Mario Wagner, BSc

Diversity-Aware Recommendations in Twitter

MASTER'S THESIS

for the attainment of the degree of

Diplom-Ingenieur (Dipl.-Ing.)

submitted to

Graz University of Technology

Supervisor:
Ass.-Prof.Dipl-Ing.Dr.techn. Elisabeth Lex

Co-advisor:
Dipl.-Ing.Dr.techn. Dominik Kowald

Institute of Interactive Systems and Data Science (ISDS)
Head: Univ.Prof.Dipl-Ing.Dr.techn. Stefanie Lindstaedt

Graz, June 2, 2020

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit/Diplomarbeit/Dissertation identisch.

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 master's thesis/diploma thesis/doctoral dissertation.

Datum / Date

Unterschrift / Signature

Abstract

Political debates today are increasingly being held online, through social media and other channels. In times of Donald Trump, the American president, who mostly announces his messages via Twitter, it is important to clearly separate facts from falsehoods. Although there is an almost infinite amount of information online, tools such as recommender systems, filters and search encourage the formation of so-called filter bubbles. People who have similar opinions on polarizing topics group themselves and block other, challenging opinions. This leads to a deterioration of the general debate, as false facts are difficult to disprove for these groups.
With this thesis, we want to provide an approach on how to propose different opinions to users in order to increase the diversity of viewpoints regarding a political topic. We classify users into a politic spectrum, either pro-Trump or contra-Trump, and then suggest Tweets from the other spectrum. We then measure the impact of this process on diversity and serendipity.
Our results show that the diversity and serendipity of the recommendations can be increased by including opinions from the other political spectrum. In doing so, we want to contribute to improving the overall discussion and reduce the formation of groups that tend to be radical in extreme cases.

Keywords. Confirmation Bias; Selective Exposure; Filter Bubbles; tweet Recommendations; Diversity; Serendipity; Polarization; Hybrid Recommendations; Topic Similarity

Zusammenfassung

Politische Debatten werden heutzutage immer mehr online, über Social Media und andere Kanäle, abgehalten. In Zeiten von Donald Trump, dem amerikanischen Präsidenten, der seine Botschaften meist über Twitter verkündet, ist es wichtig, Fakten von Unwahrheiten klar zu trennen. Obwohl es online eine fast unendliche Menge an Informationen gibt, begünstigen Tools wie Recommender Systemen, Filter und Suche von Informationen die Bildung von sogenannten Filter Bubbles. Leute, die ähnliche Meinungen zu polarisierenden Themen haben, gruppieren sich und blocken andere, fordernde Meinungen ab. Das führt zu einer Verschlechterung der allgemeinen Debatte, da falsche Fakten nur schwer für diese Gruppen widerlegt werden können. Wir wollen mit dieser Arbeit einen Ansatz liefern, wie man Benutzern unterschiedliche Meinungen vorschlägt, damit sich die Vielfalt der Ansichten zu einem Thema erhöht. Wir klassifizieren Benutzer in ein politisches Spektrum, entweder pro-Trump oder contra-Trump und schlagen ihnen dann Tweets aus dem jeweiligen anderen Spektrum vor. Anschließend messen wir den Einfluss dieses Vorgangs auf die Metriken 'Diversität' und die 'Serendipität'. Unsere Resultate zeigen, dass die Diversität und Serendipität der Vorschläge eines Recommender Systems erhöht werden kann, indem man Meinungen des jeweiligen anderen politischen Spektrums miteinbezieht. Damit wollen wir einen Beitrag zur allgemeinen Diskussionsverbesserung schaffen und die Bildung von Gruppen, die im Extremfall zu Radikalität neigen, verringern.

You have power over your mind – not outside events. Realize this, and you will find strength. -Marcus Aurelius

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Today more than ever, people are connected through the Internet and have access to vast amounts of information. The communication barrier is easy to overcome, not least because geographical distance is no longer a limit [Graells-Garrido et al., 2013].
Technology enhances access to information in a centralized manner. Internet forums, news aggregators and social media channels are widely accessible on mobile devices, wearables and computers [Liao and Fu, 2013]. Even politicians use these channels to reach voters and citizens. Twitter is a popular tool in election campaigns. Political parties, candidates and journalists actively comment and share content on Twitter [Jungherr, 2016]. Best known is that President Trump is actively engaged, sharing and tweeting most of his opinions on Twitter [Ott, 2017].
Since user experience is very important, the ability to filter for unwanted content, search for agreeable content and subscribe to feeds of people with similar opinions is ubiquitous. However, the fact that content is filtered is not always obvious to the user. Through personalization, when searching for a term on Google, the search engine automatically assumes what you like through various predetermined factors. Search results show up differently for anyone logged in. When you search for the term 'Proof of climate change, as Pariser explains it, the search results might vary depending on whether you are an executive of an oil company or a climate change activist. The same behaviour applies to recommendations. While the personalizations offer a big advantage for the users, because they have to do less and less to get more benefits, the result is that more and more unwanted, inconsistent opinions and facts are hidden [Pariser, 2011].

Problem. People prefer to interact and spend time mostly with like-minded peers. The phenomenon behind it is called 'selective exposure' - individuals tend to avoid dissonant information and embrace agreeable information. Therefore, even though the internet is filled with information and a diverse amount of beliefs regarding topics, it is not guaranteed that this leads to an equally diverse exposure to different perspectives for a user. If users share a different point of view, they tend to disconnect from a group and join another group. The term which describes this is called 'filter bubble' [Pariser, 2011]. On the other hand, exposing people to challenging views is important for decision-making and critical thinking. Dangerous radicalization or inaccurate beliefs are corrected by exposition to diverse opinions, therefore serving as a countermeasure to the generation of filter bubbles [Neisser, 2010].

Approach. This thesis investigates how recommendations can be made more diverse, thus exposing people to opposing views. Our approach considers the user's history of tweets as her preferences of topics and generates suggestions for similar tweets with a content-based recommender. By combining recommendations of similar views with recommendations of opposing views in a single set, we aim to help provide users with a broader viewpoint on issues [Lex E., 2018]. As an example, we have taken the discussion on Twitter about the election of President Trump in 2016 and have studied it. Social networks, like Twitter, have provided a platform to reach voters for the 2016 election, facilitated many discussions surrounding the election campaign and played an important part during the campaign [Graells-Garrido et al., 2013]. The hashtag #maga, which stands for 'Make America great Again', was created by the Republicans for the campaign on social media and was very popular. The hashtag was used by pro-Trump users and contra-Trump users alike to address various issues. Many issues, such as the controversial promise of President Trump to build a wall at the border of Mexico were discussed by using this hashtag and similar ones. We analysed the different viewpoints and assigned users to two stances, a pro-Trump and a contra-Trump stance. After that, we recommend to users views on pro-Trump and contra-Trump stances, based on topics in the user's history of tweets. The work is mainly based on Graells et al. [Graells-Garrido et al., 2013]. They created word clouds to unite people with conflicting opinions on the issue *abortion in Chile*.

Contributions. The key contributions of our work are: (i) we classify user, which we crawled during the election of president Trump in 2016 into two stances: pro-Trump and contra-Trump. With this, we build different variants of recommendation sets for users and calculate metrics such as diversity and serendipity. We investigate whether changing the recommendation sets results in a change of the metrics and if so, which ones. Diversity compares in pairs the similarity of items based on their content in a given list, the higher the value, the more different they are [Ziegler et al., 2005]. Serendipity, on the flip side, measures how surprising a particular content is to a user [Ge et al., 2010]. We found that the combination of opposing and like-minded views in recommendations to a user is able to boost these metrics. Furthermore, we experimented with different ratios for pro-Trump and contra-Trump tweets in these hybrid sets and found a correlation to topical similarity of each group, when measuring diversity.

1.1 Research Questions

In order to clarify the problems addressed in this thesis, a research question was stated. This question is the summary of the main problems addressed by this thesis and is explained in detail in the following:

RQ 1: Measuring diversity and serendipity of hybrid recommendation sets
How is it possible to increase the diversity in a content-based recommender?

Most recommender systems analyse content that has been consumed in the past and suggest new content that is as similar as possible, trying to optimise for accuracy metrics, which supports the fact that the user eventually ends up in a filter bubble [Pariser, 2011]. By classifying users to political stances, we can characterize tweets into a predefined stance. This enables us to recommend tweets to the user, which are similar or opposed to her viewpoints regarding a topic. By doing so, we hope to increase diversity and serendipity in order to help her escape the filter bubble. We believe that, by merging tweets of different quantities to a recommendation set, we can significantly influence diversity.

1.2 Structure of this thesis

The thesis is structured as follows:

Chapter 2 gives an overview of the related work with respect to topics like filter bubbles and recommendation engines on social media. We present recommendation engines in general and talk about their connection to social media. Additionally, we explain filter bubbles and various approaches on how to mitigate them.

Chapter 3 goes into detail of the methodology of this thesis. We explain Twitter and the terminology used on the platform. We give an overview of the statistics of the crawled dataset and how we preprocessed it. In Section 3.2, technical preliminaries are explained in order to understand the experiments conducted in this thesis. We talk about vector space models, tf-idf and cosine similarity. After that, we show the approach which was taken in order to classify the users into a political stance. Last, in Section 3.4, we go into detail on diversity and serendipity.

Chapter 4 shows the experiments and results. The first Section 4.1 of the Chapter shows the results of quantitative experiments and measurements. We computed average metrics for 1,500 users of each stance and measured diversity and serendipity for various recommendation sets. The second Section 4.2 shows the qualitative results using 2 example users, one representative for the pro-Trump stance and the other for the contra-Trump stance.

The results of the previous Chapter are discussed and interpreted in the Chapter 5. Furthermore we point out the most important limitations of our work.

The last chapter presents the results in brief and shows how we intend to carry out further research based on this work in the future.

Chapter 2

Related Work

This Chapter gives an overview of the present state research in the academic space and the previous work that has been conducted in this field. The first section explains recommendation systems and the three basic approaches to setting up recommendation systems. The next Section talks about research and existing approaches about recommending content to users on social media platforms. Next, research on the effects of filter bubbles on individuals and social groups in general is examined. In the last Section, the existing approaches of mitigating filter bubbles and ways of presenting challenging content to users are analysed.

2.1 Recommender Systems

Recommender Systems try to suggest content to users that they are very likely to find useful by predicting ratings of items. There are three general approaches to recommender systems [Ricci et al., 2011]:

1. **Collaborative Filtering.** Collaborative filtering suggests content to users that other users who are similar to them have rated positively. The biggest advantage is that with this approach texts do not need to be interpreted by the algorithms [Ricci et al., 2011]. These systems are capable of recommending complex items, without 'understanding' the items.[John S. Breese et al., 1998]

2. **Content-Based Recommender.** This type of recommender system works by taking into account the actual content of the items. This is done by analyzing the user's past and which items he has positively evaluated. Based on this, items with similar content are suggested to the user. [Ricci et al., 2011].

3. **Hybrid-Based Approach.** Hybrid-Based recommender use different recommender approaches together. Often, collaborative Filtering is used with another approach in a weighted way [Çano and Morisio, 2017].

The work of [Graells-Garrido et al., 2013], on which this thesis is based, uses a content-based recommender approach. For this reason we will go into more detail about content-based recommender systems in the following section.

2.1.1 Content-Based Recommender

Content-based recommender systems look at content that a user has positively rated in the past. They use this information as the basis for their recommendations. This system analyses documents that are available about a user, derives features from them and finally creates a profile of the user.Finally, new items with features that are as similar as possible to the user's profile are suggested. The result is a set of recommendations, that tells you how relevant a document is to a user [Ricci et al., 2011].
The architecture of a content-based recommender is explained in Figure 2.1 [Ricci et al., 2011, chap. 3]. This process is performed in three steps:

- **Content Analyzer.** The information coming from the information source is mostly unstructured information. The main purpose of this step is to convert this information into a structured representation. Various feature extraction techniques can be utilized to get relevant features from the content. In this thesis, the dataset of users and their corresponding tweets are cleaned in this step. Furthermore, the tweets are tokenized in order to enable further processing. The output of this step is a structured representation of items [Ricci et al., 2011, chap. 3].

- **Profile Learner.** The goal of this step is to collect the represented items belonging to a user and learn a generalized profile for each user. Most of the time, this generalization is done with machine learning. This thesis uses a simple approach to represent each Twitter account. Out of the last 1,000 of each user's tweets, the top trigram is found. If the frequency of the top trigrams is not distinct, the n topmost occurrences of the trigrams are concatenated [Ricci et al., 2011, chap. 3].

- **Filtering Component.** The filtering component matches the user profile with a list of items. The output is a list of recommendations, that might be interesting to the user [Ricci et al., 2011, chap. 3]. In this work, Apache Solr's more-like-this functionality is triggered to match the trigrams with tweets that are new to the user.

The record of user feedback can be distinguished between two different techniques: When feedback is obtained straight from the user, this is referred to as 'explicit feedback'. There is also a technique called 'implicit feedback'. In order to get indirect feedback from the user, the activities of the user are observed and conclusions are drawn from these. [Ricci et al., 2011, chap. 3].

