Hygiene
	Career Development	How are your career prospects for growth and professional development?	Motivation

H1 (Attention): As motivation factors positively influence employee satisfaction, we expect satisfied employees to write more in reviews of aspects assigned with motivation factors as compared to hygiene factors which, when fulfilled, are taken for granted and do not draw much attention [F. Herzberg, Mausner, and Snyderman, 1959; Gawel, 1996; Alshmemri, Shahwan-Akl, and Maude, 2017]. On the contrary, if hygiene factors are not fulfilled, dissatisfaction among employees increases and, thus, we expect that dissatisfied employees devote more attention towards hygiene factors and complain about their absence [F. Herzberg, Mausner, and Snyderman, 1959; Gawel, 1996; Alshmemri, Shahwan-Akl, and Maude, 2017]. Specifically, we operationalize attention in two different ways and define the following four hypotheses:

H1.1: *Satisfied employees write more reviews on motivation factors.*
H1.2: *Dissatisfied employees write more reviews on hygiene factors.*

Here, we compute and report ratios of the number of aspect reviews that contain optional review text in either positive and negative reviews, respectively for each country.

H1.3: *Satisfied employees write longer reviews for motivation factors.*
H1.4: *Dissatisfied employees write longer reviews for hygiene factors.*

For these two hypotheses, we define attention as aspect review lengths (i.e., how much attention was devoted to individual aspects by reviewers) and compare differences in medians between distributions of positive and negative reviews.

H2 (Sentiment): As previous studies suggest that unfavorable experiences result in negative emotions [Westbrook, 1987; Bougie, Pieters, and Zeelenberg, 2003; Mattsson, Lemmink, and McColl, 2004], we expect that reviews from dissatisfied employees about hygiene factors (as their absence should lead to dissatisfaction) convey a negative sentiment. In contrast, we expect that reviews from satisfied employees about motivation factors (as their fulfillment should lead to satisfaction) convey a positive sentiment. Specifically, we investigate:

H2.1: *Satisfied employees write more positively about motivation factors.*
H2.2: *Dissatisfied employees write more negatively about hygiene factors.*

To test for these hypotheses, we investigate the sentiment conveyed in reviews and fall back on existing translated German and English sentiment dictionaries [Yanqing Chen and Skiena, 2014] which are comparable to each other since they originate from the same dictionary. Specifically, we compute the sentiment s of an aspect review by $s = (W_p - W_n)/(W_p + W_n)$, where W_p is the number of positive words and W_n is the number of negative words in a review. Thus, s ranges from -1 to $+1$, where positive (and respectively negative) values represent a positive (resp. negative) sentiment and values close to zero indicate a neutral sentiment. We investigate the sentiment for positive and negative reviews and again compare differences in medians between the two distributions. Note that for this analysis we only consider German and English reviews (identified through automatic language detection) with at least one hundred words comprising at least one word of our sentiment dictionaries, as sentiment features are hard to interpret otherwise.

H3 (Readability): Existing research suggests that complaining is a behavioral response to dissatisfaction [J. Singh, 1988; Maute and Forrester Jr, 1993; Zeelenberg and Pieters, 2004], providing a way to cope with emotions by venting one's dissatisfaction [E. W. Anderson, 1998]. Further, previous findings suggest that negative reviews are harder to read as reviewers address a wider range of issues when describing their bad experiences [Y. Zhao, X. Xu, and M. Wang, 2019]. Similar findings report, for example, a positive correlation between satisfaction and better readability (i.e., negative reviews are harder to read) of reviews when assessing the helpfulness of online reviews [Korfiatis, Rodríguez, and Sicilia, 2008; Korfiatis, García-Bariocanal, and Sánchez-Alonso, 2012].

Following these previous observations, we expect similar behavior for online employer reviews and investigate the readability of positive and negative reviews in terms of the Two-Factor Theory. In particular, we expect:

H3.1: *Satisfied employees write more readable reviews about motivation factors.*
H3.2: *Dissatisfied employees write less readable reviews about hygiene factors.*

We test for these hypotheses by computing the *Flesch reading ease* [Flesch, 1948], providing us with a score ranging between 0 and 100, where texts with values closer to 0 are considered to be harder to read and text with values closer to 100 are considered to be easier to read. Note that we try other readability formulas[10] as well, but results are very similar as these formulas are known to have high inter-correlation [DuBay, 2004]. Similar to our analysis on sentiment, we compute readability for positive and negative reviews, compare differences in medians and only consider German and English reviews with at least 100 words as scores might result in non-interpretable results otherwise.

H4 (Content): Following the two-factor theory [F. Herzberg, Mausner, and Snyderman, 1959; Gawel, 1996; Alshmemri, Shahwan-Akl, and Maude, 2017], we expect, independent from the reviewed aspect, satisfied employees to focus on content related to motivation factors and dissatisfied employees to focus on content related to hygiene factors. In particular, we expect that satisfied employees are the only ones able to experience and write about motivation factors while they take hygiene factors for granted and neglect them. Dissatisfied employees, on the other hand, should have no experience with motivation factors to write about and only focus on not fulfilled hygiene factors. Thus, we investigate the following two hypotheses:

H4.1: *Satisfied employees use more words related to motivation factors.*
H4.2: *Dissatisfied employees use more words related to hygiene factors.*

To test for these, we adopt the method from Hofland and Johansson [1982], which is based on contingency tables and chi-squared (χ^2) tests to assess which words are characteristic for either of two corpora. More precisely, for each aspect we look at the sets of the top 100 nouns (after removing stop words and lemmatization) included in positive and negative reviews and build the union of those two sets. Next, for each aspect and each word from the union we build a 2×2 contingency table, which keeps the count of a given word, as well as the total count of all other words in both positive and negative reviews. The null hypothesis of the χ^2 test (which we perform with Yates Correction [Yates, 1934] to counteract the fact that 2×2 contingency tables are not continuous) states that the occurrence of a given word is independent of the controversy of the comment. Hence,

[10] Including the *Flesch Kincaid grade level*, the *Coleman Liau index*, the *automated readability index* and the *Dale-Chall readability formula* with an absolute (due to inverse scales for some formulas) mean Spearman rank-order correlation coefficient of 0.86 with the Flesch reading ease.

words for which we can reject this null hypothesis are used distinctively in either positive or negative reviews. For this analysis, we consider the top 20 significant words in positive and negative reviews with regard to their χ^2 values and to their relative frequencies in reviews in order to decide where is their usage significantly higher, respectively for German and English reviews. To evaluate if top words reflect hygiene and motivation factors, we compute overlaps with words that, according to literature and the theory, are mentioned frequently and of high relevance in connection with both sets of factors. More precisely, we let three independent annotators read a selection of existing English studies investigating the Two-Factor Theory [Gawel, 1996; Bassett-Jones and Lloyd, 2005; Smerek and M. Peterson, 2007; Islam and Ali, 2013; M. E. Malik and Naeem, 2013; Oladotun and Öztüren, 2013; Alshmemri, Shahwan-Akl, and Maude, 2017] with the aim to select important words related to both hygiene and motivation factors, leaving us with two sets of words. Annotators independently found 50 distinct words for hygiene factors of which we keep 22 words that were found by at least two of them. For motivation factors, annotators identified 35 distinct words of which we consider 11 words found by at least two of them. We then translate words to German and extend the sets of either language by adding synonyms manually selected by using Wiktionary.org and Thesaurus.com. We refer to Table 5.6 in the Appendix section for a complete list of extracted German and English words related to hygiene and motivation factors. To evaluate the overlap of words from our automatic subgroup discovery with words extracted by annotators, we compute the Jaccard Index[11] respectively for positive and negative words, each aspect and language.

To check our results for robustness, we repeat this experiment and compare manually extracted words with the top words form the subgroup discovery by considering their word embeddings, allowing us to incorporate semantics in this analysis. Specifically, we use pre-trained German and English vectors [Grave et al., 2018] from the fastText library[12]. We average vectors over words contained in a respective group (i.e., manually extracted Ger-

[11] The Jaccard Index is defined as the size of the intersection relative to the size of the union of two sets resulting in a value ranging between 0 (no overlap) and 1 (complete overlap).

[12] Link to website: `https://fasttext.cc`

man words, manually extracted English words, top positive German words, top negative German words, top positive English words and top negative English words) and compute cosine similarities between mean vectors from top words and manually extracted words.

H5 (Generalization): There is previous work providing evidence both for [Lodahl, 1964; Cummings, 1975] as well as against [Behling, Labovitz, and Kosmo, 1968; Furnham, Forde, and Ferrari, 1999] the generalization of the Two-Factor Theory. To shed light on how our previous findings generalize, we study the following three hypotheses:

H5.1: *Results are independent of cultural context.*
H5.2: *Results are independent of industry.*
H5.3: *Results are independent of employment status (i.e., current or former).*

To test for these hypotheses, we conduct the same analyses as for previous hypotheses respectively for each country and industry contained in our dataset as well as well for reviews from current and former employees. We quantify results by counting the number of times the two motivation factors are among the top (H1 and H2) five aspects and the bottom (H3) five aspects according to differences in median between positive and negative reviews respectively. This leaves us with 8 cases for cultural differences (2 motivation aspects times four countries), 86 cases for industrial differences (2 motivation aspects times 43 industries) and 4 cases for employment status differences (2 motivation aspects times two statuses).

Prediction

We now investigate the applicability of our findings from the hypothesis tests by conducting a prediction experiment. For that, we leverage the features computed for our previous analyses, allowing us to investigate the predictiveness of review aspects for employee satisfaction on a company level. More precisely, we want to predict whether or not a company has high or low employee satisfaction by exploiting the content as well as stylistic characteristics of reviews. Our results uncover that both content and style of aspect reviews are predictive of employee satisfaction.

Experimental Setup. We first split our dataset according to the language of reviews, leaving us with a German and English subset of reviews. Next, we aggregate review texts over individual companies and remove all companies with an aggregated review length less than 10 000 characters, respectively for the German and English subset. This leaves us with 7 148 companies located in Austria, Germany and Switzerland as well as 904 companies located in the USA. We frame our prediction task as a binary classification problem, predicting either a high or low employee satisfaction. We define the employee satisfaction (prediction target) of a company based on a majority vote, meaning that we label a company with either high or low employee satisfaction based on the majority of either positive or negative reviews written for that company. In cases of equal numbers of positive and negative reviews, we exclude companies from our prediction task, arguing that these companies have neutral employee satisfaction. After removing undecided cases and labelling companies, we remain with 2 955 companies having a high and 4 067 companies having a low employee satisfaction based on German reviews (minimum # of reviews: 1, maximum # of reviews: 7 090, mean # of reviews: 50.84 over Austrian, German and Swiss companies), as well as 450 and 430 companies, respectively for English reviews (minimum # of reviews: 1, maximum # of reviews: 1 148, mean # of reviews: 100.46 over companies located in the USA). To assess the predictive power of factors, we utilize the top positive and negative words extracted for individual aspects during the testing of H4 (Content) and count their occurrences in aggregated review texts to create numeric features for each company.

Feature Spaces. Since we are interested in the differences of significance between aspects assigned to hygiene, motivation and both factors, we accordingly separate factors into three different features spaces. In particular, we consider the total count of the top twenty positive and the top twenty negative words for each aspect and group these counts according to aspects and their assigned factor. This leaves us with three features spaces comprising counts of top words respectively for aspects assigned to *hygiene*, *motivation* and *both factors*. We complement these three features spaces by including: *textual features* comprising the mean sentiment, the mean Flesch reading ease and the mean number of words over positive and negative reviews of a respective company, and the *combination of all* of the above feature spaces. Finally, we consider *TF-IDF* (minimum document frequency:

10%; maximum document frequency: 80%; maximum number of words: 5 000; stop words removed) vectors of reviews combined with textual features, representing the upper limit and an approach dissociated from the Two-Factor Theory. This allows us to compare the predictive power of words related to hygiene and motivation factors with general bags-of-words.

Addressing Imbalanced Factor Assignment. To compensate for the imbalanced factor assignment (8 aspects assigned to hygiene factors, 2 aspects assigned to motivation factors and 3 aspects assigned to both factors), we further introduce a feature space with *subsampled hygiene factors*. More precisely, we randomly select two aspects assigned to hygiene factors for 1 000 times, allowing us to do a fair comparison of hygiene and motivation factors.

Evaluation. We conduct our prediction task using logistic regression with ℓ_2 regularization (to prevent overfitting the training data) as implemented in scikit-learn[13]. To evaluate our models, we split data into train and held out test sets (80 to 20 percent ratio) multiple times by performing stratified random sampling over 20 random runs, respectively for each feature space. We report mean balanced accuracy (defined as the average of recall obtained for both classes and suitable for imbalanced datasets) over random runs for each feature space. In case of our subsampled hygiene factors, we report mean values over the 1 000 random runs.

We compare results with an improved baseline determining the employee satisfaction of a company based on that of other companies operating in the same country and industry. For example, the employee satisfaction of a marketing company operating in Austria would be positive if the majority of all other companies operating in marketing and Austria would be positive, or negative otherwise. For that, we again conduct the 20 random train and test set splits and report the mean balanced accuracy over these random runs.

[13]Link to website: `https://scikit-learn.org`; Version used: 0.21.3

5.3.6 Empirical Results

Hypotheses

We now describe the results for individual hypotheses and provide an overview of whether or not we find support for them in Table 5.2.

H1 (Attention). We list the ratios of aspect reviews having optional review text in Table 5.5 and report that ratios are higher for negative reviews as well as all aspects. This indicates that employees with negative experiences rather tend to write reviews than satisfied employees independent from aspects and their assignments to hygiene and motivation factors. We argue that this observation could be due to the well-known *negativity bias* [Baumeister et al., 2001; Rozin and Royzman, 2001; Hilbig, 2009] which suggest a general tendency of people to focus on negative experiences.

Regarding H1.1, we observe that satisfied employees do not write more reviews for aspects related to motivation factors as compared to other aspects, suggesting a rejection of our hypothesis. However, in case of H1.2, we see that dissatisfied employees tend to write more reviews for aspects related to hygiene factors, thus, indicating support for this hypothesis.

In Figure 5.8, we illustrate the review length (in characters) distributions for positive and negative employer reviews, respectively for each of the 13 aspects. We verify the significance of differences between positive and negative distributions for each aspect by computing two-sided[14] Mann-Whitney-Wilcoxon tests at the Bonferroni corrected p-value of $\alpha = 0.004$ (corresponding to 13 aspect comparisons at $\alpha = 0.05$). We report p-values smaller than our significance level for 11 out of the 13 aspects, meaning that their difference in text length between positive and negative reviews is significant. The two aspects with non-significant differences are *company image* (p-value $= 0.24$) and *attitude towards older colleagues* (p-value $= 0.06$).

[14]Note that we use two-sided statistical tests for all hypotheses to avoid placing assumptions on the direction of differences between positive and negative reviews. In other words, we test for the existence of statistically significant distributional differences, rather than the direction (and magnitude) of such differences.

Table 5.5: **Ratios of Reviews with Optional Review Text.** The table lists the ratios of reviews with optional review text, respectively for positive and negative reviews as well as each review aspect.

Positive Reviews			Negative Reviews		
Aspect	**Assigned Factor**	**Ratio**	**Aspect**	**Assigned Factor**	**Ratio**
Company Culture	Hygiene	18.18	Support from Management	Both	33.63
Support from Management	Both	16.64	Company Culture	Hygiene	28.52
Teamwork	Hygiene	16.17	Teamwork	Hygiene	26.52
Internal Communication	Hygiene	15.19	Internal Communication	Hygiene	26.23
Freedom to work independently	Motivation	14.93	Overall compensation for your work	Hygiene	24.08
Work-Life Balance	Hygiene	14.82	Work-Life Balance	Hygiene	23.16
Office / Work Environment	Hygiene	12.87	Office / Work Environment	Hygiene	22.43
Career Development	Motivation	12.80	Freedom to work independently	Motivation	21.53
Overall compensation for your work	Hygiene	11.62	Career Development	Motivation	21.31
Company Image	Hygiene	10.34	Company Image	Hygiene	20.51
Attitude towards older colleagues	Both	9.47	Gender Equality	Both	16.39
Gender Equality	Both	9.12	Attitude towards older colleagues	Both	15.91
Environmental Friendliness	Hygiene	7.89	Environmental Friendliness	Hygiene	14.08

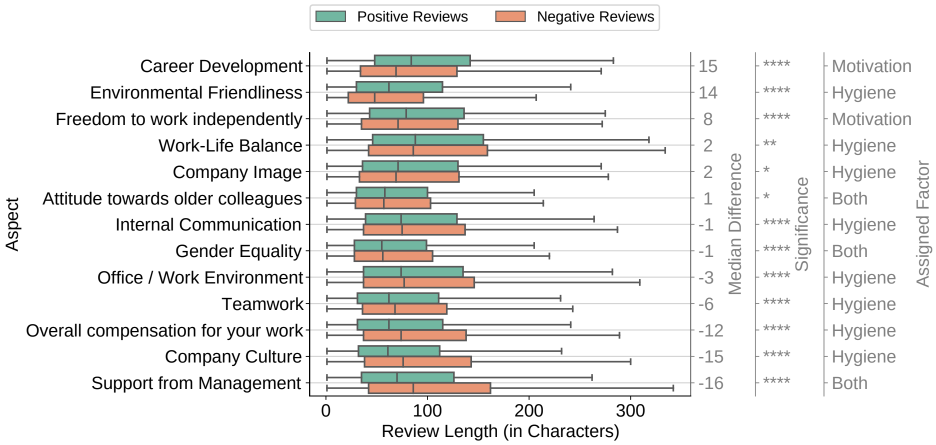


Figure 5.8: **Results for H1 (Attention).** The figure depicts the results for our two hypotheses (H1.3 and H1.4) focusing on the attention devoted by satisfied and dissatisfied employees towards review aspects. We expect that motivation factors receive more attention from satisfied employees and hygiene factors more from dissatisfied employees. With the box plot we illustrate the distributions of the number of characters in reviews and list aspects in descending order (top to bottom) by the difference in medians between positive (green color) and negative (red color) reviews. Vertical black lines indicate the median and the first and third quartile. Whiskers indicate minimum and maximum values still within 1.5 interquartile ranges. Stars indicate the significance of differences between positive and negative reviews based on two-sided Mann-Whitney-Wilcoxon tests (*: p-value ≤ 0.05, **: p-value ≤ 0.01, ***: p-value ≤ 0.001, ****: p-value ≤ 0.0001). We find motivation factors to be more relevant in positive reviews. On the contrary, for negative reviews, we observe that hygiene factors are more relevant than motivation factors, supporting both H1.3 and H1.4.

Specifically, we observe the largest positive differences in medians for the motivation factors *career development* and *freedom to work independently* as well as the hygiene factor *environmental friendliness*. This suggests that satisfied employees devote significantly more attention towards motivation factors as compared to dissatisfied employees and supports H1.3. We discuss the case of *environmental friendliness* in detail in the Discussion section of the paper. For the majority of hygiene factors, we report that they are more relevant in negative reviews which supports H1.4. Notably, the aspect *support from management* (assigned to both hygiene and motivation factors) has the largest negative difference in medians, suggesting that dissatisfied employees write more about issues related to management than satisfied employees.

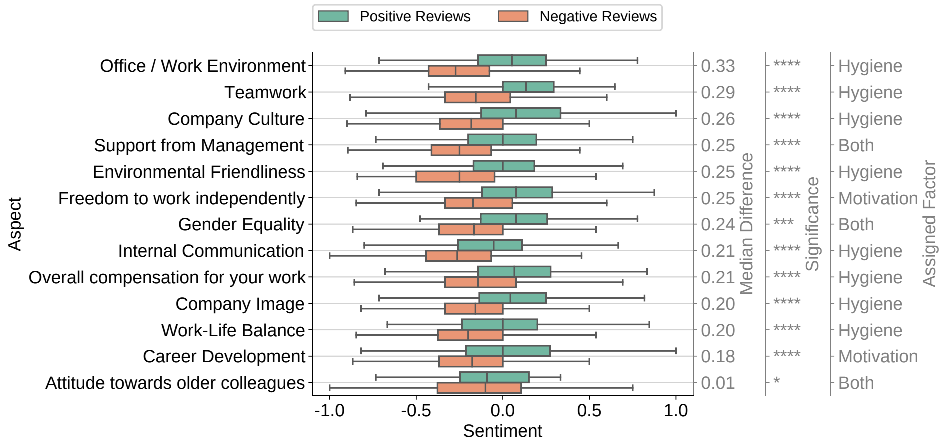


Figure 5.9: **Results for H2 (Sentiment).** The figure illustrates the results for our two hypotheses (H2.1 and H2.2) which focuses on the sentiment conveyed in aspect reviews written by satisfied and dissatisfied employees. We expect that reviews from satisfied employees on motivation factors are positive and reviews from dissatisfied employees on hygiene factors are negative. With the box plot we illustrate the distributions of sentiment conveyed in reviews. In general, we find that, as expected, satisfied employees express a more positive sentiment as compared to dissatisfied employees. While we find a more negative sentiment form dissatisfied employees towards hygiene factors as expected, we also find more positive sentiment from satisfied employees for hygiene factors instead of motivation factors. Thus, these findings provide support against H2.1 and for H2.2 and even suggest higher relevance of hygiene factors for satisfied employees than what we expected based on the Two-Factor Theory.

