Predicting the Popularity of Online Recipes

Dominik Mößlang

Dominik Mößlang B.Sc.

Predicting the Popularity of Online Recipes

Master's Thesis

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Software Engineering and Management

submitted to

Graz University of Technology

Supervisor

Dipl.-Ing. Dr.techn. Christoph Trattner, B.Sc.

Institute of Interactive Systems and Data Science

Graz, October 2017

© Copyright 2017 by Dominik Mößlang, except as otherwise noted.

Dominik Mößlang B.Sc.

Vorhersage der Popularität von Online-Rezepten

Masterarbeit

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Softwareentwicklung - Wirtschaft

eingereicht an der

Technischen Universität Graz

Betreuer

Dipl.-Ing. Dr.techn. Christoph Trattner, B.Sc.

Institute of Interactive Systems and Data Science

Graz, Oktober 2017

Diese Arbeit ist in englischer Sprache verfasst.

AFFIDAVIT

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

Eidesstattliche Erklärung

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 identisch.

Date/Datum

Signature/Unterschrift

Abstract

In the past few years, the Internet has become a popular source of cooking inspirations and ideas. A recent study showed that over 50% of all respondents consume food, based on online recipes. These numbers indicate, that the online search for recipes, is nowadays probably more popular than using traditional cooking books. However, the issue with online sourced recipes is that they are often unhealthy. Additionally, the unhealthy recipes tend to be more popular. Furthermore, latest research found possible links between popular online recipes and obesity prevalences in the US. Hence, understanding the socio-dynamic processes which make online recipes popular, could be the first step towards recommending healthier nutrition. Such recommender systems have the potential to solve many of the food related health issues, rapidly growing in modern societies. Research in the field of cognitive psychology showed that people are choosing food based on multiple aspects, but the question of which features are relevant for the popularity of online recipes remained unanswered. This thesis aims to make contributions towards a deeper understanding of such processes. The chosen approach includes a statistical analysis of datasets from two popular food community websites: Allrecipes.com and Kochbar.de. They represent two separate stereotypical food cultures of western civilization. By comparatively analysing both, a more general view is revealed. These analyses are based on characteristics of the recipes and the underlying social network of the websites. These so called features, are developed on the results of related studies regarding the popularity of online content or food preferences. In order to reveal the predictive power of the features and to test the statistical findings, predictive modelling experiments were conducted. As the experiments show, there are generally valid recipe charateristics, which have strong influence on the future popularity of online recipes. However, there are small differences: user activity features, such as written and obtained ratings/comments or the number of recipes uploaded, show reliable results for the Kochbar.de dataset. Whereas innovation features, such as recipe innovation or ingredients popularity rank and image features, for instance, saturation or image entropy seem to have more influence on the popularity of Allrecipes.com its recipes.

Kurzfassung

In den letzten Jahren hat sich das Internet zu einer beliebten Quelle für Rezeptideen entwickelt. Eine aktuelle Studie hat ergeben, dass über 50% aller Befragten Gerichte auf Basis von Online-Rezepten kochen. Dies lässt vermuten, dass die Online-Suche nach Rezepten heutzutage wahrscheinlich beliebter ist als traditionelle Kochbücher. Problematisch ist, dass Rezepte aus dem Internet häufig ungesund sind. Außerdem sind die ungesunden Rezepte in der Regel auch die Beliebteren. Neueste Forschungen fanden mögliche Verbindungen zwischen populären Online-Rezepten und der Adipositas Prävalenz in den USA. Ein besseres Verständnis über die soziodynamischen Prozesse, welche Online-Rezepte populär machen, könnten ein erster Schritt in Richtung gesünderer Ernährung sein. Ernährungsbezogene Recommender-Systeme haben das Potenzial, die stetig wachsenden Gesundheitsprobleme moderner Gesellschaften zu lösen. Die Forschung auf dem Gebiet der kognitiven Psychologie zeigte, dass Nahrung auf Grundlage mehrerer verschiedener Aspekten ausgesucht wird. Welche Aspekte allerdings genau für die Popularität von Online-Rezepten verantwortlich sind, blieb unbeantwortet. Diese Arbeit zielt darauf ab, eben diese Prozesse zu analysieren. Die gewählte Methodik beinhaltet eine statistische Auswertung von zwei populären Food-Community-Websites: Allrecipes.com und Kochbar.de. Beide repräsentieren verschiedene Esskulturen der westlichen Zivilisation. Aufgrund vergleichender Analysen, konnten allgemeine Korrelationen vorgestellt werden. Diese Analysen umfassen verschiedenste charakteristische Merkmale von Online-Rezepten und den sozialen Netzwerken, auf welchen die Websites basieren. Diese sogenannten 'Features', wurden aufgrund relevanter Studien zur Popularität von Online-Inhalten oder Lebensmittelpräferenzen entwickelt. Um die Vorhersagekraft dieser Merkmale bezüglich der Popularität von Online-Rezepten zu prüfen, wurden moderne Machine Learning Algorithmen angewandt. Die Resultate der Experimente zeigen, dass es allgemeine Rezept-Merkmale gibt, welche sich zur Vorhersage der Popularität von Rezepten eignen. Es gibt aber auch kleine Unterschiede: Die vormalige Aktivität der Benutzer, wie zum Beispiel die Anzahl der geschriebenen/erhaltenen Kommentare oder Bewertungen und die Anzahl der online gestellten Rezepte, haben einen großen Einfluss auf die Popularität von Rezepten im Kochbar.de Datensatz. Die Neuartigkeit und die Präsentation (z.B.: Rezeptbilder) von Rezepten spielen hingegen bei Allrecipes.com eine größere Rolle.

Contents

List of Figures

List of Tables

List of Listings

Acknowledgements

First of all, I would like to thank everyone who supported me. I am thankful for their aspiring advices and constructive criticism throughout the course of this master thesis.

I am indebted to my supervisor and mentor Dr. Christoph Trattner, who gave me the chance to undertake this research. Whenever necessary, he provided me invaluably advice, patient guidance and immediate attention. It was a very unique and special opportunity to be part of current research and to learn a lot about state of the art methods.

I especially wish to thank my parents who always encouraged me to pursue my own goals. Thank you for all your support and love. Without you, all this would not have been possible.

Last but not least, I would like to thank my girlfriend for her understanding and support, my friends for always lending me a sympathetic ear and my family for been always there for me.

Dominik Mößlang
Graz, Austria, October 2017

Chapter 1

Introduction

“ It is certainly important to be looking for cures to medical disorders, but it is equally important to conduct research on human health and well-being. ”

[Stephen LaBerge, American Psychologist, born 1947]

1.1 Motivation

In the past few years, the Internet has become a popular source of cooking inspirations and ideas. Online food community platforms such as Allrecipes.com or Kochbar.de have annually increasing user numbers. The US based Allrecipes.com reported[1] lately that they have 1.5 billion visits from more than 80 million global users, every year. Its European counterpart Kochbar.de, was ranked number 36 of Germany’s most visited websites in 2016[2]. Furthermore, a recent market agent[3] study, conducted in February 2017, showed that over 50% of all survey respondents consume online recipes sourced food. These numbers indicate that the online search for recipes, is nowadays probably more popular than traditional cooking books. That does not seem particularly surprising, since online food communities have many advantages for their users. These platforms can easily be accessed via computer or smartphone and feature hundreds of thousands of recipes from all possible cusines. Furthermore, they include social networks which allow user to interact with each other by becoming friends, commenting recipes or gathering in cooking groups. However, the issue with online sourced recipes is that they are often unhealthy, as Trattner and Elsweiler [2017] recently discovered. Users are usually not able to reliably distinguish between healthy and unhealthy food. Additionally, the unhealthy recipes tend to be more popular. In another recent study, Trattner, Parra, et al. [2017] found possible links between popular online recipes and obesity prevalences in the US. Hence, understanding what actually makes such online recipes popular, could be the first step to recommending

[1] http://press.allrecipes.com/allrecipes-com-celebrates-its-20th-anniversary-as-the-worlds-leading-digital-food-brand/

[2] http://meedia.de/2017/03/09/agof-chefkoch-erstmals-mit-mehr-als-20-mio-nutzern-rekorde-auch-fuer-kochbar-essen-trinken-und-lecker/

[3] http://marketagent.com/

healthier food to people. Research in the field of cognitive psychology showed that people are choosing food based on multiple aspects. Scheibehenne et al. [2007] categorised them into internal and external food aspects. Internal aspects are for example *taste, sensory appeal* or *nutritional values*. External aspects on the other hand, are related to *individual preferences*, *motives* or *information* and the *location* the food decision takes place in. However, the question which of these aspects are relevant for the popularity of online recipes remained unanswered. As such, the thesis aims to make contributions towards a deeper understanding of such processes.

1.2 Objectives

The primary goal of this master thesis is to reveal generally valid online recipe aspects that are relevant for their popularity. During the course of this thesis, the following research questions (RQ) are addressed:

- **RQ1:** To what extend can popularity patterns be found in the online food communities Allrecipes.com and Kochbar.de?
- **RQ2:** To what extent do the two online communities (Allrecipes.com and Kochbar.de) differ or are similar to each other, in respect to features induced and inspired by cognitive psychology?
- **RQ3:** To what extent can potential correlations be found that might be useful in a prediction task?
- **RQ4:** To what extent is the popularity of online recipes predictable and what are the most useful features to do so?

1.3 Contribution

The aim of the thesis is to shed light on hidden aspects of online recipe popularity. Preliminary studies such as the one of Rokicki, Herder, and Trattner [2017] already revealed correlations between popularity and social or temporal features, respectively. However, such investigations were based solely on one online food community or dataset. This thesis relies on large scale datasets from two of the biggest food community websites: Allrecipes.com (US) and Kochbar.de (EU). This way, a cross-cultural study over different countries and continents was possible. Former studies in this field, tried to capture the different recipe aspects with a small number of selected features. For the feature analysis of this thesis, over 70 features have been developed. Many of them are known to have influence either on the popularity of online content or on food choices. However, also new features describing *nutrition*, *innovation*, *healthiness* or *presentation quality* are introduced to the field of recipe popularity. Each feature has been assigned to feature sets, for example 'Nutritional information'. Using these features, a machine learning algorithm is deployed to make predictions on the future popularity of recipes. This approach is based on classification, rather than a regression analysis. The experiment analyses each feature set individually as well as combined and

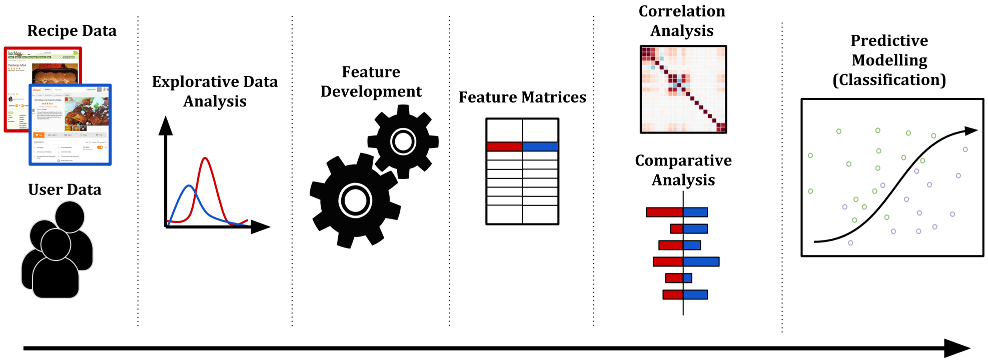


Figure 1.1: A schematic illustration of the approach to popularity prediction of online recipes.

considers different time periods of a recipe's lifespan. State-of-the-art classifiers such as Random Forest, Naive Bayes or Generalised Linear Models were utilized. To evaluate the performance of the classification models, a cross validation protocol was chosen and the performance of each feature set is reported. Figure 1.1 illustrates this process in more detail. To the best of the authors knowledge this is the first study of this kind, (i) investigating differences and communalities of two large scale online food plattforms and (ii) reavealing to what extent recipes become popular online.

1.3.1 Relevance of this Work

- The future popularity of content is a important feature for recommender systems.
- Recipes websites, such as Allrecipes.com or Kochbar.de, could boost their website traffic and visitor numbers by knowing the future food trends.
- Knowing which content will be popular in the future could be benefitial for technical problem solutions such as caching or replication.
- Allrecipes.com publishes their own recipe magazines. Always covering the future recipe trends could increase revenue.
- Human health care systems could benefit from the knowledge of future food trends by explaining potential risks.

1.4 Thesis Outline

This master thesis is split into seven Chapters. This introduction Chapter is followed by the related work (Chapter 2), which reviews appropriate and relevant background literature. The related work gives an overview of work done in the fields of online data based health studies and popularity prediction of Web content. Chapter 3 introduces the two datasets on which this thesis is based on. It explains how the datasets were collected, what are the limitations of them and why the food websites Allrecipes.com and Kochbar.de have been chosen for the purpose of this thesis. Chapter 4 addresses the methodology chosen to addresse the research questions. Furthermore, it gives insights in the data selection and pre-processing steps taken prior to any analysis and describes in detail the development of the various features that aim to capture the popularity of online recipes. Chapter 5 presents and discusses the results of the studies done to reveal possible popularity correlations between different recipe and user related aspects. Finally, the Chapter 6 discusses the conclusions and limitations of this master thesis and propose future research directions. Additionally, the Appendix includes further research findings and approaches, which have been excluded for sake of clarity and relevance.