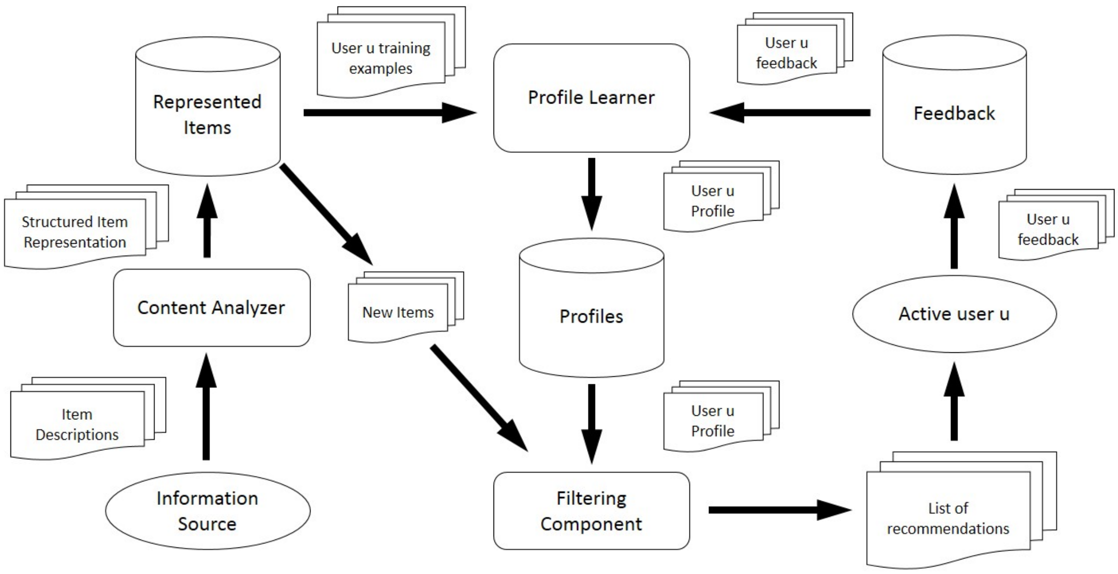

Figure 2.1: **Content-based Recommender** - This Figure shows a high level architecture of a typical Content-based recommender [Ricci et al., 2011, fig. 3.1].

2.1.2 Advantages and Disadvantages of Content-based Filtering

The principal benefits of content-based filtering recommendation engines are [Ricci et al., 2011, chap. 3]:

- **Transparency.** The content-based filtering method can be readily understood , because features of the content and the user profile can be analysed and compared . Recommendations resulting from the content-based filtering method are therefore comprehensible for the human.
- **No First-rater problem.** New items, that have not got any rating yet, can be recommended by a content-based recommender, because only the content is relevant.
- **No other users are needed.** There is no need to look at other users, only the user for which one wants to generate recommendations is needed.

The main disadvantages for these types of recommenders are [Ricci et al., 2011, chap. 3]:

- **Limitations in content analysis.** In order to extract domain knowledge from texts, you have only a limited number of features available that are present in documents. The recommender relies on the information in the content of items in order to make the predictions. If there is a lack of specific content, it is not possible to get suitable recommendations.
- **Over-specialization.** This problem is also called serendipity-problem, because content-based recommenders are not able to find surprising content for a user. For instance, a user that has only liked songs from the Beatles will only get suggestions of other Beatles songs, if the recommender works perfectly.
- **New user problem.** Users who have not yet rated items (or very few items) in the system will most certainly not get any reliable recommendations.

2.1.3 Recommender Systems for Twitter

Plenty of work has been put into researching and implementing different recommendation algorithms on Twitter and other social networks. Generally, papers divide between the above mentioned content-based approaches and collaboration-based approaches. Twitters own recommendation and search algorithm is called SALSA (Stochastic Approach for Link-Structure Analysis) and suggests Twitter accounts that users might want to follow [Gupta et al., 2013]. It is a random-walk algorithm, which constructs a bipartite graph, consisting of hubs and authorities. The graph depends on shared interests, shared connections and several other factors. It consists of vertices, representing users, connected by edges, that represent the follow relationships. Since the relationship on Twitter is one-sided, a user is able to follow another user without reciprocation. Different representations of the results might influence the users to accept a result or dismiss it. One of the metrics used for comparing the different algorithms is the 'follow-through-rate' (FTR). In order to calculate the FTR, the impact on the follows needs to be divided through the amount of views on a topic. The suggested algorithm performs well, when comparing the performance against other algorithms through FTR.. A bottleneck of the concept is the memory consumption [Gupta et al., 2013].

A different approach of recommending users to follow stems from the idea that users can be classified as 'information seekers', 'information sources' or 'friends'. Information seekers are users who follow many other users, but do not post themselves [Armentano et al., 2012]. Information sources are defined as knots of the network, which have many followers and follow less people themselves. Friends are defined as user who follow each other. Because most users in the system are information seekers, finding relevant sources is essential [Armentano et al., 2012]. The main idea of the authors is that the recommendation algorithm searches for recommended users in the vicinity of the target user, thus focusing on the topology of the network, which is build on the different relationships between each user. Once the users are found, they are weighted according to a set of rules, like the number of friends in common or the relationship between followers and people they follow [Armentano et al., 2012].

A more recent paper, with a content-based recommendation approach, was written by [H. Nidhi and Basava, 2017]. They applied two algorithms for text categorization, a noun-based detection algorithm and a naive-Bayes filtering, to obtain the content of the tweets. After that, recommendations with similar content were suggested to users and evaluated. Their experiments show that content-based recommendation systems are a feasible solution on social networks and text-based content.

2.2 Filter Bubbles

Recommendations from recommender systems are ubiquitous on the internet and have a huge influence on users. This influence is in many cases greater than recommendations from peers and experts, which underlines the importance to research filter bubbles and the influence of them on users [Senecal and Nantel, 2004].
Netflix reported in 2012, that 75% of the content that users watched came from recommendations. Information retrieval systems like recommenders are useful for users, because they provide personalized product offerings and lower the overall decision effort for them [Xiao and Benbasat, 2007].

As social beings, humans tend to form social relationships with similar, like-minded humans, a concept called homophiliy. The strongest factors for this are: Race, sex, age, religion and education. One can make observations in daily life and on social media that segregation and inequality emerge from this pattern [McPherson et al., 2001]. When political blogs were researched in the US election in 2004, the authors of a study found that most links on liberal and conservative blogs lead to pages within their separate communities and are rarely linked to sites of the other political spectrum [Adamic and Glance, 2005].
Even though the internet provides a vast amount of information, many users restrict themselves to content that they find agreeable, because it supports their beliefs. This phenomenon is called selective exposure or confirmation bias [Liao and Fu, 2013]. Selective exposure exists, because users experience a mental state called 'cognitive dissonance' when viewing content that opposes their current view regarding a specific topic. Since this effect causes mental discomfort, most people try to bypass it altogether. Therefore, they try to stay consistent with their previous viewpoints and avoid different and opposing viewpoints [Frey, 1986].
Many experts fear that selective exposure leads to social fragmentation of the internet, resulting in so called 'filter bubbles', a term which was first coined by [Pariser, 2011]. The consequences of a filter bubble are multi-faceted. Interaction with like-minded people leads to polarisation and users may get even more extreme opinions on a topic than in the beginning. Moreover, increased polarisation of the society

makes it harder to agree on solutions on important topics [Sunstein, 2002].
There are positive effects of escaping a filter bubble as well. Confrontation with diverse topics may lead to better decision-making and group problem-solving skills [Nemeth and Rogers, 1996]. Especially minorities have the natural tendency to think that their views are more common and widespread than they really are. Presenting people with the facts might lead them to more acceptance on topics where they disagree with [Sanders and Mullen, 1983].

2.3 Mitigating Filter Bubbles

Many theories in understanding filter bubbles exist. Some researchers find that users seek out items that comply with their existing viewpoint and avoid challenging content [Frey, 1986]. Other researchers dispute this theory and observe users that show diversity seeking behaviour. They found that these participants looked for challenging content and showed enjoyment in finding different opinions on a topic [Stromer-Galley, 2003]. [Munson and Resnick, 2010] studied the conflicting theories and came to the conclusion that both findings are correct. They merely describe the different preference and personality of people. Humans do not have a general trait that makes them diversity-seeking or challenge-aversive. However, people who seek a wide range of opinions appear to be in the minority.

Designing an information retrieval system that prevents filter bubbles and recommends diverse content is a challenging task. First, one must consider diversity-seeking and challenge-averse users when presenting information. Second, the interest of users in diverse content might fluctuate due to various factors like personality, knowledge and personal involvement [Fischer et al., 2011].
[Liao and Fu, 2013] researched two factors, (i) perceived threat and (ii) topical involvement, which might influence users. Perceived threat describes information seeking when tackling troubling situations, like making decisions concerning health, security or personal finance. Interestingly, people are often biased seeking information under these circumstances. On the other hand, when users are highly involved in a topic, they actively seek information to learn more about it, even if the topic is inconsistent with their views. The authors found that when presenting users with agreeable and challenging content side-by-side, they preferred the agreeable content.

News Cube is an internet news service that automatically creates multiple viewpoints on an headline of interest, trying to mitigate media bias [Park et al., 2009]. The service is split into three different functions: collection, classification and presentation. The collection service crawls news data and preprocesses it by filtering out unwanted content like advertisements, comments and meta-data. In order to classify the aspects, they used an unsupervised classification, since it is hard to develop and train pre-defined categories for news events. The extraction process extracted feature from the core parts of the article, focussing on the head, sub-head and lead part. Then, the keywords got weighted, based on the location in the text and the frequency of them. In the end, the authors surveyed a test audience, showing that presenting more perspectives on news articles can lead to more balanced views among user.

The goal of the balancer study was to research two points: (i) do some individual characteristics like demographics or political preferences predict the political bias in a users online reading behaviour? (ii) Does feedback about the bias to the user alter the behaviour of the user? In order to get answer to these questions, the study designed a browser widget that gave information to the user about the frequency of liberal and conservative pages visited in one week. The classification was simply done by classifying the URL of the visited web pages. The study found that such a browser widget influenced the behaviour of some users, who exhibited more interest in topics lying outside of their own viewpoint [Munson and Resnick, 2013].

Another study researched whether showing progress bars, which give an indication on the users particular position on an issue, to users had any effect on the reception and selection of attitude-challenging information. The study found that the indicator had no significant effect on challenge-averse participants. However, on users with information seeking motives, showing the bar decreased selective exposure. At the information reception stage, showing the bar helped participants to differentiate between moderately inconsistent views against extreme positions [Liao and Fu, 2014].

[Zhang et al., 2012] argues that focusing to much on accuracy when recommending items might generate boredom and ineffective recommendations. Furthermore, too much spotlight on personalisation might harm a users personal growth and experience. In their paper a recommender for artists, called 'Auralist', is introduced, which tries to balance the goals of accuracy, diversity, novelty and serendipity. They present three techniques for generating recommendations. (1) 'Artist-based LDA', which they describe as a recommender that uses Latent Dirichlet Allocation for computing features, (2) 'Listener Diversity' is combined with 'Artist-based LDA', to prioritise for Artists with very diverse communities. (3) 'Declustering' aims to take the existing clusters of a users history into account and recommend items outside of these clusters. Their studies show that their algoritm produces significantly more serendipitous recommendations while losing some accuracy. Nonetheless, most users in the evaluation study gave the more serendipitous recommendation algorithm a better satisfactory rating when reviewing the recommendation algorithm.

[Garimella et al., 2017] focused on controversial issues on social media and modelled re-tweets and shares of users on a graph, trying to bridge opposing views. The authors implemented an algorithm, experimented with Twitter datasets and showed, that the algorithm works efficiently. Their approach is different from ours, because they focus on who they recommend the content to instead of what content should be recommended.

The paper that we based our work on recommended tweets with similar and opposing views to the users. The goal was to present these tweets in a word cloud and hide the fact that some tweets from people with opposing views are in there as well. After obtaining vectors that describe the user stance regarding a particular topic, they computed the top n-topics that characterise a user by finding the most common n-grams in the users history. tweet recommendation happened then by recommending tweets based on the top n-grams and the stance regarding a particular subject. These recommendations are presented graphically in the form of a word cloud to mitigate the effect of cognitive dissonance. The case study showed that incorporating opposing views in an already known presentation like a word cloud reduced the effects of resistance against opposing views and led to a high overall enjoyment for the users [Graells-Garrido et al., 2013].

2.4 Summary

In this chapter, we have first introduced the different approaches of recommender systems, (i) collaborative filtering, (ii) content-based recommender and (iii) hybrid systems. Afterwards, we went into more detail about the content-based recommender and its advantages and disadvantages, as we use exactly this approach in our work. Next, we presented different approaches of recommender systems for Twitter. The algorithms build graphs that represent relationships between users and focus on the topology of the network with the goal of suggesting like-minded users.

Subsequently, we introduced the concept of the Filter Bubble, which has been extensively studied in many papers. We discussed which factors are relevant for the emergence of Filter Bubbles and why they emerge at all. We also refer to further papers that shed light on the negative effects of these filter bubbles.

Finally, we present different approaches to mitigate the formation and effects of filter bubbles. Not only the content proposed is relevant, but also how it is presented. Some papers also try to help the user to get a more balanced view on an issue by giving feedback on the bias, which was partly promising. Furthermore, it is also apparent that many user recommender systems score better if they lose some accuracy but show some unexpected content.

Our paper is based on the algorithm and approach proposed by [Graells-Garrido et al., 2013]. A major difference to the numerous papers we have discussed is that we focus on content to be proposed rather than on users to follow. Furthermore, we have extended the approach from the original paper, as we measure beyond-accuracy metrics such as diversity and serendipity and influence them by mixing tweets from different perspectives. By this, we hope to enhance these metrics, in order to generate a greater user satisfaction and extending the perspectives of a user, helping them to mitigate the effects of filter bubbles.

Chapter 3

Methodology

The first Section in this Chapter describes the dataset and the social media platform Twitter and its terminology. We explain how we have acquired the data from Twitter and the structure of the data. Further on, we describe the process of filtering non-relevant accounts in the dataset. Next, we go into detail how we preprocessed the tweets. In the end, we show the statistics of the cleaned dataset.
In the next Section, we describe the technical preliminaries for understanding this thesis. We explain concepts such as term frequency - inverse document frequency and cosine similarity.
In the following Section, we describe what our approach looked like, how we extracted the user stances and recommended the tweets to the users.
In the last Section, evaluation metrics like diversity, serendipity and topic similarity are described.

3.1 Dataset

3.1.1 Motivation

Twitter[1] is a social media network, where registered users can are able to post short messages called 'tweets'. tweets contain various types of content, ranging from simple text to videos or locations. Users on Twitter, who subscribe to other users, are called 'Followers' in Twitters terminology. If a user signs in at the Twitter website, all tweets of the followed accounts are shown on the individual main page,

[1] https://twitter.com

resulting in a mix of many different tweets. Twitter calls this individual main page 'timeline'.