Overall, we confirm H1.2, H1.3 and H1.4, suggesting that motivation factors may be more relevant to satisfied employees while hygiene factors may be more relevant to dissatisfied employees.

H2 (Sentiment). We depict sentiment distributions for positive and negative reviews respectively for each aspect in Figure 5.9. To assess the significance of differences between medians, we again compute two-sided Mann-Whitney-Wilcoxon tests with our corrected significance level $\alpha = 0.004$. We report significant differences for 12 out of 13 cases, with differences for *attitude towards older colleagues* (median of positive reviews: -0.09; median of negative reviews: -0.10) being non-significant at a p-value of 0.36, suggesting that ageism is generally perceived more negatively.

Overall, we find that, as expected, the sentiment conveyed by reviews from satisfied employees is more positive as compared to the sentiment of reviews from dissatisfied employees (all differences in medians are positive), suggesting our method of assessing sentiment is capable of capturing the overall satisfaction of employees. Further, we report that, in accordance with our hypothesis H2.2, aspects assigned to hygiene factors are more negatively perceived in reviews from dissatisfied employees (medians for aspects assigned to hygiene factors ranging from -0.26 to -0.14.) as compared to those from satisfied employees (medians for aspects assigned to hygiene factors ranging from 0 to 0.13.), which are rather neutral. However, we observe that our results regarding H2.1 are inconclusive. For *freedom to work independently* and *career development* we report a neutral sentiment in positive reviews with medians of 0.08 and 0, respectively. This indicates that satisfied employees, in contrast to our expectations, do not write more positively about aspects assigned to motivation factors. In fact, we observe that hygiene factors are perceived more positively as compared to motivation factors in reviews from satisfied employees, suggesting that the former may not only prevent the occurrence of dissatisfaction but also foster satisfaction. As such, hygiene factors may be even more important than originally envisioned by Herzberg. When considering sentiment of both positive and negative aspect reviews combined, we report that reviews on aspects assigned to motivation factors are overall perceived more positively as compared to the majority of aspects assigned to hygiene factors, as dissatisfied employers are more neutral towards motivation factors.

Overall, we reject H2.1 and find strong support for H2.2 as well as a higher relevance of hygiene factors as initially expected based on the Two-Factor Theory.

H3 (Readability). In Figure 5.10, we illustrate the Flesch reading ease distributions for positive and negative reviews, respectively for each aspect. We again check for the significance of differences in a similar fashion to the analysis of H1 (Attention) and H2 (Sentiment). Here, we find 7 out 13 cases to be non-significant (*support from management*, *gender equality*, *attitude towards older colleagues*, *environmental friendliness*, *teamwork*, *company image* and *overall compensation for your work*).

Contrary to our expectations, reviews from dissatisfied employees are in

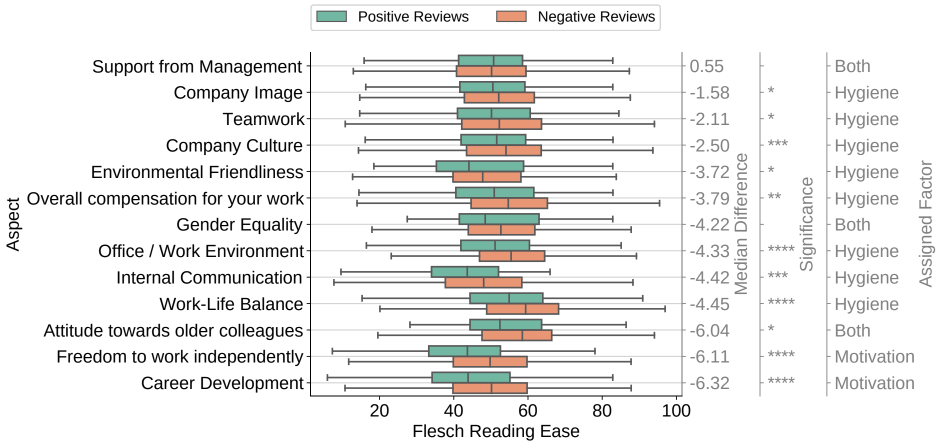


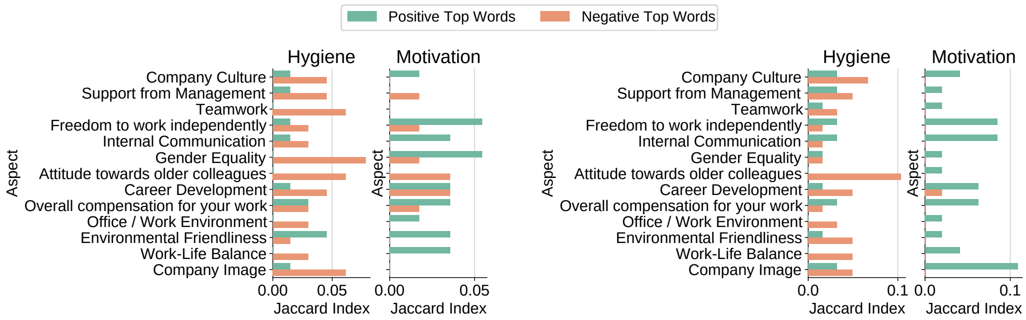
Figure 5.10: **Results for H3 (Readability).** The figure depicts the results for the two hypotheses (H3.1 and H3.2) focusing on the readability of reviews from satisfied and dissatisfied employees. We expect that reviews from satisfied employees on motivation factors are easier to read and reviews from dissatisfied employees on hygiene factors to be harder to read. With the box plot, we illustrate the distributions of the Flesch reading ease (higher values mean easier to read) of reviews. We find that reviews on motivation factors are harder to read as compared to hygiene factors. These results refute our two hypotheses and are contrary to findings in previous research which showed that reviews form dissatisfied people are harder to read.

general easier to read than reviews from satisfied employees (with the exception of *support from management*), providing different results compared to previous studies [Korfiatis, Rodríguez, and Sicilia, 2008; Korfiatis, García-Bariocanal, and Sánchez-Alonso, 2012; Y. Zhao, X. Xu, and M. Wang, 2019]. Most notably, reviews from satisfied employees on aspects assigned to motivation factors are among the top three of hard to read aspect reviews with medians ranging between 43.68 and 43.85. This suggests a rejection of our hypotheses H3.1 as well as H3.2 and indicates substantial differences between the behavior of reviewing products and reviewing employers.

Connecting these results with the ones from H2 (Sentiment), we again observe that hygiene factors may be more important than initially thought as satisfied employees write more complex reviews for both hygiene and motivation factors when compared to dissatisfied employees. Following the results from H1.1 and H1.2, which indicate that dissatisfied employees would rather write optional review text, we observe that hygiene factors may be fundamentally important for both preventing dissatisfaction and fostering satisfaction and that motivation factors become only relevant when employees have higher ambitions to develop their careers. A potential explanation for the more complex reviews on aspects assigned with motivation factors may be that the subjects related to them are more complicated to describe or that they are only relevant to formally educated employees who may engage in a more critical thinking. Further corroborating this idea, we observe the largest negative median differences for aspects assigned to motivation factors, suggesting that satisfied employees write particularly more complex about motivation factors.

H4 (Content). We depict the results for our fourth hypothesis in Figure 5.11. Overall, we observe, as expected, only minimal overlaps with a mean Jaccard index of 0.02 for German top words and a mean Jaccard index of 0.03 for English top words. When considering the differences across hygiene and motivation factors as well as positive and negative top words, we find support for both H4.1 and H4.2. In particular, we report a higher mean Jaccard index for negative top words and words related to hygiene factors in both German and English cases (German & positive: 0.01; German & negative: 0.04; English & positive: 0.02; English & negative: 0.04). On the contrary, we observe opposite behavior for words related to motivation factors, for which positive top words have a higher mean Jaccard index for both languages (German & positive: 0.02; German & negative: 0.01; English & positive: 0.05; English & negative: 0.00). Similar to the results of H2 (Sentiment) and H3 (Readability), we observe that words related to hygiene factors are also relevant in reviews from satisfied employees, though less prominent as compared to reviews from dissatisfied employees. Further strengthening these findings, we report that words related to motivation factors are also used in reviews for aspects assigned to hygiene factors, indicating a high relevance of hygiene factors in online employer reviews.

Regarding the results of word similarities based on word embeddings, we find trends which are similar to our results based on the Jaccard Index (see Figure 5.14 in the Appendix section). For German, we report higher similarities between words related to hygiene factors and negative top words for 9 of 13 aspects and higher similarities between words related to motivation factors and positive top words for 7 aspects. For English, we find



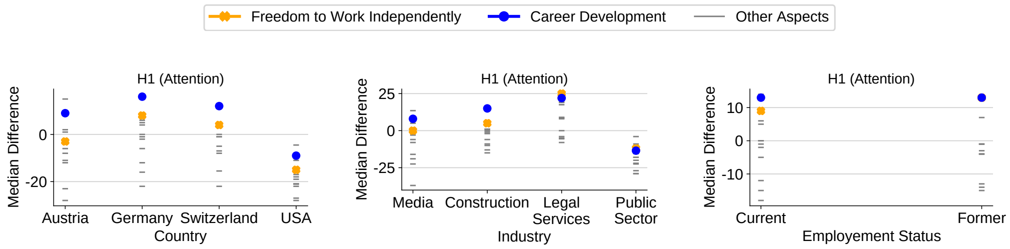
(a) German (b) English

Figure 5.11: **Results for H4 (Content).** The figure depicts the results for our two hypotheses (H4.1 and H4.2) focusing on the content of aspect reviews. We expect that the content of negative reviews reflects hygiene factors whereas the content of positive reviews reflects motivation factors. To analyse this hypothesis, we extract top words that are distinctively used in positive and negative reviews and compute the Jaccard Index to infer their overlap with manually extracted words (see Table 5.6 in the Appendix section) related to hygiene and motivation factors, respectively for positive and negative reviews as well as each aspect. In Figure 5.11a, we illustrate results for German reviews and find larger overlaps for words related to motivation factors with words from positive reviews (strong support for H4.1). On the contrary, we report larger overlaps for words related to hygiene factors and negative top words (strong support for H4.2). We find similar results for English top words (see Figure 5.11b). Note that words related to motivation factors were also used by satisfied employees for reviewing aspects related to hygiene factors, indicating their high relevance in online employer reviews.

words related to hygiene factors to be more similar to negative top words for 8 aspects and words related to motivation factors to be more similar to positive top words for 10 aspects. This suggests that our results are robust for semantics and further strengthens the support for H4.1 and H4.2.

A manual inspection of extracted top words further suggest that reviewers specifically address aspects in their reviews[15]. For example, the German top words *attitude towards older colleagues* comprise "alt" (German for "old"), "Umgang" (German for "handling") and "Altersklassen" (German for "age

[15]Extending this analysis by computing overlaps with words from Wikipedia articles related to aspects further strengthens our findings that reviewers specifically address aspects. We provide more details in the Appendix section.



(a) Cultural Differences (b) Industrial Differences (c) Employment Status Differences

Figure 5.12: **Results for H5 (Generalization).** The figure depicts the results for our hypotheses (H5.1, H5.2 and H5.3) focusing on the generalization of our previous findings (H1 to H3). We expect that the two aspects assigned to motivation factors *freedom to work independently* (orange color and × marker) and *career development* (blue color and ● marker) receive more attention (H1) and a more positive sentiment (H2) while being harder to read (H3) in positive reviews as compared to remaining aspects (gray color) and negative reviews. We find support for H1 (Attention) across cultural (a), industrial (b) and employment status (c) differences. Results for H2 (Emotions) and H3 (Readability) are inconclusive among the three comparisons.

groups"), each related with ageism in working life and clearly related to the aspect. In case of English top words we observe that, for example, words extracted from reviews on *gender equality* including "gender", "woman" and "discrimination" clearly reflect issues related to sexism.

H5 (Generalization). We depict results for our hypothesis on the generalization of previous findings for H1 (Attention) in Figure 5.12. Starting with cultural differences (H5.1; see Figure 5.12a), we report more attention devoted towards aspects assigned to motivation factors in positive reviews for Germany and Switzerland. In case of Austria, only *career development* receives more attention in positive reviews, while employees in the USA devote more attention towards all aspects in negative reviews. However, for all eight cases the median differences of aspects assigned to motivation factors is among the top five, providing similar results as observed for H1 (Attention). Regarding H2 (Sentiment), we observe that positive reviews convey a more positive sentiment for all aspects and countries. Exceptions are *Attitude towards older colleagues* for Austria (median difference $= -0.52$) and *Internal Communication* for Switzerland (median difference $= -0.03$) as these aspects received more attention in negative reviews. However, quantifying our results for H2 (Sentiment) and H3 (Readability), we find inconclusive results (3 of 8 and 5 of 8 cases respectively), indicating differences across countries when considering the sentiment conveyed by and the readability of reviews.

In Figure 5.12b, we report results for H5.2 and selected industries (based on largest positive and negative median differences for both aspects assigned to motivation factors). We observe longer positive reviews for both aspects assigned to motivation factors for industries *media*, *construction* and *legal services*, while for the *public sector* we observe longer negative reviews for these two aspects. Results for H2 (Sentiment) and H3 (Readability) are inconclusive again. Overall, we find support in 80 out of 86 cases for H1 (Attention), 41 out of 86 cases for H2 (Sentiment) and 34 out of 86 cases for H3 (Readability).

Finally, for attention differences between current and former employees (H5.3; see Figure 5.12c), we report more attention devoted towards aspects related to motivation factors in positive reviews from current and former employees (H1; 4 out of 4 cases). For H2 (Sentiment), we observe (in general) more negative reviews from former employees, suggesting that they may air their frustrations after termination. Distinguishing between hygiene and motivation factors is, similar to other comparisons, inconclusive (1 out of 4 cases). We find positive reviews on aspects assigned to motivation factors harder to read for both current and former employees (H3; 4 out of 4 cases), suggesting no difference between them.

Summary of Hypotheses Findings. We find that satisfied employees devote more attention to motivation factors (strong support for H1.3), whereas dissatisfied employees devote more attention to hygiene factors (strong support for H1.4) which reflects the original definition of Herzberg's Two-Factor Theory. Regarding sentiment conveyed in reviews, we find that dissatisfied employees write more negatively about hygiene factors (support for H2.2) and satisfied employees, contrary to our expectations, write more positively about hygiene factors instead of motivation factors (rejection of H2.1). Our results for readability of reviews refute our initial expectations of harder to read reviews from dissatisfied employees on hygiene factors and reflect the exact opposite behavior with harder to read reviews from satisfied employees on hygiene and motivation factors (rejection for H3.1 and H3.2).

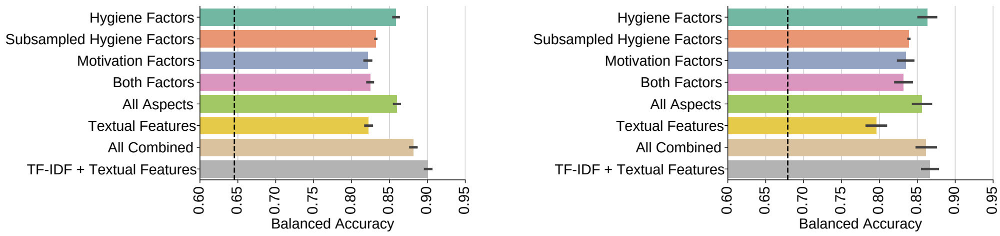
Further, we report that satisfied reviewers tend to specifically mention words related to motivation factors in reviews for aspects assigned to both hygiene and motivation factors (support for H4.1), whereas dissatisfied reviewers mostly mention words related to hygiene factors (support for H4.2). When investigating the generalization of previous hypotheses, we observe that some of our findings generalize across cultural, industrial and employment status differences (weak support for H5.1, H5.2 and H5.3). Overall, we find hygiene factors to be more relevant and important than motivation factors in the context of online employer reviews.

Prediction

We depict performance results for each feature space in Figure 5.13. In general, we report accurate prediction performance with a mean balanced accuracy of at least 0.84 and 0.85 over all models, respectively for German and English reviews. As such, we outperform our improved baseline for German (0.65) by at least 0.19 and for English (0.68) by at least 0.17.

Inspecting the predictive power of review aspects linked to either hygiene, motivation or both factors, we report highest performances for our hygiene factors feature space with a mean balanced accuracy of 0.86 for German and English reviews. We observe slightly worse performance for models using features from aspects assigned to motivation or both factors in case of both languages. However, this seems to be an artifact from the limited number of aspects assigned to these factors as our models using the subsampled hygiene factors performed insignificantly better than these two. When combining the different features from all aspects (i.e., aspects assigned to hygiene and motivation factors), we do not see any further improvements compared to the performance of hygiene factors, signaling yet again the importance of hygiene factors in online employer reviews.

When we consider textual features for prediction models, we report lower performances with a mean balanced accuracy of 0.82 for German and 0.80 for English. However, when combining all features from all aspects as well as textual features, we could improve the mean balanced accuracy to 0.88 for German and 0.86 for English. In particular, for German reviews, note that the performance of the "All Combined" feature space is significantly



(a) German Reviews

(b) English Reviews

Figure 5.13: **Prediction Task Results.** The figure illustrates the results from our prediction task aiming to predict employee satisfaction on a company level. Balanced accuracy is measured over twenty random train and test splits, respectively for each feature space and German (left-hand side) and English (right-hand side) reviews. The vertical dashed black lines indicate results for our improved baseline and error bars indicate bootstrapped 95% confidence intervals. Overall, prediction performance is good with varieties across feature spaces for both languages. Regarding German reviews (a), we observe that features from aspects assigned to hygiene factors achieve best performance which is slightly outperformed by the combination of features from all aspects. In case of English reviews (b), we report similar results, except that the combination of all aspect feature spaces does not yield better prediction performance as compared to the model with aspects linked to hygiene factors only. Even though textual features (i.e., text length, readability and sentiment) perform worse compared to features from aspects, the combination of all aspect and textual features achieves the best performance, respectively for German and English reviews. Note that for both languages performance of models with words related to the Two-Factor Theory perform almost similar to general bag-of-words models (TF-IDF + Textual Features) comprising a multiple of words and, hence, more information. This further highlights the predictive strengths of theory related words.

better than the "All Aspects" one (bootstrapped $p < 0.001$), suggesting that capturing textual content and sentiment contained in reviews results in the best performance when predicting employee satisfaction for a company. Comparing the models based on the Two-Factor Theory to our dissociated TF-IDF approach including textual features, we observe only small deficits of 0.2 for German and 0.01 for English. This indicates that a small number of words related to hygiene and motivation factors can describe employee satisfaction in online reviews almost as well as all words.

Overall, we note that review content of aspects related to hygiene factors has high predictive power (equal to considering content of all review aspects) for employee satisfaction, further suggesting their high relevance in online employer reviews.

5.3.7 Discussion

With our empirical study of employer reviews found on kununu through the lens of Herzberg's Two-Factor Theory, we shed light on what information to expect from them in terms of employee satisfaction. Our prediction experiment demonstrates the high predictive power of review content, enabling us to accurately predict employee satisfaction on a company level. We now connect our findings to our initial research question and discuss their implications for employers.

Online Employer Reviews Through the Lens of Herzberg's Two-Factor Theory. Overall, considering the results for H1 to H5 and keeping in mind the vast number of different employers and industries contained in our dataset as well as the limitation discussed in this section, we observe that hygiene factors are more relevant to reviewers than motivation factors. Thus, when analyzing such reviews in future works, we suggest to focus on hygiene factors as motivation factors seem to be only of incidental relevance to reviewers. In particular, our analysis revealed that, as according to the original theory, hygiene factors attract more attention of dissatisfied employees while motivation factors attract more attention from satisfied employees (see Figure 5.8). This observation reflects the connection of the Two-Factor Theory to Maslow's hierarchy of needs [Gawel, 1996], suggesting that fundamental needs have to be satisfied in order to become motivated to strive for greater things. However, we also found that hygiene factors have the potential to increase satisfaction although, according to the original theory, they should only prevent dissatisfaction. Most notably, we depicted that hygiene factors are perceived more positively by satisfied employees as compared to motivation factors (see Figure 5.9) and that they use terms related to motivation factors in reviews of aspects assigned to hygiene factors (see Figure 5.11). To us, online employer reviews on kununu suggest that employers can satisfy the majority of their employees through fulfilling hygiene factors, whereas motivation factors may need to be fulfilled only for the minority of employees who want to climb the career ladder. Another possible explanation could be that some hygiene factors became more important over time and transitioned into motivation factors as current circumstances are very different to what we had in 1959. Howsoever, it is clear that hygiene factors are more important and relevant to online reviewers of employers and that motivation factors are considered incidental by them or only relevant to a minority of reviewers. This may only be a phenomenon in the context of online employer reviews or it may unfold a new (modern) interpretation of the theory. We suggest further (offline) studies to make more precise assumptions about how to conclude our observations.