Chapter 2

Related Work

“ It is better to fail in originality than to succeed in imitation. ”

[Herman Melville, American Novelist, 1819–1891]

This Chapter tries to give an overview of the relevant work and literature in the context of this thesis. It highlights different methods that solve comparable problems and work that proposed related methods to solve the same problem. The Chapter is split into two parts: Part one (i) deals with the prediction of popularity on the World Wide Web, social networks and in particular online recipes. Part two (ii) reviews work on data science in the domain of dietary behavior and their social implications, which significantly influenced this thesis.

2.1 Popularity Prediction of Web Content

In their survey about Web content popularity prediction, Tatar et al. [2014] argue that the easy access to social media platforms has resulted in a massive increase of information. A transition from bare consumers to creators, has led to an overwhelming and almost impossible to overlook, amount of content. However, Web content is a broad and generic term. Tatar et al. [2014] define it as

> "any individual item (in the form of text, image, audio, or video), publicly available on a website, which contains a measure that reflects a certain level of interest showed by an online community".

For example, users write Twitter tweets, upload images to Instagram, make Youtube video logs, sell their things on Ebay or just read news articles online. Furthermore, they stated three types of actors: Online users, content creators and content providers. Knowing what makes content popular and interesting for others, would be benificial for each one of them. For example, prediction models could help to filter, deliver and better organize huge amounts of information. Algorithms could automatically measure the interest in news articles, then select and organize them according to the predicted future popularity, instead of relying on human editors. Online marketing maybe profit from better and faster adapted advertising strategies. Traditionally, the success of a marketing campaign was measured afterwards, or when interest fades away. Popularity driven strategies might predict it beforehand. Content delivery networks would benefit from a more accurate demand prediction. This may save money, as an optimal amount and local placement of replications and copies results in less storage demand. Furthermore, the performance of proxies and webservers could be improved. Such services buffer information that is frequently used. The two main approaches for cache replacement are Least Recently Used (LRU) and Least Frequently Used (LFU). Famaey et al. [2013] stated a new approach call P-LFU. It integrates popularity predictions in the cache replacement decision, for future demand calculation. Similarly, the mobile Internet trend often results in low data rates. Internet Service Providers (ISP) search intensively for methods to reduce load on their infrastructure. Therefore, mobile data offloading sounds like a promising approach. It offloads data, that might be popular in the future in a specific region (mobile network cell), to the devices of users. When needed, the information gets transferred over device to device communication (peer to peer). Also, the accuracy of search engine results could be increased be predicting the future popularity of content, and substitute older information. Nevertheless, these prediction approaches have drawbacks. Inaccurate predictions can lead to higher network and infrastructure load. Also, the problem of the "self-fulfilling prophecies" need to be solved. A good example for that are popularity based ranking algorithms. News articles that are classified as "popular in the future", are be placed in the upper part of a website. This way it will gain more attention and fulfills the prophecy.

Popularity prediction approaches and methods have been an active research field for at least a decade. In order to give an overview, Tatar et al. [2014] grouped prediction methods, which have been applied in the literature to predict popularity of Web content, corresponding to granularity and data used. Figure 2.1 reflects this grouping of methods. First, they distinguish between the source of information. *Single domain* methods use only information available on one particular website. Regardless of the origin of the data. On the contrary, *cross domain* exploit information from multiple website and overcome the limitation of a single information source. Secondly, the temporal dimension of information is used for categorisation. *Before publication* is a very difficult task, since the prediction task can only rely on information before uploading content, such as metadata or the publisher's social connections. Depending on the moment (time) of the prediction, *after publication* methods can include information about the early attention phase of Web content. Usually, this gives an advantage over before publication. *After publication* methods are then further split into *aggregate behavior* and individual behavior. The approach of *aggregate behavior* treats the attention of all users equally. It either uses the *cumulative growth* of the popularity values until the prediction moment, a *temporal analysis* of the popularity evolution (how) until the prediction or it clusters similar Web content and compares their popularity evolution trends. The *individual behavior* methods stands in contrast to this. It treats every users individually and analyses their behavior. User behavior is described by states, each state represents a possible activity a users is able to do on a website (e.g, searching, rating, commenting, reading etc.). Based on the activity history of all users, probabilistic Bayesian inference models are calculated and can be used to predict popularity.

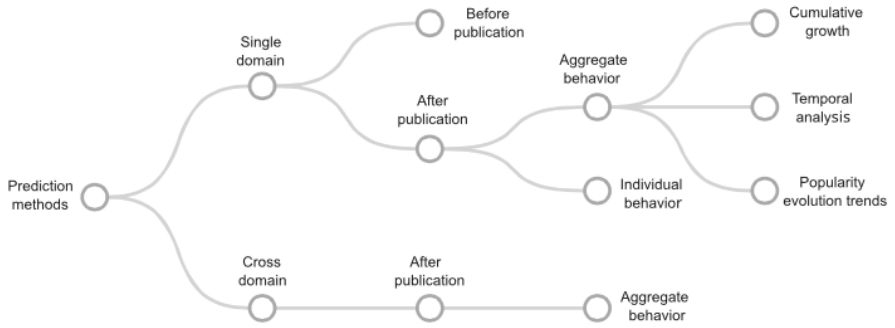


Figure 2.1: The classification of popularity prediction methods based on the granularity of data. Taken from Tatar et al. [2014].

This thesis tries to give insights into the predictability of online food recipes popularity. Special interest lies in the predictability of recipes before their publication, since this would be beneficial to the content creator's reputation. For example, the upload view of the website could support the user by providing hints on what words a recipes title should contain or how the recipes title image should look like, in order to generate more attention. Additionally, the domain of popularity prediction of Web content is a very broad research field with a big number of studies. Reviewing all possible approaches and methods would exceed the scope of this master thesis in terms of methodology and time constraints. Therefore, this Section will only cover selected and representative *single domain* and *before publication* approaches in more detail.

The identification of the driving factors, which influence the popularity of online content, is crucial for an accurate prediction. Tatar et al. [2014] has summarized various types predictive features that have shown good results in describing popularity and are broadly used in the literature:

- **Content creator features**: These features describe the characteristics of the creator itself. There are many types of creators, for example bloggers, news agencies or artists. Such features could be the social connections within a network or his fame in the real world.

- **Textual features**: One of the main consumed content forms is text. Features that describe the characteristics of text, especially emotions provoked, are strong predictors. Examples are key phrases, specific words or controversial topics. All of them can produce attention.

- **Categorical features**: Predictive models can also be based on categorical features. They describe the relations of different Web content items to each other. They allow to derive popularity from similar items. The problem with categorical features is, that items often belong to multiple categories.

- **Named entity features**: Popular items in real life, often are popular in the virtual world as well. People such as celebrities or important people can gain attention just by mentioning them. An example could be the sports article about a famous race driver.

- **Sentiment analysis features**: Emotion triggering Web content, is also considered to attract attention. The extraction of the sentiment value is considered a very difficult task. However, for text analysis there are many tools (such as SentiStrength[1]) available.
- **Attention in social media**: Social media is a good indicator for popularity of content on websites. The number of likes or comments on Facebook, are strong predictors for the popularity of articles on news websites.
- **Sharing behavior**: It has been observed that the sharing behavior on social platforms are useful as predictors of popularity (for example of Youtube videos).
- **Real world features**: Real life events, such as sports events or natural disasters, have influence on social media published content. However, it is very challenging to exploit this information and create predictors.

With their popularity experiment, Tsagkias et al. [2009] they tried to predict the comment volume of news articles from different online news organisations. They collected over 290 thousand articles with 1.89 million comments. The dataset was contained articles from November 2008 until April 2009. Their goal was to identify the characteristics of news article, that make them popular. 24 different features were extracted and grouped into one of five features sets. Table 2.1 lists all features in detail. For comparison reasons, all features were calculated for each news agent separately. The five feature sets are:

- **Surface features:** Mainly describing quantitative characteristics of the article such as text length, whether it has a summary, how many categories are assigned, date of publication, how many authors etc.
- **Cumulative features:** Often newspapers use stories from news providers, since they do not have the resources to cover all relevant stories by themselves. When a story occurres multiple times for all news organisation in a defined time window, it can be an indicator that a news is relevant for different user groups. Tsagkias et al. [2009] measure how much identical articles are published in the same hour.
- **Textual features:** They extracted the 100 most discriminative words within the article text body. This was done by a log-likelihood score. This feature describes differences between types or categories of articles and news sources.
- **Semantic features:** Are primarily named entities such as locations, persons, organisations etc. they counted the occurrence of various types within an article.
- **Real world features:** This feature set reveals correlations between real world phenomena and article popularity. They assigned the median temperature in the Netherlands from the day of publication to each article.

[1] http://sentistrength.wlv.ac.uk/

Table 2.1: The Table lists the features used by the classification experiments. Taken from Tsagkias et al. [2009].

Feature	Description	Type
Surface features		
month	Month(1-12)	Nom
wom	Week of the month (1-4)	Nom
dow	Day of the week (1-7)	Nom
day	Day of the month (1-31)	Nom
hour	Hour of the day (0-23)	Nom
first_half_hour	Publication in the first 30 minutes of the hour	Nom
art_char_length	Article content length	Int
category_count	Number of categories it is published on	Int
has_summary	Article has summary	Int
has_content	Article has content (HTML incl.)	Int
has_content_clean	Article has content (only text)	Int
links_cnt	Number of out-links	Int
authors_cnt	Number of authors	Int
Cumulative features		
art_same_hr	Published articles in same hour for source	Int
dupes_int_cnt	Near-duplicates in same source	Int
dupes_ext_cnt	Near-duplicates in other sources	Int
Tetual features		
	tf of top-100 terms ranked by their log-likelihood score for each source	Int
Semantic features		
ne_loc_cnt	Number of location-type entities	Int
ne_per_cnt	Number of person-type entities	Int
ne_org_cnt	Number of organisation-type entities	Int
ne_misc_cnt	Number of miscellaneous-type entities	Int
has_local	Any entities referring to the Netherlands	Int
	tf of top-50 entities from each entity type, ranked by their log-likelihood score for each source	Int
Real-world Features		
temperature	Temperature in Celsius at publication time	Num

They conducted two different experiments: First, they tried to predict whether an article will receive comments or not, using an RandomForest classification algorithm. They assigned a binary feature to each article, describing if the article received at least one comment. As their performance baseline the used models consisting of six temporal features (month, week of the month, day of the week, day of the month, hour, and first half hour). For each news agent, they created a training and a test set, containing articles published between November 2008 until February 2009 and articles from March 2009, respectively. As a performance measure, they used the F1-score and calculated the significance of the results via a Kappa-statistic. The results of this first experiment showed that textual and semantic features worked the best for the most news sources. Other features were very near the baseline classification models and the combination of all features did not improve the results significantly. Secondly, they tried to classify articles that will become a high or low volume of comments. The experimental setup was almost the same, only the predictor variable was different. Articles that were classified as "getting comments" in the first experiment, got binary marked as high volume. This experiment again showed very good results for the textual and semantic features. Another interesting observation was that there are performance variations for the different news agents and features. Showing that not all features are generally applicable.

Bandari et al. [2012] tried to answer similar questions regarding popularity of news articles. In contrast to Tsagkias et al. [2009], they measured popularity by article mentioning in tweets. This way, they can use Twitter as a proxy. Their dataset consisted of 44 thousand articles collected from the news feed aggregator Feedzilla. Feedzilla was an online news distributor started in 2005. They primarily used articles from the 8th until 16th of August 2011. Because of redundancies, naming variants and spam, the data was cleaned and filtered, resulting in a discard of about 2 thousand articles.

Bandari et al. [2012] conducted two prediction experiments. The features used strongly relate to their research questions:

- Do news categories play a role in terms popularity?
- To what extend does emotional language influence popularity?
- Are named entities a driving factor?
- Does it matter which news agency publishes an article?

In order to answer these questions, they used the following features:

- **Categorical score:** The Feedzilla articles were assigned to categories (e.g. sports or politics). They introduced a metric which describes the prior popularity of each category based on the number of tweets and the number of links in each category:

$$t - density = \frac{Number\ of\ Tweets}{Number\ of\ Links} \tag{2.1}$$

- **Subjectivity:** This feature describes how factual articles are. In other words, the emotional charge of the articles language. They used the LingPipe natural language toolkit for the sentiment extraction and binary classified each article from 0 to 1 (0 for factual and 1 for subjective).

- **Named entities:** Similar to Tsagkias et al. [2009], they used locations, persons, organisations. The extraction was done using the Stanford-NER entity extraction tool.

- **Source score:** The dataset contained 1350 unique news agencies. They ranked every one with the historical precedence and popularity in the collected data with two different scores: The number of articles of each source that were shared on twitter and the t-density over all articles of each source.

Their first experiment was a regression analysis with the goal of identifying the best features for the prediction of the articles future tweet volume. They used linear, k-nearest neighbors and support vector machine regression. Because of the long tailed *number of tweets* distribution, they logarithmically transformed the values for the regression calculation. The linear regression analysis revealed the following relationships:

$$T = S^{1.24} C^{0.45} e^{-(0.1 Ent_{max} + 3)}, \; with\ R^2 = 0.258 \tag{2.2}$$

$$or\ alternatively$$

$$T^{0.45} = (0.2S - 0.1Ent_{ct} - 0.1Ent_{avg} + 0.2Ent_{max})^2, \; with\ R^2 = 0.34 \tag{2.3}$$

where Ent_{max} denotes the maximum t-density, T represents the number of tweets. C and S are the t-densities of the categories and sources, respectively. R^2 is the coefficient of determination for the regression experiments. The SVM regression reached similar R^2 values as the linear regression. For the K-Nearest Neighbor algorithm predicted the number of tweets of an articles based on the Euclidean distance of articles in the multidimensional feature space. They tuned the algorithm with parameter $k = 7$ and $k = 3$, which states the number of considered neighbors. Using a 10 thousand articles test set the K-Nearest Neighbor reached $R^2 = 0.05$ and a mean squared error of 5101.695. Furthermore, they observed that the performance decreases with more neighbors to consider. Secondly, they conducted a classification experiment based on three classes: A (1 - 20 tweets, about 7.600 articles), B (20 - 100 tweets, about 1.800 articles) and C (100 tweets and more, about 600 articles). For the classification the they used Decision Tree, Bagging, SVM and Naive Bayes on a dataset of 10 thousand articles and a 10-fold cross validation approach. With this setup, they were able to predict the class of an article with an max. accuracy of about 84%. (Bagging 83.96%, J48 Decision Tress 83.75%, SVM 81.54% and Naive Bayes 77.79%) By testing each feature individual, they were able identify the source of the article (news agent) as the strongest predictor for the classification experiment.