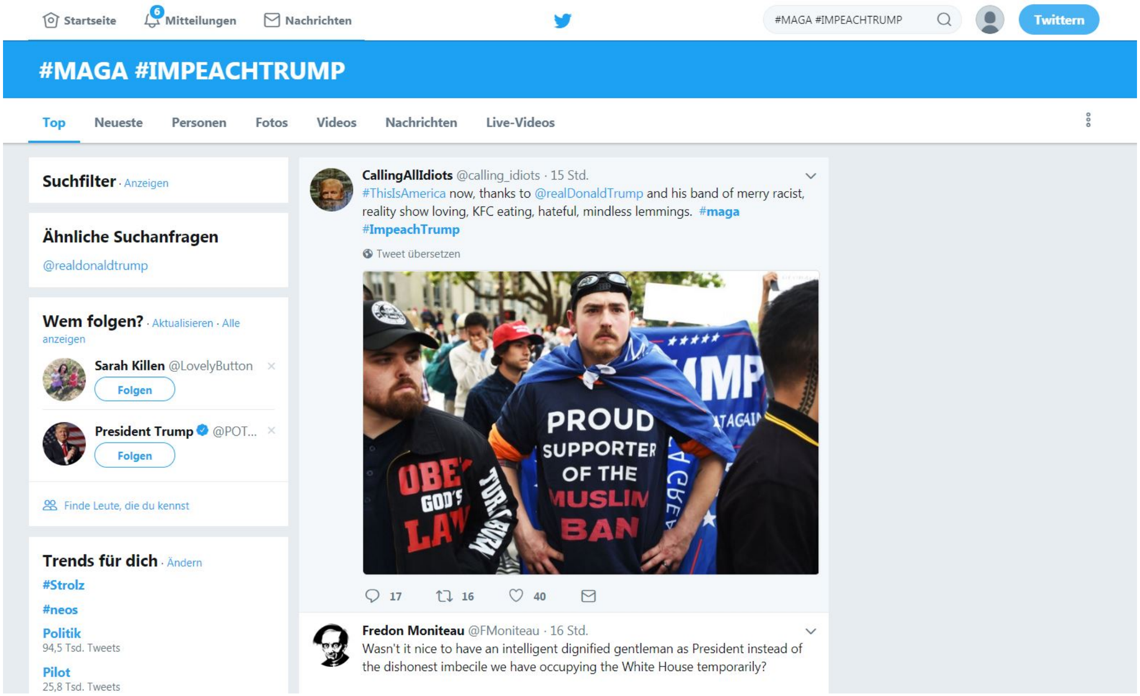

Figure 3.1: **Twitter Interface** - This Figure depicts the Twitter interface after a search for the hashtags *#maga* and *#impeachtrump*. On the right side, the latest and most popular tweets in the search results are shown. The top navigation bar allows to filter for different kinds of results. The left side shows the current filter criteria, similar search criteria and suggests who to follow.

Twitter provides us with the following information:

tweets tweets are texts which are limited to a certain character length (280 characters since 2017). Anything a user posts on Twitter is considered a tweet. tweets are made publicly available in the standard setting, meaning that even unregistered readers are able to read the tweets of accounts they choose to watch. tweets are composed of:

- Hashtags (indicated by an #-character)
- Links (URL)
- References to other Twitter profiles (indicated by an @-character)

- Images
- Locations

Registered users are able to react to tweets in different ways:

Likes Likes can be used to show appreciation for a tweet. If a user wants to like a tweet, she clicks on a heart-shaped icon depicted on the bottom of the tweet. Moreover, some users use this feature to 'bookmark' tweets.

Retweets tweets can reposted to a user's own timeline. Therefore, a retweet is a way of sharing information across the personal network.

Hashtags Hashtags are a type of metadata tag, which allow users to apply dynamic, user-generated tagging of tweets. It is defined by a prefix, the #-character and some text. Twitter hashtags permit grouping of tweets by facets and categories and can assist in providing different visual representations of tweets. Furthermore, users can search for a hashtag to retrieve all tagged tweets [Chang and Iyer, 2012]. Examples for hashtags, which are also used in this thesis, are: *#maga*, *#impeachtrump*, *#nobannowall*, *#trump*

Replies A reply is a comment to a tweet. When clicking on a tweet, a window pops up and the user can navigate through all replies to the tweet. Additionally, the user is able to reply to the tweet as well. Because a reply is also a tweet, each reply has all functionalities a standard tweet has, including the ability to reply to it.

Accounts Accounts are the profiles of the registered users. The registration enables the user to interact with other users and their tweets. The ability to crawl for accounts offers a lot of additional metadata, including information such as the user screen name, the user id, the language of the user, her number of followers and her location.

3.1.2 Dataset description

The dataset was crawled by Twitter using the standard Twitter API[2] in February of 2017. Our goal was to map the 2 different groups with 2 stances (i) pro-Trump and (ii) contra-Trump as closely as possible in the dataset. We used the following hashtags to get an initial sample of users and tweets for the two opposing stances:

- *#maga* - the motto "Make America Great Again" of the republican party in the election of 2016, which was used to acquire users for the pro-Trump stance.
- *#impeachtrump* - opposing groups of president Trump want to impeach him. This hashtag was used to acquire users for the contra-Trump stance.
- *#nobannowall* - a hashtag used to speak against president Trumps executive orders targeting immigrants, refugees and muslims in the beginning of the year 2017 [Silard, 2017], which was used to acquire users for the contra-Trump stance.

Characteristics of the Dataset

We have crawled 73,868 tweets in total, posted by 39,698 different accounts. The dataset statistics of the results are shown in Table 3.1. Initially, we used only two hashtags, *#maga* to get pro-Trump users and *#nobannowall* for contra-Trump users. However, after the initial crawling phase we found out, that the amount of users that posted tweets related to *#maga* far exceeded the amount of users, who posted content containing *#nobannowall*, in the same time period. We decided to add another hashtag, *#impeachtrump*, which we suspected to be related to contra-Trump content. We are aware that with this method only a small sample of all tweets regarding these topics are acquired.

[2] https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html

Total number of users	39,698
Total number of tweets	73,868
Number of tweets containing #maga	34,743
Number of tweets containing #nobannowall	17,423
Number of tweets containing #impeachtrump	21,702

Table 3.1: **Initial dataset statistics** - This Table provides a statistic of the initially crawled tweets.

Attributes of tweets Twitter provides a lot of meta data for each tweet. In order to save memory and disk space, we deliberately defined the following attributes to be relevant for this thesis:

- **id** - This is the integer representation of the unique identifier of each tweet.
- **created_at** - UTC time
- **text** - the content of the status update
- **text_cleaned** - this is the preprocessed text of the tweet.
- **user** - a complex object, containing various user-related metadata, such as:
  - **id** - the unique id of each user
  - **name** - the name of the user
  - **screen_name** - the name that is actually displayed in the Twitter interface
  - **verified** - whether the user is verified or not
  - **followers_count** - the amount of followers this user has
  - **friends_count** the amount of friends this user has
  - **statuses_count** - the amount of tweets this user has posted
  - **lang** - the language of the user
- **entities** - entities which have been parsed out of the text automatically by the Twitter API, like hashtags, URLs and images.
- **lang** - the language of the tweet, automatically detected by Twitter.
- **user_stance** - one of the two stances (i) pro-Trump and (ii) contra-Trump, which we evaluated in this thesis.

3.1.3 Preprocessing and Statistics

In the following Subsection the preprocessing steps for accounts are explained and justified. After removing all statistical outliers and non-English speaking accounts from the initial dataset, we downloaded the most recent 1,100 tweets of each account. We could not download more because 1,100 tweets is the limit of the free version of the Twitter API. Since tweets contain a lot of unnecessary data like emoticons and stop words, we preprocessed the texts of the users in the next step. In the end of this Subsection, the final dataset statistics are presented.

Preprocessing of Accounts

Once the initial dataset was crawled, statistical outliers were found and removed with the help of boxplot diagrams. Boxplot diagrams can be used to show variation in samples of a statistical population without making any assumptions of the underlying distribution. Outliers lie far from the majority of the other data points in a distribution of variables. By filtering these outliers, bots and managed accounts will be removed from the dataset. These accounts are causing biases in the acquired data, therefore their removal is essential before applying any algorithms on the data [Kwak and Kim, 2017]. Boxplot diagrams are generated for the following statistics:

- Number of tweets an account has posted, see Figure 3.2
- Number of favorites an account has, see Figure 3.2
- Number of followers an account has, see Figure 3.3
- Number of friends an account has, see Figure 3.3

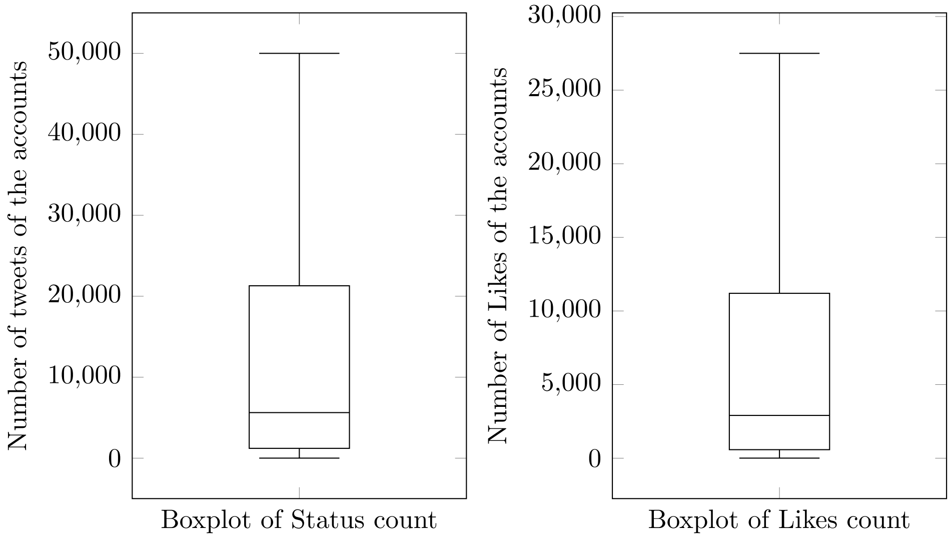

Figure 3.2: **Boxplot diagrams for number of tweets and counts of likes - Left:** This Figure depicts a boxplot diagram showing the number of status an account has posted. The median is 5,627.5, the upper quartile is 21,292.5, the lower quartile is 1,212. **Right:** This Figure depicts a boxplot diagram showing the number of tweets an account has liked. The median is 2,906.5, the upper quartile is 11,200.5, the lower quartile is 579.

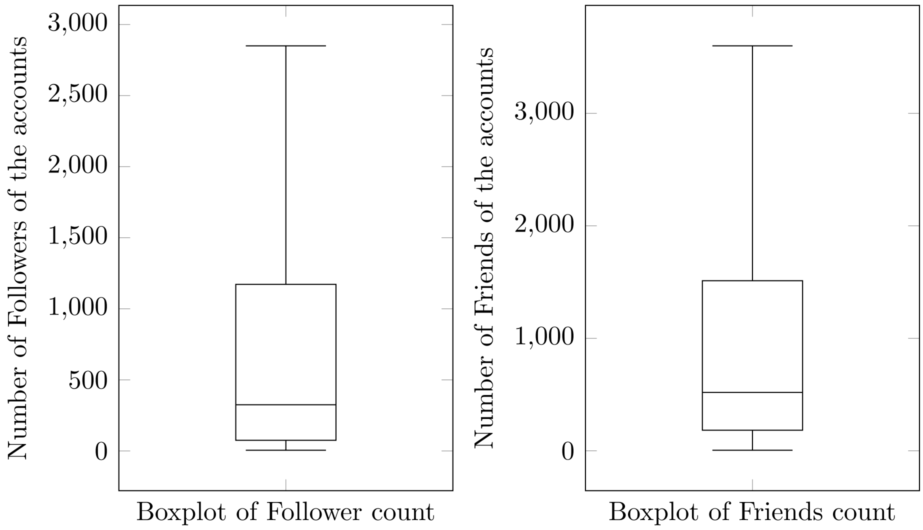

Figure 3.3: **Boxplot diagrams for number of followers and count of friends** - **Left:** This Figure depicts a boxplot diagram showing the number of followers the accounts have. The median is: 325, the upper quartile is: 1,172, the lower quartile is: 75. **Right:** This Figure depicts a boxplot diagram showing the number of friends an account has. The median is 519, the upper quartile is 1,512, the lower quartile is 183.

We accpted all accounts that were between the first and third quartile. This left us with 6,913 accounts.

After removing all non-English accounts, we consider 5,672 accounts for further evaluation, see Table 3.2.

Total number of users after crawling	39,698
Total number of users after trimming of outliers	6,913
Total number of users after trimming of non-english users	5,672

Table 3.2: **Preprocessed dataset** - This Table depicts the amount of users after each preprocessing step. We removed a total of 34,026 accounts from the dataset. This results in 5,672 accounts.

Preprocessing of tweets

For the trimmed number of accounts we downloaded the latest 1,100 tweets. We chose the number slightly below the lower quartile of 1,212 tweets, which each user in the resulting dataset has posted, according to the boxplot presented in Section 3.2. However, it is more than enough to get a good indication of the political stance of the user. We had to remove about 500 accounts from the final dataset, because the Twitter API returned a privacy error when crawling for the individual user's tweet history. The statistics of the final dataset are shown in Table 3.3.

Total number of accounts	5,172
Total number of tweets	6,468,035
Total number of pro-Trump accounts	2,150
Total number of pro-Trump tweets	2,615,140
Total number of contra-Trump accounts	3,522
Total number of contra-Trump tweets	3,852,895

Table 3.3: **Statistics of the final dataset** - This Table shows the final statistics of the dataset. We aquired more contra-Trump accounts than pro-Trump accounts. The total number of tweets under consideration is 6,468,035.

At first, we decided to clean the original data from Twitter, in order to achieve better insights and prepare it for further analysis [Batrinca and Treleaven, 2015]. While doing that, we removed capital letters, got rid of all kinds of punctuation, performed tokenization as well as stop word removal using the Python NLTK framework[3]. The most common hashtags in the resulting dataset for contra-Trump stances and pro-Trump stances are shown in Table 3.4. Even though we had trouble finding the same amount of users posting the #impeachtrump hashtag compared to the #maga hashtag, which was more common among different users, the dataset shows that individually, users who belong to the contra-trump stance, use the #impeachtrump more frequently than the pro-trump group the #maga hashtag.