Regarding the three review aspects that our annotators assigned to both hygiene and motivation factors, our results suggest that they are more similar to aspects assigned to hygiene factors than aspects assigned to motivation factors, further supporting a higher relevance of hygiene factors. For example, by manually inspecting selected aspect reviews of *support form management*, we see that reviewers mostly use this rating aspect to negatively comment on their supervisors or bosses and that they do not specifically address the points mentioned in the description provided by kununu. As such, reviewers neglect the part of the involvement in decision making processes, potentially explaining why this aspect is not related to motivation factors.

We now briefly discuss the aspect *Environmental Friendliness*, which receives the second most attention in positive reviews (see Figure 5.8). Here, we argue that this may be a reflection of high environmental standards and awareness in European countries. Since more than 67% of reviews in our dataset originate from these countries, it explains the general relevancy in positive reviews contained in our dataset. Further, a recent increase in media coverage of climate change may also add to that observation. To test for this assumption, we investigate the ratio of reviews with optional text for *Environmental Friendliness* over the years. We find that, while in 2015 only 6% of reviews included dedicated texts for this aspect, in 2019 it were already 17%, strengthening our explanation of increased environmental awareness among reviewers and indicating the importance for employers to take action.

Thus, this example highlights the presence of longitudinal effects in shaping individual aspects related to factors that influence employee satisfaction.

Finally, we discuss discrepancies regarding readability between previous studies on product reviews [Korfiatis, Rodríguez, and Sicilia, 2008; Korfiatis, García-Bariocanal, and Sánchez-Alonso, 2012; Y. Zhao, X. Xu, and M. Wang, 2019] (dissatisfied reviewer write harder to read reviews) and our results (satisfied reviewers write easier to read). Here, we argue that the harder to read positive reviews on aspects assigned to motivation factors could be due to the fact that such factors are more likely to be granted to higher employee positions or only in specific industries which have to deal with more complex matters. To test for this assumption, we assess the mean Flesch reading ease of reviews for aspects related to motivation factors, respectively for each employee position as well as each industry. We find reviews from *co-ops* (i.e., employees who simultaneously study and work part-time) and *managers* to be hardest to read and reviews from *apprentices* and *temporaries* to be easiest to read. In the case of industries, we find harder to read reviews for sectors that may require formal education, such as *tax consulting and auditing* and *software engineering*. Further, we observe easier to read reviews for *health, wellness & fitness* or *food production & farming*, which may have less formal requirements. These results indicate that harder to read reviews from satisfied employees may indeed be due to a more advanced critical thinking and language through formal education. An in-depth study of this observation might be a promising research avenue for future work.

Predictiveness of Employee Satisfaction. We demonstrated the predictiveness of employee satisfaction based on a logistic regression model achieving a maximum mean balanced accuracy score of 0.87. By creating different feature spaces, we uncover that review content of aspects linked to either hygiene and motivation factors are equally predictive for employee satisfaction and that only half of the aspects already yield best performance. Only when integrating textual features we could further increase prediction performance, suggesting that not only the content but also stylistic devices should be considered in prediction models, further corroborating similar findings [Siering, Muntermann, and Rajagopalan, 2018]. Our model has the potential to support employers in better understanding the needs of their employees and in helping them to take optimized measures in order to achieve higher efficiency.

Ethical Implications. While our work solely intends to learn about online employer reviews in order to benefit employees as well as employers, performing such analyses may still put both of them at risk. For example, employers may attempt to identify reviewers (despite the fact that reviews on kununu are anonymous), as demonstrated in existing works, such as those of Almishari and Tsudik [2012] or Goga et al. [2013], who correlated texts of users to those posted in a non-anonymous context on other online social platforms. This has the potential to negatively impact the careers of both current (e.g., disciplinary transfer through offended supervisor) and former (e.g., negative reference letters from former employer) employees. Further, employers may misinterpret the general mood of their employees by relying too strongly on their reviews (e.g., because not all employees are aware of such platforms) and, in doing so, adjust their managerial decisions in a way that may create dissatisfaction among their employees.

Analyses of employer reviews may be used for company valuations, as recent research suggests that online employer reviews may relate to stock returns [Green et al., 2019]. Thus, employees may negatively influence valuations by intentionally writing bad reviews ("review bombing" through trolls or bots) or, conversely, employers may trick future investors by whitewashing themselves through faked positive reviews. The manipulation of online reviews to harm or embellish reviewed entities is already a subject of research [Hu, Bose, Gao, et al., 2011; Mayzlin, Dover, and Chevalier, 2014]. Finally, existing research highlights the importance of employer branding for job seekers [Cable and Yu, 2006; Melián-González and Bulchand-Gidumal, 2016]. When building recommender systems based on analyses like ours, one must account for biased reviews to prevent discrimination against employers, as such biases may lead to reduced opportunities for employers to recruit well-educated and talented employees.

Limitations. In our work, we explore reviews found on kununu, one platform among a variety of others providing the possibility to review employers on the Web. While we strongly believe that the amount of data and its variety (i.e., different countries and languages, multiple industries and hundreds of thousands companies) is appropriate for an analysis like this, we still acknowledge a potential sample and selection bias in the type of people that write reviews on kununu. This bias includes different interpretations of review aspects, for example, across countries. Grasping the full extent of cross-cultural differences calls for further qualitative and quantitative research. As such, the inclusion of other platforms, such as glassdoor.com, may help to generalize our study.

Further, the definition of employee satisfaction (i.e., positive and negative reviews) is based on a threshold and, thus, results may change according to it. However, slightly adjusting this threshold or using an alternative definition based on rating stars (i.e., reviews with less than or equal to two stars represent negative reviews and reviews with equal to or more than four stars represent positive reviews) did not noticeably alter our results. Similarly, the input from our annotators is depending on their individual opinions and results may differ if consulting other experts.

5.3.8 Conclusions

Summary. In this work, we conducted a large-scale analysis of online employer reviews through the lens of Herzberg's Two-Factor Theory and investigated characteristics of reviews from satisfied and dissatisfied employees. Overall, we reported that hygiene factors are more relevant to reviewers and that motivation factors are considered incidental or only relevant to a minority of reviewers. While we expected and found that dissatisfied employees devote more attention towards hygiene factors and satisfied employees devote more attention to motivation factors, other experiments suggested a higher importance of hygiene factors contrary to our expectations based on the theory. For example, satisfied employees write more positively about hygiene factors as compared to motivation factors which contradicts with the definition of the theory. Finally, we inspected the generalization of our findings across cultural, industrial and employment status differences and demonstrated their applicability for predicting employee satisfaction on a company level.

Implications. The results of our work highlight the importance of hygiene factors in online employer reviews. Researchers as well as employers who want to leverage such reviews to learn more about employee satisfaction and motivation should lay their focus on hygiene factors as informational content of reviews about motivation factors might be limited. These observations indicate potentially necessary adjustments of the theory's factor assignments due to temporal changes since the introduction of the theory. Further, our analysis distilled deficiencies in some countries and industries with regards to employee satisfaction, highlighting an opportunity to counteract appropriately based on our results and, thus, create better working conditions for employees. Our prediction experiment uncovers the predictive powers of reviews related to hygiene and motivation factors for employee satisfaction, demonstrating how employers could use such models to complement other feedback channels from their employees.

Future Work. A more detailed investigation of certain aspects, including *gender equality* and *handicapped accessibility*, might help to achieve more fair conditions at work. Another promising idea is an in-depth analysis of the temporal component of reviews, including the potential to develop tools that help in better understanding other longitudinal trends in employer satisfaction and the Herzberg theory. Further, our analysis provides insights into the needs of individual employees as well as what is offered by industries and companies, opening up possibilities to support individuals' career choices, similarly to the work of Kern, McCarthy, et al. [2019].

5.3.9 Appendix

Preprocessing

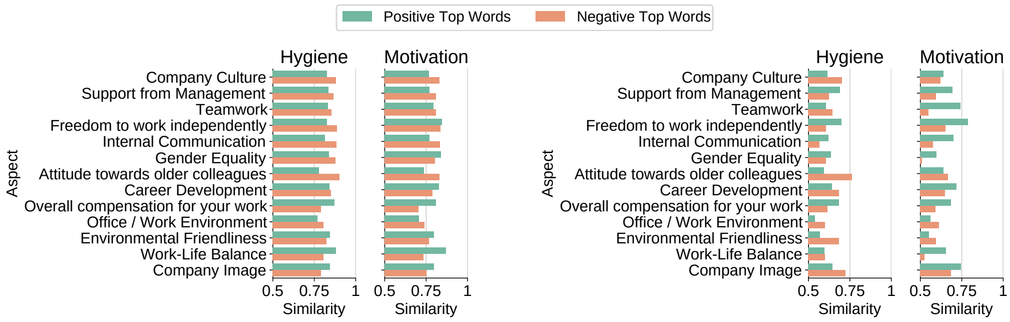
We used the sitemap of respective kununu versions (Austria, Germany, Switzerland, USA) to automatically (via Python' Requests library) gather hyperlinks to all employers found on the platform. Using multiprocessing and multiple web proxies, we crawled and collected all reviews of these employers in less than 48 hours. As such, we could avoid any changes in the website of kununu. We then parsed the aggregated HTML text (via Python's Beautiful Soup library) for each company and extracted individual reviews, always asserting the integrity of data (we did not find any anomalies in HTML texts). After extracting reviews, we checked for missing or duplicate data and found no such instances. Additionally, we manually verified the integrity of selected reviews and found no indications of an incorrect extraction.

Annotation Process

Before presenting the review aspects to the three annotators, we first briefed them about the general principles of the Two-Factors Theory as well as about the platform kununu. We then instructed annotators to read works about the theory [Gawel, 1996; Bassett-Jones and Lloyd, 2005; Smerek and M. Peterson, 2007; Islam and Ali, 2013; M. E. Malik and Naeem, 2013; Oladotun and Öztüren, 2013; Alshmemri, Shahwan-Akl, and Maude, 2017], for which they could take as much time as needed. To make sure they understood the

Table 5.6: **Extracted Words Related to Hygiene and Motivation Factors.** In this table we list the words related to hygiene and motivation factors (F) extracted from English research articles by three independent experts. We translated words to German and extended respective lists by adding synonyms for both languages (L).

L	F	Extracted Words
German	Hygiene	Supervision, Management, Geld, Bedingung, Beziehung, Administration, Firma, Richtlinie, Gehalt, zwischenmenschlich, persönlich, Leben, Kommunikation, Kollegen, Nebenleistungen, Supervisor, Führung, Organisation, Sicherheit, Aufsicht, Überwachung, Vorstand, Geschäftsführung, Leitung, Geschäftsleitung, Zustand, Verhältnis, Bindung, Relation, Verwaltung, Leitung, Unternehmen, Regeln, Vorgehensweise, Bezahlung, Lohn, privat, Mitteilung, Nachricht, Kolleginnen, Mitarbeiter, Mitarbeiterinnen, Vorteile, Betreuer, Vorgesetzter, Leitung, Anführung, Einrichtung, Garantie, Schutz
	Motivat.	Anerkennung, Erreichen, Fortschritt, Arbeit, Verantwortung, Wachstum, Möglichkeit, Inhalt, Mission, Karriere, Bestätigung, Anrechnung, Leistung, Ergebnis, Erfolg, Förderung, Weiterentwicklung, Aufstieg, Entwicklung, Anstellung, Beschäftigung, Tätigkeit, Zuständigkeit, Kompetenz, Pflicht, Zuwachs, Steigerung, Ausbau, Zunahme, Gelegenheit, Chance, Aufgabe, Auftrag, Einsatz, Berufung, Laufbahn, Beruf, Aufstieg
English	Hygiene	supervision, management, money, condition, relationship, administration, company, policy, salary, interpersonal, personal, life, communication, peers, benefits, supervisors, leadership, organization, security, control, guidance, instruction, oversight, authority, situation, status, relation, corporation, guideline, pay, wage, remuneration, social, conversation, co, workers, bonus, assistance, aid, protection, safty, guarantee
	Motiv.	recognition, achievement, advancement, work, responsibility, growth, opportunity, content, mission, career, Synonyms, acceptance, , accomplishment, progression, improvement, progress, development, labour, occupation, effort, accountability, expansion, increase, chance, occasion, objective, assignment, task, goal



(a) German

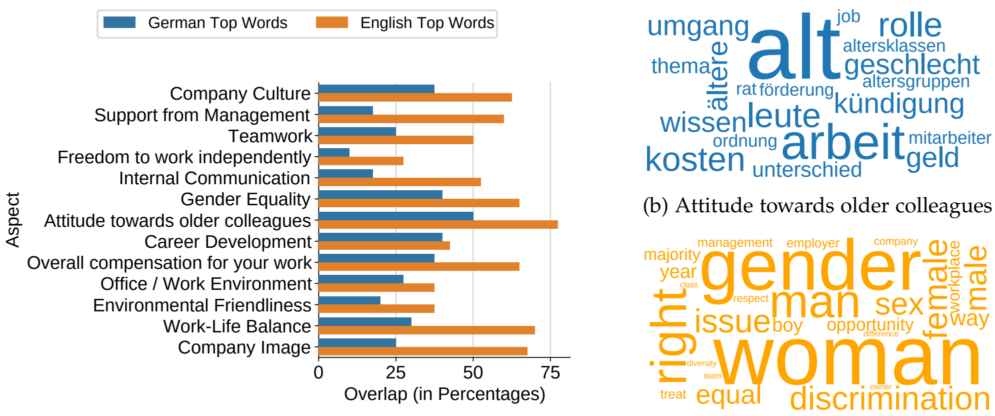
(b) English

Figure 5.14: **Word Embedding Similarity.** The figure depicts supplementary results for our fourth hypothesis focusing on the content of aspect reviews. Similarly to Figure 5.11, we see higher similarities between words related to hygiene factors and words extracted from negative reviews as well as higher similarities between words related to motivation factors and words extracted from positive reviews, respectively for German and English reviews.

underlying principles of hygiene and motivation factors, we separately let the three annotators explain the two factors in their own words, with which we did not find any ambiguity. We then presented the task to the annotators, comprising a table with all 13 review aspects in separate rows. For each aspect, we included the description provided by kununu (see Table 5.4). We then asked annotators to decide whether a review aspect can be assigned to hygiene or motivation factors by considering what they have learned about the theory as well as the description from kununu. Note that we specifically instructed annotators to assign either hygiene or motivation to aspects, even in cases where it was not completely distinguishable for them. Further, we did not limit the task by time and allowed annotators to resort to provided research if they needed to.

Extracted Words Related to Hygiene and Motivation Factors

In Table 5.6, we list the words related to hygiene and motivation factors extracted from existing research by our three independent annotators. In Figure 5.14, we illustrate supplementary results for H4 (Content) based on word embeddings.



(a) Top Words Overlap

(b) Attitude towards older colleagues

(c) Gender Equality

Figure 5.15: **Overlap With Words Extracted From Aspect Related Wikipedia Articles.** The figure depicts the results for experiment on the overlap of words extracted from aspect reviews with words extracted from Wikipedia articles. In Figure 5.15a, we show the overlap of extracted top words with words found in related Wikipedia articles. While we find larger overlaps for English top words, the overlap for German top words is smaller, probably due to shorter article lengths in the German version of Wikipedia. In Figure 5.15b, we depict the overlapping German top words for the aspect *attitude towards older colleagues*, where word sizes represent the number of occurrences in the Wikipedia article. After manually inspecting these words, we argue that reviewers address the aspect in their texts. Similarly, in Figure 5.15c, we illustrate the overlapping English words for the aspect *gender equality*. Again, we can clearly observe that reviewers focus on the aspect.

Wikipedia Articles

To further evaluate if top words reflect aspects, we compute the overlap with words contained in related Wikipedia articles, respectively for German and English language. For example, for the aspect *company culture* we crawl the article *Organizational culture* and investigate the overlap of words (also lemmatized). In Table 5.7, we list the related Wikipedia articles, respectively for the German and English version and in Figure 5.15 we depict the results of this experiment.

Table 5.7: **Assigned Wikipedia Articles.** This table lists the related German and English Wikipedia articles we assigned to the 13 aspects in order to evaluate the content of reviews.

Aspect	**Assigned German Article**	**Assigned English Article**
Company culture	Organisationskultur	Organizational culture
Internal Communication	Interne Kommunikation	Internal communications
Teamwork	Gruppenarbeit (Arbeitsorganisation)	Teamwork
Work-Life Balance	Work-Life-Balance	Work-life balance
Support from Management	Management	Management
Freedom to work independently	Eigenverantwortung	Social responsibility
Gender Equality	Gleichstellung der Geschlechter	Gender equality
Attitude towards older colleagues	Altersdiskriminierung	Ageism
Office / Work Environment	Arbeitsplatz	Workplace
Environmental Friendliness	Umweltschutz	Environmental protection
Overall compensation for your work	Arbeitsentgelt	Salary
Company Image	Corporate Identity	Public relations
Career Development	Personalentwicklung	Organization development

6 Conclusions

Quantitative text analysis for digital humanities holds great potential to assist traditional research in a plethora of disciplines related to humanities and social sciences. However, the related methods typically stem from statistics and computer science and, thus, guidelines and demonstrations of how to apply these methods are needed to aid the researchers working in the two areas. In this thesis, I demonstrated how to conduct various quantitative text analyses in three different case studies, one related to literary studies, one related to communication studies and one related to management sciences. To investigate how existing sentiment analysis methods can be applied on historic texts, in the first case study, I conducted such a sentiment analysis on multilingual Spectator periodicals originating from the Age of Enlightenment. The proposed approach to this sentiment analysis was three-fold. First, I studied sentiment with regards to additional meta information that was provided by the curators of the dataset. Second, I created sentiment networks that depict the sentiment relation between different entities. Third, I constructed sentient word networks to learn more about the usage patterns of words that convey sentiment in these historic texts. In each of these parts, the results were linked to close-reading experiences. As demonstrated, the existing sentiment analysis methods can successfully be used on historic texts and provide new insights into these periodicals. However, colleagues and I also became aware that the creation of dedicated methods for earlier language forms may further increase the quality of results. In the second case study, I used quantitative text analysis to study cultural and linguistic discrepancies in controversies occurring on the social media platform Reddit. In combination with additional non-linguistic features, I revealed that, despite the topics being discussed, controversy is universal across languages. The outlined approach is a valuable example of how to employ quantitative text analysis in multilingual settings and illustrates how to conduct comparisons between multiple languages. Further, the presented findings are of use to computer scientists as well as administrators of such platforms as they indicate that existing controversy prediction models may be transferable to other languages. In the third case study, I focused on employee satisfaction expressed in online employer reviews. At first, I quantitatively analyzed additional information that was provided by reviewers, including the benefits they had received, their position as well as their employment status. I then investigated whether the influence of these aspects on employee satisfaction is similar to what we have seen from previous studies conducted in an offline context. The quantified features have been used to accurately predict employee satisfaction expressed in such reviews. Subsequently, I conducted a quantitative text analysis on the same dataset and utilized the different review aspects to link the presented results to Herzberg's Two-Factor Theory. The related findings highlighted that hygiene factors (i.e., factors that should prevent dissatisfaction) are more important to reviewers and that they even can foster employee satisfaction. Based on words that were distinctively used in review aspects, I could accurately predict the satisfaction of employees on a company level. Finally, both analyses related to this case study exemplify how to connect novel computational methods with findings and theories from traditional research.

In this chapter, I review the individual findings and contributions of respective case studies in Section 6.1, highlight potential implications of this work in Section 6.2, acknowledge potential limitations in Section 6.3 and finally recommend future work in Section 6.4.

6.1 Findings & Contributions

Case Study I: Historic Sentiment

Researchers in digital humanities cautiously approach computational methods to complement their close reading experiences. A promising tool among these computational methods to learn more about historic literature is sentiment analysis, allowing us to understand emotions, opinions and attitudes captured in a written form. In the first article presented in this thesis [Koncar, Fuchs, et al., 2020], I proposed one way to analyze sentiment in multilingual periodicals published during the Age of Enlightenment. At first, I computed sentiment based on existing dictionaries originally created for data from the Web. The fallback on these dictionaries was due to a lack of dedicated and comparable dictionaries designed for the languages of the time. Despite my expectations, these dictionaries performed arguably well on the historic data. The extension of the analysis to other sentiment dictionaries designed for today's texts, reported similar findings and, hence, I argue that the collected results are robust. Once the sentiment was computed, I exploited the meta information included in the dataset and analyzed how sentiment evolved over time, how different topics, such as politics or religion, have been perceived back then and how different narrative forms influenced sentiment. Subsequently, I created sentiment networks and studied the relation between a plethora of entities. In the final part of the article, I constructed sentiment word networks to infer usage patterns of words related to sentiment. The results of these quantitative analyses were then interpreted by experts working in the fields of humanities (i.e., Alexandra Fuchs and Elisabeth Hobisch). In doing so, the experts could find support for many of their hypothesis stemming from close-reading experiences, but also provoked eye-opening experiences in which they had to reevaluate their original thoughts about the data. Overall, the presented article successfully demonstrated one way to apply existing sentiment analysis methods to historic texts.