Hong et al. [2011] conducted a study about the popularity predictability of twitter posts. Therefore, they crawled Twitter tweets from November until December of 2009. In contrast to Bandari et al. [2012] and Tsagkias et al. [2009], they also gathered graph data spanned by the social network of the users (following paradigm). The dataset contained about 10 million messages from 2.5 million different users. They tried to capture popularity characteristics with a variety of features:

- **Content features:** Term frequency–inverse document frequency (TF-IDF) scores and Latent Dirichlet Allocation (LDA) was used to obtain the topics of the tweets.

- **Graph topological features** They used standard graph metrics such as Pagerank, degree distribution, local clustering coefficient and reciprocal links to describe the characteristics of the social network.

- **Temporal features:** Hong et al. [2011] argue that user may lose interest in topics very fast, so the tried to capture the time dimension of tweets (and retweets) too. They measured the time difference between origin messages and retweets in the same chain, the average time between all retweets, the time between retweet message and the previous tweet and the average time until posts of a specific user get retweeted.

- **Meta-data features:** Furthermore, they captured whether a post has been retweeted (Retweet Before) and how many tweets of a specific users have been retweeted (User Retweet).

They did two separate experiments both with logistic regression. The first was a binary classification, predicting whether a tweet post will be retweeted in the future. The second one was a multi-class classification experiment predicting the volume of retweets for a particular messages before upload. The setup for both experiments used one week of the dataset as training-set and the next week as test-set. This was repeatedly done for all weeks, starting with the first week of the dataset. The results they presented are the average values over four such experiments. The classification performance was evaluated with precision, recall and F1-score. In order to compare the results, they used two baseline models: one model based on "has retweeted or not in the past" feature and the second one based on the TF-IDF topic scores. Best performance was achieved by using the content and meta-data features. By individually testing the performance of every feature, they found that "retweet before are" is a strong predictor for the classification of "retweeted in the future". The multi-class prediction is based on four different classes: 0: No retweets, 1: less than hundred retweets, 2: less than thousand retweets, 3: more than ten thousand retweets. The features that worked the best in the binary classification showed very good results for prediction class zero and three. However, they did not work for classes 1 and two. This indicates that posts with low number of retweets have very different characteristics in terms of prediction. When adding temporal features, classes one and two gained a bit accuracy and adding user activity features improved it further, but only very little.

The so far reviewed literature and methods are dedicated to the popularity predictability of Web content and in particular tweets and online news articles. The following paragraphs cover work that may be the most relevant related work for this master thesis and reveal to what extent recipe popularity could potentially be predicted.

Rokicki, Herder, and Trattner [2017] tried to reveal the bias of online recipe popularity. To the best knowledge of the author, this is the first and only study explicitly researching popularity prediction in the field of online recipes until now. They used data collected from the German online food community Kochbar.de. This dataset with 400 thousand recipes from 2008 till 2014 is the very same than that used for this thesis. Furthermore, it contains 2.7 million comments and 7.7 million ratings assigned to the online recipes. Popularity of online content can be defined in various ways. They used the number of ratings, comments and views of the recipes as a measure of popularity and comment sentiment as measure for appreciation. In their first study, they conducted a statistical analysis of possible popularity biases. They analysed 4 different bias types, which they believed to have an influence on the popularity of recipes:

- **Social bias:** Their data analysis revealed a high correlation between the popularity of a recipes (ratings,comments) and the number of friends of a user. This stays true until a user has uploaded 300 recipes. They call this the completion of the reputation. The number of views of a recipe, remains stable during this period. Furthermore, friends are responsible for about 50% of all ratings and write longer comments in comparison to other users. Female friends also tend to comment in a more positive fashion, than male friends. Figure 2.2 shows the spearmanś rank correlation matrix of the social bias variables.

- **Gender bias:** Due to the results of the social bias section, which revealed differences in gender behavior, Rokicki, Herder, and Trattner [2017] expected that user of the same gender rate each other more. However, this is not the case. Recipes uploaded by male users are more rated by females. Furthermore, female write longer comments on men recipes and write longer comments than men in general. On the other hand, man do write longer comments on recipes uploaded by men. In terms of sentiment, female comment more positive in general and on same gender recipes. Complementary to this, men write more positive on female recipes.

- **Temporal bias:** They found differences in terms of ratings, comments and sentiment. The most recipes are uploaded on Sunday. Also most ratings and comments are written on Sunday. In terms of comment sentiment, users tend to be more positive on Sundays and Mondays. Over the course of a year, they found significant sentiment variations with a peak in summer (most positive).

- **Editorial bias:** Mechanisms such as "recommended recipes" or "recipes of the week", recommend recipes that are selected by the editorial staff for different (unspecified) reasons. Such a bias can of course be influential for the popularity and appreciation of recipes. The analysis showed that the editorial bias is strong. Recipes reach a lot of ratings after being picked as recipes of the week. Furthermore, the number of ratings falls much slower than that of not featured recipes.

  In a regression analysis, they used 3 different regression models and multiple features in order to find the best predictive recipe characteristics. Therefore, they used the statistical computing engine R and dropped stepwise individual features. They predicted the mean comment sentiment with the ordinary least squares (OLS) method, number of ratings and the number of views with Poisson models. They found that for example "being overly productive" doesn't bring advantage in terms of appreciation. Furthermore, author popularity features such as "number of friends" or "number of guest book entries" have correlations with the recipe popularity.

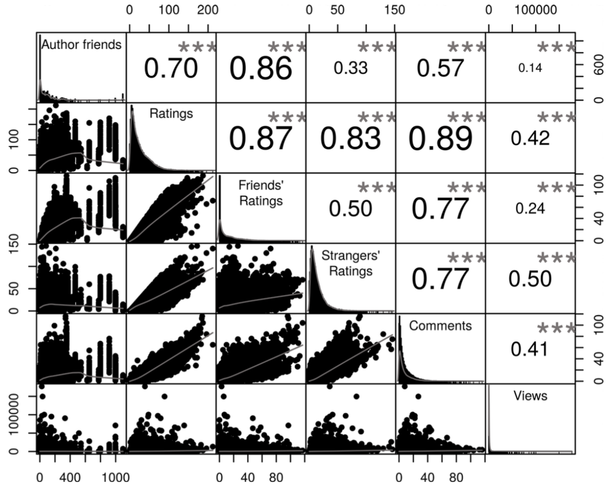


Figure 2.2: The spearman's rank correlation matrix, showing that the number of ratings strongly correlates with ratings from friends. Taken from Rokicki, Herder, and Trattner [2017].

The study of Elsweiler et al. [2017] is the second one in the context of popularity prediction of online recipes. However, it does not actualy predict the popularity, but exploits the popularity of recipes to nudge people towards more healthier food. Their idea was to find very similar pairs of a popular and a more healthy recipe and trick users into choosing the healthier one over the unhealthy one.

They defined 5 research questions in order to solve this problem:

- **RQ1:** Is it possible to replace unhealthy recipes with similar healthy ones based on online food databases?
- **RQ2:** Can users differentiate healthy from unhealthy recipes?
- **RQ3:** To what extend do cues about the healthiness of a recipe influence this estimation?
- **RQ4:** What are the biases that influence the recipe choices of users?
- **RQ5:** Can these biases be exploited to trick people into healthier recipe selection?

The experiments of this master thesis are based on the same dataset as the study of Elsweiler et al. [2017] does. It contains about 240.000 recipes from the online food community Allrecipes.com. After preprocessing and filtering, about 58.000 recipes were accepted for the experiments. They only used recipes that had sufficient nutrient information. In order to measure the healthiness of recipes, they following the approach of Howard et al. [2012] and calculated the FSA score of recipes, based on the nutritional information available.

For RQ1 they calculated the cosine similarity of recipes based on their assigned ingredients. They replace unhealthy recipes with healthier ones that have comparable ratings. In order to test this approach, they conducted a survey with 107 participants, asking them to identify the recipe with the highest fat content. The participants were not able to reliable distinguish between the recipe pairs. They answered RQ2 and RQ3 by analysing the cues that can mislead users judgement about healthiness. The experiments showed that a lack of information, in terms of recipe title or recipe image and lack of knowledge about healthy food can mislead user's to guess wrong fat content of food. This paper is primarily analysed for this Section because of their approach for RQ4. They used machine learning techniques to estimate which recipe of a pair (healthy, unhealthy) will be selected by a user and to identify the characteristics that influenced this decision. Therefore, they derived 96 features that relate to the different cue types, found in the previous experiments. These features were assigned to 5 feature sets:

- **Title features:** Since they observed that recipe titles can mislead users judgment, they extracted features such as length, entropy or words appearing in a dictionary, in order to capture this bias.
- **Image features:** Also, image features seemed to influence user recipe choices. Because of this, they derived features that describe low level characteristics of images such as sharpness, contrast or brigthness.
- **Ingredient features:** They also derived ingredient related features such as the number of ingredients used in a recipe or the text features of the ingredient name.
- **Popularity & Appreciation features:** Since the replacement of similar recipe only will work if both recipes have comparable popularity and appreciation, they derived features such as average rating or the sentiment of comments, in order to describe these biases.
- **Nutrition features:** They also extracted features describing the nutritional values, such as calories, fat or sugar of a recipe.

The regression experiment was performed with the R statistical computing engine. Furthermore, the Weka machine learning suite was used for feature selection, measured via Information Gain (IG), in order to reduce the feature space. Table 2.2 shows the result of the IG feature selection. They used 3 different algorithms: RandomForest, Logistic Regression and Naive Bayes and evaluated the results with a 10-fold cross validation approach. Besides the feature sets described above, they also tested all features combined and the best ten features, identified by IG, only. The response variable of this experiment was a binary indicator that states true when the second recipe of a pair, will be chosen over the first one. Additionally, they conducted multiple classification experiments with varying test and training datasets. The results show that the Random Forest classifier has the best performance (about 66%, using 10.000 random recipes) and the best working feature set was the image set. Like in the previous studies about the cues that influence recipe selection, this experiment showed that the title features are no good predictors, because they are often misleading. When using all recipe instances, they even reached about 84% accuracy. Again, the image features were the best.

Table 2.2: This Table shows the 10 best features, for each study, identified by the Information Gain (IG) feature selection. Taken from Elsweiler et al. [2017]

	Study 1		Study 2		Rand. Sample (rating)	
Rank	IG	Feature	IG	Feature	IG	Feature
1	.0933	IMG:contrast1	.0743	NUT:fat1	.1018	POP:sent2
2	.0829	IMG:brigthness1	.0634	IMG:contrast2	.1016	POP:sent1
3	.0719	IMG:entropy1	.0573	IMG:colorfullness1	.0679	IMG:colorfullness1
4	.0707	POP:rating2	.0568	NUT:cal1	.0609	NUT:fat2
5	.0703	IMG:entropy2	.0542	NUT:satfat1	.0605	NUT:cal1
6	.065	POP:sent2	.0512	NUT:fat2	.0562	POP:book1
7	.0612	POP:book2	.0484	NUT:salt2	.0549	POP:book2
8	.0568	NUT:cal2	.0454	IMG:entropy1	.0430	IMG:sharpness1
9	.0551	IMG:colorfullness2	.0417	ING:charCount2	.0361	POP:ratings2
10	.055	POP:ratings1	.0390	IMG:entropy2	.0344	NUT:satfat2

The overall findings of the classification experiments are that users tend to choose recipes with more fat and recipes that are popular among others and choice also seems to be visually biased. Finally, they successfully tried to nudge people into selecting more healthy food. This was done with the findings of RQ5 and the features that seemed to influence the food choice. This experiment, which answered RQ6, was tested via a survey with 135 participants.

2.2 Studies Employing the Web as Mean to Investigate Human Food Behavior

There is an increasing trend of people that use the Internet to help them decide what to cook. The reasons for that are manifold. It could partly be explained with the ever rising amount of Internet users and devices, and maybe also with the trend that more and more millennials like to cook. Research [Cooper, 2015] showed that the 25- to 34-year-olds like the preparation of the meal as much as eating it. 59% of these people use the Internet and their devices as cooking books. This could be an indication of the resurrection of home cooking. However, many of these recipes can be found on online food websites: Social networks for accessing, searching and sharing recipes to the Internet. Millions of hobby cooks gather in these food communities and interact with others that have the same interests. One of the most popular in the US is Allrecipes.com. It has about 7 Million subscribers, 750.000 uploaded recipes and 180 million recipe views [2]. Besides that, here are many other food websites such as Kochbar.de, Ichkoche.at, Chefkoch.de (all mainly German) and Food.com or Cucumbertown, to mention just a few. This online recipe trend has recently become a topic of interest for researchers and is now an active field with a growing body of work. This Section presents selected work that uses this food platforms to understand the dietary behavior, social and cultural influence or health related issues within these communities. Furthermore it details work on simmilar studies based on other online data sources like Twitter.com or search logs. Some of the papers have significantly influenced this Thesis in terms of relevance and domain.