[3]https://github.com/nltk/nltk

Contra-Trump		Pro-Trump	
Hashtag	count	Hashtag	count
#impeachtrump	21,552	#maga	17,088
#theresistance	8,394	#trump	2,206
#nobannowall	4,615	#tcot	1,419
#trumprussia	3,117	#americafirst	1,220
#russiagate	2,704	#trumptrain	1,040
#Trump	1,799	#presidenttrump	829
#impeach45	1,655	#draintheswamp	728
#trumpleaks	1,616	#fakenews	691
#nottheenemy	1,354	#trumpimpeachmentparty	689

Table 3.4: **Hashtags in the final dataset** - This Table shows the most frequent hashtags in the resulting dataset. On the left side, the most frequent contra-Trump hashtags and on the right side the most frequent pro-Trump hashtags are shown.

3.2 Technical Preliminaries

This Section explains the theoretical concepts used in the experimental part of the thesis. The most important concepts like term frequency (tf), inverse document frequency (idf), term frequency - inverse document frequency (tf-idf) and cosine similarity are introduced.

Vector Space Model (VSM) Each document is expressed as a vector in a multi-dimensional space. In this space called VSM, each dimension represents a term that is part of the entire collection of documents [Salton et al., 1975]. Every document is a vector represented by term-weights, where each weight states how strong a term and a document correlate. Suppose we have a set of documents $D = \{d_1, d_2, ..., d_N\}$. All terms in a document collection are represented as $T = \{t_1, t_2, ..., t_n\}$. Each document is then represented as a vector $d_j = \{w_{1j}, w_{2j}, ..., d_{nk}\}$ where w_{kj} is the weight for each term t_k in document d_j. Tf-idf is a common way to weight these terms in the document vector d_j [Salton et al., 1975].

Term Frequency The value of term frequency (tf) grows with the frequency of words in the document. For instance, if a corpus of documents includes six documents, each containing the word 'tree' twice , the term frequency of the word 'tree' equals to 12. The challenge is that the words appearing most frequently in documents are often not the most important ones (such as 'the' or 'and') and need to be removed before classification. Terms like these are called stop words [Rajaraman and Ullman, 2011].

The term frequency is computed as follows. Suppose we have a collection of N documents in total. F_{td} is the frequency of word t in document d, $max_k f_{kd}$ is the maximum occurrence of any word in document d. For instance, the tf of the term with the highest frequency in document d equals to 1. With this, the term frequency is normalized, which helps when dealing with documents with various lengths [Rajaraman and Ullman, 2011].

$$tf_{td} = \frac{f_{td}}{max_k f_{kd}} \tag{3.1}$$

Inverse Document Frequency Inverse document frequency (idf), decreases the importance of the term proportionally to the frequency of the term in the corpus. A corpus is defined as a large and structured set of documents. With this value, words that occur more often in the corpus are weighted less heavily, relative to words which appear only in a few documents making them more important for a specific document. Suppose we have M documents in the corpus where the term x appears n_x-times, then *idf* is calculated with [Baeza-Yates and Ribeiro, 1999].

$$idf_x = log_2(M/n_x) \tag{3.2}$$

Term Frequency - Inverse Document Frequency Term frequency - inverse document frequency(tf-idf) indicates how important a word in a document is. Words that occur very frequently in the corpus of documents are given less weight [Baeza-Yates and Ribeiro, 1999]. When attempting to classify documents to a certain topic, special words can be found that characterize the text about that topic. This can be done by analysing the documents with tf-idf and weighting these special words for each document. One advantage of tf-idf is that the metric is easy to compute. One of the major disadvantages is that the location of the terms in the text is not considered at all in the tf-idf [Ricci et al., 2011].

The tf-idf score evaluated as follows:

$$w_{td} = tf_{td} \cdot idf_t \tag{3.3}$$

A high score is reached by a term which occurs very frequently in document d, but has a low frequency in the whole collection of documents.

Cosine Similiarty Cosine similarity measures the similarity of two vectors. In order to gauge how similar they are, the cosine angle between the two vectors is calculated. This measure is used for analysing the orientation of a vector compared to another vector. When two vectors have the same orientation (when they are parallel), the cosine similarity equals to 1, indicating that they are similar. On the other hand, when the angle between the vectors is 90° (orthogonal), the cosine similarity evaluates to 0. The outcome is usually bounded in the positive space [0,1] [Ricci et al., 2011].
The cosine similarity between two documents d_j and q can be calculated as [Singhal, 2001]:

$$sim(d_j, q) = \frac{d_j \cdot q}{||d_j|| \cdot ||q||} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i_q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \tag{3.4}$$

Cosine similarity is used in conjunction with tf-idf in this thesis for classifying users into one of the two user stances (i) pro-Trump and (ii) contra-Trump. Tf-idf is a weighting schema for vectors, where the value of each dimension corresponds to the tf-idf values for the respective terms. These vectors can be used to calculate the pairwise cosine similarities and thus indicating, how much the documents correlate with each other [Singhal, 2001]. Furthermore, users get their individual tweet recommendation based on cosine similarities, calculated with taking their history of tweets into account.

3.3 Approach

In this Section, we describe the general approach used in this thesis. First, we cover how we extracted user stances, which we used to classify users into one of the two issue stances. Figure 3.4 gives a good overview of the process. Next, we explain how we recommended tweets to a specific account. In the end, some performance metrics like diversity, topic similarity and serendipity are explained.

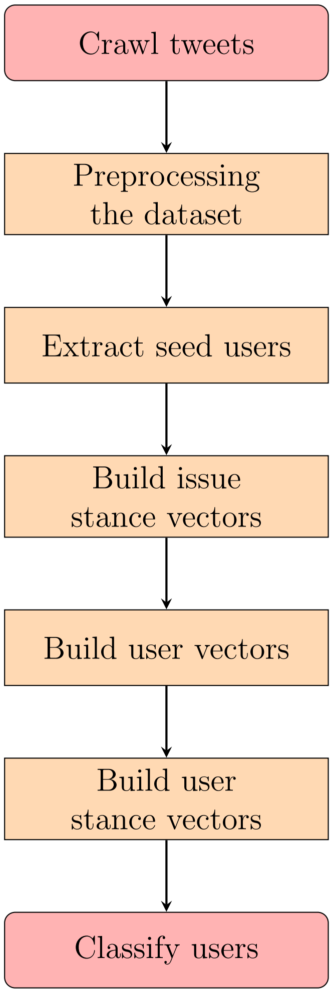

Figure 3.4: **User Stance Flowchart** - The flowchart shows the process for creating the user stance vectors.

3.3.1 Extraction of User Stances

Extraction of seed users After we crawled and performed preprocessing on the dataset, we extracted seed users from it. Seed users are users that have hashtags distinctly from one stance of a sensitive issue and no hashtags from another stance. By defining more than 10 hashtags for each stance we make sure that the user has a high probability of belonging to the assigned stance, following the approach that [Graells-Garrido et al., 2013] have used.

We selected the hashtags for the classification manually with a tool called 'hashtag-analytics'[4]. Hashtag-Analytics enables us to see which hashtags are connected and commonly used together - for example, *#americafirst* is often used in conjunction with *#buildthewall*, *#potus* and *#maga*, thus indicating a strong connection to pro-Trump stances. After gaining insight on the meaning of the hashtags, we inspected the selected hashtags by using this tool. Only hashtags that showed a word cloud which was consistent with the corresponding stance were selected. For example, the hashtag #impeachtrump has a high correlation with other hashtags associated with stances against president Trump. Taking this into account, we concluded, that this hashtag can be taken as an indicator for a contra-Trump user. Since some hashtags are used by both parties and their followers, taking into account several hashtags, which highly correlate to one of the stances, greatly improves the chances that the political view of the user aligns with the stance that we assigned. Fig. 3.5 shows a screen-shot of the tool, after searching for the term *#americafirst*. A Table with the hashtags, which we used to classify the seed users, is shown in Table 3.5 [Lex E., 2018].

Because the criteria for selecting seed users strict, we found 290 seed users for the pro-Trump stance and 237 seed users for the contra-Trump stance, out of the 5,672 users in total. Nonetheless, downloading more than 1,000 tweets for each of these users left us with a dataset large enough to create the corresponding stance vectors.

issue stance	hashtags used for seed users
pro-Trump	*'maga','tcot', 'americafirst', 'trumptrain', 'presidenttrump', 'draintheswamp', 'fakenews', 'potus','buildthewall', 'presidentelecttrump'*
contra-Trump	*'impeachtrump', 'theresistance', 'nobannowall', 'resist', 'trumprussia', 'impeach45', 'nottheenemy', 'resistance', notmypresident', 'iamamuslimtoo', 'nobannowallnoraids', 'fakepresident', 'dumptrump', 'trumplies'*

Table 3.5: **Hashtags used for classifying seed users** - This Table shows hashtags, which we used to classify seed users. Seed users for the pro-Trump stance have strictly hashtags from the pro-Trump Section in their tweets and no hashtags of the contra-Trump Section and vice versa.

[4]http://keyhole.co/hashtag-analytics

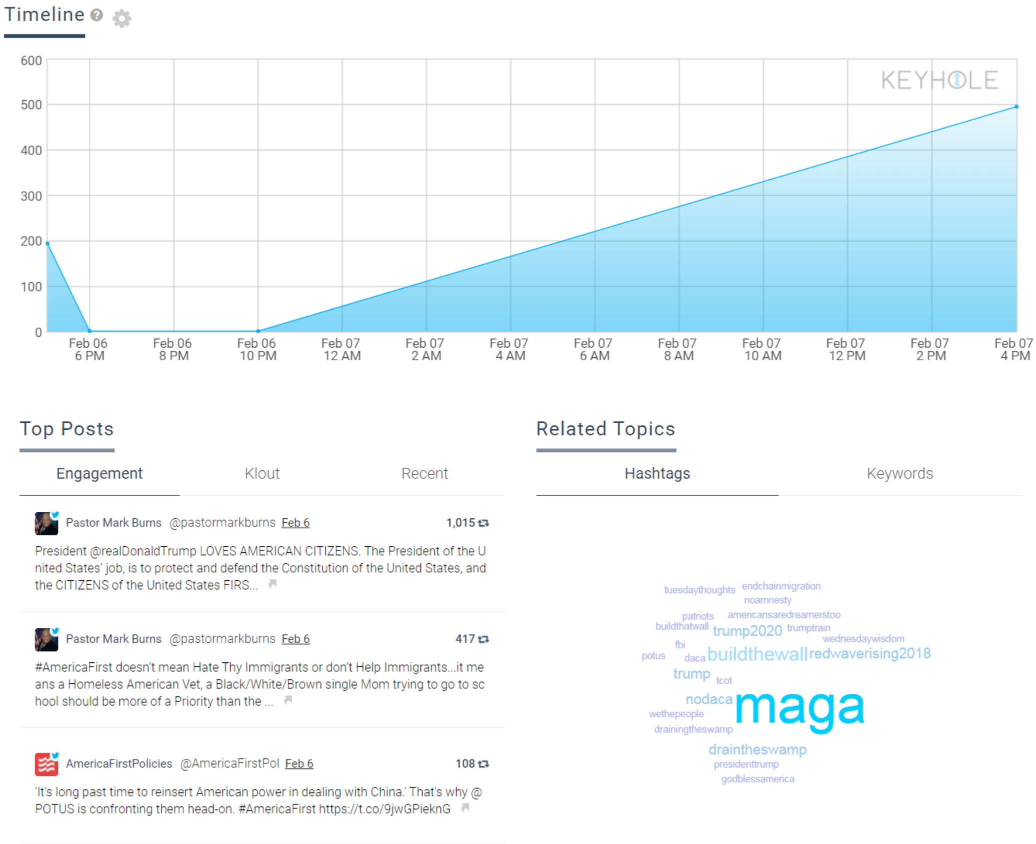

Figure 3.5: **Hashtag analytics** - This Figure shows an example screenshot for the hashtag analytics tool. The screen depicts a search for *#americafirst*. The timeline shows the number of tweets associated with the hashtag per day. On the left side, the top posts are shown. On the right side, a wordcloud together with related hashtags is depicted.

Extraction of issue stances Next we created issue stances with the help of the latest 1,100 tweets of each of the seed users. Since we made sure to pick only users, which strongly belong to a stance, we have a good basis for our issue stance [Graells-Garrido et al., 2013]. For each of the two stances, all tweets of the corresponding seed users were concatenated, leaving us with 2 documents. We computed tf-idf for these 2 documents, the outcome was the so-called **issue vectors**. In the issue vectors, each dimension refers to the importance of the word for each specific stance [Graells-Garrido et al., 2013]. The feature space used was the same as the one we used for the user vectors, in order to calculate cosine similarity for them.

Extraction of user vectors Furthermore, we defined a user vector, similar to the issue stance vectors. We concatenated all tweets of all users minus the seed users to documents and calculated TF-IDF for them. Each dimension in the user vector reflects the significance of a word of the tweets from the users [Graells-Garrido et al., 2013].

Generation of user stance vectors Each dimension in the user stance vector correlates to the result of the cosine similarity between the user vector and the issue stance vectors [Graells-Garrido et al., 2013]. This allows us to classify users into one of the issue stances (i) pro-Trump or (ii) contra-Trump. The magnitude of the vector demonstrates the opinion of the user regarding these stances. In total, we identified the following amount of users and tweets:

- 2,150 pro-Trump users with 2,615,140 tweets
- 3,522 contra-Trump users with 3,852,895 tweets

Having the user stance vectors enables us to recommend tweets of users with different opinions regarding the topic.

3.3.2 Recommending User tweets

To gain more knowledge about a user, we use the 15 most frequent trigrams from her tweets. These tweets are considered the basis for the recommendations. We decided to use trigrams in order to gain some semantic context. The preferences are the baseline for recommending tweets to the user. An example for the user 'FxgFx' is shown in Table 3.6.