Case Study II: Multilingual Controversy

Controversy in online social media has been quantitatively analyzed in the past, but our community still lacks a comprehensive comparison of controversy on such websites across different cultures and languages. To fill this gap, the second article presented in this thesis [Koncar, Walk, and Helic, 2021] focused on multilingual controversy on Reddit, which was characterized through quantitative text analysis and other structural features that have already been investigated separately in existing research. Further, the demonstrated analysis comprised more than 120 million user comments posted in various Subreddits dedicated to arbitrary topics. The amount of data exceeds previous works and allows for better generalization of results. To achieve these results, the presented study again relied on existing text analysis methods and focused on basic linguist features, text readability, sentiment and other structural features that have been investigated in previous research. What separates this article from other studies is the focus on linguistic differences regarding controversy. For that, I had to consider text analysis methods that are specifically tailored to the six languages allowing for all results to be easily comparable across languages. As as result, I observed that controversy on such platforms is independent from language, with the exception of topics being discussed in the respective countries. The quantified features proved useful in the prediction of controversy in online social media. The automatic prediction of controversy can, for example, help administrators of such websites to identify potential sources of conflicts as well as sociologists in uncovering the obstinate aspects of the public discourse. Overall, the demonstrated analysis of multilingual controversy on social media serves as an example for other researchers that want to compare textual features across multiple languages.

Case Study III: Employee Satisfaction

Employee satisfaction has been studied extensively in the past, typically through qualitative interviews with employees. The manual collection of survey data proved useful for a plethora of studies in different countries, industries and companies. However, a possible drawback of these studies is the challenging inter-comparability due to the differences in methodologies of respective studies. Online employer reviews hold great potential in filling this gab because a multitude of data from employees working in a variety of industries all around the globe can be analyzed. Thus, the third [Koncar and Helic, 2020] and fourth [Koncar, T. Santos, et al., 2021] article presented in this thesis demonstrated how quantitative analyses applied to such reviews can help to learn more about the influencing factors of employee satisfaction. In the former article, I studied the influence of employee benefits and employee position on employee satisfaction as well as the influence thereof on employment status. I specifically focused on these factors because they were thoroughly analyzed in the past and, thus, I was able to investigate whether or not employee satisfaction is reflected similarly by such reviews when compared to the manually collected data. Further, I found that these factors are predictive of employee satisfaction in such reviews, with benefits and employment status having the most predictive strengths. In the latter article, I focused on the textual content of reviews and applied quantitative text analysis methods to gain further insights. Specifically, I considered the same dataset of employer reviews through the lens of Herzberg's Two-Factor Theory and linked review aspects to the corresponding factors described by the theory. This allowed for the investigation of how online employer reviews reflect one of the most used theories to explain employee satisfaction. Similar to Case Study II, the extracted textual features proved useful for the prediction of employee satisfaction in such reviews. Contrary to the prediction of controversy, the proposed model solely relies on textual content (i.e., words that are distinctively used in positive and negative review aspects). As such, the presented text analysis methods could help employers to predict the satisfaction expressed by employees independent from characteristics of the underlying reviewing platform. Overall, this case study highlighted how one can connect traditional theories as well as results from traditional data with findings from novel computational and quantitative methods.

6.2 Implications

Lessons Learned. During the conduction of the analyses for the respective case studies, colleagues and I could distill the following three major lessons learned. The first would be that your analyses can be as perfectly conducted as possible, you will not make any real use of them if you apply them on low-quality data. In hindsight, I probably spent at least a third of the time in order to clean and preprocess the data thoroughly while conducting these analyses. So if you were to follow these guidelines, please make sure to prepare your data accordingly and do not take eye-opening results for granted without these preparations.

The second lesson learned was that extensive expertise and background knowledge can substantially positively impact the interpretations of results generated by such quantitative methods. I fully became aware of this during my work for the first case study presented in this thesis. Since I was responsible for conducting the different analyses and preparing the results, I off course knew how to interpret the outcome. However, it was only the input of Alexandra Fuchs and Elisabeth Hobisch (both working in the fields of Romance studies) that enabled the detailed and in-depth interpretations the work presents. To me, this is evidence for the usefulness of interdisciplinary research, such as the digital humanities, and it underlines that the use of computational methods always requires further analysis. Hence, do not blindly use and trust such methods without proper background knowledge.

Regarding the third lesson learned, colleagues and I noticed that the quality of results presented in the first case study may be further improved by implementing dedicated methods for the languages of the 18$^{\text{th}}$. For that, we already introduced an approach to computationally create sentiment dictionaries specifically designed for the particularities of such languages. A preliminary version of this work can be found on GitHub[1] and the actual publication was under review for the vDHd'21 at the time of writing this thesis. In a nutshell, the proposed method comprises three parts: (i) the annotation and selection of sentiment seed words, (ii) the computation of word embeddings on a historic corpus, and (iii) a classification task to transfer sentiment of seed words to other words in the text corpus based on the word embeddings. The necessary code to conduct these steps is already publicly available (see the GitHub link) and contained in Jupyter Notebooks (including descriptions of respective methods) to make it easily accessible for any researcher.

Universal Methods for Empirical Analyses. Throughout all the case studies, I used the exact same or similar methods, demonstrating that the presented guidelines can be applied in various contexts of text analyses. The showcased procedure of how such an empirical analysis needs to be conducted is also the same across the respective case studies. At the beginning of each analysis, you have to prepare, clean and preprocess the data. For example, assume you want to analyze the sentiment of historic correspondences between lawyers and their clients in German. Despite being manually curated in a digital edition, you may encounter some noise in the data, such as postal addresses, that add no valuable information to the analysis. Hence, you remove them and other anomalies from the texts. Then you compute preliminary statistics of your data to acquire a first overview of it. For our historic correspondences, you could, for example, investigate the numbers of letters

[1] Link to GitHub repository: `https://github.com/philkon/sentiment-tool-chain`

per lawyer or the number of letters between any unique pair of senders and receivers. In doing so, you may observe outliers in both directions, meaning that there may be lawyers who wrote very small numbers of letters (so they are not significant for the analysis) or who wrote substantially high numbers of letters (so they skew your results). Based on the observations you make in this preliminary analysis, you may decide to remove the outliers for the remaining analysis. In the next step, you apply different methods—depending on your interest—and compute textual features. For our example, let us say we are interested in differences in sentiment and readability between letters written by lawyers and letters written by clients. Hence, we rely on the respective methods presented in this thesis and compute the sentiment based on German sentiment dictionaries [Yanqing Chen and Skiena, 2014] as well as the readability based on the German version of the Flesch Reading Ease [Amstad, 1978], respectively for each letter. According to your hypothesis, letters from lawyers to clients should be more complex and convey a neutral sentiment because they may be very factual. Contrary, letters from clients to lawyers should be emotionally charged and easier to read as you expect them to be less technically written by upset individuals. To assert your assumptions, you create plots (e.g., kernel density estimations, cumulative distribution functions, box plots) and observe your results. Your first results are promising, however, up until now you do not know whether these observations are significant or rather coincidental. Thus, you finally conduct statistical hypothesis testing to gain confidence in the observations you made. The above approach and methods can virtually be applied to any textual phenomenon and, thus, may inspire a plethora of researchers to conduct similar analyses.

Sentiment Analysis in a Historic Context. The approach to analyze sentiment of historic texts based on methods intended for modern Web data as proposed in Koncar, Fuchs, et al. [2020] indicates that dedicated methods are not strictly necessary. The successful application of such existing methods may open up opportunities for many other researchers of historic data and may go beyond the analysis of sentiment. Considering the plausible interpretations by experts working on these periodicals for years, I argue that such analyses can be a valuable input to close-reading experiences. This holds especially true in cases where one has no time to read through each individual issue and must rely on a comprehensive analysis of sentiment to learn how emotions and opinions spread back in the time. Thus, quantitative text analysis proves useful in capturing the bigger picture of textual data and other digital editions may benefit from it in a similar fashion.

Controversy on Reddit. The comparison of controversy between different languages as presented by Koncar, Walk, and Helic [2021] is likely the first of its kind. Further, the conducted analysis considers a variety of features that have only been used separately in existing research. Hence, the findings that controversy is, except for the topics involved, universal across languages as well as that user involvement features are most predictive of it, open up many opportunities in research but also in real-world applications. Regarding researchers, they are encouraged to transfer their English models to other languages and investigate whether or not they observe similar results across languages and, further, researchers should try to focus on and to refine user involvement features to increase prediction performance. Administrators and moderators of non-English social media platforms perhaps gain new possibilities to automatically aid their moderations as the universal characteristics of controversy would allow for transfers of existing prediction models that normally target the English language.

Improvement of Employee Satisfaction. The analysis of employee satisfaction expressed in online reviews has great potential to better the lives of many individuals spending most of their time at work. In particular, the identified factors that positively influence employee satisfaction in Koncar and Helic [2020] as well as in Koncar, T. Santos, et al. [2021] include pointers for employers to better understand what matters to their employees. The novel and previously unexplored dataset, which comprises more than 2 million multi-aspect reviews of employers located in four different countries and operating in 43 industries, adds an up-to-date and never seen before input to the discussion of these influencing factors. By using these new insights, employers could specifically target what is most important to their employees working in the respective country, industry and position. The combination of the quantitative text analyses with the Two-Factor Theory allows for the interpretation of such reviews in a usual and familiar setting. From the management scientist's perspective, the conducted analysis provides a great opportunity to revisit the theory and adjust it to today's circumstances if necessary. I address the potential to evaluate the theory in more detail in the Future Work section (see Section 6.4) of this thesis.

Ethical Implications. While the aim of the conducted quantitative text analyses was to demonstrate useful applications for the digital humanities, conducting these kinds of analyses also comprises risks. In the case of the Spectator periodicals analyzed in Koncar, Fuchs, et al. [2020], this risk may not be as significant as compared to the risk emerging from the other analyses conducted on current Web data. However, the presented results could still question the previous assumptions of literati about this period and, hence, put their reputation in the community at risk. At this point, I want to specifically mention that computational analyses should ever only be understood as complementary and the subsequent interpretation and evaluation of results through experts who are familiar with the data is inevitable. Regarding the analysis and the automatic prediction of controversy on Reddit presented in Koncar, Walk, and Helic [2021], the risk can be considered much higher as present-day data of individuals is utilized. Besides using these methods to help administrators in moderating their platforms, they could also be exploited to enforce censorship of controversially discussed issues. Another possible negative application of text analysis methods in the context of controversy is to employ incentivized users to steer such discussions in their favor. For example, weapon enthusiasts could specifically target discussions about gun laws to spread their political agenda. Finally, the analysis of employer reviews conducted in Koncar and Helic [2020] as well as in Koncar, T. Santos, et al. [2021] comprises risks for employers as well as employees. Employers, who may be evaluated and considered by job seekers solely based on the results of such computational analyses, could be the target of review bombing (i.e., intentionally harming the employer through writing overwhelmingly negative reviews) and may lose their appeal on the job market and, thus, miss out on talented and well-educated personnel. On the other hand, employers may use such analyses to identify the individuals behind the reviews which can have serious consequences for employees, such as termination or other defamatory actions preventing future job opportunities.

6.3 Limitations

Computational Methods and the Languages of the 18^{th} Century. The sentiment analysis conducted on Spectator periodicals in the first article of this thesis relies on methods created for Web data. My colleagues and I fell back on these methods since our community lacks approaches that are specifically designed for the languages of the 18^{th} century. The fact that our analysis required methods that are comparable across the different languages of the periodicals contained in our dataset added to this problem. Creating dedicated methods was out of scope at the time of writing the paper. However, since the aim of this work was to explore the possibilities of how to conduct a sentiment analysis on such previously computationally unexplored data, I am still confident that the proposed method provides promising results. Considering the conducted comparison of the used sentiment dictionary with other publicly available dictionaries as well as the repetition of experiments without lemmatization, which both revealed stable results, further adds to my confidence.

General Shortcomings of Computational Methods to Analyze Natural Language. The majority of analyses conducted in the respective publications heavily rely on computational methods to extract various features, such as readability or sentiment, from natural languages. Even though such methods have been proven to be accurate in existing research, they may still be erroneous in particular cases. For example, to accurately compute the readability of a text, it requires a minimum length as the returned value may be uninterpretable otherwise. To counteract these limitations, I only considered texts with at least 100 words for such analyses in the respective case studies. Another example for which computed results may be misleading is sentiment analysis. While the utilized dictionary based approaches correctly compute sentiment based on the words in a text, they are not capable of detecting certain linguistic and stylistic devices, such as sarcasm, irony or satire. While such cases are most certainly present in the respective datasets, I argue that their occurrences are vanishingly small. However, specifically filtering these cases could further increase the quality of results and may be conducted in future work.

Correlation Versus Causality. Any of the presented empirical studies investigated correlations and not causalities. For example, a comment posted in a Subreddit may receive higher levels of attention due to other reasons and not solely because it is controversial. Similarly, a negative review created on kununu may not necessarily stem from dissatisfaction experienced at a workplace, but may be the result of trolling or other intentionally damaging behavior. Vice versa, a positive review could be the result of review manipulation through incentives and not solely of high employee satisfaction.

Significance Levels. The analyses conducted in the four publications all rely on significance levels (e.g., quartiles of the overall rating distribution to define employee satisfaction) or similar thresholds. Naturally, these thresholds affect the outcome of the different quantitative analyses, which can have consequences for the derived conclusions and implications. To counteract this limitation, small adjustments were made to all such thresholds in the respective publications. These alterations did not significantly impact the presented findings in any case. However, using considerably deviating thresholds or completely different definitions (e.g., defining employee satisfaction based on overall rating stars of reviews) may still lead to different findings.

Dataset Selection. The conducted analyses demonstrated quantitative text analyses applied on three different datasets. However, I acknowledge certain limitation regarding the used datasets as well as the generalizability connected therewith. First, the findings on the sentiment expressed during the Age of Enlightenment stem from Spectator periodicals. Even though these periodicals enjoyed great popularity, they were still only a small part of the important literature back in the time. Further, the proposed sentiment analysis methods require sufficient amounts of data (which is not always given with historic data) and, thus, they cannot be easily applied to other datasets. This, however, would be necessary to make more general assumptions about sentiment in the Age of Enlightenment. Second, for the analysis of controversy on Reddit only comments posted in 50 Subreddits during 2019 were considered, which is only a small subset of all comments posted on Reddit. Still, the variety of Subreddits and their respective topics and languages allow for a generalization of results, though the adaption of our analysis methods to other Subreddits or even other online social media platforms may further increase the quality of the made assumptions.

Third, regarding the analysis of employee satisfaction in online reviews, I acknowledge a certain bias of the proposed methods towards kununu. While the number of reviews and the diversity of involved entities is sufficient, especially compared to traditional data originating from surveys, the gained results are specific to the platform. Thus, extending the analysis to other platforms, such as Glassdoor, is still necessary to better generalize results. Note that this extension cannot be conducted easily as other platforms have different reviewing mechanisms (e.g., different rating scales or different review aspects), however.

6.4 Future Work

Dedicated Computational Methods for the Languages of the 18^{th} Century. As mentioned before in Section 6.3, the sentiment analysis conducted on the Spectator periodicals relied on methods primarily created for modern data originating from the Web. Designing and implementing methods for the languages of the 18^{th} century would allow to further increase the quality of results. While colleagues and I already took a first step in this direction with the creation of a sentiment tool chain for the languages of the 18^{th} century, there are numerous other areas that need to be addressed. Most importantly for this process is the manual annotation of data to use in various machine learning tasks. However, this annotation requires extensive background knowledge about the data and, thus, is something which should only be considered to be done by experts in the respective fields. Based on the experiences I made during my doctoral research, I assume that dedicated methods for stemming as well as lemmatization and the creation of further sentiment dictionaries are most desperately needed for the languages of earlier times.

Improvements to the Automatic Prediction of Controversy in Online Social Media. Another promising aspect for future work is the improvement of the automatic prediction of controversy in online social media. While the presented straightforward approach based on gradient boosted decision trees achieved moderate results, more sophisticated approaches, such as neural networks, could further improve the prediction performance. For that, it may be useful to consider additional textual features, such as named entities or word embeddings, as well as user-dependent characteristics, such as their history of previous posts in respective communities. The latter may be essential for the early prediction of controversy, which would be even more useful to moderators and administrators as it would allow them to react more promptly.

Optimization of Employee Satisfaction Analysis and Prediction. Similar to the improvements to the automatic prediction of controversy in online social media, the proposed methods to analyze and predict employee satisfaction expressed in online reviews could further be extended. In the particular case of the presented dataset, one could exploit additionally provided texts by reviewers, such as the pros and cons they listed as well as the suggestions they mentioned. Analyzing this additional information could further clarify what aspects are most important to employees. Especially the suggestions stated by employees may prove useful for managers to adapt their business strategies as employees have the hands-on experiences in the day to day business. Further, the additional aspects of reviews could be analyzed in combination with the provided position of employees to eliminate inequalities. Depending on the features one can extract from this additional review characteristics, they may also prove useful to achieve better prediction performance. For the latter, one could also consider more sophisticated machine learning approaches to predict employee satisfaction more accurately and, thus, better understand the influential factors of it.

Revision of the Two-Factor Theory. The insights gained in the fourth article presented in this thesis [Koncar, T. Santos, et al., 2021] hold great potential in revisiting and reevaluating the Two-Factor Theory. Since the introduction of this theory, it has been criticized by some for the strict factor assignment [e.g., Parsons and Broadbridge, 2006; Y. Li, 2018], arguing that individuals of different professions also have different needs. While the corresponding findings in this thesis suggested a general support for the theory based on online employer reviews, there also seem to be shifts for some of the factors. For example, *work-life-balance* as well as *environmental friendliness* seem to be very important in satisfied reviews, however, according to the original theory, they would only be considered as hygiene factors which prevent dissatisfaction. This may indicate that the influential factors for employee satisfaction assessed back in 1959 changed over the time and, thus, a reconsideration of factor assignments may be required. Further, the variety of countries and industries in the analyzed dataset provides a great opportunity to further discuss the generalization of the theory, which was also criticized in the past [e.g., Furnham, Forde, and Ferrari, 1999].

Extension to Further Datasets. To produce more generalizable results, it is necessary to adapt the proposed methods and extend them to other datasets. For the first part of this thesis, further literature could be considered to better understand sentiment conveyed during the Age of Enlightenment. For example, additional translations and imitations of the Spectator periodicals can be considered. In particular, conducting the proposed analysis on the original English issues should be considered as it laid the foundation for the genre of Spectator periodicals. Other types of prominent literature of that time could be analyzed as well, including popular plays, such as the ones from Gotthold Ephraim Lessing or William Congreve, as well as novels, for example, from Daniel Defoe. Regarding controversy on social media platforms, further multilingual datasets should be analyzed to make more generalizable assumptions. A fairly simple solution to that would be to extend the selection of Subreddits and to increase the timespan of the extracted data. Another possibility would be to investigate Tweets written in different languages. Using data from Twitter would provide a great way to study linguistic and cultural differences of controversy, as it not only allows to focus on textual aspects but also on underlying network structures of the various communities. Finally, for the analysis of online employer reviews, the study of other platforms through the lens of Herzberg's Two-Factor theory would allow for even more possibilities to evaluate it as well as to gain more insights for employers to improve employee satisfaction. The most obvious suggestion in this regard would be Glassdoor, which was already in the focus of researchers to learn about employee satisfaction in the past. Other less known platforms that provide possibilities to extract multi-aspect employer reviews include *Indeed*[2] or *RateMyEmployer*[3].

[2] Link to website: `https://indeed.com`
[3] Link to website: `https://ratemyemployer.ca`

List of Figures

List of Tables

Bibliography

Abdaoui, Amine et al. (2017). "Feel: a french expanded emotion lexicon." In: *Language Resources and Evaluation* 51.3, pp. 833–855 (cit. on p. 95).

Abedi, Jamal et al. (2011). "Accessible Reading Assessments for Students with Disabilities: The Role of Cognitive, Grammatical, Lexical, and Textual/Visual Features. CRESST Report 785." In: *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)* (cit. on p. 33).

Abel, Jennifer et al. (2017). "Sentiment-analysis for German employer reviews." In: *International conference on business information systems*. Springer, pp. 37–48 (cit. on p. 35).

Acerbi, Alberto et al. (2013). "The expression of emotions in 20th century books." In: *PloS one* 8.3, e59030 (cit. on pp. 1, 22).