[2] http://expandedramblings.com/index.php/business-directory/38543/allrecipes-stats-and-facts/

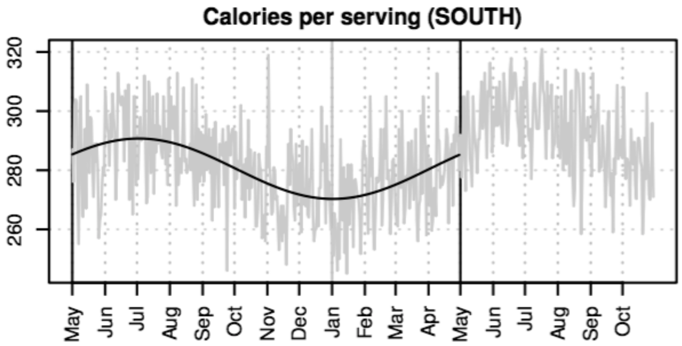


Figure 2.3: Example chart of the periodic fluctuations of calorific food content (Southern Hemisphere). Chart by West et al. [2013].

Studies employing online data to study aspects of human nutrition is a relatively new field. Compared to traditional research in nutrition science it has many advantages. These studies are less intrusive or biased than questionnaires, have higher population sizes and can be scaled globally [De Choudhury et al., 2016]. Arguably they have also drawbacks, for example the assumption that searching for a specific recipe counts as cooked and consumed. However, this Section will present successful studies and highlight the importance of this field.

In their spatiotemporal analysis of Web search and browsing logs, West et al. [2013] argued that nutrition is an elementary part of human health. Therefore, it would be advantageous for the public health, to understand population wide dietary behaviors. For their study, they used anonymized weblog data of an 18 month period, from May 2011 until October 2012. The data was collected via the popular bing browser search plugin and contained billions of page views and search queries. Based on this, West et al. conducted three studies. First, they used the online recipe search behavior of the users and revealed periodic variation in terms of recipe and ingredient preferences. They found large yearly and weekly components as well as regional differences within the data. The time analysis was done via a discrete fourier transformation (DTF) on the time component of the data. Figure 2.3 shows an example of such a fluctuation. The second study concentrated on understanding changing search behavior of users. They were able to identify users that had decided to lose weight, based on different search patterns within a couple of weeks. In the third and last study, they found correlation between the average sodium content of online recipes searched and the number of patients admitted to the emergency department of a hospital in Washingtion DC with a chief complaint linked to congestive heart failure (CHF).

Abbar et al. [2015] conducted a study on Twitter tweets and successfully answered the question if this would reveal dietary behavior. They argued that social media could be predestinated for such purposes, because users upload permanently all aspects of their daily lives and eating habits. Like West et al., Abbar et al. also emphasized the integral part of food in our cultures and lives as well for

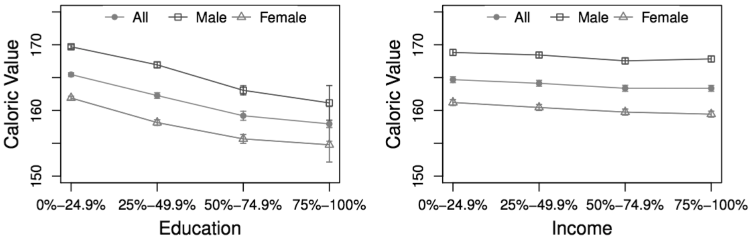


Figure 2.4: The chart shows the relevance of income and education on the caloric value of foods consumed. Chart by Abbar et al. [2015].

the healthiness. For the study, they collected 892.000 tweets that contained a predefined food-related set of keywords. That way it was possible to use the tweeted dining experiences of 210.000 distinct users. Furthermore, they linked this dietary data to the user's interests, demographics and social networks. To estimate the calorific content of tweets, they searched nutritional information websites for a specific keyword (e.g. pizza) and averaged the associated values per serving. In the first study, they found correlations between the calorific value of tweeted food and the state-wide obesity rates, measured by the Centers for Disease Control and Prevention (CDC). The correlation was 77% across 50 US states. They used these results to build prediction models for obesity rates based on the combination of demographic features and food names mentioned in tweets. With this approach, they were able to gain better results as previous CHI (ACM Conference on Human Factors in Computing Systems) studies likes the one of Culotta [2013] which were also using twitter as data source. In a second study, they linked this data to societal and economical facts such as education and income. Showing that a higher educated people, for example, with a bachelors degree or above, tweet or eat significant less unhealthy in the terms of calories. Figure 2.4 shows the relationship between these factors. Furthermore, they tried to shed light on the discussion about the social nature of obesity. Using two social networks, namely a friendship network (approx. 84.000 users) and a mention network (approx. 85.000 users). The friendship graph was spanned by the "following" paradigm of twitter, whereas the mention network captures the mentioning of a user. The similarity was measured via the jaccard similarity of the tweeted foods. They found that it is more likely that friends share similar interests toward food.

Fried et al. [2014] followed an similary approach as Abbar et al. [2015]. They studied the potential of Twitter posts for food related population characteristics prediction. Comparably, the data contained only tweets related to hashtags such as "#dinner" "#breakfast" "#lunch" "#brunch" "#snack" "#meal" and "#supper", resulting in 3,5 million tweets from a period of October 2013 until May 2014. They successfully implemented several prediction tasks based on 30 million words of the tweets and their models were outperforming most baseline studies. These predictive models are based on multiple text features and predicted successfully the locations (city, region, state) of tweet authors, as well as state-level characteristics such as overweight, diabetes and political leaning (party preferences). With a certain combination of features, they predicted state-level overweight with an

Table 2.3: The 20 best working features for two of the prediction classes. The plus symbol indicates features that work for classifying "having diabetes". This Table was taken from Fried et al. [2014]

Class	Highest-weighted features
diabetes: +	Mexican (mexican, tacos, burrito), American Diet (chicken, baked, beans, fried), #food, After Work (time, home, after, work), #pdx, my, lol, #fresh, Delicious (foodporn, yummy, yum), #fun, morning, special, good, cafe, #nola, fried, bacon, #cooking, all, beans
diabetes: -	#dessert, Turkish (turkish, kebab, istanbul), #foodporn, #paleo, #meal, Paleo Diet (paleo, chicken, healthy), i, Giveaway (win, competition, enter), I, You (i, my, you, your), your, new, today, #restaurant, Japanese (ramen, japanese, noodles), some, jerk, #tapas, more, Healthy DIY (salad, chicken, recipe), You, We (you, we, your, us)

accuracy of 80% and diabetes with 68%. They used two sets of features: On the one hand lexical features, such as food related keywords from food glossaries and on the other hand on topic modeling. In order to infer topics from tweets, Latent Dirichlet Allocation (LDA) and an unsupervised learning algorithm was chosen. LDA is a generative statistics model often used to describe hidden relations between different text documents. Such a document can be seen as a mixture of hidden sub-topics, called latent topics. Each topic represents a probability distribution over words. They used the MALLET language toolkit [3] for the training task and predefined the resulting number of topics as 200. In the next step, they labeled the automatic found topics by hand, for example "American diet", based on the words describing it. The lexical and topical features were than fed to a support vector machine (SVM) classification algorithm. Table 2.3 shows the 20 best working keywords, hashtags and LDA topics for prediction (italic) from the classification experiment on diabetes. Furthermore, they used this data and created multiple visualizations such as geo-referenced heatmaps, wordclouds or temporal histograms. This allows to discover complex global patterns in the terms of food consumption.

De Choudhury et al. [2016], like others before, pointed out that social media has become a promising data source for public health studies and in particular for analysing inequalities in food and health access. This study focuses on census tracts that are categorized as "food deserts". These geographic regions, defined for census purpose, are indicated by inadequate access to affordable healthy food. Furthermore, they are known to be associated with poor diet and diet-related health outcomes such as obesity, diabetes and cardiovascular diseases. Considering this, there is an obvious public interest in precisely identifying such regions and the challenges they are facing. The authors argued that former studies on food deserts mainly relied on surveys, self-reported information and lacked thoroughly conducted research methods and sufficient population sizes. This study tries to tackle the mentioned problems with the help of the social media platform Insta-

[3] http://mallet.cs.umass.edu/

Table 2.4: Example posts from Instagram. User often assign the tags to describe a meal. Taken from De Choudhury et al. [2016].

Post	Post tags	Canonical name(s)	Energy	Sugar	Fat	Chol.	Fiber	Protein
	butter, cakes, peanut, jelly, kellylou- cakes, tea, decorating, cupcake	butter, cake, peanut, jelly, tea, cupcake	436,26	95,9	25,3	256	3,08	9,401
	healthyfood, meal, goodfood, foodgasm, carrots, vitamin, cucum- ber, veggies, foodisfuel, corns, lime, beetroot, nofilter, salad, potato, instafood, eatcleanmenu, apples, rich, fruits	carrot, cucumber, corn, lime, beetroot, salad, potato, apple	206,01	81,1	16,8	86	79	25,4

gram[4]. Instagram is a fast-growing social network where users can upload photos and share it with the world. The dataset was collected via the official API of Instagram, which provides methods for fetching public images and the corresponding metadata (e.g, hashtags to describe food content). They downloaded 14 million posts from 8 million different users. By filtering with a popular online food vocabulary, they were able to identify 3 million food related posts. In the first study, they estimated the nutrients of the food mentioned with hashtags. Therefore, they used the National Nutrient Database for Standard Reference of the US Department of Agriculture (USDA) that contains nutrient data for over 8000 food-items. Table 2.4 shows two example posts with the associated tags and the approximated nutrients. They observed that food deserts have a 5-17% higher consumption in fat, cholesterol and sugar. Furthermore, food deserts have low fruit and vegetable consumptions. In a second study, they were able to predict the USDA defined nutritional and food deprivation status of different census tracts. They used a LDA topic model, revealing the ingestion language of food deserts. This way, >80% accuracy of predicting diet behavior and food deprivation status of food deserts were achieved.

Wagner and Aiello [2015] understand food not only as the important factor for living but also as an identity conveying medium for modern societies. They did a quantitative study about gender-differences in the terms of food and media transported stereotypes based on social media data. Flickr[5], a online social network for sharing images and small videos, was crawled for this purpose. Wagner and Aiello argued that this approach could reveal interesting factors, since users recreate their identity by sharing content online and the so expose information very likely correlates with their real interessts and likes. Primarily, they wanted to answer the questions if gender specific upload behavior can be observed and what the driving factors are. In order to shed light on that, they collected about 15 million Flickr images from 1 million users in the time of 2005 until 2014.

[4] https://www.instagram.com/?hl=de
[5] https://www.flickr.com/

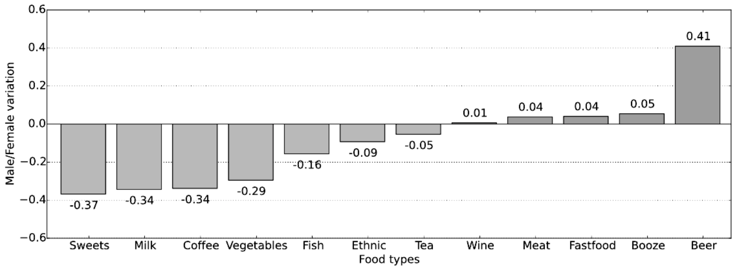


Figure 2.5: Food category image upload differences between the two genders. 37% of the women like to upload pictures with sweets, whereas 41% of the men upload content with beer. Taken from Wagner and Aiello [2015].

41% of the users were female. They filtered out non-food related posts by using a online vocabulary that contained popular food words. Posts with at least one food related tag (metadata) and public gender information of the user was kept and anonymized. They found statistically evidence that there are specific food types that are related and will be posted primarily of either one of the genders. For example, 24% of the men will post at least once a picture with beer. This is true only for 17% of the females. Considering the ratio, this means that beer is 41% more popular among men. Figure 2.5 shows the food category preferences of the genders based on the Flickr data. However, they argued that popularity and preferences are not the only driving factors (e.g., culture, cost or health, availability, etc.) for uploading images of specific food types. It can also be explained with medial presentation of food consumption. In a second study they collected the top 100 images returned by a search engine when querying for terms such as "eating meat" or "eating fish". In the next step, crowd workers were asked to determine who is more likely to eat what is shown on the images (male or female, adult or younger). This survey showed that alcohol seems to be more popular among men although it is frequently promoted with women. Whereas, milk and fast-food can be seen as gender neutral and sweets, milk, coffee are more female to the media. In summary, Wagner and Aiello argue that their approach can not completely substitute classical surveys about dietary preferences and food consumption. However, it highlights the potentials of this technique.