Username	#-tweets	Issue stance	User preferences
FxgFx	1,100	contra-Trump	'trump say mexico', 'mexico would pay', 'pay wall meant', "trump's tax returns", 'showing true face', '#trumprussia #russiagate #resist' 'rep devin nunes', 'health care plan', 'house oversight committee', "can't wait til", 'get new orders', 'defund planned parenthood', 'make health insurance', 'bring candles back', 'gop members congress'

Table 3.6: **User preferences for user 'FxgFx'** - This Table shows the most common trigrams for the user 'FxgFx'. We have analysed the 1,100 latest tweets of her and have already classified her as contra-Trump user. The colum 'User preferences' lists the 15 most common trigrams of the user, which we used to recommend tweets.

We used Apache Solr's MoreLikeThis feaure in order to recommend the tweets to the user [5], which is a content-based recommender engine.
We used a random function to create a set of 100,000 tweets. The tweets of the target user were excluded. The set was created out of all the tweets that we crawled. We did this for each recommendation separately. After that, we queried SOLR with the MoreLikeThis functionality for the top 15 trigrams of the user. We then took the recommendations that SOLR returned, filtered them for their user stances, taking into account their priority and assigned them to several recommendation variants. The goal of the variants is to measure how the composition of the first 10 recommendations influences our metrics.

[5] http://lucene.apache.org/solr/

Variant Nr.	Description
1	The top 10 recommendations
2	The top 10 recommendations from contra-Trump users
3	The top 10 recommendations from pro-Trump users
4	Variant with 1 contra-Trump tweet and 9 pro-Trump tweets
5	Variant with 2 contra-Trump tweets and 8 pro-Trump tweets
6	Variant with 3 contra-Trump tweets and 7 pro-Trump tweets
7	Variant with 4 contra-Trump tweets and 6 pro-Trump tweets
8	Variant with 5 contra-Trump tweets and 5 pro-Trump tweets
9	Variant with 6 contra-Trump tweets and 4 pro-Trump tweets
10	Variant with 7 contra-Trump tweets and 3 pro-Trump tweets
11	Variant with 8 contra-Trump tweets and 2 pro-Trump tweets
12	Variant with 9 contra-Trump tweets and 1 pro-Trump tweet

Table 3.7: **Recommendation variants** - This Table shows the recommendation variants, which we used for the evaluation. The size of the variant is always 10, however, the amount of contra-Trump and pro-Trump tweets in the variant varies.

After creating these recommendation variants we performed further analysis on them, as explained in the following subsections.

3.3.3 Topic Similarity

In order to gain a more in depth understanding about the diversity of our dataset,especially between the two stances, the average topic similarity per user stance was defined and calculated.
Since we work with only two issue stances and we want to measure diversity, we are also interested in how diverse the topics within the stances are and whether there is an influence between topic similarity and diversity. It is reasonable to assume that stances that talk about many different topics have a low affinity to the formation of filter bubbles and vice versa. This helps us to make more detailed conclusions about our results. For the comparison we use tf-idf and the cosine similarity, as with the previous metrics.

Suppose S is the set of pro-Trump users and D_u is the document containing all tweets of user u, then the topic similarity for the pro-Trump stance is defined as follows:

$$TopicSimilarity = Avg(\sum_{u \epsilon S} \sum_{j \epsilon S, j>u} CosSim(D_u, D_j)) \tag{3.5}$$

3.3.4 Botdetection with Botometer

In this thesis, a tool called 'Botometer' is used in order to measure the amount of accounts that are not controlled by humans in the dataset. A significant amount of bot accounts in the dataset can lead to unwanted distortions in the result of our metrics. Even after preprocessing and filtering out statistical outliers, as explained in Section 3.1.3, we could still encounter bots in our dataset. Botometer is a public service, which leverages more than a thousand features. It uses these features to calculate a score, which indicates whether an account is a bot or not [Davis et al., 2016].

Botometer can be used by passing a user screen name to an API or to a user interface. The service has grouped it's features into main categories: **Network** features are based on retweets, hashtags and mentions. **User** related features contain metadata of the account. Social contacts are considered under the category **Friends**. Another feature group is called **Temporal**, which capture content related timing patterns. **Content** features use part-of-speech tagging to get more insight on the content. The last category is called **sentiment**, which groups sentiment analysis algorithms.

The purpose of evaluating these different categories of features is to evaluate the quality of our dataset. Having too many bots in our dataset distorts the quality of the several metrics, which we calculated before. Therefore, we provide bot scores for the dataset.

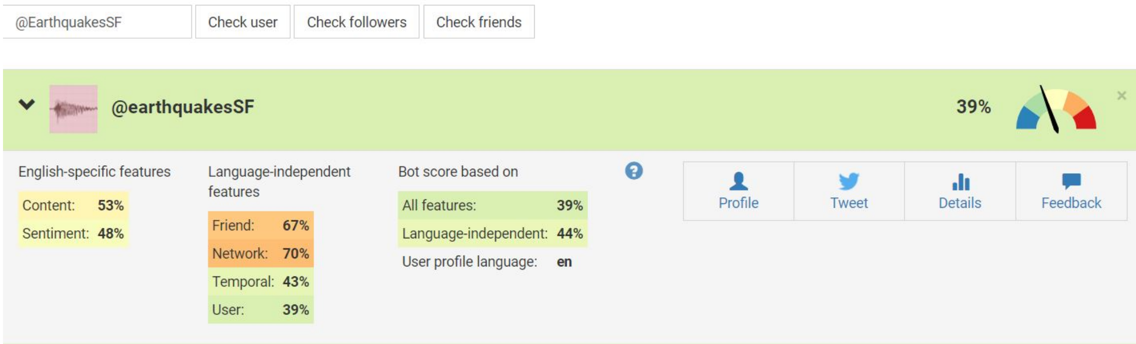

Figure 3.6: **Botometer Interface** - This Figure depicts the Botometer interface after a search for the user screen name 'earthquakesSF'. The 6 categories of features are shown including the calculated values of botometer regarding each feature group.

3.4 Evaluation Metrics

This Section talks about the evaluation metrics which are used in this thesis. In order to evaluate the quality of the recommended tweets with consideration of the user stances, we chose the metrics (i) diversity, (ii) serendipity, (iii) topical diversity and (iv) the botometer score.

3.4.1 Diversity

Diversity is usually defined as the opposite of similarity. Traditionally, accuracy metrics are used to measure recommendation quality, but there is a growing argument that other factors than accuracy also influence recommendation quality [McNee et al., 2006]. A recommendation set for music, which recommends only songs of one artist for instance, won't give a user the opportunity to explore different kinds of music from different artists. In the case of this thesis, the effects of different views on a sensitive topic and its impact on the diversity of the recommended content is explored. The most common method for measuring diversity uses item-item similarity, typically based on item content [Ricci et al., 2011]. We use intra-list similarity metric, as defined by [Ziegler et al., 2005]. The score is high, if a given set has a lot of items that are similar. Conversely, we get a low value if we have very dissimilar items. In order to calculate the metric, first we have to calculate all pairwise cosine similarities of the items. Then we have to calculate the average of these values

[Lex E., 2018].
Suppose S is the set of all users and R_u gives the top-10 recommended items for user u, then intra-list similarity is calculated with:

$$IntraListSimilarity = \frac{1}{|S|} \sum_{u \epsilon S} \sum_{i,j \epsilon R_{u,j<u}} CosSim(i,j) \tag{3.6}$$

$$Diversity = 1 - IntraListSimilarity \tag{3.7}$$

3.4.2 Serendipity

Serendipity measures the degree of surprise a content brings to a user The goal of a serendipitous recommender is that users find new topics and explore new content of the system, leading to greater recommendation satisfaction [Zhang et al., 2012]. We need to differentiate between novel content and serendipitous content. For example, if a user has watched a lot of episodes of a TV show and a new episode of the same TV show is recommended, the content will be novel but might not be serendipitous. Depending on the goal of the recommendation system, a balance between serendipity and accuracy might be considered. A random recommendation for an episode of any TV show might be more serendipitous, but not very accurate [Ricci et al., 2011].
Suppose there are 10 recommendations in the recommendation variant for each user. S again is the set of users, H_u is the history of user u and R_u is the recommendation variant of user u. The metric is defined as follows [Ricci et al., 2011]:

$$Familiarity = \sum_{u \epsilon S} \frac{1}{|S||H_u|} \sum_{h \epsilon H_u} \sum_{i \epsilon R_u} \frac{CosSim(i,h)}{10} \tag{3.8}$$

$$Serendipity = 1 - Familiarity \tag{3.9}$$

Chapter 4

Experiments and Results

This Chapter explains the experiments and results of this master thesis.
First, we evaluate serendipity and diversity for each user individually by taking 4 different recommendation sets into account. The (i) standard recommendation set, which consists of the first 10 recommendations found by the recommender. The (ii) Contra-Trump and (iii) Pro-Trump sets consists of only Contra or Pro-Trump tweets respectively. The (iv) hybrid recommendation consists of 5 pro-Trump tweets and 5 contra-Trump tweets. We then take 1,500 pro-Trump and contra-Trump users and calculate the average for each of the two stances. At the end of this Section, we introduce two sample users, one for each stance and present the most important metrics and findings in order to underline our quantitative findings.

4.1 Quantitative Results

The quantitative experiments we evaluated are designed to show how much recommending different sets of tweets of the same or opposing views to a user influences diversity and serendipity measures. We computed average metrics for 1,500 users of each stance. Furthermore, we also show that the results are statistically significant.

4.1.1 Topical Similarity

We calculated the average topical similarity per user stance. We noticed, that there are significant differences between the contra-Trump group and the pro-Trump group. The first has an average topic similarity of **44.6%**. The latter has a average topic similarity of **27.7%**. In other words, contra-Trump accounts in our sample talk about very similar topics, whereas pro-Trump accounts talk about a wider range of topics.

4.1.2 Average Contra-Trump Diversity

The average diversity results for the 1,500 randomly selected contra-Trump users are given in Table 4.1.

Recommendation variant	Diversity
Standard	.4516*
Contra-Trump	.4937*
Pro-Trump	.7369*
Hybrid	.7081*

Table 4.1: **Contra-Trump evaluation results for diversity** - This Table shows the average diversity calculated for 1,500 contra-Trump users. The hybrid set consists of half pro-Trump and half contra-Trump tweets. Interestingly, the Pro-Trump recommendation set produces a higher diversity than the hybrid set. Based on a t-test, the symbol *($\alpha = .05$) indicate statistically significant differences between the recommendation variants.

As expected, with respect to diversity, the lowest result is achieved with the standard recommendation set. Since the standard set consists of the 'best' recommendations regarding the content, it makes sense that the topics are similar to the topics that the user has already discussed. Likewise, the contra-Trump set diversity is low as well. However, the pro-Trump recommendation set achieves a higher diversity than the hybrid set. This is rather surprising to us. We suspect that this might be due to the higher topic similarity of the contra-Trump group, as shown in Section 4.1.1. We conclude, that if a group exhibits high average topic similarity and we mix the tweets of the group into the recommendation set, the overall diversity of the set becomes lower. Therefore, when mixing tweets into the recommendation set, we must account for the topic diversity of the group. In this case, if we want to achieve higher diversity results for contra-Trump users, we need to mix more of the diverse pro-Trump tweets into the set instead of more contra-Trump tweets, which exhibit higher topic similarity.
Additionally, we calculated different pro-Trump and contra-Trump recommendations in the hybrid set, as shown in Figure 4.1.

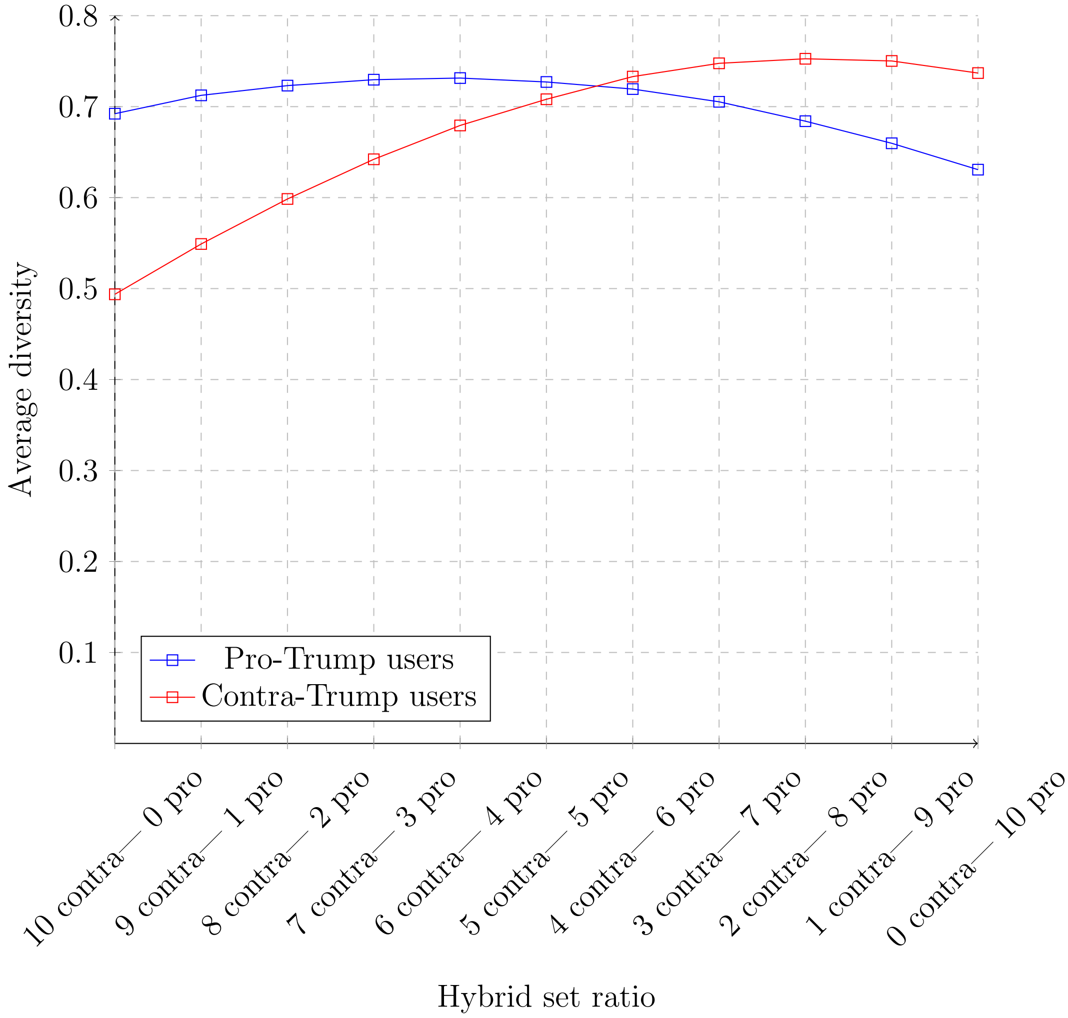

Figure 4.1: **Contra-Trump/Pro-Trump evaluation results for diversity with different hybrid set ratios** - This Figure shows the average diversity calculated for 1,500 contra-Trump and pro-Trump users for different ratios in the hybrid recommendation set.