Adamic, Lada A and Natalie Glance (2005). "The political blogosphere and the 2004 US election: divided they blog." In: *Proceedings of the 3rd international workshop on Link discovery*. ACM, pp. 36–43 (cit. on pp. 1, 21, 40, 42, 103).

Aguirre, Patricia Estefania Ayala et al. (2017). "Evaluating the dental caries-related information on Brazilian websites: qualitative study." In: *Journal of medical Internet research* 19.12, e415 (cit. on p. 112).

Ahire, Sagar (2014). "A survey of sentiment lexicons." In: *Computer Science and Engineering IIT Bombay, Bombay* (cit. on p. 34).

Akoglu, Leman (2014). "Quantifying Political Polarity Based on Bipartite Opinion Networks." In: *ICWSM* (cit. on p. 44).

Alaei, Ali Reza, Susanne Becken, and Bela Stantic (2019). "Sentiment analysis in tourism: capitalizing on big data." In: *Journal of Travel Research* 58.2, pp. 175–191 (cit. on p. 35).

Alfayad, Zaid and Lily Suriani Mohd Arif (2017). "Employee voice and job satisfaction: An application of Herzberg two-factor theory." In: *International Review of Management and Marketing* 7.1, pp. 150–156 (cit. on pp. 48, 145, 151).

Allan, David and Learning Virtue (1993). *The Scottish Enlightenment*. Edinburgh, Scotland: Edinburgh University Press (cit. on pp. 37, 58).

Allport, Gordon W and Janet M Faden (1940). "The psychology of newspapers: Five tentative laws." In: *Public opinion quarterly* 4.4, pp. 687–703 (cit. on p. 24).

Almeida, Felipe and Geraldo Xexéo (2019). "Word embeddings: A survey." In: *arXiv preprint arXiv:1901.09069* (cit. on pp. 29, 30).

Almishari, Mishari and Gene Tsudik (2012). "Exploring linkability of user reviews." In: *European Symposium on Research in Computer Security*. Springer, pp. 307–324 (cit. on p. 180).

AlMousa, Mohannad, Rachid Benlamri, and Richard Khoury (2021). "Exploiting non-taxonomic relations for measuring semantic similarity and relatedness in WordNet." In: *Knowledge-Based Systems* 212, p. 106565 (cit. on p. 25).

Alshmemri, Mohammed, Lina Shahwan-Akl, and Phillip Maude (2016). "Job satisfaction of Saudi nurses working in Makkah region public hospitals, Saudi Arabia." In: *Life Science Journal* 13.12, pp. 22–33 (cit. on p. 47).

Alshmemri, Mohammed, Lina Shahwan-Akl, and Phillip Maude (2017). "Herzberg's two-factor theory." In: *Life Science Journal* 14.5, pp. 12–16 (cit. on pp. 47, 158, 160, 161, 183).

Altarawneh, Rasha (2017). "Spelling detection errors techniques in NLP: A survey." In: *Int. J. Comput. Appl* 172, pp. 1–5 (cit. on p. 26).

Amancio, Diego R, Osvaldo N Oliveira Jr, and Luciano da F Costa (2012). "Unveiling the relationship between complex networks metrics and word senses." In: *EPL* 98.1, p. 18002 (cit. on pp. 29, 59).

Amancio, Diego Raphael (2015). "A complex network approach to stylometry." In: *PloS one* 10.8, e0136076 (cit. on pp. 29, 59).

Amancio, Diego Raphael et al. (2011). "Using metrics from complex networks to evaluate machine translation." In: *Physica A: Statistical Mechanics and its Applications* 390.1, pp. 131–142 (cit. on p. 60).

Amstad, T (1978). "Wie verständlich sind unsere Zeitungen?(unveröffentlichte Dissertation)." In: *Zürich: Universität* (cit. on pp. 32, 112, 193).

An, Jisun, Daniele Quercia, and Jon Crowcroft (2014). "Partisan sharing: facebook evidence and societal consequences." In: *Proceedings of the second ACM conference on Online social networks*. ACM, pp. 13–24 (cit. on p. 44).

Anderson, Eugene W (1998). "Customer satisfaction and word of mouth." In: *Journal of service research* 1.1, pp. 5–17 (cit. on p. 159).

Anderson, Richard Chase and Alice Davison (1986). "Conceptual and empirical bases of readability formulas." In: *Center for the Study of Reading Technical Report; no. 392* (cit. on p. 31).

Antiqueira, Lucas et al. (2009). "A complex network approach to text summarization." In: *Information Sciences* 179.5, pp. 584–599 (cit. on pp. 29, 59).

Arıcan, Bilge Nas et al. (2021). "Creating domain dependent Turkish wordnet and sentinet." In: *Proceedings of the 11th Global Wordnet Conference*, pp. 243–251 (cit. on p. 25).

Artese, Maria Teresa and Isabella Gagliardi (2019). "Preprocessing pipeline for Italian Cultural Heritage multimedia datasets." In: *Archiving Conference*. Vol. 2019. 1. Society for Imaging Science and Technology, pp. 81–85 (cit. on p. 27).

Artz, Benjamin (2010). "Fringe benefits and job satisfaction." In: *International journal of manpower* (cit. on pp. 13, 52, 127–129, 134, 135, 140).

Astigarraga, Jesús (2015). *The spanish enlightenment revisited*. Oxford, United Kingdom: Voltaire Foundation (cit. on p. 67).

Ayanouz, Soufyane, Boudhir Anouar Abdelhakim, and Mohammed Benhmed (2020). "A smart chatbot architecture based NLP and machine learning for health care assistance." In: *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, pp. 1–6 (cit. on pp. 1, 26).

Aydogdu, Sinem and Baris Asikgil (2011). "An empirical study of the relationship among job satisfaction, organizational commitment and turnover intention." In: *International review of management and marketing* 1.3, p. 43 (cit. on pp. 127, 129).

Bahadori, M et al. (2015). "Confirmatory factor analysis of the herzberg job motivation model for workers in the military health organizations of Iran." In: (cit. on pp. 45, 47).

Balakrishnan, Vimala and Ethel Lloyd-Yemoh (2014). "Stemming and lemmatization: A comparison of retrieval performances." In: (cit. on p. 28).

Balmer, Sylvia and Tom Baum (1993). "Applying Herzberg's hygiene factors to the changing accommodation environment." In: *International Journal of Contemporary Hospitality Management* 5.2 (cit. on pp. 48, 145, 151).

Banks, George C et al. (2018). "A review of best practice recommendations for text analysis in R (and a user-friendly app)." In: *Journal of Business and Psychology* 33.4, pp. 445–459 (cit. on p. 27).

Barbosa, Jardeson LN, Raimundo Santos Moura, and Roney L de S Santos (2016). "Predicting Portuguese Steam Review Helpfulness Using Artificial Neural Networks." In: *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*. ACM, pp. 287–293 (cit. on p. 131).

Barkur, Gopalkrishna and Giridhar B Kamath Vibha (2020). "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India." In: *Asian journal of psychiatry* 51, p. 102089 (cit. on p. 35).

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 238–247 (cit. on p. 29).

Bassett-Jones, Nigel and Geoffrey C Lloyd (2005). "Does Herzberg's motivation theory have staying power?" In: *Journal of management development* (cit. on pp. 161, 183).

Baumeister, Roy F et al. (2001). "Bad is stronger than good." In: *Review of general psychology* 5.4, pp. 323–370 (cit. on p. 165).

Becker, Howard (1932). "Space apportioned forty-eight topics in The American Journal of Sociology, 1895-1930." In: *American Journal of Sociology* 38.1, pp. 71–78 (cit. on p. 24).

Beevers, David (2008). *Chinese Whispers: Chinoiserie in Britain, 1650-1930*. Brighton, United Kingdom: Royal Pavilion Libraries & Museums (cit. on p. 78).

Behling, Orlando, George Labovitz, and Richard Kosmo (1968). "The Herzberg controversy: A critical reappraisal." In: *Academy of Management Journal* 11.1, pp. 99–108 (cit. on p. 162).

Behzad, Shabnam and Amir Zeldes (2020). "A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging." In: *arXiv preprint arXiv:2004.14312* (cit. on p. 29).

Bengio, Yoshua, Réjean Ducharme, et al. (2003). "A neural probabilistic language model." In: *The journal of machine learning research* 3, pp. 1137–1155 (cit. on p. 30).

Bengio, Yoshua, Jean-Sébastien Senécal, et al. (2003). "Quick Training of Probabilistic Neural Nets by Importance Sampling." In: *AISTATS*, pp. 1–9 (cit. on p. 30).

Berelson, Bernard (1952). "Content analysis in communication research." In: (cit. on pp. 22, 24).

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation." In: *the Journal of machine Learning research* 3, pp. 993–1022 (cit. on p. 51).

Blombach, Andreas et al. (2020). "A corpus of German Reddit exchanges (GeRedE)." In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 6310–6316 (cit. on p. 27).

Blood, Milton R (1969). "Work values and job satisfaction." In: *Journal of Applied psychology* 53.6, p. 456 (cit. on pp. 3, 45, 127, 129).

Bodmer, Johann Jakob (1728). *Anklagung Des verderbten Geschmackes...* (Cit. on p. 39).

Bojanowski, Piotr et al. (2017). "Enriching word vectors with subword information." In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146 (cit. on p. 30).

Bonferroni, Carlo E (1935). "Il calcolo delle assicurazioni su gruppi di teste." In: *Studi in onore del professore salvatore ortu carboni*, pp. 13–60 (cit. on pp. 85, 111).

Bougie, Roger, Rik Pieters, and Marcel Zeelenberg (2003). "Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services." In: *Journal of the academy of marketing science* 31.4, pp. 377–393 (cit. on p. 158).

Boulard, Claire (2000). *Presse et socialisation féminine en Angleterre de 1690 à 1750, conversations à l'heure du thé: étude du Gentleman's Journal, du Spectator et du Female Spectator*. Paris, France: Editions L'Harmattan (cit. on p. 62).

Brill, Eric (1992). *A simple rule-based part of speech tagger*. Tech. rep. PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER and INFORMATION SCIENCE (cit. on p. 29).

Brown, Mark Graham (2006). *The pocket guide to the Baldrige award criteria*. CRC Press (cit. on pp. 49, 50).

Brown, Steven P and Robert A Peterson (1993). "Antecedents and consequences of salesperson job satisfaction: Meta-analysis and assessment of causal effects." In: *Journal of marketing research* 30.1, pp. 63–77 (cit. on pp. 13, 52, 128, 138, 140).

Budi, Indra et al. (2005). "Named entity recognition for the Indonesian language: combining contextual, morphological and part-of-speech features into a knowledge engineering approach." In: *International Conference on Discovery Science*. Springer, pp. 57–69 (cit. on p. 28).

Cable, Daniel M and Kang Yang Trevor Yu (2006). "Managing job seekers' organizational image beliefs: The role of media richness and media credibility." In: *Journal of applied psychology* 91.4, p. 828 (cit. on p. 180).

Cantwell, George T and MEJ Newman (2019). "Mixing patterns and individual differences in networks." In: *Physical Review E* 99.4, p. 042306 (cit. on p. 92).

Carr, Rosalind (2014). *Gender and Enlightenment culture in eighteenth-century Scotland*. Edinburgh, Scotland: Edinburgh University Press (cit. on pp. 37, 58).

Chall, Jeanne Sternlicht (1958). *Readability: An appraisal of research and application*. 34. Ohio State University (cit. on p. 31).

Chandra, Vallury (2012). "Work–life balance: Eastern and western perspectives." In: *The International Journal of Human Resource Management* 23.5, pp. 1040–1056 (cit. on pp. 51, 130, 149).

Chatterjee, Patrali (2001). "Online reviews: do consumers use them?" In: (cit. on p. 130).

Chen, Pei-Yu, Shin-yi Wu, and Jungsun Yoon (2004). "The impact of online recommendations and consumer feedback on sales." In: *ICIS 2004 Proceedings*, p. 58 (cit. on pp. 130, 149).

Chen, Qiang, Chenliang Li, and Wenjie Li (2017). "Modeling language discrepancy for cross-lingual sentiment analysis." In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. New York, United States: ACM, pp. 117–126 (cit. on p. 59).

Chen, Shun-Hsing et al. (2006). "The development of an employee satisfaction model for higher education." In: *The TQM Magazine* (cit. on p. 144).

Chen, Yanqing and Steven Skiena (2014). "Building sentiment lexicons for all major languages." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 383–389 (cit. on pp. 10, 22, 29, 36, 59, 62, 114, 159, 193).

Chen, Yubo, Scott Fay, and Qi Wang (2003). "Marketing implications of online consumer product reviews." In: *Business Week* 7150, pp. 1–36 (cit. on pp. 22, 130).

Chi, Christina G and Dogan Gursoy (2009). "Employee satisfaction, customer satisfaction, and financial performance: An empirical examination." In: *International Journal of Hospitality Management* 28.2, pp. 245–253 (cit. on p. 50).

Chiang-Soong, Betty and Robert E Yager (1993). "Readability levels of the science textbooks most used in secondary schools." In: *School Science and Mathematics* 93.1, pp. 24–27 (cit. on p. 33).

Choi, Yoonjung, Yuchul Jung, and Sung-Hyon Myaeng (2010). "Identifying controversial issues and their sub-topics in news articles." In: *Intelligence and Security Informatics*, pp. 140–153 (cit. on pp. 11, 40, 41, 103).

Choo, Stephen and Christine Bowley (2007). "Using training and development to affect job satisfaction within franchising." In: *Journal of Small Business and Enterprise Development* (cit. on pp. 45, 46).

Chu, Hui-Chin and Tsui Yang Kuo (2015). "Testing Herzberg's two-factor theory in educational settings in Taiwan." In: *The Journal of Human Resource and Adult Learning* 11.1, pp. 54–65 (cit. on p. 48).

Church, Kenneth Ward (1989). "A stochastic parts program and noun phrase parser for unrestricted text." In: *International Conference on Acoustics, Speech, and Signal Processing,* IEEE, pp. 695–698 (cit. on p. 28).

Civre, Žana, Nina Lovec, Daša Fabjan, et al. (2013). "Herzberg's two factor theory of work motivation: The case of tourism employees." In: *Management* 8.3, pp. 219–232 (cit. on p. 47).

Collins-Thompson, Kevyn (2014). "Computational assessment of text readability: A survey of current and future research." In: *ITL-International Journal of Applied Linguistics* 165.2, pp. 97–135 (cit. on p. 31).

Cong, Jin and Haitao Liu (2014). "Approaching human language with complex networks." In: *Physics of life reviews* 11.4, pp. 598–618 (cit. on p. 59).

Conover, Michael et al. (2011). "Political polarization on twitter." In: *ICWSM* 133, pp. 89–96 (cit. on pp. 21, 42).

Consortium, TEI (2020). *Guidelines for Electronic Text Encoding and Interchange*. URL: `http://www.tei-c.org/P5/` (visited on 02/16/2020) (cit. on p. 61).

Cornelißen, Thomas (2009). "The interaction of job satisfaction, job search, and job changes. An empirical investigation with German panel data." In: *Journal of Happiness Studies* 10.3, pp. 367–384 (cit. on pp. 13, 14, 52, 127–129, 136, 141).

Cummings, Paul W (1975). "Does Herzberg's theory really work." In: *Management Review* 64.2, pp. 35–37 (cit. on p. 162).

Dabirian, Amir, Jan Kietzmann, and Hoda Diba (2017). "A great place to work!? Understanding crowdsourced employer branding." In: *Business Horizons* 60.2, pp. 197–205 (cit. on pp. 3, 45, 50, 127, 130, 145, 149).

Dale, Edgar and Jeanne S Chall (1948). "A formula for predicting readability: Instructions." In: *Educational research bulletin*, pp. 37–54 (cit. on p. 32).
Dale, Edgar and Jeanne S Chall (1949). "The concept of readability." In: *Elementary English* 26.1, pp. 19–26 (cit. on p. 31).
Daraz, Lubna et al. (2018). "Readability of online health information: a meta-narrative systematic review." In: *American Journal of Medical Quality* 33.5, pp. 487–492 (cit. on pp. 1, 31, 32).
Darling, Kimberly, Joseph Arn, and Rebecca Gatlin (1997). "How to effectively reward employees." In: *Industrial Management-Chicago then Atlanta-* 39, pp. 1–4 (cit. on pp. 3, 127, 129).
Das Swain, Vedant et al. (2020). "Modeling organizational culture with workplace experiences shared on glassdoor." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15 (cit. on pp. 145, 149).
De Choudhury, Munmun and Scott Counts (2013). "Understanding affect in the workplace via social media." In: *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 303–316 (cit. on p. 149).
De Cremer, David (2003). "How self-conception may lead to inequality: Effect of hierarchical roles on the equality rule in organizational resource-sharing tasks." In: *Group & Organization Management* 28.2, pp. 282–302 (cit. on pp. 13, 52, 127, 128, 136, 137, 140).
De Cremer, David, Eric van Dijk, and CP Reinders Folmer (2009). "Why leaders feel entitled to take more." In: *Psychological perspectives on ethical behavior and decision making*, pp. 107–119 (cit. on pp. 13, 52, 127, 128, 136, 137, 140).
DeCenzo, David A, Stephen P Robbins, and Susan L Verhulst (2016). *Fundamentals of human resource management*. John Wiley & Sons (cit. on pp. 13, 52, 127, 128, 134, 140).
Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis." In: *Journal of the American society for information science* 41.6, pp. 391–407 (cit. on p. 30).
Dell'Orletta, Felice (2009). "Ensemble system for Part-of-Speech tagging." In: *Proceedings of EVALITA* 9, pp. 1–8 (cit. on p. 28).
Dellarocas, Chrysanthos, Xiaoquan Michael Zhang, and Neveen F Awad (2007). "Exploring the value of online product reviews in forecasting sales: The case of motion pictures." In: *Journal of Interactive marketing* 21.4, pp. 23–45 (cit. on p. 22).

Deprez, Annelore and Karin Raeymaeckers (2012). "A longitudinal study of job satisfaction among Flemish professional journalists." In: *Journalism and Mass Communication* 2.1, pp. 1–15 (cit. on p. 141).

DeRose, Steven J (1988). "Grammatical category disambiguation by statistical optimization." In: *Computational linguistics* 14.1 (cit. on p. 28).

DeShields Jr, Oscar W, Ali Kara, and Erdener Kaynak (2005). "Determinants of business student satisfaction and retention in higher education: applying Herzberg's two-factor theory." In: *International journal of educational management* 19.2, pp. 128–139 (cit. on pp. 48, 145, 151).

Deshmukh, Rushali Dhumal and Arvind Kiwelekar (2020). "Deep learning techniques for part of speech tagging by natural language processing." In: *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, pp. 76–81 (cit. on p. 29).

Deutsch, Morton (1973). *The resolution of conflict: Constructive and destructive processes*. Yale University Press (cit. on p. 40).

Diaz, Gerardo Ocampo and Vincent Ng (2018). "Modeling and Prediction of Online Product Review Helpfulness: A Survey." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 698–708 (cit. on pp. 22, 130).

Dienhart, John R and Mary B Gregoire (1993). "Job satisfaction, job involvement, job security, and customer focus of quick-service restaurant employees." In: *Hospitality research journal* 16.2, pp. 29–43 (cit. on pp. 13, 14, 52, 127–129, 136, 140).

Dill, Hans-Otto (2015). *Aufklärung als Weltprojekt*. Bern, Switzerland: Peter Lang. ISBN: 978-3-653-97062-3 (cit. on p. 67).

DiMicco, Joan et al. (2008). "Motivations for social networking at work." In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 711–720 (cit. on pp. 49, 50, 149).

Dobre, Ovidiu-Iliuta (2013). "Employee motivation and organizational performance." In: *Review of applied socio-economic research* 5.1 (cit. on p. 144).

Dori-Hacohen, Shiri and James Allan (2013). "Detecting controversy on the web." In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, pp. 1845–1848 (cit. on pp. 11, 40, 42, 101–104, 114, 121, 122).

Dori-Hacohen, Shiri and James Allan (2015). "Automated controversy detection on the web." In: *European Conference on Information Retrieval*. Springer, pp. 423–434 (cit. on pp. 40, 42, 101, 103).

Dori-Hacohen, Shiri, Elad Yom-Tov, and James Allan (2015). "Navigating Controversy as a Complex Search Task." In: *SCST@ ECIR* (cit. on pp. 40, 42, 103).

Dos Rieis, Julio Cesar Soares et al. (2015). "Breaking the news: First impressions matter on online news." In: *Ninth International AAAI conference on web and social media*. California, United States: The AAAI Press (cit. on p. 93).

Dotoli, Giovanni (1983). "Il Mercure galant di Donneau de Visé in L'information en France au XVII siècle." In: *Quaderni del Seicento Francese* 5, pp. 219–282 (cit. on p. 37).