Chunara et al. [2013] try to reveal the hidden relationship of social networks and obesity prevalence. They conducted a cross-sectional study to explore the predictive capabilities of user uploaded or posted interests on Facebook[6] to the obesity prevalence in the United states. Facebook is a very popular social network with about 1,94 billion users worldwide (March 2017). Users can present themselves by uploading content and creating their profile pages. Furthermore, users can like the postings of others (users, companies, products, tv shows etc.) and express their preferences in that way. Facebook is a big advertisement company. It provides detailed information and numbers for interested companies. For example, the Facebook advertisement platform offers a mechanism to query for target users with specific interests, different age or geographical location, based

[6] https://www.facebook.com/

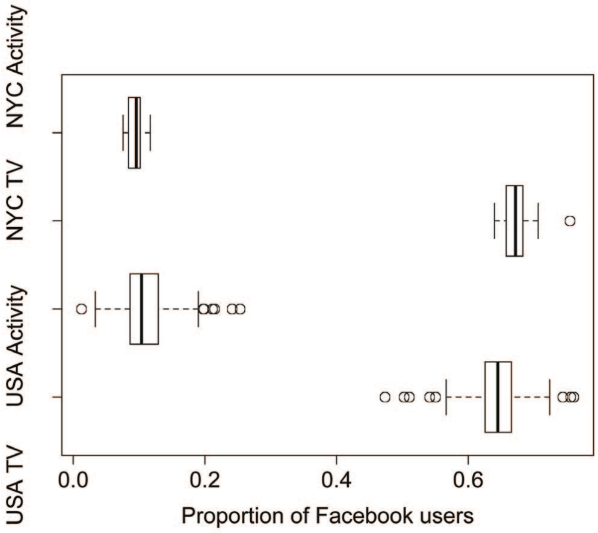


Figure 2.6: The chart shows the proportions of Facebook users that have active-related or non-active-related interests. Taken from Chunara et al. [2013].

on the profile data. Chunara et al. selected those users that have interests in activities, that are either positive or negative related to obesity. For example, "Watching television" can be categorized as a sedentary activity and lead to obesity. On the contrary, "outdoor fitness activities" is an indicator for an active and a healthy life style. Figure 2.6 shows the proportions of either group. They used linear regression and k-fold cross validation to model the activeness and sedentary of the users. K-fold cross validation is primarily used when the dataset is very small and the experiment will lose statistical relevance, when dividing the data in training and test data. Additionally, it helps to derive a better prediction model in terms of accuracy. They mentioned that the 10-fold cross validation error was comparable to the one of a model, developed using the entire dataset. The outcome variables for this experiment were validated with data from two different health control and surveillance systems. First, the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS). A telephone survey based system by the CDC for measuring prevalence of obese and overweight people and primarily conducted on US state-level. Second, the New York City EpiQuery Community Health Survey, also based on an annual telephone survey of the Community Health Survey (CHS), exclusively conducted in New York city. Both systems target to provide crucial information for identifying, tracking health related problems and for evaluation of public health countermeasures. However, Chunara et al. point out that they were able to gather datasets, many times bigger then the BRFSS or CHS data. They found that Facebook users with activity-related interests have a 12% lower predicted prevalence of obesity across the United States of America and about 7.2% in New York areas. On the other hand, a 27.5% prevalence of obesity in New York City neighborhoods with interests in activities such as watching tv. In summary, their experiment showed significantly association of non-active hobbies to obesity. However, they argued that more research need to be made, in order to fully understand this connections.

In their study about human food preferences, Wagner, Singer, et al. [2014] where one of the first to analyse data from online recipe platforms. They argued that groups do sometimes define themselves by the characteristics of their food. For instance after migrations, such as the European immigration to the USA, the cuisines and food preferences survived, but the languages did not. Additionally, they mentioned that this domain is not only relevant for the anthropological or sociological research, but also for the field of food related health issues. Wagner, Singer, et al. [2014] wrote:

> "...food preferences are amongst others manifestations of social, cultural and economic forces that influence the way we view, prepare and consume food"

For their research, they used server logs from ichkoche.at[7], the largest online recipe platform in Austria at that time. They see the recipe access behavior of the users as a proxy for food preferences. From these logs, they extracted how frequently a user access different recipes and the ip addresses, which reveal the approximated location of a user. They concentrated only on user from Austria and German-speaking countries such as Germany and Switzerland. Furthermore, the logs contained information about the ingredients used in recipes. All in all, they had a dataset of 184 thousand recipes which were visited about 24 million times from 1,6 thousand varying regions. In order to answer their questions, they focused on 4 different dimensions of food preferences:

1. Recipe preferences: What are the properties of food popularity and how general can they be described?
2. Ingredient preferences: Do the regions have diverging ingredients distribution and what can they reveal about the users?
3. Spatial food preferences: What is the impact of geographical distance on food preferences? Are close regions more similar?
4. Temporal food preferences: Do food preferences change in time and how often?

They found that the popularity function of recipes and ingredients are heavy tailed distributions, following a truncated power law function. In Austria and Germany, recipe preferences of close regions tend to be more similar. This was measured with cosine similarity of the recipe and ingredient frequency vectors of the regions. Furthermore, the ingredients preferences are similar for Austrian and German regions, whether they are close or far away. However, this cannot be observed for Switzerland, which is a small but very manifold country. When taking all regions of all countries into account, they show a tendency for similar near regions. The temporal dimension was measured with a rank biased overlap of the ingredient frequency vectors of the regions. This approach focuses on the top ranked ingredients. This makes sense, since the top ranked ones can be seen as the important ingredients of the region. They found that there are changing preferences during the week. This can be explained mainly because of ingredient

[7] https://www.ichkoche.at/

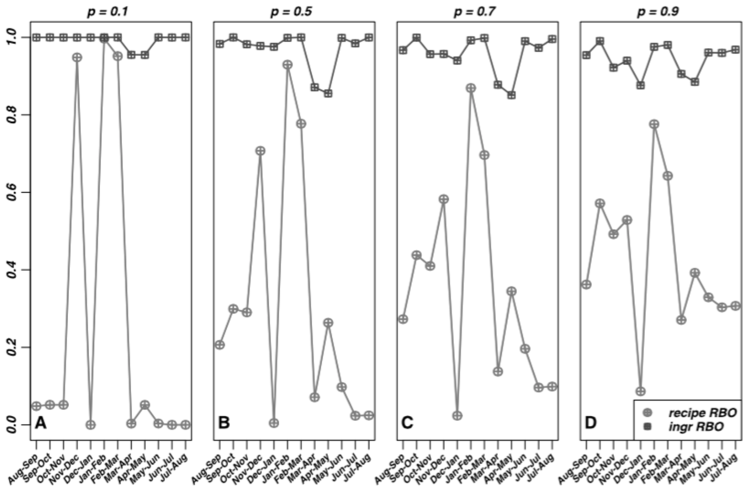


Figure 2.7: This chart shows the recipe and ingredient changes measured with rank biased overlap. Taken from Wagner, Singer, et al. [2014].

preferences on specific weekdays, for example meat on weekends or fish on Friday. Furthermore, selected ingredients showed a prevalence for specific seasons ("asparagus season") and weekdays. Figure 2.7 shows recipe and ingredient preference fluctuation over the course of a year. This can be explained with real world phenomena, as mentioned before.

Trattner, Elsweiler, and Howard [2017] made a study about the healthiness of online sourced recipes in comparison to ready meals and recipes from cooking books. They pointed out that many studies found relations between health issues and as bad considered nutritional behavior. Two diet programs, namely ChooseMyPlate (US) and Change4Life (UK) try to encourage and promote home cooking, since it is seen as healthier. However, Trattner et al. argue that the healthiness, regardless of self cooked or ready meal, depends on what is cooked and how. Therefore, they made a statistical analysis of three different meal types that are regularly eaten in modern societies. These three are Internet sourced recipes, ready meals and, meals from popular cooking books. They compared 100 recipes from books, 100 ready meals and online recipes from the online food platform Allrecipes.com. They crawled 5,237 online recipes from the years between 2000 and 2010. For this comparison, only main dishes with enough information of the nutritional values (carbohydrates, sodium, energy content, fat, etc.) were taken. Allrecipes.com estimates this values via the ESHA research database, based on the ingredients that user can assign to their recipes per free form text. Furthermore, they calculated the expected healthiness of all recipes with two known international standards. One the one hand the guidelines of the World Health Organization (WHO), on the other hand the "traffic light" system of the FSA (UK). The FSA system is mainly used for nutrient labeling on food packages. They derived two, so called, health scores from these standards. The WHO score was claculated with the content of the 7 most important nutrients, that should be in a daily meal plan. The score ranges from 0 - 7 (0 for unhealthy, 7 for all criteria

Table 2.5: The Table shows the three different meal types and the percentage of recipes that meet WHO criterias. Taken from Trattner, Elsweiler, and Howard [2017].

	Percentage (total)		
Number of WHO criteria fulfilled	Internet recipes (N = 5,237)	TV chef recipes (N = 100)	Ready meals (N = 100)
0	5.94 (311)	7	1
1	46.27 (2,423)	42	27
2	27.63 (1,447)	28	30
3	11.34 (594)	14	24
4	4.98 (261)	8	13
5	3.04 (159)	1	4
6	0.69 (36)	0	1
7	0.11 (6)	0	0

fulfilled). The FSA score considers the amount 4 major macronutrients. Green is considered healthy and red unhealthy. In their first study, they found that only 6 online recipes fulfilled the WHO guidlines totally. Table 2.5 shows how much recipes of either category meet a WHO criteria. The recipes from Allrecipes.com were evaluated to be less healthy overall, because they often do not meet the norm for fat, saturated fat and fiber content. However, these recipes often do meet the considered protein amount. Another observation was, that recipes from cooking books are the healthiest by sodium amount, followed by Internet recipes and ready meals. When considering the sugar amount, recipes from cooking books and Internet equally often meet the criteria, but ready meals are the best of all three. The second study about the temporal dimension of these findings shows,that the results do not vary over time. Trattner et al. concluded that Internet recipes may not be as healthy as expected, but there are limitations to their approach. There are very likely variations in what user really consume. They may do not follow the exact ingredients and instructions mentioned by recipes. Furthermore, there is the possibility of variations of the nutrient values stated by the of the ready meals labels and by the nutrient calculation approach of Allrecipes.com

Kusmierczyk, Trattner, et al. [2015a] research in the field of food and recipe innovation of online food communities. They argue that innovation is important for the long-term success of restaurants and chefs. However, there has not been much research that focused on the virtual perspective of this domain. For their study, they used a dataset collected from Kochbar.de, which contained more than 400 thousand recipes from 2008 until 2014. Furthermore, preparation instructions and categories are assigned to the recipes. They counted 230 distinct recipe categories, 200 thousand different users and about 7 million recipe ratings. Only about 5000 users are regularly uploading recipes (more than 10 recipes). Kusmierczyk, Trattner, et al. [2015a] chose Kochbar.de mainly because of their various metadata and additional recipe information such as ingredients, nutrient values etc. This allowed them to use an approach that heavily relied on ingredient combinations of recipes. A major problem was that the ingredients are only available as free form text. This made pre-processing and filtering necessary. They have come up with a simple approach, mainly based on statistical filtering. They only used ingredient names that occurred more than 100 times. Ingredients occurring less than 200 times were replaced with more popular ones. For example, "the glass of salted

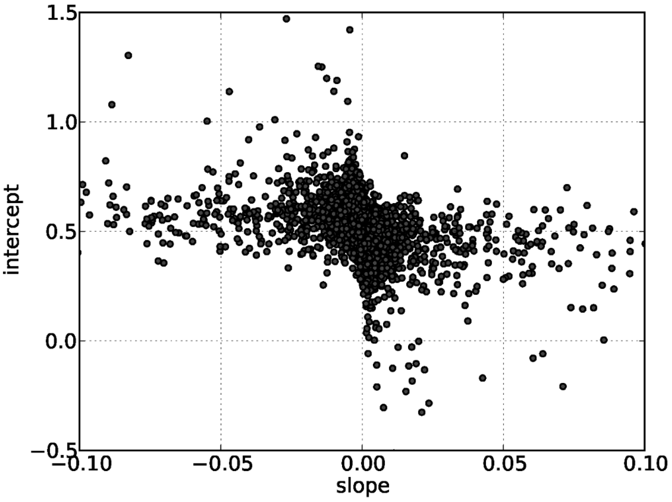


Figure 2.8: The graph shows the parameters of the linear regression of the user innovation. The vast majority of the users do not change really in time. Taken from Kusmierczyk, Trattner, et al. [2015a].

water" got replaced with "water" and "salt". This way they were able to remove disambiguation and identified 2208 distinct ingredients (334 thousand before filtering). Their first study concentrated on exploring community patterns. They tried to measure innovation and complexity with 3 different features. Two of them used entropy and conditional entropy, respectively. The third one was an innovation factor metric based on jaccards similarity. All metrics consider the recipes ingredients for comparison. They found that although the number of ingredients used stays constant, the innovation within the community increases continously. This can be explained by the hypothesis, that user keep combining the known ingredients to new and novel recipes. However, the grow ratio of this observation is falling, meaning that at some point in the future, the innovation will fade out. The innovation in the community also has seasonal and time depending patterns. It slightly varies over the course of a year with two peaks, at the beginning of the year and after the summer. They argue that this can be explained with the natural assumption of people starting creative in the new year or when their back from holidays. The categorical dimension shows that, for example meat dishes have one of the highest innovation factors. In the second study, they analyzed the innovation patterns on user level. They filtered out users with less than 10 recipes, arguing that this way the results are more reliable. This resulted in a user set of about 5000. The main observation was that two different types of users exits. User with a smaller innovation factor and the more innovative ones. They fitted a linear regression to each user innovation factors (similar to the one of the recipes, but on user level) over time. Finding that the innovation of most users does not change over the years. Figure 2.8 gives an visual impression of this observation. Furthermore, they explored the factors that drive this innovation. The best explaining feature is the location of the user (measured with information gain). They mentioned that this is an unexpected finding and should be investigated in detail.