As depicted in Fig.4.1, the best diversity values for contra-Trump users can be accomplished by mixing 8 pro-Trump tweets and 2 contra-Trump tweets into the recommendation set.
With reference to our research question (**RQ 1**) we receive the following answer: By including opposing tweets regarding a view stance in the user's recommendation set, we have shown that diversity can be significantly influenced, as shown in Fig.4.1. However, depending on the stance, the number of these conflicting tweets you add makes a difference in how diversity is affected. We will try to explain why this

difference arises in the following sections.

4.1.3 Average Pro-Trump Diversity

The average diversity results measured for the 1,500 randomly selected pro-Trump users are shown in Table 4.2.

Recommendation variant	Diversity
Standard	.5552*
Contra-Trump	.6922*
Pro-Trump	.6307*
Hybrid	.7271*

Table 4.2: **Pro-Trump evaluation results for diversity** - This Table shows the average diversity calculated for 1,500 pro-Trump users. The hybrid set consists of half pro-Trump and half contra-Trump tweets. Based on a t-test, the symbol *($\alpha = .05$) indicate statistically significant differences between the recommendation variants.

As expected, the best results with respect to diversity are achieved by the hybrid set, which consists of 5 pro-Trump and 5 contra-Trump tweets. The standard set achieves the lowest diversity results.
Additionally, we calculated different pro/contra-Trump recommendations in the hybrid variant, as shown in Figure 4.1.
Unlike the recommendation set for contra-trump users, we get the highest diversity value with a composition of 6 contra-trump tweets and 4 pro-trump tweets. The curve in Fig. starts with a relatively high diversity and drops off after a short increase by adding opposite tweets.

4.1.4 Average Contra-Trump Serendipity

The average serendipity results for the 1,500 randomly selected contra-Trump users are shown in Table 4.3.

Recommendation variant	Serendipity
Standard	.9229*
Contra-Trump	.9251*
Pro-Trump	.9571*
Hybrid	.9372*

Table 4.3: **Contra-Trump evaluation results for serendipity** - This Table shows the average serendipity calculated for 1,500 contra-Trump users. The hybrid set consists of half pro-Trump and half contra-Trump tweets. The highest score regarding serendipity is achieved with the hybrid set. Based on a t-test, the symbol *($\alpha = .05$) indicate statistically significant differences between the recommendation variants.

The highest serendipity scores are achieved by recommending the pro-Trump recommendation set to the contra-Trump user. In other words, recommending tweets of the opposing view increases serendipity in our setting. A mixture of pro-Trump and contra-Trump tweets has lower serendipity than the hybrid set.
Figure 4.2 shows the linear relation for different ratios of contra-Trump to pro-Trump tweets in the hybrid recommendation set. The least surprising recommendations are the tweets from the same view stance, the most surprising tweets are the tweets from the opposing view stance.

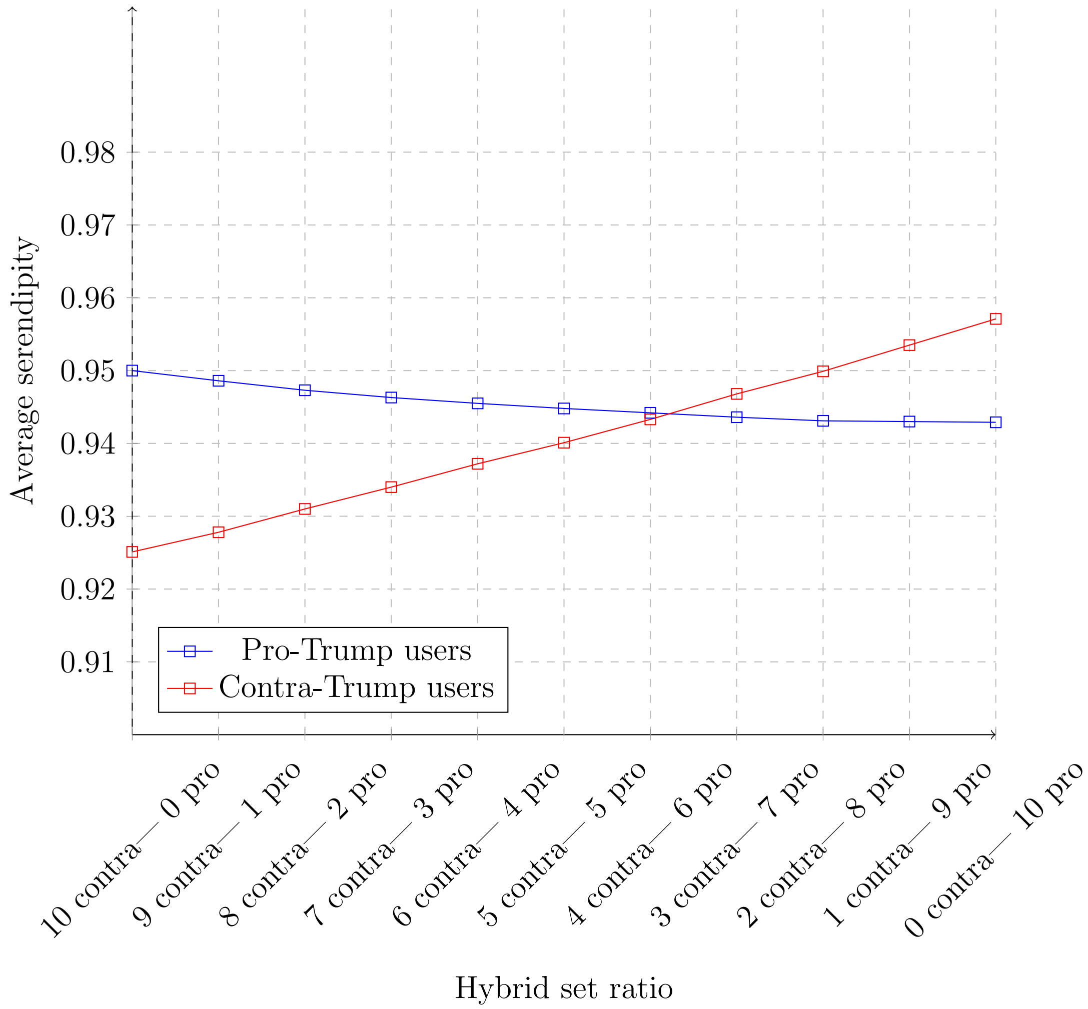

Figure 4.2: **Contra-Trump/Pro-Trump evaluation results for serendipity with different hybrid set ratios** - This Figure shows the average serendipity calculated for 1,500 contra-Trump and pro-Trump users for different ratios in the hybrid recommendation set.

4.1.5 Average Pro-Trump Serendipity

Similar to the contra-Trump serendipity calculation, we calculated the average serendipity results for the 1,500 randomly selected pro-Trump users. The results are shown in Table 4.4. The general trend of the serendipity results is similar to the contra-Trump results.

Recommendation variant	Serendipity.
Standard	.9332*
Contra-Trump	.9500*
Pro-Trump	.9429*
Hybrid	.9455*

Table 4.4: **Pro-Trump evaluation results for serendipity** - This Table shows the average diversity calculated for 1,500 pro-Trump users. The hybrid set consists of half pro-Trump and half contra-Trump tweets. Based on a t-test, the symbol *($\alpha = .05$) indicate statistically significant differences between the recommendation variants.

Figure 4.2 shows the evaluation results for serendipity with different hybrid set ratios.

4.1.6 Bot-Score

We computed the average values for bot scores in our sample set and depicted the result in Fig.4.3. Scores close to 0% indicate that the account is a human, scores close to 100% indicate that the account is a bot. The scores indicate that pro-Trump and contra-Trump users have no bias towards bot accounts, as expected by manually removing outliers while preprocessing the dataset.

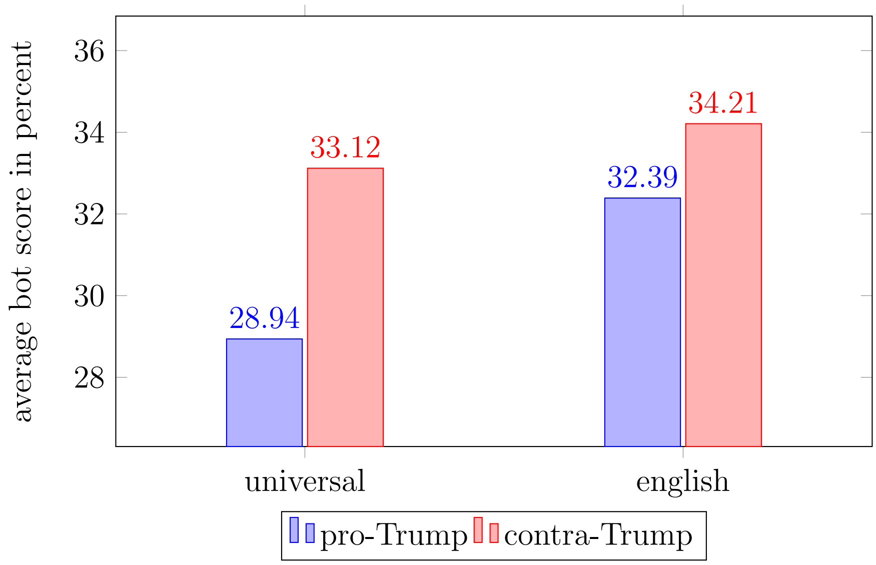

Figure 4.3: **Bot-Score values for pro-Trump and contra-Trump users** - This Figure shows the average bot score for 1,853 pro-Trump users and 3,327 contra-Trump users. Scores based on english features and scores based on language-independent features named 'universal' are depicted separately in this Figure.

4.1.7 Significance

Since we want to make sure that our results have statistical significance and are not caused by a sampling error, we calculated the Wilcoxon Rank-Sum non-parametric test for our results [Neuhäuser, 2011]. For each of the 1,500 recommendation sets for either pro-Trump or contra-Trump users, we can test the following:

Null hypothesis H0 If you randomly observe samples from populations and the populations of each sample have the same medians.
We use a $p - value <= 0.05$ as a strong indicator against the null-hypothesis.

Diversity P-Values

Tables 4.5 and 4.6 depict the p-values calculated for the various diversity sets.

	Normal set	Contra-Trump set	Pro-Trump set	Hybrid set
Normal set	1	3,8E-36	6,41E-07	9,48E-32
Contra-Trump set		1	3,91E-0,7	1,00E-05
Pro-Trump set			1	1,28E-09
Hybrid set				1

Table 4.5: **P-values for the 1,500 diversity metrics for pro-Trump users** - This table shows the p-values calculated for the diversity values for the recommendations sets for pro-Trump users. The different sets are compared pairwise - therefore, only half of the values need to be calculated. Since we used $p <= 0.05$, all the values prove to be significant regarding the null-hypothesis.

	Normal set	Contra-Trump set	Pro-Trump set	Hybrid set
Normal set	1	1,44E-05	1,08E-15	5,18E-18
Contra-Trump set		1	9,94E-10	2,01E-21
Pro-Trump set			1	1,84E-25
Hybrid set				1

Table 4.6: **P-values for the 1,500 diversity metrics for contra-Trump users** - This table shows the p-values calculated for the diversity values for the recommendations sets for contra-Trump users. The different sets are compared pairwise - therefore, only half of the values need to be calculated. Since we used $p <= 0.05$, all the values prove to be significant regarding the null-hypothesis.

Serendipity P-Values

Tables 4.7 and 4.8 depict the p-values calculated for the various serendipity sets.

	Normal set	Contra-Trump set	Pro-Trump set	Hybrid set
Normal set	1	5,42E-26	1,24E-06	1,17E-07
Contra-Trump set		1	1,67E-09	4,4E-09
Pro-Trump set			1	1,04E-8
Hybrid set				1

Table 4.7: **P-values for the 1,500 serendipity metrics for pro-Trump users** - This table shows the p-values calculated for the serendipity values for the recommendations sets for pro-Trump users. The different sets are compared pairwise - therefore, only half of the values need to be calculated. Since we used $p <= 0.05$, all the values prove to be significant regarding the null-hypothesis.

	Normal set	Contra-Trump set	Pro-Trump set	Hybrid set
Normal set	1	1,27E-11	4,3E-14	3,06E-18
Contra-Trump set		1	1,5E-15	2,7E-14
Pro-Trump set			1	1,93E-58
Hybrid set				1

Table 4.8: **P-values for the 1,500 serendipity metrics for contra-Trump users** - This table shows the p-values calculated for the serendipity values for the recommendations sets for contra-Trump users. The different sets are compared pairwise - therefore, only half of the values need to be calculated. Since we used $p <= 0.05$, all the values prove to be significant regarding the null-hypothesis.

4.2 Qualitative Results

In order to better understand the quantitative results, we evaluated a sample pro-Trump and a sample contra-Trump user and explained their results. In the beginning, we give a general description of the user, followed by the diversity and serendipity results for the different recommendation variants.

4.2.1 Sample Pro-Trump User 'brucejwicks'

We have picked a random pro-Trump user from the dataset and give an overview in Table 4.9. In this example, the user description also validates the classification into the pro-Trump stance, since it includes 'republican conservative', which leads us to conclude that he is indeed a pro-Trump user. The bot score strongly indicates that this account is a 'real' user and not a bot of any sort.