Dovring, Karin (1954). "Quantitative semantics in 18th century Sweden." In: *Public Opinion Quarterly*, pp. 389–394 (cit. on p. 23).

Downie, James Allan (1993). "Periodicals and politics in the reign of Queen Anne." In: *Serials and their Readers*, pp. 1620–1914 (cit. on p. 38).

Duan, Wenjing, Bin Gu, and Andrew B Whinston (2008). "Do online reviews matter?—An empirical investigation of panel data." In: *Decision support systems* 45.4, pp. 1007–1016 (cit. on p. 149).

DuBay, William H (2004). "The Principles of Readability." In: *Online Submission* (cit. on p. 160).

Eberhard, Lukas et al. (2018). "Investigating Helpfulness of Video Game Reviews on the Steam Platform." In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, pp. 43–50 (cit. on pp. 22, 131, 149).

Ellickson, Mark C and Kay Logsdon (2002). "Determinants of job satisfaction of municipal government employees." In: *Public Personnel Management* 31.3, pp. 343–358 (cit. on p. 46).

Ernst Kossek, Ellen and Cynthia Ozeki (1998). "Work–family conflict, policies, and the job–life satisfaction relationship: A review and directions for organizational behavior–human resources research." In: *Journal of applied psychology* 83.2, p. 139 (cit. on pp. 49, 127).

Ertler, Klaus-Dieter (2003). *Moralische Wochenschriften in Spanien: José Clavijo y Fajardo: El Pensador*. Tübingen, Germany: Gunter Narr Verlag (cit. on p. 62).

Ertler, Klaus-Dieter (2004). *Tugend und Vernunft in der Presse der spanischen Aufklärung–El Censor*. Tübingen, Germany: Gunter Narr Verlag (cit. on pp. 61, 74).

Ertler, Klaus-Dieter et al. (2011-2020). *The Spectators in the international context*. URL: `https://gams.uni-graz.at/spectators` (visited on 02/16/2020) (cit. on p. 61).

Fabian, Benjamin, Tatiana Ermakova, and Tino Lentz (2017). "Large-scale readability analysis of privacy policies." In: *Proceedings of the International Conference on Web Intelligence*, pp. 18–25 (cit. on pp. 32, 33).

Fan, David P (1988). *Predictions of public opinion from the mass media: Computer content analysis and mathematical modeling*. Vol. 12. Greenwood Publishing Group (cit. on p. 25).

Fang, Bin et al. (2016). "Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics." In: *Tourism Management* 52, pp. 498–506 (cit. on p. 32).

Fanoon, ARFS and GAI Uwanthika (2019). "Part of speech tagging for Twitter conversations using Conditional Random Fields model." In: *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*. IEEE, pp. 108–112 (cit. on p. 29).

Feltovich, Nick (2003). "Nonparametric tests of differences in medians: comparison of the Wilcoxon–Mann–Whitney and robust rank-order tests." In: *Experimental Economics* 6.3, pp. 273–297 (cit. on p. 111).

Fenton, Frances (1910). "The influence of newspaper presentations upon the growth of crime and other anti-social activity." In: *American journal of sociology* 16.3, pp. 342–371 (cit. on p. 23).

Fernández Huerta, José (1959). "Medidas sencillas de lecturabilidad." In: *Consigna* 214, pp. 29–32 (cit. on p. 112).

Ferraz de Arruda, Henrique et al. (2018). "Representation of texts as complex networks: a mesoscopic approach." In: *Journal of Complex Networks* 6.1, pp. 125–144 (cit. on p. 97).

Feyel, Gilles (2000). *L'Annonce et la nouvelle: La presse d'information en France sous l'Ancien Régime, 1630-1788*. Voltaire Foundation (cit. on p. 37).

Fischer-Pernkopf, Michaela, Veronika Mussner, and Klaus-Dieter Ertler (2018). *Die "Spectators" in Frankreich: "Le Nouveau Spectateur" und "Le Monde comme il est" von Jean-François de Bastide*. Bern, Switzerland: Peter Lang. ISBN: 978-3-631-75989-9. URL: `https://www.peterlang.com/view/title/68171` (cit. on p. 73).

Flaxman, Seth, Sharad Goel, and Justin M Rao (2016). "Filter bubbles, echo chambers, and online news consumption." In: *Public Opinion Quarterly* 80.S1, pp. 298–320 (cit. on p. 44).

Fleiss, Joseph L (1971). “Measuring nominal scale agreement among many raters.” In: *Psychological bulletin* 76.5, p. 378 (cit. on p. 155).
Flesch, Rudolph (1948). “A new readability yardstick.” In: *Journal of applied psychology* 32.3, p. 221 (cit. on pp. 27, 32, 112, 159).
Flory, Steven M, Thomas J Phillips Jr, and Maurice F Tassin (1992). “Measuring readability: A comparison of accounting textbooks.” In: *Journal of Accounting Education* 10.1, pp. 151–161 (cit. on p. 33).
Franchina, Valerio and Roberto Vacca (1986). “Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages.” In: *Linguaggi* 3, pp. 47–49 (cit. on pp. 32, 112).
Fuchs, Alexandra and Klaus-Dieter Ertler (2014). *Die Moralischen Periodika in Italien: L'Osservatore veneto*. Hamburg, Germany: Verlag Dr. Kovač. isbn: 978-3-8300-7680-3 (cit. on p. 70).
Fuchs, Alexandra and Klaus-Dieter Ertler (2018). *Kommunikative Inszenierung und Migration der Mikroerzählungen in der Gazzetta urbana veneta von Antonio Piazza*. Hamburg, Germany: Verlag Dr. Kovač. isbn: 978-3-8300-9735-8 (cit. on p. 72).
Fuchs, Alexandra, Klaus-Dieter Ertler, and Jürgen Holzer (2019). *Mikroerzählungen in den Spectators*. Hamburg, Germany: Verlag Dr. Kovač. isbn: 978-3-339-10670-4 (cit. on p. 55).
Furnham, Adrian, Liam Forde, and Kirsti Ferrari (1999). “Personality and work motivation.” In: *Personality and individual differences* 26.6, pp. 1035–1043 (cit. on pp. 52, 162, 200).
Galli, Beatriz (2020). “Challenges and opportunities for access to legal and safe abortion in Latin America based on the scenarios in Brazil, Argentina, and Uruguay.” In: *Cadernos de Saúde Pública* 36, e00168419 (cit. on p. 111).
Garimella, Kiran et al. (2016). “Quantifying controversy in social media.” In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pp. 33–42 (cit. on pp. 11, 44).
Garimella, Kiran et al. (2017). “Reducing controversy by connecting opposing views.” In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, pp. 81–90 (cit. on p. 44).
Gawel, Joseph E (1996). “Herzberg's theory of motivation and Maslow's hierarchy of needs.” In: *Practical Assessment, Research, and Evaluation* 5.1, p. 11 (cit. on pp. 158, 160, 161, 177, 183).
George, Alexander L (1959). “Propaganda analysis.” In: *Evanston und New York 1959* (cit. on p. 24).

Gerlach, Martin, Tiago P Peixoto, and Eduardo G Altmann (2018). "A network approach to topic models." In: *Science advances* 4.7, eaaq1360 (cit. on pp. 22, 29, 60).

Gesmundo, Andrea and Tanja Samardzic (2012). "Lemmatisation as a tagging task." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 368–372 (cit. on p. 28).

Gharge, Sagar and Manik Chavan (2017). "An integrated approach for malicious tweets detection using NLP." In: *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE, pp. 435–438 (cit. on p. 26).

Ghazi, Safdar Rehman, Gulap Shahzada, and M Shuaib Khan (2013). "Resurrecting Herzberg's two factor theory: An implication to the university teachers." In: *Journal of Educational and Social Research* 3.2, pp. 445–445 (cit. on pp. 48, 49).

Gilot, Michel and Jean Sgard (1981). "Le journalisme masqué." In: *Le journalisme d'ancien régime (Lyon)*, pp. 285–313 (cit. on pp. 38, 58).

Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena (2007). "Large-Scale Sentiment Analysis for News and Blogs." In: *Icwsm* 7.21, pp. 219–222 (cit. on p. 93).

Godes, David and Dina Mayzlin (2004). "Using online conversations to study word-of-mouth communication." In: *Marketing science* 23.4, pp. 545–560 (cit. on p. 22).

Goga, Oana et al. (2013). "Exploiting innocuous activity for correlating users across sites." In: *Proceedings of the 22nd international conference on World Wide Web*, pp. 447–458 (cit. on p. 180).

Gomes, Fernando Barbosa, Juan Manuel Adán-Coello, and Fernando Ernesto Kintschner (2018). "Studying the Effects of Text Preprocessing and Ensemble Methods on Sentiment Analysis of Brazilian Portuguese Tweets." In: *International Conference on Statistical Language and Speech Processing*. Springer, pp. 167–177 (cit. on p. 27).

Gómez Trueba, Teresa (1999). "El sueño literario en España." In: *Madrid, Cátedra* (cit. on p. 70).

Grave, Edouard et al. (2018). "Learning Word Vectors for 157 Languages." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (cit. on p. 161).

Green, T Clifton et al. (2019). “Crowdsourced employer reviews and stock returns.” In: *Journal of Financial Economics* (cit. on pp. 3, 50, 51, 127, 130, 145, 149, 180).
Greenhaus, Jeffrey H and Nicholas J Beutell (1985). “Sources of conflict between work and family roles.” In: *Academy of management review* 10.1, pp. 76–88 (cit. on pp. 49, 127).
Grevet, Catherine, Loren G Terveen, and Eric Gilbert (2014). “Managing political differences in social media.” In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, pp. 1400–1408 (cit. on pp. 21, 44).
Griffeth, Rodger W, Peter W Hom, and Stefan Gaertner (2000). “A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium.” In: *Journal of management* 26.3, pp. 463–488 (cit. on pp. 13, 52, 128, 138, 140).
Groth, Otto (1948). *Die Geschichte der deutschen Zeitungswissenschaft: Probleme und Methoden*. K. Weinmayer (cit. on pp. 1, 23).
Gruzd, Anatoliy and Jeffrey Roy (2014). “Investigating political polarization on Twitter: A Canadian perspective.” In: *Policy & Internet* 6.1, pp. 28–45 (cit. on pp. 1, 3, 21, 40, 43, 101, 103, 123).
Guerra, Pedro Henrique Calais et al. (2013). “A Measure of Polarization on Social Media Networks Based on Community Boundaries.” In: *ICWSM* (cit. on pp. 3, 40, 42, 101, 103, 123).
Guinard, Paul-J (1973). “La presse espagnole de 1737 à 1791.” In: *Formation et signification d’un genre, op. cit*, p. 344 (cit. on pp. 67, 74).
Gunlu, Ebru, Mehmet Aksarayli, and Nilüfer Şahin Perçin (2010). “Job satisfaction and organizational commitment of hotel managers in Turkey.” In: *International Journal of Contemporary Hospitality Management* (cit. on p. 129).
Gunning, Robert (1952). “The technique of clear writing.” In: (cit. on p. 32).
Gustafson, Walter W (1932). “The Influence of the" Tatler" and" Spectator" in Sweden.” In: *Scandinavian studies and notes* 12.4, pp. 65–72 (cit. on pp. 38, 58).
Guy, Ido et al. (2016). “What is your organization’like’? A study of liking activity in the enterprise.” In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3025–3037 (cit. on p. 149).

Habash, Nizar Y (2010). "Introduction to Arabic natural language processing." In: *Synthesis Lectures on Human Language Technologies* 3.1, pp. 1–187 (cit. on p. 26).

Harmon, Mark D (2001). "Poll question readability and 'don't know'replies." In: *International Journal of Public Opinion Research* 13.1, pp. 72–79 (cit. on p. 33).

Harter, James K, Frank L Schmidt, and Theodore L Hayes (2002). "Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis." In: *Journal of applied psychology* 87.2, p. 268 (cit. on pp. 45, 49).

Hatzivassiloglou, Vasileios and Kathleen McKeown (1997). "Predicting the semantic orientation of adjectives." In: *35th annual meeting of the association for computational linguistics and 8th conference of the european chapter of the association for computational linguistics*, pp. 174–181 (cit. on p. 33).

Hausknecht, John P, Julianne Rodda, and Michael J Howard (2009). "Targeted employee retention: Performance-based and job-related differences in reported reasons for staying." In: *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management* 48.2, pp. 269–288 (cit. on pp. 13, 52, 128, 138, 140).

Hays, David G (1960). *Automatic Content Analysis: Some Entries for a Transformation Catalog*. Rand Corporation (cit. on p. 25).

Hearst, Marti A (1992). "Direction-based text interpretation as an information access refinement." In: *Text-based intelligent systems: current research and practice in information extraction and retrieval*, pp. 257–274 (cit. on p. 33).

Henny-Krahmer, Ulrike (2018). *Exploration of Sentiments and Genre in Spanish American Novels*. URL: `https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/` (visited on 02/16/2020) (cit. on pp. 36, 59).

Herbermann, Charles George (1913). *The Catholic encyclopedia: An international work of reference on the constitution, doctrine, discipline, and history of the Catholic Church*. Vol. 9. Encyclopedia Press (cit. on pp. 1, 23).

Herzberg, Frederick et al. (1968). *One more time: How do you motivate employees* (cit. on p. 47).

Herzberg, Frederick (2017). *Motivation to work*. Routledge (cit. on pp. 142, 150).

Herzberg, Frederick, Bernard Mausner, and Barbara Snyderman (1959). "The motivation to work." In: (cit. on pp. 46, 47, 142, 144, 150, 158, 160).
Herzberg, Frederick I (1966). "Work and the nature of man." In: (cit. on pp. 46, 47, 142, 150).
Hessel, Jack and Lillian Lee (2019). "Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features." In: *arXiv preprint arXiv:1904.07372* (cit. on pp. 3, 6, 40, 43, 45, 101–103, 109–111, 121, 122).
Hickman, Louis et al. (2020). "Text preprocessing for text mining in organizational research: Review and recommendations." In: *Organizational Research Methods*, p. 1094428120971683 (cit. on pp. 27, 28).
Hilbig, Benjamin E (2009). "Sad, thus true: Negativity bias in judgments of truth." In: *Journal of Experimental Social Psychology* 45.4, pp. 983–986 (cit. on p. 165).
Hofland, Knut and Stig Johansson (1982). *Word frequencies in british and american english*. Norwegian computing centre for the Humanities (cit. on pp. 109, 160).
Holberg, Ludvig (1748). *Pensées morales*. Vol. 2. Aux Depens Du Traducteur Et Compagnie (cit. on p. 39).
Holland, Peter, Brian K Cooper, and Rob Hecker (2016). "Use of social media at work: a new form of employee voice?" In: *The International Journal of Human Resource Management* 27.21, pp. 2621–2634 (cit. on p. 127).
Holmberg, Christopher, Jino Caro, and Iwona Sobis (2018). "Job satisfaction among Swedish mental health nursing personnel: Revisiting the two-factor theory." In: *International journal of mental health nursing* 27.2, pp. 581–592 (cit. on pp. 45, 47, 145, 151).
Honegger, Claudia (1991). *Die Ordnung der Geschlechter: Die Wissenschaften vom Menschen und das Weib, 1750-1850*. Frankfurt, Germany: Campus (cit. on p. 72).
Honnet, Pierre-Edouard et al. (2017). "Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German." In: *arXiv preprint arXiv:1710.11035* (cit. on p. 26).
Hoppock, Robert (1935). "Job satisfaction." In: (cit. on pp. 3, 45).
Hösle, Johannes (1993). *Carlo Goldoni*. Munich, Germany: Piper Verlag GmbH (cit. on p. 80).
Howard, Jack L and Dwight D Frink (1996). "The effects of organizational restructure on employee satisfaction." In: *Group & Organization Management* 21.3, pp. 278–303 (cit. on pp. 49, 50).

Hu, Nan, Indranil Bose, Yunjun Gao, et al. (2011). "Manipulation in digital word-of-mouth: A reality check for book reviews." In: *Decision Support Systems* 50.3, pp. 627–635 (cit. on p. 180).
Hu, Nan, Indranil Bose, Noi Sian Koh, et al. (2012). "Manipulation of online reviews: An analysis of ratings, readability, and sentiments." In: *Decision support systems* 52.3, pp. 674–684 (cit. on p. 32).
Hur, Yongbeom (2018). "Testing Herzberg's two-factor theory of motivation in the public sector: is it applicable to public managers?" In: *Public Organization Review* 18.3, pp. 329–343 (cit. on pp. 49, 145, 151).
Hutto, Clayton J and Eric Gilbert (2014). "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In: *Eighth international AAAI conference on weblogs and social media*. California, United States: The AAAI Press (cit. on pp. 22, 27, 34, 35, 58).
Ignatow, Gabe and Rada Mihalcea (2017). *An introduction to text mining: Research design, data collection, and analysis*. Sage Publications (cit. on p. 21).
Islam, Saif ul and Nazim Ali (2013). "Motivation-Hygiene Theory: Applicability on Teachers." In: *Journal of Managerial Sciences* 7.1 (cit. on pp. 161, 183).
Jacobson, Susan, Eunyoung Myung, and Steven L Johnson (2016). "Open media or echo chamber: The use of links in audience discussions on the Facebook pages of partisan news organizations." In: *Information, Communication & Society* 19.7, pp. 875–891 (cit. on p. 44).
Jang, Myungha and James Allan (2018). "Explaining controversy on social media via stance summarization." In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1221–1224 (cit. on pp. 3, 40, 43, 101–103, 111, 121).
Janis, Irving L and Raymond H Fadner (1943). "A coefficient of imbalance for content analysis." In: *Psychometrika* 8.2, pp. 105–119 (cit. on p. 24).
Jannidis, Fotis et al. (2016). "Analyzing features for the detection of happy endings in german novels." In: *arXiv preprint arXiv:1611.09028* (cit. on pp. 36, 59).
Janssen, Onne (2001). "Fairness perceptions as a moderator in the curvilinear relationships between job demands, and job performance and job satisfaction." In: *Academy of management journal* 44.5, pp. 1039–1050 (cit. on p. 45).

Jasser, Jasser et al. (2020). "Controversial information spreads faster and further in Reddit." In: *arXiv preprint arXiv:2006.13991* (cit. on pp. 3, 6, 40, 43).

Jelodar, Hamed et al. (2019). "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey." In: *Multimedia Tools and Applications* 78.11, pp. 15169–15211 (cit. on p. 26).

Joulin, Armand et al. (2016). "Bag of tricks for efficient text classification." In: *arXiv preprint arXiv:1607.01759* (cit. on p. 30).

Jung, Yeonjae and Yongmoo Suh (2019). "Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews." In: *Decision Support Systems* 123, p. 113074 (cit. on p. 51).

Juršic, Matjaz et al. (2010). "Lemmagen: Multilingual lemmatisation with induced ripple-down rules." In: *Journal of Universal Computer Science* 16.9, pp. 1190–1214 (cit. on p. 28).

Kandel, Liliane and Abraham Moles (1958). "Application de l'indice de Flesch à la langue française." In: *Cahiers Etudes de Radio-Télévision* 19.1958, pp. 253–274 (cit. on p. 112).

Karimi, Fariba et al. (2018). "Homophily influences ranking of minorities in social networks." In: *Scientific reports* 8.1, p. 11077 (cit. on p. 89).

Kasper, Patrick et al. (2019). "On the Role of Score, Genre and Text in Helpfulness of Video Game Reviews on Metacritic." In: *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, todo (cit. on p. 131).

Kern, Margaret L, Paul X McCarthy, et al. (2019). "Social media-predicted personality traits and values can help match people to their ideal jobs." In: *Proceedings of the National Academy of Sciences* 116.52, pp. 26459–26464 (cit. on p. 182).

Kern, Margaret L, Gregory Park, et al. (2016). "Gaining insights from social media language: Methodologies and challenges." In: *Psychological methods* 21.4, p. 507 (cit. on p. 27).

Khanna, Virender (2017). "Measuring job satisfaction of academicians using Herzberg theory." In: *Delhi Business Review* 18.2, pp. 75–86 (cit. on p. 48).

Kim, Evgeny, Sebastian Padó, and Roman Klinger (2017). "Prototypical Emotion Developments in Literary Genres." In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 17–26 (cit. on p. 36).

Kim, Soo-Min and Eduard Hovy (2004). "Determining the sentiment of opinions." In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1367–1373 (cit. on p. 34).

Kim, Soo-Min, Patrick Pantel, et al. (2006). "Automatically assessing review helpfulness." In: *Proceedings of the 2006 Conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 423–430 (cit. on pp. 22, 130).

Kincaid, J Peter et al. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch (cit. on pp. 31, 33).

King, Rachael Scarborough (2018). "All the News that's Fit to Write: The Eighteenth-Century Manuscript Newsletter." In: *Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century*. "Leiden, The Netherlands": Brill, pp. 95–118 (cit. on p. 58).