Rokicki, Herder, Kuśmierczyk, et al. [2016] did a study about the differences in cooking behavior of men and women. They argue, that in western culture, such strong differences mostly have gone. However, there are still this stereotypes such as women are responsible for every day and men love meat and hefty food. In order to investigate these differences they stated 6 hypotheses, that need to be checked:

- Men cook better
- Men try to impress with cooking
- Mean cook meat and women mostly sweet recipes
- Women to not season that much
- Men use tools and gadgets more often
- Men are more cutting edge regarding new recipes

As the source of data, they choose the online recipe platform Kochbar.de. They crawled 400 thousand recipes that where published between 2008 and 2014. These recipes where uploaded by about 200 thousand users. 2,7 million comments and 7,7 million ratings where assigned to the recipes. In order to proof or reject the hypotheses, they needed the ingredient data of the recipes. When uploading a recipe to Kochbar.de, user can assign ingredients in freeform text. This made it very difficult to automatically process and identify ingredients. Rokicki et al. used the USDA nutrition database to find and match food items. Furthermore, cooking utensils are often mentioned in the preparation instructions of the recipes and identified via a list of 250 cooking utensils from the German version of Wikipedia. Sweet dishes were marked with a the help of a co-occurrence network of all found ingredients. In such a network, ingredients that are considered to be used in e.g. sweet recipes, are clustered together and are topographically nearby. With this approach, the could cluster sweet or hefty ingredients and in turn successfully label 57 thousand recipes containing sweet ingredients. In a first study, they statistically analyzed the differences between the two genders. The results showed that there are indeed still differences:

- Men prepare dishes with more ingredients and longer preperation time
- Women do not as often prepare meat dishes and season more subtle
- Women receive better, but not as much feedback as men

In their second, they tried to classify gender by stereotypical features. This classification experiment was done with the Weka[8] data mining tool. The feature selection of the 88 features were done with Information Gain (IG) and Random Forest (RF) accuracy loss. For the classification experiment, they used RF, Logistic Regression and AdaBoost. The best working features were for example sweet dishes for woman and meat dishes for men. Furthermore, the use of spices and

[8]

url http://www.cs.waikato.ac.nz/ml/weka/

Table 2.6: The Table shows a selection of the 10 best working features of the feature selection process. Taken from Rokicki, Herder, Kuśmierczyk, et al. [2016].

feature name	IG	rank	RF	rank	H
sweet recipes	.058	1	9.325	4	H3
'forms' gadgets	.045	2	16.122	1	H5
spices per recipe	.043	3	11.019	3	H4
'pots & pans' gadgets	.039	4	3.358	18	H5
red meat recipes	.034	5	7.335	6	H3
'coffee & cake' recipes	.027	6	13.567	2	H3
bacon recipes	.025	7	2.260	30	H3
distinct spices count	.023	8	1.376	49	H4
preparation time	.021	9	4.181	12	H2
international category	.018	10	1.521	43	H3

the use of tools. Table 2.6 shows 10 of the 20 best working features and the IG compared to the mean accuracy loss of RF. With this approach, they were able to predict the gender with an accuracy of 75%. Their third study tried to answer how much these approaches can help to improve food recommenders. They successfully tested this gender filtering approach in their experiment and showed that it improves similar recommenders, by restricting recommendations to authors of the same gender. However, they argued that they probably can improve this approach by adding more features such as rating behavior etc.

Kusmierczyk and Nørvåg [2016] tried to reveal the patterns of online recipe titles and enabling practical applications from these results. Their main goal was to understand the relations of title words and the nutritional values of a recipe. They argued that although users primarily interact in the form of text, there is not much information about the associations between textual content and health related issues. For this reason, they did a study on 204 thousand online recipes from Allrecipes.com. In the pre-processing phase, they filtered out recipes with insufficient nutrient information, which resulted in a dataset of about 58 thousand recipes. The recipe titles are short free form text and also needed to be pre-processed. For example, they removed all special characters, numbers and stop words. Because of the free form character (possible ambiguities and misspellings) of the titles, they applied stemming and took only words that occurred more than 2 times. This approach resulted in 4,679 unique words. In their first experiment, they did a statistical analysis of the nutrient value distributions for each word of the recipe titles. The influence of individual food words on nutrients were calculated via information gain (IG). They discovered a correlations between distinct food words and nutrients, as well as correlations among the nutrients itself. Table 2.7 shows the most important words for each nutrient fact. In a second experiment, they combined LDA and linear regression in a novel approach, to create a low level and interpretable model of the findings of the first experiment. The variables for the LDA are ingredients of the recipes, the dependent variables of the linear regression are the nutritional values. Their validation of different approaches showed that, this model works offers the best results. The third and last experiment tried to predict the nutritional values of recipes based on the title words.

Table 2.7: The Table states the most important food words for describing the nutrient values. Taken from Kusmierczyk and Nørvåg [2016].

Nutrient Fact	Important Words
kcal	chicken, cooki, pie, dip, pasta
fat	cooki, pie, chicken, casserol, sausage
carbos	cake, pie, dip, pasta, cooki
proteins	chicken, cooki, cake, chocol, pork
sugars	cake, chocol, pie, cooki, apple
sodium	cooki, chicken, cake, chocol, casserol
cholesterol	chicken, cooki, shrimp, pork, egg

They did this by considering three different information circumstances:

- No information about the recipe
- Only one nutrient is known
- All but one nutrients are known

All circumstances were tested with different heuristics. For example, linear regression or gradient boosted regression trees. The resulting prediction quality was measured with symmetric Mean Absolute Percentage Error (sMAPE). It is interesting to note that the calories of a recipe were determined precise, when other nutrients were known.

Ahn et al. [2011] studied the underlying ingredient combination patterns of cuisines all over the world. Try to solve the question, if there are

> "quantifiable and reproducible principles behind our choice of certain ingredient combinations and avoidance of others?"

They followed the idea of shared flavor compounds of combined ingredients. This is a hypothesis, stating that ingredients, which share flavor components, are more likely to taste well together. For example, there are restaurants that combine white chocolate and caviar based on the organic compound trimethylamine, which is contained in both ingredients. The approach of Ahn et al. [2011] is based on a bipartite graph spanned by ingredients and their flavor components. Most ingredients have 51 of this compounds on average. Such an ingredient-compound network allows them to formulate and test their hypothesis as topological properties. Figure 2.9 gives an visual impression of this bipartite graph. This resulting graph can then be explored for similar ingredients. In order to test their hypothesis, they gathered data from 3 different online recipes websites, namely Allrecipes.com, Epicurious.com (both US) and Menupan.com (Korean). They chose Korean website to avoid a to western biased result. Their dataset consists of 1021 different flavor compounds and 381 distinct ingredients. Each recipe has 8 ingredients on the average assigned to it. In their first experiment, the found statistically evidence that north american and western European cuisine combine more ingredients that

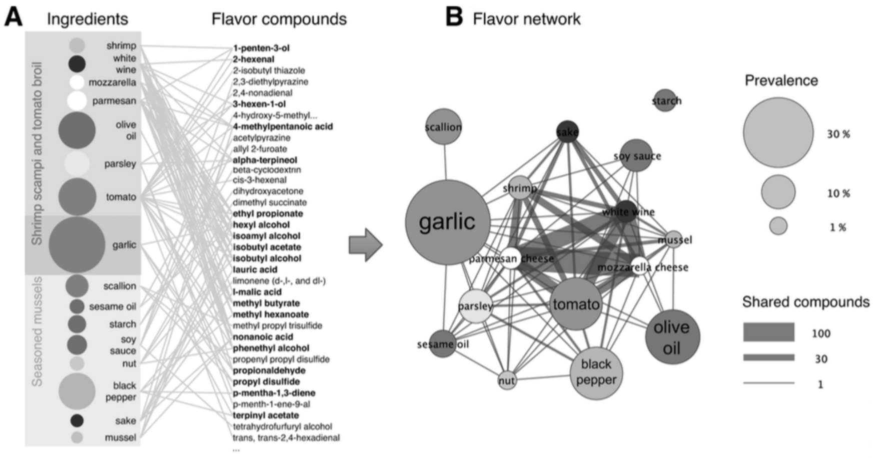


Figure 2.9: The left side of the Figure shows the bipartite graph of ingredients and their assigned flavor compounds. The right side shows the resulting flavor network. Taken from Ahn et al. [2011].

share flavor compounds. This is in contrast to the Asian cuisine, that mainly uses contrary ingredients. In addition, they tested the likelihood that ingredients, which share more compounds, are common in specific cuisines. This experiment confirmed the results of the first experiment. As a next step, they tried to find reasons for their observations. The found that only a few ingredients (but used very often in the specific cuisines) are responsible for that effect. For example, egg, cream, cacao, butter and milk are very popular in the north American cuisine. On the contrary, onion, ginger, pork and chicken are common in the east Asian one. In a third experiment they compared different cuisines and showed that the south European and Latin cuisine are much more similar to the asian cuisine than to the western european. The Latin and south European cuisines also use ingredients that do not share much compounds.

Said and Bellogìn [2014] investigate in their study how social interaction, regarding online recipes, can help to improve possible food recommender systems. They argue that, in contrast to other real-world product recommendations (such as for videos or music etc.), the health aspect is essential. A system with such an influence on the user's health needs to be aware of the potential danger. This stays true, regardless of business outcome for the vendor. Furthermore, they argue that the location of the user also needs to be considered, since there are countries with a higher risk of food related issues. Their study was conducted on a dataset crawled from Allrecipes.com during Octorber 2013. It contained data of 170 thousand users, 54 thousand recipes, 8400 ingredients and 17 million ratings. The health data was collected from County Health Rankings [9] and is mainly focused on obesity data. This dataset contained more than 3400 US counties and health related data, such as the obesity rates per county. In order to conduct their experiment, they need map the users of Allrecipes.com to the counties. A major problem was that users can state their location in freeform text. This made it difficult to auto-

[9] http://www.countyhealthrankings.org/

Table 2.8: This Table shows how often ingredients appear in counties with a high (↑) or low (↓) obesity rate. Taken from Said and Bellogìn [2014].

No.		Salt	Butter	Sugar	Eggs	Flour	Onions	Garlic	Water	Pepper	Milk
1-10	↑ *Obesity*	51.04%	33.72%	30.67%	27.25%	26.14%	23.93%	22.79%	21.96%	20.65%	14.96%
	↓ *Obesity*	55.30%	32.92%	31.01%	26.77%	25.68%	24.86%	27.31%	21.54%	21.42%	13.23%
No.		Vanilla	Olive Oil	Brown Sugar	Chicken	Cinnamon	Parmesan	Baking Soda	Veg. Oil	Cheddar Ch.	Cream Ch.
11-20	↑ *Obesity*	14.85%	14.07%	12.54%	10.20%	9.81%	7.96%	7.89%	7.29%	6.81%	6.79%
	↓ *Obesity*	14.52%	18.04%	12.56%	8.70%	10.00%	8.25%	8.75%	7.41%	5.35%	5.21%

matic map users to their counties They solved this by manual matching and tuning of the dataset. As a first step, they wanted to test the feasibility of their approach. Therefore, they choose only 10 counties for this experiment. Most interesting were these with the lowest and highest obese rates. For the analysis they calculated the ingredients frequencies for each county based on the ratings. They selected the recipe ratings as a proxy for "being cooked/consumed by a user". Then they analysed the top 110 most used ingredients in both county groups. They found, that the groups can in fact be identified by the frequencies of the ingredients used. Table 2.8 shows the occurrence percentage of specific ingredients in either county group. This observation was tested by a t-test on the frequency vectors. Said and Bellogìn [2014] argue that this information can be very useful for personalized food recommenders in the future. Such recommenders could either mitigating recipes that deemed unhealthy or create personalized recipes for a user based on their obesity risks. The authors are well aware of the limitations of their approach. For example, they do not really know how much of the ingredients are used for a meal. Moreover, the identification of ingredients is difficult, because they are also assigned in free form text (ambiguous and non-standardized) by the user. However, they see this early work as a promising start of future research.

2.3 Differences to previous research

In summary, the review of the related work shows that there has been several studies about online food and diatary behavior in general. However, is still little known about why certain online recipes get popular and others do not. None of the reviewed studies actually searched for generally valid characteristics that are relevant for recipes popularity. Additionally, the presented studies only relied on a small number of selected features in order to describe the popularity of recipes. This master thesis tries to close that gap by using two large scale datasets from international food community websites and a large variety of acknowledged or novel features.

Chapter 3

Materials

> *“ Data, I think, is one of the most powerful mechanisms for telling stories. I take a huge pile of data and I try to get it to tell stories. ”*
>
> [Steven Levitt, American economist, born 1967]

The materials Chapter gives insightes into data used by the statistical analysis and predictional experiments. Part one (i) describes Allrecipes.com and part two (ii) highlightes Kochbar.de. The two parts explain why these online food community websites were chosen for this purpose and how the recipes data were collected. Furthermore, they describe the structure of the data and what useful information it provides.

3.1 Allrecipes.com

The first dataset used in this thesis and that shall be described is Allrecipes.com. Since this master thesis tries to find general properties that make recipes popular, data about international and wide spread food databases is needed. One of the most popular online food websites is Allrecipes.com. It is found[1] to be the most popular food community websites by Alexa.com, an Web analytics online service. It has an global websites rank[2] of 885 and 246 inside the USA (July 2017). The data set was obtained by Trattner, Elsweiler, and Howard [2017] during the summer of 2015 and contains recipes uploaded between 1998 and 2015. Trattner, Elsweiler, and Howard [2017] chose the global version of Allrecipes.com, because the British version contained too few recipes for a substantial data analysis. This data set containes over 60.000 recipes published by over 25.000 unique user. The basic statistics can be found in Table 3.1. Since Allrecipes.com is also a social network, the data includes information about the user follower network (a simmilar approach as Twitter.com uses) and the user profiles. Furthermore, data about user comments and ratings of recipes are included. The ratings are displayed as stars and range from 1-5 (Likert scale[3]). Comments and ratings are coupled together, since a user needed to provide both, when judging a recipe. A very important

[1] http://www.ebizmba.com/articles/recipe-websites

[2] https://www.alexa.com/siteinfo/allrecipes.com

[3] https://en.wikipedia.org/wiki/Likert_scale

feature of Allrecipes.com is the nutritional Information of the recipes uploaded. This information gets calculated based on the ingredients by Allrecipes.com via the ESHA Research's nutrient databases[4].Allrecipes.com offers information for 18 different nutrients like calories, fat, protein, fiber and sodium. Furthermore, each recipe is labled with at least with one of the 939 categories like 'Main Dish' or 'Desserts'.