Username	brucejwicks
Userstance	Pro-Trump
Top trigrams	retweet think jeff, think jeff sessions, president trump exonerated
User description	NY giants fan, jersey shore, republican conservative
User friends counter	665
User bot score english	0.28

Table 4.9: **User 'brucejwicks' Pro-Trump description** - This Table shows the description of the sample user for the pro-Trump stance named 'brucejwicks'. The top 3 trigrams, which we tagged 'user preferences' in this thesis, the user description of the Twitter account and the user bot score value are shown.

Next, we evaluated the diversity values for the different recommendation variants for these users and presented them in Table 4.13. The standard set and the pro-Trump set offer the lowest diversity, whereas the hybrid set offers the highest diversity, more than the contra-trump set. In this specific example, if we wanted to maximize the diversity of the recommendations, the hybrid set should be recommended to the user.

Recommendation variant	Diversity
Standard	.3930
Contra-Trump	.5412
Pro-Trump	.3828
Hybrid	.6503

Table 4.10: **User 'brucejwicks' diversity measures** - This Table shows the results for the diversity metric for the pro-Trump stance named 'brucejwicks'. We evaluated 4 different variants with a size of 10 recommendations per set.

In reference to our research question (**RQ 1**), we can show in Fig.4.13 as an example, that adding tweets of an opposing stance increases diversity compared to the standard recommendation set. In this case, the hybrid set gives the best results if diversity should be increased.

We repeated our experiments and computed the serendipity metrics, as shown in Table 4.14. The highest serendipity values are achieved with the contra-Trump set, whereas recommending the standard recommendation set to the user results in the lowest serendipity.

Recommendation variant	Serendipity
Standard	.9436
Contra-Trump	.9638
Pro-Trump	.9449
Hybrid	.9556

Table 4.11: **User 'brucejwicks' serendipity measures** - This Table shows the results for the serendipity metric for the pro-Trump stance named 'brucejwicks'. We evaluated 4 different variants with a size of 10 recommendations per set.

4.2.2 Sample Contra-Trump User 'johnnyatab'

Next, we picked a random sample contra-Trump user and show the results for the user in this Section. The user was classified into the contra-Trump stance. We presented a general description along with the top trigrams of the user 'johnnyatab' in Table 4.12. In this case, the trigrams already indicate a strong contra-Trump view, which validates our algorithm and the resulting classification into the contra-Trump stance. Interestingly, the standard set achieves the highest diversity values in this case, whereas the pro-Trump set achieves the lowest diversity values, as shown in Tab 4.13. Even though he might be disagreeing with them, his tweets still consist mostly of pro-Trump topics. Therefore, contra-Trump topics are very diverse to him. Serendipity results seem to lead to the same conclusion 4.14.

Username	johnnyatab
Userstance	contra-Trump
Top Trigrams	'#impeachtrump #resistance #resist'
User description	I am a dark comedy. I write, film things, do photography, dj, make music and try to find the answers.
User friends counter	1,144
User bot score english	0.26

Table 4.12: **Contra-Trump Sample User description** - This Table shows the description of the sample user for the contra-Trump stance named 'johnnyatab'.

Recommendation variant	Diversity
Standard	.7243
Contra-Trump	.7243
Pro-Trump	.3903
Hybrid	.6682

Table 4.13: **Pro-Trump Sample User Diversity measures** - This Table shows the results for the diversity metric for the pro-Trump stance named 'johnnyatab'. We evaluated 4 different variants with a size of 10 recommendations per set.

Recommendation variant	Serendipity
Standard	.9436
Contra-Trump	.9638
Pro-Trump	.9449
Hybrid	.9556

Table 4.14: **Contra-Trump Sample User Serendipity measures** - This Table shows the description of the sample user for the pro-Trump stance named 'johnnyatab'. We evaluated 4 different variants with a size of 10 recommendations per set.

4.2.3 Most common trigrams per stance

The following Figures Fig.4.4 and Fig.4.5 show the most common trigrams per user-stance and their frequency. They show the most important topics in the issue stances.

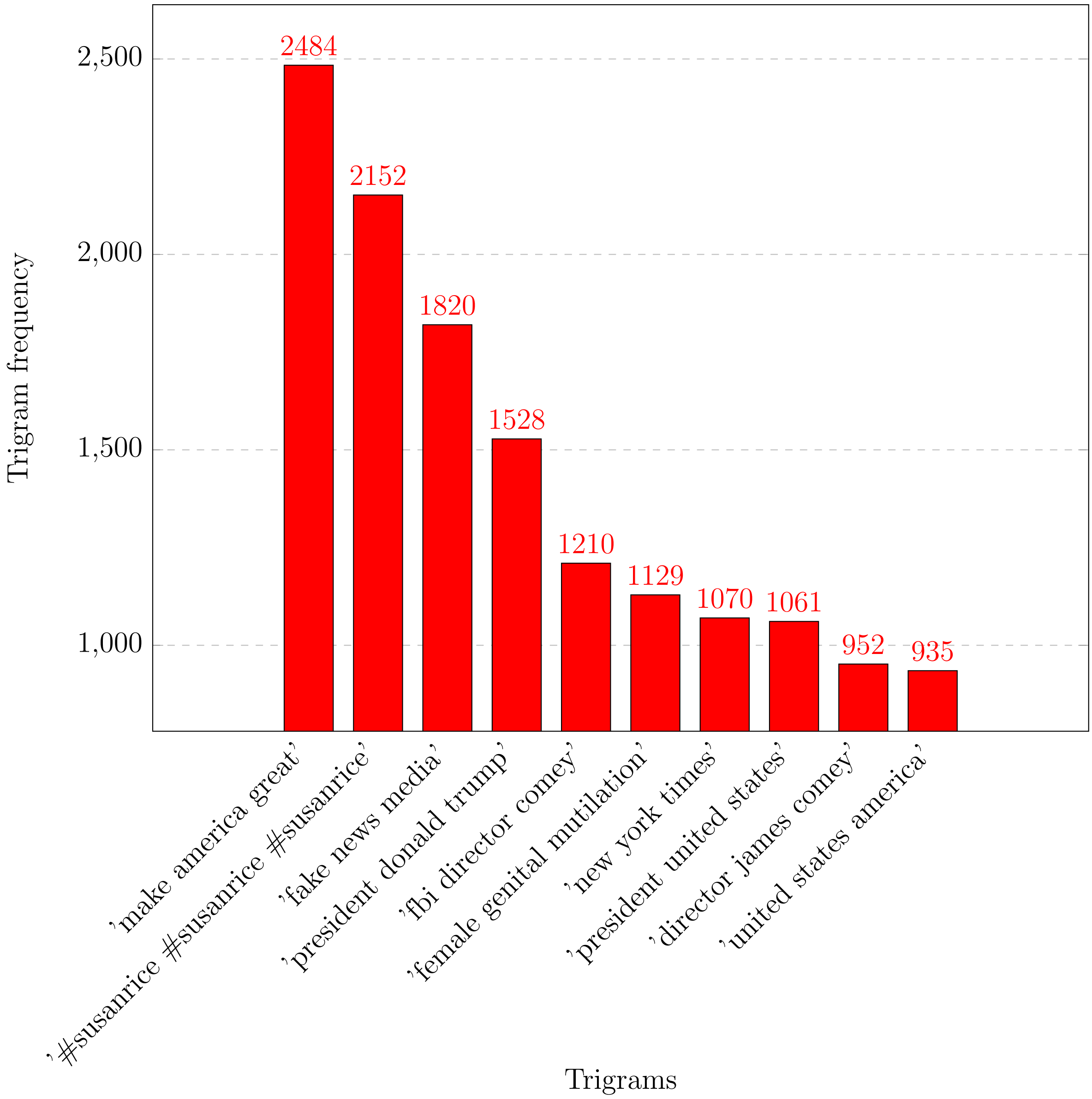

Figure 4.4: **Pro-Trump stance - Most common trigrams** This Figure shows the most common trigrams in the pro-Trump stance.

Fig.4.4 depicts that pro-Trump seed users focus on Donald Trump and his campaign slogan 'make america great'. Most hashtags deal with the US election campaign and election results. One can assume that the followers are patriotic (much focus on the term 'america' in various forms). One can see that the hashtags are also strongly influenced by events that have been heavily covered in the media, e.g. *#susan* *rice*. She was a White House security advisor at the time and was a strong voice against Donal Trump when the Twitter crawl was performed. This hashtag is closely followed by *#fake news media*, which tries to present the media reports as fake.

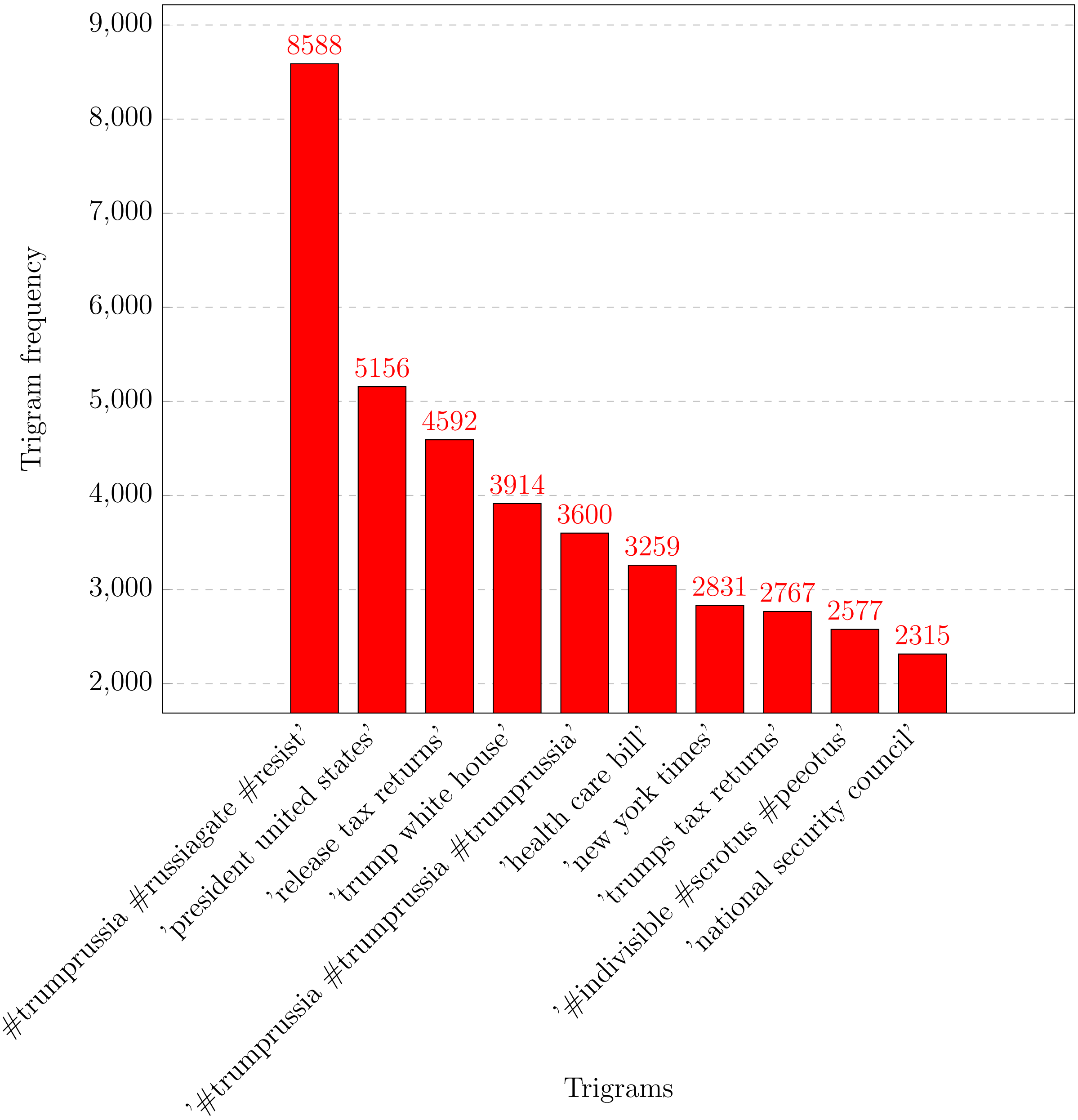

Figure 4.5: **Contra-Trump stance - Most common trigrams** This Figure shows the most common trigrams in the contra-Trump stance.

Contra Trump seed users try to establish a connection between Russia and Donald

Trump in Fig.4.5. They imply that Donald Trump has received help from Russia during the election campaign. Other topics are the tax returns of Donald Trump, where he refused to publish them and his health care bill. In addition, even the hashtags *#indivisible #scrotus #peeotus* made it into the top 10, which shows bad language and absolute dislike for the newly elected president.

Chapter 5

Discussions and Limitations

In this chapter, we discuss the results and the extent to which they provide an adequate answer to the research question. Furthermore, we mention some limitations of this work.

5.1 Discussion of the Research Question

To answer the research question (**RQ 1**), we used a content-based recommender and suggested tweets for the individual users. We defined different sets of recommendations for the user we were investigating in order to examine the influence of various compositions of tweets created by users with different views. The impact on the diversity was measured for each user per political stance and then the average over the 1,500 users per stance was calculated. We could see that we were able to influence diversity with our approach and that the different, hybrid recommendation sets have varying degrees of influence.

We discovered that if you want to increase the diversity of recommendations for a group with our approach, you also have to consider the topic similarity of the group. The optimal composition of recommendation sets with respect to the different viewpoints depends on how similar the views in a group already are. Furthermore, the diversity can be impacted to a varying extent with our approach, depending on the topic similarity of the examined group. In the case of the contra-Trump stance we were able to increase the diversity from 0.5 using the standard recommendation set (10 contra-Trump recommendations) to 0.75 with the optimal hybrid set (8 pro-Trump and 2 contra-Trump recommendations), which corresponds to an improvement of 50%. In the case of the pro-Trump stance, we were able to increase diversity from 0.63 with the standard recommendation set (10 pro-Trump recommendations) to 0.73 with the optimal hybrid set (6 contra-Trump and 4 pro-Trump recommendations), an increase of 16%.