Klein, Sheldon and Robert F Simmons (1963). "A computational approach to grammatical coding of English words." In: *Journal of the ACM (JACM)* 10.3, pp. 334–347 (cit. on p. 29).

Kobayashi, Vladimer B et al. (2018). "Text mining in organizational research." In: *Organizational research methods* 21.3, pp. 733–765 (cit. on p. 27).

Koncar, Philipp, Alexandra Fuchs, et al. (2020). "Text sentiment in the Age of Enlightenment: an analysis of Spectator periodicals." In: *Applied Network Science* 5.1, pp. 1–32 (cit. on pp. 9, 10, 15, 19, 53, 188, 193, 195).

Koncar, Philipp and Denis Helic (2019). "Text Sentiment in the Age of Enlightenment." In: *International Conference on Complex Networks and Their Applications*. Cham, Switzerland: Springer, pp. 350–362. ISBN: 978-3-030-36683-4 (cit. on pp. 53, 56, 88, 92, 94).

Koncar, Philipp and Denis Helic (2020). "Employee Satisfaction in Online Reviews." In: *International Conference on Social Informatics*. Springer, pp. 152–167 (cit. on pp. 13–15, 19, 125, 149, 190, 194, 195).

Koncar, Philipp, Tiago Santos, et al. (2021). "What Herzberg's Two-Factor Theory Reveals About Employee Satisfaction in Online Employer Reviews." In: *Business & Information Systems Engineering* 5.1, pp. 1–32 (cit. on pp. 13–15, 19, 125, 190, 194, 195, 199).

Koncar, Philipp, Simon Walk, and Denis Helic (2021). "Analysis and Prediction of Multilingual Controversy on Reddit." In: *13th ACM Conference on Web Science*, pp. 21–28 (cit. on pp. 11, 12, 14, 15, 19, 99, 189, 194, 195).

Kondratyuk, Daniel et al. (2018). "LemmaTag: Jointly tagging and lemmatizing for morphologically-rich languages with BRNNs." In: *arXiv preprint arXiv:1808.03703* (cit. on p. 28).

Könsgen, Raoul et al. (2018). "Finding meaning in contradiction on employee review sites—effects of discrepant online reviews on job application intentions." In: *Journal of Interactive Marketing* 43, pp. 165–177 (cit. on pp. 51, 149).

Korfiatis, Nikolaos, Elena García-Bariocanal, and Salvador Sánchez-Alonso (2012). "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content." In: *Electronic Commerce Research and Applications* 11.3, pp. 205–217 (cit. on pp. 146, 159, 170, 179).

Korfiatis, Nikolaos, Daniel Rodríguez, and Miguel-Angel Sicilia (2008). "The impact of readability on the usefulness of online product reviews: a case study on an online bookstore." In: *World Summit on Knowledge Society*. Springer, pp. 423–432 (cit. on pp. 32, 146, 159, 170, 179).

Kotni, VV Devi Prasad and Venkateswarlu Karumuri (2018). "Application of Herzberg two-factor theory model for motivating retail salesforce." In: *IUP Journal of Organizational Behavior* 17.1, pp. 24–42 (cit. on pp. 49, 145, 151).

Koutra, Danai, Paul N Bennett, and Eric Horvitz (2015). "Events and controversies: Influences of a shocking news event on information seeking." In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 614–624 (cit. on pp. 40, 42, 103).

Krefting, Ellen (2018). "News versus Opinion: The State, the Press, and the Northern Enlightenment." In: *Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century*. Leiden, The Netherlands: Brill, pp. 299–318 (cit. on pp. 38, 58).

Krefting, Ellen, Aina Nøding, and Mona Ringvej (2015). *Eighteenth-century periodicals as agents of change: perspectives on northern Enlightenment*. Leiden, The Netherlands: Brill (cit. on pp. 38, 58).

Krippendorff, Klaus (2018). *Content analysis: An introduction to its methodology*. Sage publications (cit. on pp. 23–25).

Kulig, Andrzej et al. (2015). "Modeling the average shortest-path length in growth of word-adjacency networks." In: *Physical Review E* 91.3, p. 032810 (cit. on p. 59).

Kumar, V and Anita Pansari (2015). "Measuring the benefits of employee engagement." In: *MIT Sloan Management Review* 56.4, p. 67 (cit. on pp. 49, 127, 144).

Lacic, Emanuel, Dominik Kowald, and Elisabeth Lex (2016). "High enough? explaining and predicting traveler satisfaction using airline reviews." In: *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, pp. 249–254 (cit. on p. 149).

Lamprecht, Daniel et al. (2017). "How the structure of wikipedia articles influences user navigation." In: *New Review of Hypermedia and Multimedia* 23.1, pp. 29–50 (cit. on p. 122).

Lasswell, Harold Dwight (1971). *Propaganda technique in world war I*. Vol. 170. MIT press (cit. on p. 24).

Lauw, Hady W, Ee-Peng Lim, and Ke Wang (2012). "Quality and leniency in online collaborative rating systems." In: *ACM Transactions on the Web (TWEB)* 6.1, pp. 1–27 (cit. on p. 142).

Le, Hang et al. (2020). "FlauBERT: Unsupervised Language Model Pre-training for French." In: *LREC* (cit. on p. 27).

Lenzner, Timo (2014). "Are readability formulas valid tools for assessing survey question difficulty?" In: *Sociological Methods & Research* 43.4, pp. 677–698 (cit. on pp. 31, 33).

Lévrier, Alexis (2007). *Les journaux de Marivaux et le monde des "spectateurs"*. Paris, France: Presses Paris Sorbonne (cit. on p. 62).

Li, Xinxin and Lorin M Hitt (2008). "Self-selection and information role of online product reviews." In: *Information Systems Research* 19.4, pp. 456–474 (cit. on pp. 130, 149).

Li, Yaokui (2018). "Insight into Herzberg's Two-factor Incentive Theory." In: *2018 2nd International Conference on Economic Development and Education Management (ICEDEM 2018)*. Atlantis Press (cit. on pp. 52, 156, 199).

Liu, Bing (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press (cit. on p. 33).

Liu, Bing and Lei Zhang (2012). "A survey of opinion mining and sentiment analysis." In: *Mining text data*. Bosten, Massachusetts, United States: Springer, pp. 415–463 (cit. on pp. 34, 55).

Liu, Zhiyuan et al. (2010). "Automatic keyphrase extraction via topic decomposition." In: *Proceedings of the 2010 conference on empirical methods in natural language processing*. Cambridge, Massachusetts, United States: Association for Computer Linguistics, pp. 366–376 (cit. on pp. 29, 60).

Löbl, Emil (1903). *Kultur und Presse*. Duncker & Humblot (cit. on p. 24).
Locke, Edwin A (1976). "The nature and causes of job satisfaction." In: *Handbook of industrial and organizational psychology* (cit. on pp. 45, 127, 129).
Lodahl, Thomas M (1964). "Patterns of job attitudes in two assembly technologies." In: *Administrative Science Quarterly*, pp. 482–519 (cit. on p. 162).
López, Teófanes Egido (2002). *Prensa clandestina española del siglo XVIII:" El Duende Crítico"*. Vol. 24. Valladolid, Spain: Universidad de Valladolid (cit. on p. 67).
Lovins, Julie Beth (1968). "Development of a stemming algorithm." In: *Mech. Transl. Comput. Linguistics* 11.1-2, pp. 22–31 (cit. on p. 28).
Lumley, EJ et al. (2011). "Exploring the job satisfaction and organisational commitment of employees in the information technology environment." In: *Southern African Business Review* 15.1 (cit. on p. 129).
Lund, Kevin and Curt Burgess (1996). "Producing high-dimensional semantic spaces from lexical co-occurrence." In: *Behavior research methods, instruments, & computers* 28.2, pp. 203–208 (cit. on p. 30).
Lundberg, Christine, Anna Gudmundson, and Tommy D Andersson (2009). "Herzberg's Two-Factor Theory of work motivation tested empirically on seasonal workers in hospitality and tourism." In: *Tourism management* 30.6, pp. 890–899 (cit. on pp. 45, 47, 145, 151).
Luo, Ning, Yilu Zhou, and John Shon (2016). "Employee satisfaction and corporate performance: Mining employee reviews on glassdoor.com." In: (cit. on pp. 49, 50, 130, 145, 149).
Maas, Andrew L et al. (2011). "Learning word vectors for sentiment analysis." In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 142–150 (cit. on pp. 34, 59).
Makrehchi, Masoud (2014). "The correlation between language shift and social conflicts in polarized social media." In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Vol. 2. New Jersey, United States: IEEE Computer Society, pp. 166–171 (cit. on p. 93).
Malik, M and Khalid Iqbal (2018). "Review helpfulness as a function of Linguistic Indicators." In: *Int J Comput Sci Netw Secur* 18, pp. 234–240 (cit. on p. 130).

Malik, MSI and Ayyaz Hussain (2018). “An analysis of review content and reviewer variables that contribute to review helpfulness.” In: *Information Processing & Management* 54.1, pp. 88–104 (cit. on pp. 22, 130).
Malik, Muhammad Ehsan and Basharat Naeem (2013). “Towards understanding controversy on Herzberg theory of motivation.” In: *World Applied Sciences Journal* 24.8, pp. 1031–1036 (cit. on pp. 161, 183).
Malik, Muhammad Ehsan, Samina Nawab, et al. (2010). “Job satisfaction and organizational commitment of university teachers in public sector of Pakistan.” In: *International journal of business and management* 5.6, p. 17 (cit. on pp. 127, 129).
Mann, Michael (1996). “The emergence of modern European nationalism.” In: *Poznan Studies in the Philosophy of the Sciences and the Humanities* 48, pp. 147–170 (cit. on p. 94).
Marinescu, Ioana et al. (2018). *Incentives can reduce bias in online reviews*. Tech. rep. National Bureau of Economic Research (cit. on pp. 51, 130, 149).
Martens, Wolfgang (2017). *Die Botschaft der Tugend: die Aufklärung im Spiegel der deutschen moralischen Wochenschriften*. Heidelberg, Germany: Springer-Verlag (cit. on pp. 38, 58).
Martin, Helen (1936). “Nationalism in children’s literature.” In: *The Library Quarterly* 6.4, pp. 405–418 (cit. on pp. 1, 24).
Massey, Brian L and Cindy J Elmore (2011). “Happier working for themselves? Job satisfaction and women freelance journalists.” In: *Journalism Practice* 5.6, pp. 672–686 (cit. on p. 141).
Mathews, Byron C (1910). “A study of a New York daily.” In: *Independent* 68, pp. 82–86 (cit. on p. 23).
Mattsson, Jan, Jos Lemmink, and Rod McColl (2004). “The effect of verbalized emotions on loyalty in written complaints.” In: *Total Quality Management & Business Excellence* 15.7, pp. 941–958 (cit. on p. 158).
Maute, Manfred F and William R Forrester Jr (1993). “The structure and determinants of consumer complaint intentions and behavior.” In: *Journal of economic psychology* 14.2, pp. 219–247 (cit. on p. 159).
Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014). “Promotional reviews: An empirical investigation of online review manipulation.” In: *American Economic Review* 104.8, pp. 2421–55 (cit. on p. 180).
Mc Laughlin, G Harry (1969). “SMOG grading-a new readability formula.” In: *Journal of reading* 12.8, pp. 639–646 (cit. on p. 32).

McAuley, Julian, Jure Leskovec, and Dan Jurafsky (2012). "Learning attitudes and attributes from multi-aspect reviews." In: *2012 IEEE 12th International Conference on Data Mining*. IEEE, pp. 1020–1025 (cit. on p. 149).
McCullough, Michael E, Frank D Fincham, and Jo-Ann Tsang (2003). "Forgiveness, forbearance, and time: the temporal unfolding of transgression-related interpersonal motivations." In: *Journal of personality and social psychology* 84.3, p. 540 (cit. on p. 141).
Mcteer, M, R Schwartz, and R Weischedel (1991). "Empirical studies in part-of-speech labeling." In: *Proceedings Of the 4th DARPA Workshop on Speech and Natural Language*, pp. 331–336 (cit. on p. 28).
Mejova, Yelena et al. (2014). "Controversy and sentiment in online news." In: *arXiv preprint arXiv:1409.8152* (cit. on pp. 11, 40, 41, 101–104, 109, 110, 114, 121, 122).
Melián-González, Santiago and Jacques Bulchand-Gidumal (2016). "Worker word of mouth on the internet." In: *International Journal of Manpower* (cit. on p. 180).
Melton, James Van Horn (2001). *The rise of the public in Enlightenment Europe*. Cambridge, United Kingdom: Cambridge University Press (cit. on p. 58).
Merchant, Kaiz and Yash Pande (2018). "Nlp based latent semantic analysis for legal text summarization." In: *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, pp. 1803–1807 (cit. on p. 26).
Messbarger, Rebecca (1999). "Reforming the Female Class:" Il Caffè"'s" Defense of Women"." In: *Eighteenth-century studies* 32.3, pp. 355–369 (cit. on pp. 37, 58).
Meyer, Erin (2018). "Being the Boss in Brussels, Boston, and Beijing." In: *Harvard Business Review* (cit. on p. 141).
Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient estimation of word representations in vector space." In: *arXiv preprint arXiv:1301.3781* (cit. on p. 30).
Mikolov, Tomas, Jiri Kopecky, et al. (2009). "Neural network based language models for highly inflective languages." In: *2009 IEEE international conference on acoustics, speech and signal processing*. IEEE, pp. 4725–4728 (cit. on p. 30).
Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed representations of words and phrases and their compositionality." In: *arXiv preprint arXiv:1310.4546* (cit. on p. 30).

Mikolov, Tomáš, Martin Karafiát, et al. (2010). "Recurrent neural network based language model." In: *Eleventh annual conference of the international speech communication association* (cit. on p. 30).
Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic regularities in continuous space word representations." In: *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751 (cit. on p. 30).
Miles, Sandra Jeanquart and W Glynn Mangold (2014). "Employee voice: Untapped resource or social media time bomb?" In: *Business Horizons* 57.3, pp. 401–411 (cit. on pp. 3, 50, 127, 145).
Miller, George A et al. (1990). "Introduction to WordNet: An on-line lexical database." In: *International journal of lexicography* 3.4, pp. 235–244 (cit. on pp. 25, 34).
Milliken, Frances J, Elizabeth W Morrison, and Patricia F Hewlin (2003). "An exploratory study of employee silence: Issues that employees don't communicate upward and why." In: *Journal of management studies* 40.6, pp. 1453–1476 (cit. on p. 127).
Milne, George R, Mary J Culnan, and Henry Greene (2006). "A longitudinal assessment of online privacy notice readability." In: *Journal of Public Policy & Marketing* 25.2, pp. 238–249 (cit. on pp. 32, 33).
Mishne, Gilad, Natalie Glance, et al. (2006). "Leave a reply: An analysis of weblog comments." In: *Third annual workshop on the Weblogging ecosystem*. Edinburgh, Scotland (cit. on pp. 12, 40, 41, 43, 101–104, 115, 121, 122).
Misra, Amita and Marilyn A Walker (n.d.). "Topic Independent Identification of Agreement and Disagreement in Social Media Dialogue." In: *Evolution* 460.460 (), p. 920 (cit. on p. 100).
Mitra, Tanushree et al. (2017). "Spread of Employee Engagement in a Large Organizational Network: A Longitudinal Analysis." In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW, pp. 1–20 (cit. on p. 149).
Mnih, Andriy and Geoffrey Hinton (2007). "Three new graphical models for statistical language modelling." In: *Proceedings of the 24th international conference on Machine learning*, pp. 641–648 (cit. on p. 30).
Montemurro, Marcelo A and Damián H Zanette (2013). "Keywords and co-occurrence patterns in the Voynich manuscript: An information-theoretic analysis." In: *PloS one* 8.6 (cit. on p. 82).

Morales, AJ et al. (2015). "Measuring political polarization: Twitter shows the two sides of Venezuela." In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.3, p. 033114 (cit. on pp. 3, 21, 40, 43, 101, 103, 123).

Moreno-Ortiz, Antonio (2017). "Lingmotif: Sentiment analysis for the digital humanities." In: *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, pp. 73–76 (cit. on p. 55).

Morgan, Betty Trebelle (1929). *Histoire du Journal des Sçavans depuis 1665 jusqu'en 1701*. Les presses universitaires de France (cit. on p. 38).

Morin, Frederic and Yoshua Bengio (2005). "Hierarchical probabilistic neural network language model." In: *Aistats*. Vol. 5. Citeseer, pp. 246–252 (cit. on p. 30).

Mornet, Daniel (1910). "Les enseignements des bibliothèques privées (1750-1780)." In: *Revue d'histoire littéraire de la France* 17.3, pp. 449–496 (cit. on p. 39).

Murray, Henry Alexander (1943). "Thematic apperception test." In: (cit. on p. 25).

Muse, Lori et al. (2008). "Work-life benefits and positive organizational behavior: is there a connection?" In: *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 29.2, pp. 171–192 (cit. on p. 141).

Nacos, Brigitte L et al. (1991). "Content analysis of news reports: Comparing human coding and a computer-assisted method." In: *Communication (New York)* 12.2, pp. 111–128 (cit. on p. 25).

Nasukawa, Tetsuya and Jeonghee Yi (2003). "Sentiment analysis: Capturing favorability using natural language processing." In: *Proceedings of the 2nd international conference on Knowledge capture*, pp. 70–77 (cit. on p. 33).

Newman, Jerry M, Barry Gerhart, and George T Milkovich (2016). *Compensation*. McGraw-Hill Higher Education (cit. on pp. 13, 52, 127, 128, 134, 140).

Oladotun, Kolawole Julius and Ali Öztüren (2013). "Motivational Factors Of Hospital Employees: Evidence From North Cyprus." In: *Interdisciplinary journal of contemporary research in businesses (IJCRB)* 4.12, pp. 106–123 (cit. on pp. 161, 183).

Orellana, Marcos, Andrea Trujillo, and Priscila Cedillo (2020). "A Comparative Evaluation of Preprocessing Techniques for Short Texts in Spanish."

In: *Future of Information and Communication Conference*. Springer, pp. 111–124 (cit. on p. 27).

Oshikawa, Ray, Jing Qian, and William Yang Wang (2018). "A survey on natural language processing for fake news detection." In: *arXiv preprint arXiv:1811.00770* (cit. on p. 26).

Özçelik, Merve et al. (2021). "HisNet: A Polarity Lexicon based on WordNet for Emotion Analysis." In: *Proceedings of the 11th Global Wordnet Conference*, pp. 157–165 (cit. on p. 25).

Pallares-Burke, Maria Lâucia G (2002). "The New History Confessions and Conversations." In: (cit. on p. 39).

Pallares-Burke, Maria Lúcia (1996). "The spectator abroad: The fascination of the mask." In: *History of European Ideas* 22.1, pp. 1–18 (cit. on p. 39).

Pallares-Burke, Maria Lúcia (2007). "The Spectator, or the metamorphoses of the periodical: a study in cultural translation." In: *Cultural translation in early modern Europe. Cambridge*, pp. 142–159 (cit. on pp. 38, 39, 58).

Pallares-Burke, Maria Lúcia G (1994). "A spectator in the tropics: a case study in the production and reproduction of culture." In: *Comparative Studies in Society and History*, pp. 676–701 (cit. on p. 38).

Pancer, Ethan et al. (2019). "How readability shapes social media engagement." In: *Journal of Consumer Psychology* 29.2, pp. 262–270 (cit. on pp. 31, 32).

Pang, Bo and Lillian Lee (2008). "Opinion mining and sentiment analysis." In: *Foundations and Trends in Information Retrieval* 2.1–2, pp. 1–135 (cit. on pp. 22, 33, 58).

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). "Thumbs up?: sentiment classification using machine learning techniques." In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Pennsylvania, United States: Association for Computational Linguistics, pp. 79–86 (cit. on pp. 34, 59).

Parsons, Elizabeth and Adelina Broadbridge (2006). "Job motivation and satisfaction: Unpacking the key factors for charity shop managers." In: *Journal of Retailing and Consumer Services* 13.2, pp. 121–131 (cit. on pp. 52, 156, 199).

Pennacchiotti, Marco and Ana-Maria Popescu (2010). "Detecting controversies in Twitter: a first study." In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*. ACL, pp. 31–32 (cit. on pp. 11, 123).