In the following paragraphs, the main views of the Allrecipes.com website get described in a bit more detail. Allrecipes.com consists of 3 major views with various features and informations:

- **Homepage**
  This view acts as the entry point of the website. It displays editorial featured recipes and recommended recipes. A search bar is located at top the view that allows the user to search recipes and recipes with specific ingredients, see Figure 3.1a.

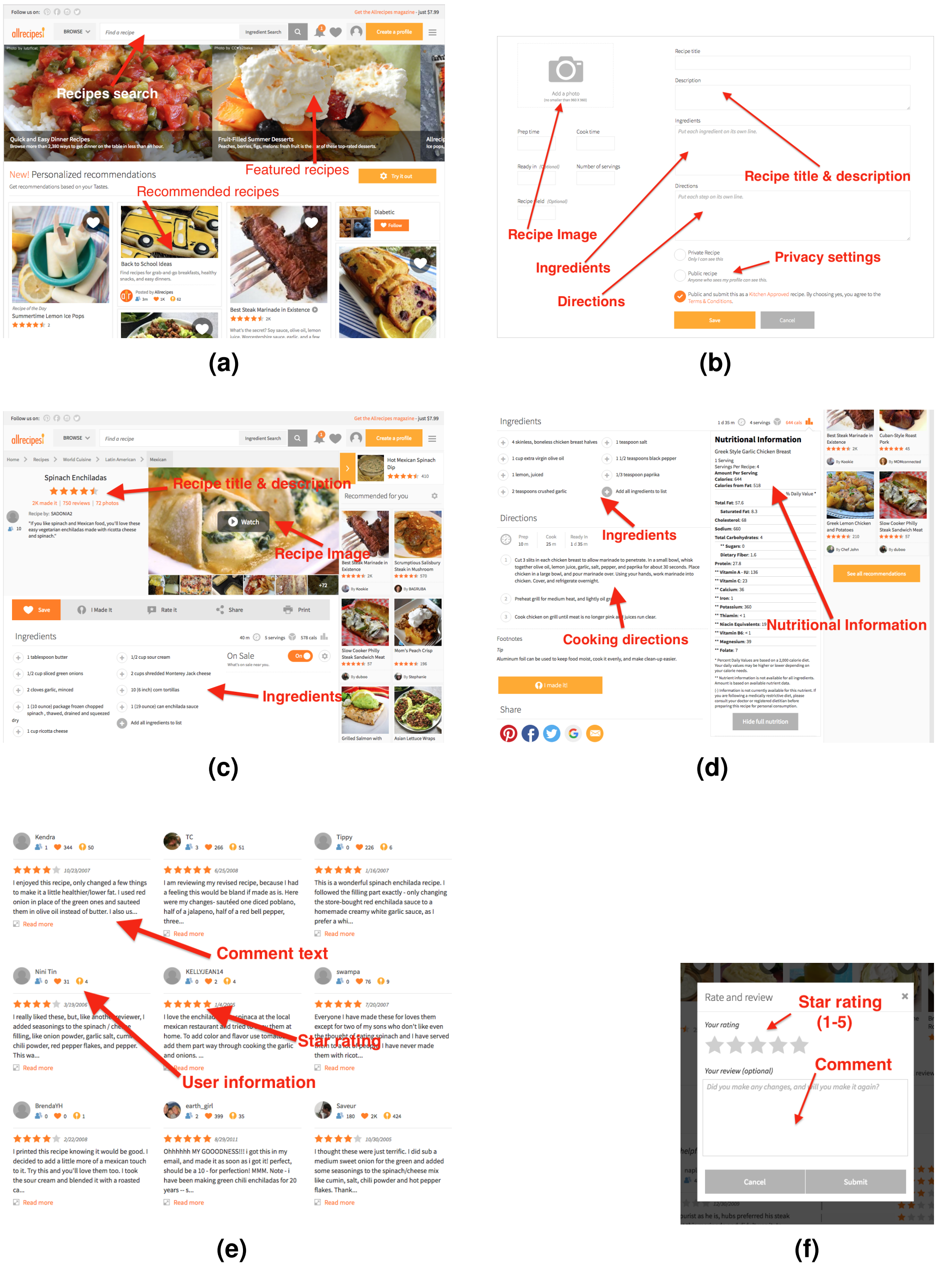
- **Recipe Upload**
  In order to upload a recipe to Allrecipes.com's database, one has to provide multiple recipe parameters. These include the recipes title, a descripion, the preparation steps, a list of ingredients, preparation and cooking time and the number of servings. Also, the privacy level hast to be defined. When submitted as public recipe and "Kitchen Approved", the editorial staff will review the recipe to ensure quality and calcuate the nutritional information based on the ingredients. The picture of the recipe is optional. Furthermore, it is important to note that the ingredients are captured in a free form text with no spellcheck or guiding feature whatsoever. This is the reason for the many misspellings and word variants of the ingredients, found in the data set. See Figure 3.1b

- **Recipe View**
  The central point of Allrecipes.com is of course the recipes view. It features all information about the recipe and allows the users to rate and comment it. Besides the typical recipe ingredients and cooking directions it also shows the nutritional information based on 100g of the prepared meal. As mentioned before, this information is only available for public and reviewed recipes. User comments and ratings are displayed at the bottom of the view. See figures 3.1c, 3.1d and 3.1e. Additionally, there is the "rate and review" button, when pressed it shows a window with the rating feature (based on 5 stars) and a text area for the comment. Important to note is that this screenshots were made in August of 2017. However, the data set was cralwed in 2015 and adding a comment, when leaving a rating, was mandatory. This is not the case anymore. See Figure 3.1f.

3.2 Kochbar.de

The second data set, this thesis relies on, is from the online food website Kochbar.de. It was crawled by Kusmierczyk, Trattner, et al. [2015a] in 2014. This food community is considered as one of the largest in Europe. Kochbar.de hosts about half a

[4] http://dish.allrecipes.com/customer-service/nutrition-information/



(a) (b)

(c) (d)

(e) (f)

Figure 3.1: The main views of Allrecipes.com. (a) shows the homepage with recommended recipes and search functionality (b) presents the recipe upload form (c) shows the top of the recipe detail view with the image and ingredients (d) shows the bottom of the recipe detail view with the nutritional information and the cookig directions (e) presents the comments section, located at the bottom most part of the recipe view (f) highlights the commenting and rating window.

Table 3.1: Basic statistics of the Allrecipes.com and Kochbar.de dataset.

Feature	Allrecipes.com	Kochbar.de
Total published recipes	60,983	405,864
Recipes containing nutrition information	58,263	309,360
Users with published recipes	25,037	18,212
Recipes rated/commented	46,713†	400,155 / 360,668
Ratings/comments	1,032,226†	7,796,004 / 2,751,820
Users who provided ratings/comments	125,762†	19,444 / 21,951
Distinct Ingredients	3842	2028
Distinct recipe categories	939	246

† Number of ratings and comments are the same per recipe in Allrecipes.com as one can only rate, if a comment is provided.

million recipes by users from all over the world. Furthermore, it provides rich social network community features and meta-data like a user friendship paradigms and cooking-groups. Kusmierczyk, Trattner, et al. [2015a] crawled over 400.000 recipes uploaded in the years of 2008 until 2014. These recipes were uploaded by about 200.000 unique users. Kochbar.de has very simmilar features and recipe parameters such as Allrecipes.com. Comparatively, recipes are also labeled with at least one of the 230 distinct categories, basic cooking instrutions, ingredients and nutrient informations are provided. In contrast to Allrecipes.com, the nutrient informations are calculated via the German Nutrient Data Base[5]. However, Kochbar.de does not present as many nutrients as Allrecipes.com. They only offer data for calories, protein, carbohydrates and fat. However, the data set contains information of about 1 million comments and 7 million ratings, provided by the community users. Ratings are also display as stars, ranging from 1-5. Basic data set statistics can also be found in Table 3.1. Generally speaking, the two food community websites are pretty simmilar and offer basically the same informations, which is ideal for the intention of this master thesis.

Kochbar.de, like Allrecipes.com, has 3 major website views that offer the primiarily used features:

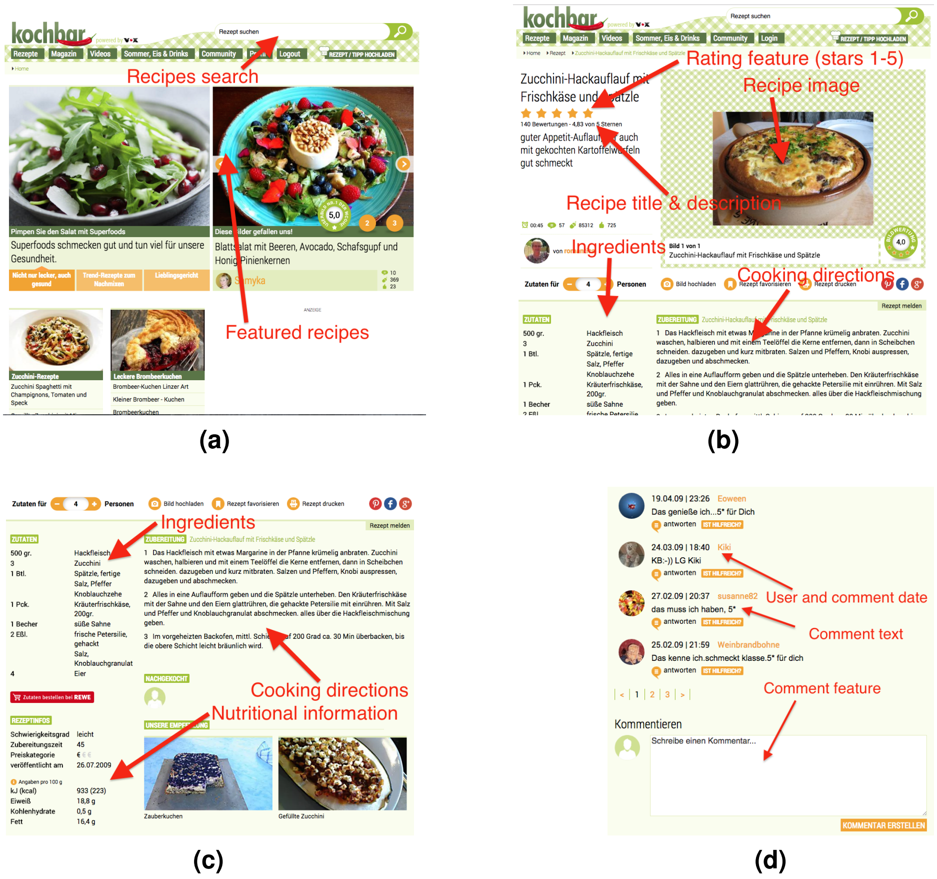
- **Homepage**
  The homepage of Kochbar.de shows recommended recipes, either based on the users preferences or editorial features. It also provides a recipe search feature.

- **Recipe Upload**
  The upload view is more complex, compared to Allrecipes.com. Besides the standard recipe parameters like title, description and preparation time, difficulty level or price level, it features also a very elaborated ingredients definition widget. The ingredients widget supports the user with ingredients suggests, based on the input provided. Besides the ingredient name, it has two separate input fields for the amount and the unity. Presumably, this improves the automatical ingredient parsing of Kochbar.de, on which the nutrient information is based. However, there are also many misspellings and

[5] https://www.blsdb.de/

word variants in this data set. This could either be explained with wrong usage of the ingredient widget, misspelled ingredients that are already in the Kochbar.de database, or that the widget has not been always part of the upload form. Furthermore, the cooking instructions are captured in a more structed manner. The form provides separat text fields for every step. Recipes can also be uploaded with an sample image. In contrast to Allrecipes.com, the Kochbar.de upload form allows the user to manually overwrite the automatically calculated nutritional information. It is also possible to assign describing tags and categories to the recipe, which is not possible in Allrecipes.com. See figures 3.3a and 3.3b

- **Recipe View**
  The recipe site shows almost the same information as Allrecipes.com. It displays the main recipe information like title, description, preparation steps, and ingredients. As already mentioned, there are only 4 nutrients, but the recipes are assigned to difficulty and price levels. Whats different to Allrecipes.com, is the rating and commenting functionality. Kochbar.de has a convenient rating feature right at the top of the recipe view. Likewise, the commenting feature is lcoated beneath the comments section of the recipe. See figures 3.2b, 3.2c and 3.2d



Pimpen Sie den Salat mit Superfoods

Superfoods schmecken gut und tun viel für unsere Gesundheit.

Nicht nur lecker, auch gesund | Trend-Rezepte zum Nachmixen | Lieblingsgericht

Diese Bilder gefallen uns!