In any case, the presented method has shown that it is a possibility to influence the diversity of the recommendations in a sustainable way. However, the increase depends strongly on the topic similarity within the group. Whether we can repeat this result for other topics must be clarified in further work.

5.2 Limitations

Crawling of tweets. We started by crawling tweets of partisan groups, i.e. contra-Trump and pro-Trump groups, in 2017, shortly after the election of president Trump. The hashtags we chose to crawl for potential users were #maga, #impeachtrump and #nobannowall. We chose these hashtags because we observed Twitter streams and found them to be valuable indicators for either one of the two groups. By doing so, we restricted us to users who used these hashtags and ignored others, which might also belong to either of the two partisan groups. Another limitation of the crawling process is limited size of the samples. Our dataset might be prone to certain biases that result of taking samples during a short timeframe. In addition, we have divided the users in our dataset according to the content of the tweets into 2 extreme opposites and have not considered less extreme opinions, therefore, a generalization cannot be made at this point without further research.

Hashtag analysis of the dataset. We then researched hashtags regarding the election of president Trump and used them to classify users into one of the two stances. We chose the hashtags for seed user classification manually by examining the context of these hash tags through an external tool. Even though we studied the chosen hashtags carefully, this selection is subject to subjective biases and might influence the results of this thesis. The final dataset, depicted in table 3.3 shows, that the dataset is composed of 2/3 contra-Trump users and only 1/3 of pro-Trump users. However, for the evaluation, we made sure to always pick the same amount of pro-Trump and contra-Trump users. In table 3.4 we show the most common hashtags in the dataset. The table already gives a good indication about the difference in topical diversity of the two stances - the second most used hashtag in the contra-Trump set has a frequency of 38% of the most common used hashtag in the same stance. When we compare the pro-Trump stance, the second most used hashtag is only used 12% as often as the most common hashtag.

Recommendation approach. We used a combination of TF-IDF and cosine similarity and the issue and user stance vectors to generate the user stance vectors, which describe how much a user belongs to either the pro-Trump or contra-Trump stance. We took a simple approach to weigh the words in the tweets because this was sufficient for demonstrating our results [Graells-Garrido et al., 2013]. In order to improve recommendation quality, many other techniques for improving recommendations exist, as shown in [H. Nidhi and Basava, 2017]. As opposed to most papers referenced in Chapter 2, we decided to recommend tweets directly to users instead of finding users to follow. We did this because we wanted to measure the influence of recommending tweets to a user of an opposing stance. The recommendations are found by taking the 15 most-common trigrams of the user into account. There are certainly more sophisticated ways to find recommendations for a user. One approach for improvement would be to use a hybrid recommendation approach, as explained in Section 2.1. Furthermore, we did not filter any recommendations in the recommendation set. Very similar recommendations in the recommendation set are not very valuable for humans and should be filtered or ranked lower.

Discussion of Metrics. We measured topical diversity in the pro-Trump and contra-Trump stance. The diversity differs a lot - pro-Trump users in our dataset have a lot more diverse topics than contra-Trump users. Whether this finding applies only to this dataset or can be generalized for pro-Trump and contra-Trump users is subject to research. Next, we created mixed sets of opposing beliefs of partisan groups and recommended these sets to users. In order to measure the effect of including opposing views in the recommendation set, we used serendipity and diversity metrics. To our knowledge, this is a new approach that has not been researched before.

However, many more metrics like precision, recall and novelty could be measured in this context, as shown in [Zhang et al., 2012]. One of the biggest weaknesses is that we do not measure prediction accuracy, as it is the most discussed property of recommender systems. The underlying assumption is that recommender systems that provide more accurate results are preferred by users. Prediction accuracy can be measured in an offline experiment. This experiment can be performed by recording historical data and hiding certain interactions to simulate the next proposed interactions and then comparing them with the actual ones (training set and test set) [Ricci et al., 2011].
Next, we averaged the results of 1,500 users of each stance. The reasons behind this number are computational. However, we chose these users randomly out of a larger set, therefore lessening the influence of a small sample. For measuring, the order in which the contra-Trump and pro-Trump tweets are arranged is not important. When we would show these tweets to users, the order might be a big influence and should therefore be considered in the research.

Chapter 6

Conclusion

This thesis proposed a content-based recommendation approach. With the help of diversity and serendipity metrics, we constructed recommendation sets, which help groups of users to escape the filter bubble and gain more exposure to a wider area of views regarding a certain topic.

We crawled tweets on Twitter shortly after the election of Donald Trump in 2017 and restricted them with corresponding hashtags. We cleaned up the data set and excluded bot accounts. After that we downloaded the last 1,100 tweets for the remaining accounts.

Having built the dataset, we extracted seed users and built issue stance vectors, which represent the two different stances, pro-trump and contra-trump. These operations were performed using tf-idf and cosine similarity. With the help of these we could classify each user in our dataset to one of these 2 issue stances.

Following this, we recommended 10 additional tweets from our dataset for a specific user by means of a content-based recommender. In order to measure the quality of the results of the recommender, we used specific metrics, such as diversity and serendipity. Additionally, we defined a metric called topic similarity to determine how diverse the content within the two different stances is.

Our goal was to see if we could influence diversity through our approach (RQ1). We generated various recommendation sets, which differ in the composition of pro-

Trump and contra-Trump tweets, and measured the diversity. We came to the following conclusion:

Suggesting tweets of the opposing stance can increase the diversity of the recommendations. By mixing the tweets from both stances (hybrid recommendation set) the diversity can be increased even further.

We then tested different compositions of pro-Trump and contra-Trump recommendations. We noticed that the optimal composition of pro-Trump and contra-Trump recommendations changes depending on the topic similarity of the respective stance. If too many tweets from the opposing group are added, the diversity might even decrease. When using hybrid recommendation sets, i.e. tweets from both stances, depending on the goal and the various input factors, we can conclude that it is important to engineer an optimal ratio for this set.

The results were quantitatively evaluated over 1,500 pro-Trump and 1,500 contra-Trump users. Furthermore, we also presented a user from each stance, showing the content of the recommendation sets and the quantitatively calculated metrics for each user. The results were also checked for significance.

6.1 Future Work

For future work, we want to continue with our findings and focus our research on an optimal strategy for combining tweets of opposing beliefs, taking the inherent topic similarity of the different stances into account. Additionally, we want to test the recommendations on users and get their feedback regarding the usefulness of the recommendations. In order to understand partisan actions, we want to increase our study on communication patterns. While doing so, we must be aware of the backfire effect, which reinforces the views of people when faced with other opinions [Nyhan and Reifler, 2010]. In order to mitigate this effect, we need to show the recommendations in the right context [Lex E., 2018]. In order to verify our findings, we plan to use a larger Twitter dataset and different polarizing topics. Furthermore, in order to get more insight on the performance of our recommender, we plan to measure additional metrics such as prediction and recall.

Bibliography

[Adamic and Glance, 2005] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 u.s. election. In Adibi, J., Grobelnik, M., Mladenic, D., and Pantel, P., editors, *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, pages 36–43, New York, New York, USA. ACM Press.

[Armentano et al., 2012] Armentano, M. G., Godoy, D., and Amandi, A. (2012). Topology-based recommendation of users in micro-blogging communities. *Journal of Computer Science and Technology*, 27(3):624–634.

[Baeza-Yates and Ribeiro, 1999] Baeza-Yates, R. and Ribeiro, B. d. A. N. (1999). *Modern information retrieval*. Addison-Wesley, Harlow.

[Batrinca and Treleaven, 2015] Batrinca, B. and Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1):89–116.

[Çano and Morisio, 2017] Çano, E. and Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent Data Analysis*, 21(6):1487–1524.

[Chang and Iyer, 2012] Chang, H.-C. and Iyer, H. (2012). Trends in twitter hashtag applications: Design features for value-added dimensions to future library catalogues. *Library Trends*, 61(1):248–258.

[Davis et al., 2016] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., and Menczer, F. (2016). Botornot. In Bourdeau, J., Hendler, J. A., Nkambou, R. N., Horrocks, I., and Zhao, B. Y., editors, *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pages 273–274, New York, New York, USA. ACM Press.

[Fischer et al., 2011] Fischer, P., Kastenmüller, A., Greitemeyer, T., Fischer, J., Frey, D., and Crelley, D. (2011). Threat and selective exposure: the moderating role of threat and decision context on confirmatory information search after decisions. *Journal of experimental psychology. General*, 140(1):51–62.

[Frey, 1986] Frey, D. (1986). Recent research on selective exposure to information. volume 19 of *Advances in Experimental Social Psychology*, pages 41–80. Elsevier.

[Garimella et al., 2017] Garimella, K., de Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2017). Reducing controversy by connecting opposing views. In de Rijke, M., Shokouhi, M., Tomkins, A., and Zhang, M., editors, *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining - WSDM '17*, pages 81–90, New York, New York, USA. ACM Press.

[Ge et al., 2010] Ge, M., Delgado-battenfeld, C., and Jannach, D. (2010). Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *In RecSys '10*, page 257.

[Graells-Garrido et al., 2013] Graells-Garrido, E., Lalmas, M., and Quercia, D. (2013). Data portraits: Connecting people of opposing views. *CoRR*, abs/1311.4658.

[Gupta et al., 2013] Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., and Zadeh, R. (2013). Wtf. In Schwabe, D., Almeida, V., Glaser, H., Baeza-Yates, R., and Moon, S., editors, *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, pages 505–514, New York, New York, USA. ACM Press.

[H. Nidhi and Basava, 2017] H. Nidhi, R. and Basava, A. (2017). Twitter-user recommender system using tweets: A content-based approach. pages 1–6.

[John S. Breese et al., 1998] John S. Breese, David Heckerman, and Carl Kadie (1998). Empirical analysis of predictive algorithm for collaborative filtering. In *Proceedings of the 14 th Conference on Uncertainty in Artificial Intelligence*, pages 43–52.

[Jungherr, 2016] Jungherr, A. (2016). Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 13(1):72–91.

[Kwak and Kim, 2017] Kwak, S. K. and Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, 70(4):407–411.

[Lex E., 2018] Lex E., Wagner M., K. D. (2018). Mitigating confirmation bias on twitter by recommending opposing views.

[Liao and Fu, 2013] Liao, Q. V. and Fu, W.-T. (2013). Beyond the filter bubble. In Mackay, W. E., Brewster, S., and Bødker, S., editors, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, page 2359, New York, New York, USA. ACM Press.

[Liao and Fu, 2014] Liao, Q. V. and Fu, W.-T. (2014). Can you hear me now? In Fussell, S., Lutters, W., Morris, M. R., and Reddy, M., editors, *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pages 184–196, New York, New York, USA. ACM Press.

[McNee et al., 2006] McNee, S. M., Riedl, J., and Konstan, J. A. (2006). Being accurate is not enough. In Olson, G. and Jeffries, R., editors, *CHI '06 extended abstracts on Human factors in computing systems - CHI EA '06*, page 1097, New York, New York, USA. ACM Press.

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

[Munson and Resnick, 2013] Munson, S. and Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget.

[Munson and Resnick, 2010] Munson, S. A. and Resnick, P. (2010). Presenting diverse political opinions. In Mynatt, E., Schoner, D., Fitzpatrick, G., Hudson, S., Edwards, K., and Rodden, T., editors, *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, page 1457, New York, New York, USA. ACM Press.

[Neisser, 2010] Neisser, P. (2010). Going to extremes: How like minds unite and divide - by cass r. sunstein just how stupid are we? facing the truth about the american voter - by rick shenkman. *Political Psychology*, 31(3):482–489.

[Nemeth and Rogers, 1996] Nemeth, C. and Rogers, J. (1996). Dissent and the search for information. *British Journal of Social Psychology*, 35(1):67–76.

[Neuhäuser, 2011] Neuhäuser, M. (2011). Wilcoxon–mann–whitney test. In Lovric, M., editor, *International Encyclopedia of Statistical Science*, pages 1656–1658. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Nyhan and Reifler, 2010] Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

[Ott, 2017] Ott, B. L. (2017). The age of twitter: Donald j. trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1):59–68.

[Pariser, 2011] Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press, New York, NY.

[Park et al., 2009] Park, S., Kang, S., Chung, S., and Song, J. (2009). Newscube. In Olsen, D. R., Arthur, R. B., Hinckley, K., Morris, M. R., Hudson, S., and Greenberg, S., editors, *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, page 443, New York, New York, USA. ACM Press.

[Rajaraman and Ullman, 2011] Rajaraman, A. and Ullman, J. D. (2011). Data mining. In Rajaraman, A. and Ullman, J. D., editors, *Mining of Massive Datasets*, pages 1–17. Cambridge University Press, Cambridge.

[Ricci et al., 2011] Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B. (2011). *Recommender Systems Handbook*. Springer US, Boston, MA.

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Sanders and Mullen, 1983] Sanders, G. S. and Mullen, B. (1983). Accuracy in perceptions of consensus: Differential tendencies of people with majority and minority positions. *European Journal of Social Psychology*, 13(1):57–70.

[Senecal and Nantel, 2004] Senecal, S. and Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing*, 80(2):159–169.

[Silard, 2017] Silard, T. (2017). No ban, no wall: Standing with immigrant communities. `https://www.huffingtonpost.com/timothy-p-silard/no-ban-no-wall-standing_b_14520530.html`. Accessed: 2017-09-30.

[Singhal, 2001] Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43.

[Stromer-Galley, 2003] Stromer-Galley, J. (2003). Diversity of political conversation on the internet: Users' perspectives. *Journal of Computer-Mediated Communication*, 8(3):0.

[Sunstein, 2002] Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195.

[Xiao and Benbasat, 2007] Xiao and Benbasat (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, 31(1):137.

[Zhang et al., 2012] Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., and Jambor, T. (2012). Auralist. In Adar, E., Teevan, J., Agichtein, E., and Maarek, Y., editors, *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, page 13, New York, New York, USA. ACM Press.

[Ziegler et al., 2005] Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In Ellis, A. and Hagino, T., editors, *Proceedings of the 14th international conference on World Wide Web - WWW '05*, page 22, New York, New York, USA. ACM Press.