Pennebaker, James W, Ryan L Boyd, et al. (2015). *The development and psychometric properties of LIWC2015*. Tech. rep. (cit. on p. 27).
Pennebaker, James W, Martha E Francis, and Roger J Booth (2001). "Linguistic inquiry and word count: LIWC 2001." In: *Mahway: Lawrence Erlbaum Associates* 71.2001, p. 2001 (cit. on pp. 35, 58).
Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543 (cit. on p. 30).
Podsakoff, Philip M and Larry J Williams (1986). "The relationship between job performance and job satisfaction." In: *Generalizing from laboratory to field settings* 207, p. 253 (cit. on pp. 3, 127).
Porter, Martin F (1980). "An algorithm for suffix stripping." In: *Program* (cit. on p. 28).
Porter, Martin F (2001). *Snowball: A language for stemming algorithms* (cit. on p. 28).
Prettenhofer, Peter and Benno Stein (2010). "Cross-language text classification using structural correspondence learning." In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computer Linguistics, pp. 1118–1127 (cit. on p. 59).
Rad, Hoda Sepehri and Denilson Barbosa (2012). "Identifying controversial articles in Wikipedia: A comparative study." In: *Proceedings of the eighth annual international symposium on wikis and open collaboration*. ACM, p. 7 (cit. on pp. 3, 6, 40, 41, 101, 103).
Rainoff, TJ (1929). "Wave-like fluctuations of creative productivity in the development of West-European physics in the eighteenth and nineteenth centuries." In: *Isis* 12.2, pp. 287–319 (cit. on p. 24).
Ramayah, T and AM Nasurdin (2006). "Integrating importance into the relationship between job satisfaction and commitment: A conceptual model." In: *The ICFAI Journal of Organization Behavior* 5.2, pp. 20–27 (cit. on pp. 3, 127).
Rathi, Neerpal and Ronu Rastogi (2008). "Job satisfaction and psychological well-being." In: *The Icfai University Journal of Organizational Behavior* 7.4, pp. 47–57 (cit. on pp. 3, 127).
Rau, Fritz (1980). *Zur Verbreitung und nachahmung des tatler und spectator*. 145. C. Winter (cit. on pp. 38, 62).

Ravasz, Erzsébet and Albert-László Barabási (2003). "Hierarchical organization in complex networks." In: *Physical review E* 67.2, p. 026112 (cit. on p. 91).

Reagan, Andrew J, Lewis Mitchell, et al. (2016). "The emotional arcs of stories are dominated by six basic shapes." In: *EPJ Data Science* 5.1, p. 31 (cit. on pp. 36, 97).

Reagan, Andrew J, Brian Tivnan, et al. (2015). "Benchmarking sentiment analysis methods for large-scale texts: a case for using continuum-scored words and word shift graphs." In: *arXiv preprint arXiv:1512.00531* (cit. on p. 95).

Reinstein, David A and Christopher M Snyder (2005). "The influence of expert reviews on consumer demand for experience goods: A case study of movie critics." In: *The journal of industrial economics* 53.1, pp. 27–51 (cit. on pp. 1, 22, 149).

Rello, Luz et al. (2012). "Graphical schemes may improve readability but not understandability for people with dyslexia." In: *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pp. 25–32 (cit. on pp. 31, 33).

Ries, Paul (1977). "The Anatomy of a Seventeenth-Century Newspaper." In: *Daphnis* 6.1, p. 171 (cit. on p. 37).

Rohde, Douglas LT, Laura M Gonnerman, and David C Plaut (2006). "An improved model of semantic similarity based on lexical co-occurrence." In: *Communications of the ACM* 8.627-633, p. 116 (cit. on p. 30).

Rosa, Colombo (1966). "Lo Spectator ei giornali veneziani del settecento." In: *Bari, Adriatica editrice* (cit. on p. 62).

Roxas, Ranzivelle Marianne and Giovanni Tapang (2010). "Prose and poetry classification and boundary detection using word adjacency network analysis." In: *Int. J. Mod. Phys. C* 21.04, pp. 503–512 (cit. on pp. 29, 60).

Rozin, Paul and Edward B Royzman (2001). "Negativity bias, negativity dominance, and contagion." In: *Personality and social psychology review* 5.4, pp. 296–320 (cit. on p. 165).

Russo, Irene, Francesca Frontini, and Valeria Quochi (2016). *OpeNER Sentiment Lexicon Italian - LMF*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa. URL: `http://hdl.handle.net/20.500.11752/ILC-73` (cit. on p. 95).

Ryan, Kathleen M (2009). "The performative journalist: Job satisfaction, temporary workers and American television news." In: *Journalism* 10.5, pp. 647–664 (cit. on p. 141).

Saha, Koustuv et al. (2019). "Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–30 (cit. on pp. 50, 150).

Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975). "A vector space model for automatic indexing." In: *Communications of the ACM* 18.11, pp. 613–620 (cit. on p. 29).

Santos, Tiago et al. (2019). "What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, p. 140 (cit. on pp. 131, 149).

Sarvis, Betty and Hyman Rodman (1973). *The abortion controversy.* Columbia U. Press (cit. on p. 41).

Schlesinger, Arthur M and Arthur Walworth (1938). "School Histories at War: A Study of the Treatment of Our Wars in the Secondary School History Books of the United States and in Those of its Former Enemies." In: (cit. on pp. 22, 24).

Schmidt, Anna and Michael Wiegand (2017). "A survey on hate speech detection using natural language processing." In: *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10 (cit. on p. 26).

Schmidt, Thomas and Manuel Burghardt (2018). "Toward a Tool for Sentiment Analysis for German Historic Plays." In: pp. 46–48 (cit. on p. 55).

Schmidt, Thomas, Manuel Burghardt, and Christian Wolff (2018). "Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten." In: *INF-DH-2018* (cit. on pp. 36, 59).

Schneider, Benjamin and Neal Schmitt (1976). *Staffing organizations*. Goodyear Publishing Company (cit. on pp. 3, 45, 127, 129).

Schnurr, Paula P, Stanley D Rosenberg, and Thomas E Oxman (1992). "Comparison of TAT and free speech techniques for eliciting source material in computerized content analysis." In: *Journal of personality assessment* 58.2, pp. 311–325 (cit. on p. 25).

Scholger, Martina (2018). “"Spectators" in the International Context - A Digital Scholarly Edition.” In: *Discourses on Economy in the Spectators*. Hamburg, Germany: Verlag Dr. Kovač, pp. 229–247 (cit. on p. 61).

Schommer-Aikins, Marlene and Rosetta Hutter (2002). “Epistemological beliefs and thinking about everyday controversial issues.” In: *The journal of Psychology* 136.1, pp. 5–20 (cit. on pp. 2, 100).

Schröder, Thomas (1995). *Die ersten Zeitungen: Textgestaltung und Nachrichtenauswahl*. Gunter Narr Verlag (cit. on p. 37).

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge (cit. on p. 28).

Sebeok, Thomas A and Valdis J Zeps (1958). “An Analysis of Structured Content, with Application of Electronic Computer Research, in Psycholinguistics.” In: *Language and Speech* 1.3, pp. 181–193 (cit. on p. 25).

Sedelow, Sally Yeates (1989). “The Interlingual Thesaurus Model for Global Technical Communication: Research Results.” In: (cit. on p. 25).

Segarra, Santiago, Mark Eisen, Gabriel Egan, et al. (2016). “Attributing the authorship of the Henry VI plays by word adjacency.” In: *Shakespeare Quarterly* 67.2, pp. 232–256 (cit. on p. 29).

Segarra, Santiago, Mark Eisen, and Alejandro Ribeiro (2015). “Authorship attribution through function word adjacency networks.” In: *IEEE Transactions on Signal Processing* 63.20, pp. 5464–5478 (cit. on p. 29).

Seguin, Jean-Pierre (1963). “L’Information en France avant le périodique: 500 canards imprimés entre 1529 et 1631.” In: *Arts et traditions populaires* 11.1, pp. 20–32 (cit. on p. 37).

Shami, N Sadat et al. (2014). “Understanding employee social media chatter with enterprise social pulse.” In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pp. 379–392 (cit. on p. 149).

Shaver, Kelly G (2012). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, United States: Springer Science & Business Media (cit. on p. 71).

Shi, Feng et al. (2019). “The wisdom of polarized crowds.” In: *Nature Human Behaviour*, p. 1 (cit. on p. 100).

Shin, Jieun and Kjerstin Thorson (2017). “Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media.” In: *Journal of Communication* (cit. on p. 44).

Siering, Michael, Amit V Deokar, and Christian Janze (2018). "Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews." In: *Decision Support Systems* 107, pp. 52–63 (cit. on p. 149).

Siering, Michael, Jan Muntermann, and Balaji Rajagopalan (2018). "Explaining and predicting online review helpfulness: The role of content and reviewer-related signals." In: *Decision Support Systems* 108, pp. 1–12 (cit. on p. 179).

Siersdorfer, Stefan et al. (2014). "Analyzing and mining comments and comment ratings on the social web." In: *ACM Transactions on the Web (TWEB)* 8.3, pp. 1–39 (cit. on pp. 40, 41, 101–103, 109–111, 121).

Silva, Thiago C and Diego R Amancio (2012). "Word sense disambiguation via high order of learning in complex networks." In: *EPL (Europhysics Letters)* 98.5, p. 58001 (cit. on pp. 29, 59).

Simpson, George Eaton (1936). "The negro in the Philadelphia press." In: (cit. on pp. 1, 24).

Singh, Jagdip (1988). "Consumer complaint intentions and behavior: definitional and taxonomical issues." In: *Journal of marketing* 52.1, pp. 93–107 (cit. on p. 159).

Singh, Vivek Kumar et al. (2013). "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification." In: *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*. IEEE, pp. 712–717 (cit. on p. 35).

Smerek, Ryan E and Marvin Peterson (2007). "Examining Herzberg's theory: Improving job satisfaction among non-academic employees at a university." In: *Research in higher education* 48.2, pp. 229–250 (cit. on pp. 161, 183).

Smith, Patricia Cain et al. (1969). "The measurement of satisfaction in work and retirement: A strategy for the study of attitudes." In: (cit. on pp. 3, 127).

Solangi, Yasir Ali et al. (2018). "Review on Natural Language Processing (NLP) and its toolkits for opinion mining and sentiment analysis." In: *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. IEEE, pp. 1–4 (cit. on p. 26).

Spector, Paul E (1985). "Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey." In: *American journal of community psychology* 13.6, pp. 693–713 (cit. on pp. 3, 46).

Spector, Paul E (1997). *Job satisfaction: Application, assessment, causes, and consequences*. Vol. 3. Sage (cit. on p. 46).

Speed, John Gilmer (1893). "Do newspapers now give the news." In: *Forum*. Vol. 15, pp. 705–711 (cit. on p. 23).

Spitzer, Robert J (2020). *The politics of gun control*. Routledge (cit. on p. 41).

Sprugnoli, Rachele et al. (2016). "Towards sentiment analysis for historical texts." In: *Digital Scholarship in the Humanities* 31.4, pp. 762–772 (cit. on p. 55).

Srujan, KS et al. (2018). "Classification of amazon book reviews based on sentiment analysis." In: *Information Systems Design and Intelligent Applications*. Springer, pp. 401–411 (cit. on p. 35).

Stamatatos, Efstathios (2017). "Authorship attribution using text distortion." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1138–1149 (cit. on p. 26).

Stamolampros, Panagiotis et al. (2019). "Job satisfaction and employee turnover determinants in high contact services: Insights from Employees' Online reviews." In: *Tourism Management* 75, pp. 130–147 (cit. on p. 50).

Steinbrügge, Lieselotte (2016). *Das moralische Geschlecht: Theorien und literarische Entwürfe über die Natur der Frau in der französischen Aufklärung. Ergebnisse der Frauenforschung*. Vol. 11. Heidelberg, Germany: Springer-Verlag (cit. on p. 72).

Steinhaus, Carol S and James L Perry (1996). "Organizational commitment: Does sector matter?" In: *Public Productivity & Management Review*, pp. 278–288 (cit. on pp. 49, 127).

Stello, Christina M (2011). "Herzberg's two-factor theory of job satisfaction: An integrative literature review." In: *Unpublished paper presented at The 2011 Student Research Conference: Exploring Opportunities in Research, Policy, and Practice, University of Minnesota Department of Organizational Leadership, Policy and Development, Minneapolis, MN* (cit. on p. 46).

Stephen, Leslie (1920). *English Literature and Society in the Eighteenth Century: Ford Lectures, 1903*. Vol. 28. Duckworth and Company (cit. on p. 38).

Stone, Philip J, Dexter C Dunphy, and Marshall S Smith (1966). "The general inquirer: A computer approach to content analysis." In: (cit. on pp. 25, 35).

Sumi, Róbert, Taha Yasseri, et al. (2011). "Edit wars in Wikipedia." In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, pp. 724–727 (cit. on pp. 6, 101).

Symitsi, Efthymia et al. (2021). "The informational value of employee online reviews." In: *European Journal of Operational Research* 288.2, pp. 605–619 (cit. on p. 51).

Tan, Chenhao et al. (2016). "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions." In: *Proceedings of the 25th international conference on world wide web*, pp. 613–624 (cit. on pp. 3, 6, 40, 43, 45, 121).

Tannenbaum, Percy H and Bradley S Greenberg (1961). "'JQ'References: A Study of Professional Change." In: *Journalism Quarterly* 38.2, pp. 203–207 (cit. on p. 24).

Temnikova, Irina, Sarah Vieweg, and Carlos Castillo (2015). "The case for readability of crisis communications in social media." In: *Proceedings of the 24th international conference on world wide web*, pp. 1245–1250 (cit. on pp. 31, 32).

Tessema, Mussie T, Kathryn J Ready, and Abel B Embaye (2013). "The effects of employee recognition, pay, and benefits on job satisfaction: Cross country evidence." In: *Journal of Business and Economics* 4.1, pp. 1–12 (cit. on pp. 13, 52, 127–129, 134, 140).

Thede, Scott M and Mary Harper (1999). "A second-order hidden Markov model for part-of-speech tagging." In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pp. 175–182 (cit. on p. 28).

Thelwall, Mike (2017). "The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength." In: *Cyberemotions*. Springer, pp. 119–134 (cit. on p. 35).

Thelwall, Mike et al. (2010). "Sentiment strength detection in short informal text." In: *Journal of the American Society for Information Science and Technology* 61.12, pp. 2544–2558 (cit. on pp. 35, 58).

Thet, Tun Thura, Jin-Cheon Na, and Christopher SG Khoo (2010). "Aspect-based sentiment analysis of movie reviews on discussion boards." In: *Journal of information science* 36.6, pp. 823–848 (cit. on p. 35).

Thompson, Cynthia A, Laura L Beauvais, and Karen S Lyness (1999). "When work–family benefits are not enough: The influence of work–family culture on benefit utilization, organizational attachment, and work–family conflict." In: *Journal of Vocational behavior* 54.3, pp. 392–415 (cit. on p. 141).
Thorndike, Edward L (1921). "The teacher's word book." In: (cit. on p. 31).
Tichaawa, Tembi Maloney and Love Odion Idahosa (2020). "Adapting Herzberg: Predicting Attendees' Motivation, Satisfaction, and Intention to Revisit a Festival in Cameroon Using an Ordered Logit Approach." In: *Tourism Review International* 24.2-3, pp. 91–107 (cit. on pp. 45, 48).
Tietjen, Mark A and Robert M Myers (1998). "Motivation and job satisfaction." In: *Management decision* 36.4, pp. 226–231 (cit. on pp. 145, 150).
Tjosvold, Dean (1985). "Implications of controversy research for management." In: *Journal of management* 11.3, pp. 21–37 (cit. on pp. 3, 40).
Vaussard, Maurice (2001). *L'Italia nel Settecento*. Italy: Fabbri (cit. on p. 77).
Vellusig, Robert Heinz (2000). *Schriftliche Gespräche: Briefkultur im 18. Jahrhundert*. Vol. 54. Vienna, Austria: Böhlau Verlag Wien (cit. on p. 70).
Vermeulen, Ivar E and Daphne Seegers (2009). "Tried and tested: The impact of online hotel reviews on consumer consideration." In: *Tourism management* 30.1, pp. 123–127 (cit. on pp. 130, 149).
Véronis, Jean (2004). "Hyperlex: lexical cartography for information retrieval." In: *Computer Speech & Language* 18.3, pp. 223–252 (cit. on pp. 29, 59).
Vijayarani, S, Ms J Ilamathi, Ms Nithya, et al. (2015). "Preprocessing techniques for text mining-an overview." In: *International Journal of Computer Science & Communication Networks* 5.1, pp. 7–16 (cit. on p. 28).
Vogel, Mabel and Carleton Washburne (1928). "An objective method of determining grade placement of children's reading material." In: *The Elementary School Journal* 28.5, pp. 373–381 (cit. on p. 31).
Von Tschilschke, Christian (2009). *Identität der Aufklärung–Aufklärung der Identität: Literatur und Identitätsdiskurs im Spanien des 18. Jahrhunderts*. Vol. 9. Madrid, Spain: Iberoamericana Editorial (cit. on pp. 68, 73, 78).
Voutilainen, Atro (2003). "Part-of-speech tagging." In: *The Oxford handbook of computational linguistics*, pp. 219–232 (cit. on p. 28).
Vroom, Victor Harold (1964). "Work and motivation." In: (cit. on p. 45).
Vydiswaran, VG et al. (2012). "BiasTrust: Teaching biased users about controversial topics." In: *Proceedings of the 21st ACM international conference on*

Information and knowledge management. ACM, pp. 1905–1909 (cit. on pp. 3, 100).

Wegerif, Rupert and Neil Mercer (1997). "Using computer-based text analysis to integrate qualitative and quantitative methods in research on collaborative learning." In: *Language and Education* 11.4, pp. 271–286 (cit. on pp. 1, 22).

Weiner, Bernard (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, United States: Guilford Press (cit. on p. 71).

Weiss, Howard M (2002). "Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences." In: *Human resource management review* 12.2, pp. 173–194 (cit. on pp. 49, 127).

Westbrook, Robert A (1987). "Product/consumption-based affective responses and postpurchase processes." In: *Journal of marketing research* 24.3, pp. 258–270 (cit. on p. 158).

White, Paul W (1924). "Quarter century survey of press content shows demand for facts." In: *New York* (cit. on p. 23).

Widdows, Dominic and Beate Dorow (2002). "A graph model for unsupervised lexical acquisition." In: *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. United States: Association for Computational Linguistics, pp. 1–7 (cit. on pp. 29, 59).

Wilcox, Delos F (1900). "The American newspaper: A study in social psychology." In: *The Annals of the American Academy of Political and Social Science* 16.1, pp. 56–92 (cit. on p. 23).

Wong, Kam-Fai et al. (2009). "Introduction to Chinese natural language processing." In: *Synthesis Lectures on Human Language Technologies* 2.1, pp. 1–148 (cit. on p. 26).

Xiao, Min and Yuhong Guo (2013). "Semi-supervised representation learning for cross-lingual text classification." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, United States: Association for Computer Linguistics, pp. 1465–1475 (cit. on p. 59).

Yates, Frank (1934). "Contingency tables involving small numbers and the χ 2 test." In: *Supplement to the Journal of the Royal Statistical Society* 1.2, pp. 217–235 (cit. on pp. 109, 160).

Ye, Qiang, Rob Law, and Bin Gu (2009). "The impact of online user reviews on hotel room sales." In: *International Journal of Hospitality Management* 28.1, pp. 180–182 (cit. on pp. 130, 149).

Ye, Qiang, Ziqiong Zhang, and Rob Law (2009). "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." In: *Expert systems with applications* 36.3, pp. 6527–6535 (cit. on pp. 34, 59).

Yom-Tov, Elad, Susan Dumais, and Qi Guo (2014). "Promoting civil discourse through search engine diversity." In: *Social Science Computer Review* 32.2, pp. 145–154 (cit. on pp. 40, 42, 103).

Zakaluk, Beverly L and S Jay Samuels (1988). *Readability: Its Past, Present, and Future.* ERIC (cit. on p. 31).

Zeelenberg, Marcel and Rik Pieters (2004). "Beyond valence in customer dissatisfaction: A review and new findings on behavioral responses to regret and disappointment in failed services." In: *Journal of business Research* 57.4, pp. 445–455 (cit. on p. 159).

Zeldow, Peter B and Dan P McAdams (1993). "On the comparison of TAT and free speech techniques in personality assessment." In: *Journal of personality assessment* 60.1, pp. 181–185 (cit. on p. 25).

Zhang, Lei, Shuai Wang, and Bing Liu (2018). "Deep learning for sentiment analysis: A survey." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, e1253 (cit. on pp. 34, 59).

Zhao, Yabing, Xun Xu, and Mingshu Wang (2019). "Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews." In: *International Journal of Hospitality Management* 76, pp. 111–121 (cit. on pp. 146, 159, 170, 179).

Ziegele, Marc, Timo Breiner, and Oliver Quiring (2014). "What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items." In: *Journal of Communication* 64.6, pp. 1111–1138 (cit. on pp. 1, 2, 6, 12, 22, 40, 41, 43, 100–104, 115, 121, 122).

Zielinski, Kazimierz et al. (2018). "Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries." In: *Information Processing & Management* 54.1, pp. 14–36 (cit. on pp. 3, 6, 40–42, 101–104, 114, 121, 122).

Zuo, Yuan, Jichang Zhao, and Ke Xu (2016). "Word network topic model: a simple but general solution for short and imbalanced texts." In: *Knowledge and Information Systems* 48.2, pp. 379–398 (cit. on pp. 22, 29, 60).