Blattsalat mit Beeren, Avocado, Schafsgupf und Honig Pinienkernen

Samyka

Zucchini-Rezepte: Zucchini Spaghetti mit Champignons, Tomaten und Speck

Leckere Brombeerkuchen: Brombeer-Kuchen Linzer Art; Kleiner Brombeer - Kuchen; Brombeerkuchen

ANZEIGE

Zucchini-Hackauflauf mit Frischkäse und Spätzle

140 Bewertungen - 4,83 von 5 Sternen

guter Appetit-Auflauf, der auch mit gekochten Kartoffelwürfeln gut schmeckt

00:45 57 85312 725

Bild 1 von 1
Zucchini-Hackauflauf mit Frischkäse und Spätzle

500 gr.	Hackfleisch
3	Zucchini
1 Btl.	Spätzle, fertige
	Salz, Pfeffer
	Knoblauchzehe
1 Pck.	Kräuterfrischkäse, 200gr.
1 Becher	süße Sahne
2 Eßl.	frische Petersilie, gehackt
	Salz, Knoblauchgranulat
4	Eier

1 Das Hackfleisch mit etwas Margarine in der Pfanne krümelig anbraten. Zucchini waschen, halbieren und mit einem Teelöffel die Kerne entfernen, dann in Scheibchen schneiden. dazugeben und kurz mitbraten. Salzen und Pfeffern, Knobi auspressen, dazugeben und abschmecken.

2 Alles in eine Auflaufform geben und die Spätzle unterheben. Den Kräuterfrischkäse mit der Sahne und den Eiern glattrühren, die gehackte Petersilie mit einrühren. Mit Salz und Pfeffer und Knoblauchgranulat abschmecken. alles über die Hackfleischmischung geben.

3 Im vorgeheizten Backofen, mittl. Schiene auf 200 Grad ca. 30 Min überbacken, bis die obere Schicht leicht bräunlich wird.

Schwierigkeitsgrad	leicht
Zubereitungszeit	45
Preiskategorie	€
veröffentlicht am	26.07.2009

Angaben pro 100 g

kJ (kcal)	933 (223)
Eiweiß	18,8 g
Kohlenhydrate	0,5 g
Fett	16,4 g

Zauberkuchen | Gefüllte Zucchini

19.04.09 | 23:26 Eoween
Das genieße ich...5* für Dich

24.03.09 | 18:40 Kiki
KB:-)) LG Kiki

27.02.09 | 20:37 susanne82
das muss ich haben, 5*

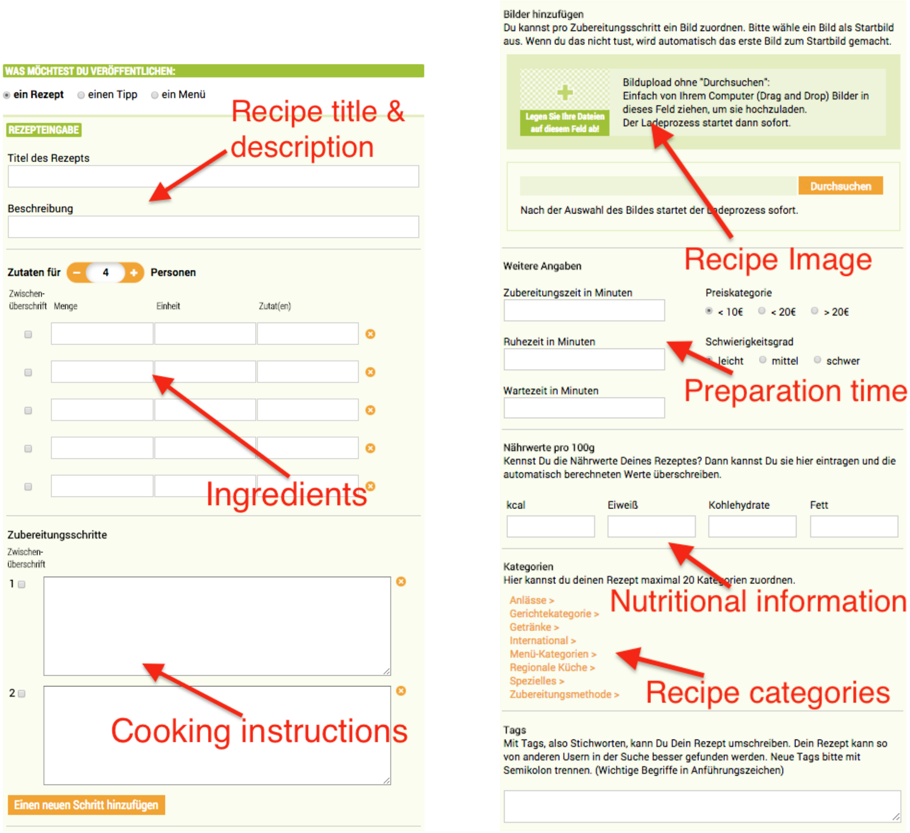
25.02.09 | 21:59 Weinbrandbohne
Das kenne ich.schmeckt klasse.5* für dich

| < | 1 | 2 | 3 | > |

Kommentieren

Schreibe einen Kommentar...

Figure 3.2: The main views of Kochbar.de. (a) shows the homepage with recommended recipes and a searchbar (b) highlights the top part of the recipe deatil view, showing the main parameters of the recipe and the rating feature (c) presents the bottom part of the recipe detail view with the nutritional information on the lower left side (d) shows the comments section located at the bottom part of the detail view.



WAS MÖCHTEST DU VERÖFFENTLICHEN:

● ein Rezept ○ einen Tipp ○ ein Menü

REZEPTEINGABE

Titel des Rezepts

Beschreibung

Zutaten für − 4 + Personen

Zwischen-überschrift Menge Einheit Zutat(en)

Zubereitungsschritte

Zwischen-überschrift

1 □

2 □

Einen neuen Schritt hinzufügen

(a)

Bilder hinzufügen

Du kannst pro Zubereitungsschritt ein Bild zuordnen. Bitte wähle ein Bild als Startbild aus. Wenn du das nicht tust, wird automatisch das erste Bild zum Startbild gemacht.

Legen Sie Ihre Dateien auf diesem Feld ab!

Bildupload ohne "Durchsuchen":
Einfach von Ihrem Computer (Drag and Drop) Bilder in dieses Feld ziehen, um sie hochzuladen.
Der Ladeprozess startet dann sofort.

Durchsuchen

Nach der Auswahl des Bildes startet der Ladeprozess sofort.

Weitere Angaben

Zubereitungszeit in Minuten

Preiskategorie: ● < 10€ ○ < 20€ ○ > 20€

Ruhezeit in Minuten

Schwierigkeitsgrad: ● leicht ○ mittel ○ schwer

Wartezeit in Minuten

Nährwerte pro 100g

Kennst Du die Nährwerte Deines Rezeptes? Dann kannst Du sie hier eintragen und die automatisch berechneten Werte überschreiben.

kcal Eiweiß Kohlehydrate Fett

Kategorien

Hier kannst du deinen Rezept maximal 20 Kategorien zuordnen.

Anlässe >
Gerichtekategorie >
Getränke >
International >
Menü-Kategorien >
Regionale Küche >
Spezielles >
Zubereitungsmethode >

Tags

Mit Tags, also Stichworten, kann Du Dein Rezept umschreiben. Dein Rezept kann so von anderen Usern in der Suche besser gefunden werden. Neue Tags bitte mit Semikolon trennen. (Wichtige Begriffe in Anführungszeichen)

(b)

Figure 3.3: The upload form of Kochbar.de (a) shows the top part of the form with the structured ingredients and cooking instructions input widgets (b) presents the bottom part of the upload form, allowing the user to upload an sample image. This part also offers the possibility to assign the recipe to different categories, difficulties and price levels.

Chapter 4

Methodology

> *“ Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom. ”*
>
> [Clifford Stoll, American astronomer, born 1950.]

This Chapter describes the approaches taken, to understand which aspects are relevant for the popularity of online recipes. Section one (i) explains the data selection and Section two (ii) describes pre-processing steps. Then the popularity analysis (iii) is described in detail. Section four (iv) highlights the feature engineering processes. The chapter concludes with the explanation of (v) the comparative statistical analysis and the (vi) preditive modelling, which aims to identify the best predictive features and models.

4.1 Data Selection

Allrecipes.com features 3 different recipe upload types: ‘Private Recipe’, ‘Public recipe’ and ‘Public and Kitchen Approved’. Only reviewed and editorial approved recipes show calculated nutritional information and are published on the main website. Because of this, only ‘Public and Kitchen Approved’ recipes are used for this study. Kochbar.de however, does not feature similary processes. Nutritional information is always calculated automatically, based on the ingredients of the recipe. Therefore, all recipes of Kochbar.de are used.

4.2 Data Pre-Processing

Since both datasets were crawled and utilized in previous related studies, the data was already structured and partly cleaned. However, the ingredients play a important role in cooking recipes and therefore need to be almost noise free. Furthermore, normalized ingredients are crucial for the sake of comparability between the two datasets. During the recipe upload process, both food community platforms allow the assignment of ingredients as a free form text. This led to noisy ingredient data with misspellings and word variants. Although, Kochbar.de has a ingredient completion widget which suggests ingredients based on the input, data cleaning and noise reduction was necessary. To normalize the ingredients, several NLP processing possibilities exist, such as described in Kusmierczyk, Trattner, et al. [2015a]. However, in order to extract the ingredients of a recipe, an other approach was used. A Web research showed two ingredient parsing alternatives: The New York Times[1] (blog entry of Erica Greene [2015]) solved this problem when launching their online food community NYT Cooking[2]. They even made their "CRF Ingredient Phrase Tagger" [3] open source. The second possibility was Spoonacular.com, a recipe search engine and online meal planer website, with a comprehensive API[4] for analysing food and recipes. Since Spoonacular offered that their API could be used for free (only for this master thesis) under their special education/research pricing plan and the fact that the API supports everything needed for ingredient recognition, it was the solution of choice. Besides the free text parsing, another major problem were non english ingredient names in Kochbar.de. Neither the NYT tagger, nor the Spoonacular API support the german language, which made it necessary to translate them. This step of preprocessing was done with the Google Translation API [5], a online service that can be accessed via REST and produces good results. Proper names of ingredients are kept as they are (e.g. "Kasseler"), but others get translated (e.g. "spring onions"). In the next step, the ingredient texts were extracted with the "parse-ingredients"[6] Spoonacular API call. The API returns the extrated ingredients as a JSON formated string. Listing 4.1 shows the capabilities of the API: If the ingredient text contains quantities, it is able to extract them. Furthermore, it classifies the course (aisle) and returns a sample image of the ingredient.

```json
[
  {
    "id": 10072,
    "original": "2 oz pork shoulder",
    "name": "pork shoulder",
    "amount": 2,
    "unitShort": "oz",
    "unitLong": "ounces",
    "aisle": "Meat",
    "image": "https://spoonacular.com/cdn/ingredients_100x100/pork-shoulder.jpg",
    "meta": []
  }
]
```

Listing 4.1: Example JSON result of the "parse-ingredients" API call.

[1] https://www.nytimes.com/

[2] https://cooking.nytimes.com/

[3] https://github.com/NYTimes/ingredient-phrase-tagger

[4] https://spoonacular.com/food-api

[5] https://cloud.google.com/translate/

[6] https://market.mashape.com/spoonacular/recipe-food-nutrition#parse-ingredients

Because of misspellings or word variants, not all ingredients were successfully extracted by the API. To tackle this problem, unidentified ingredients were replaced with the most similar extracted ingredient. This was done by first splitting the orignal ingredient text at space characters. Secondly, calculating the string similiarities of all ingredient text parts to all known ingredients that occured more than twenty times. The ingredient with the highest similarity was accepted as the matching ingredient. For the similiarity caluclation, the "SIFT4" algorithm of the "java-string-similarity" [7] library was used. It is an experimental approach to reproduce the human perception of distance, based on longest common subsequence, character substitution and character distance. Since this approach does not solve the problem of misspelled german ingredients as they do not get translated correctly by the Google translate API, an additional processing step was needed. When a ingredient occured not more than 10 times (assuming that they are misspelled), its orignal ingredient text (bevor translation) was compared to the original texts of correctly parsed ingredients. The ingredient with the hightest similarity was taken.

The preprocessing approach outlined above is not satisfying and would need future improvement. However, it was possible to reduce the number of distinct ingredients significantly: From 723.911 to 3842 for Allrecipes.com and from 302.126 to 2028 for Kochbar.de.

4.3 Popularity Analysis

To show whether or not popularity patterns exist in the datasets, a graphical investigation of recipe popularity over time as well as the investigation over different recipe categories were performed. There is a variety of possible measurements or proxies for popularity, such as the number of visits or the mentionings in social media. Studies (for example Bandari et al. [2012]) have even used the number Twitter tweets for this purpose. However, in this work, as proxies for popularity, *number of comments, number of ratings* and *number of bookmarks* were chosen. Other proxies such as number of views were not available in the dataset.

To compare popularities also across different types of categories, the categorical popularity and appreciation per category was compared. For each category, the mean number of ratings, comments and the mean rating and sentiment got calculated. Statistical significancy tests were deployed to proof differences (See Section 4.5 for further details).

4.4 Feature Engineering

As mentioned previously (See Chapter 2) there many possibilities to induce features for the sake of popularity prediction of food and recipes. In their paper about food choices Scheibehenne et al. [2007] pointet out that there are many factors influencing our decisions such as taste, texture, nutritional content, physical environment, attitudes, motives, individual preferences and information. Therefore, serveral predictive features were deployed, based on aspects from cognitive psychology, features that were previously used by other researchers (like Elsweiler

[7] https://github.com/tdebatty/java-string-similarity

et al. [2017] or Rokicki, Herder, and Trattner [2017]) and the knowledge gained from the popularity analysis. This thesis takes advantage of all these previous attempts and defines 7 feature sets with serveral different features that are aimed to capture the popularity aspects of online recipes. For the creation of this features, multiple tools like MySQL, Java or Python got utilized.

4.4.1 Feature Sets

The predictive features are assigned to one of the seven feature sets listed below. Each feature set will be described in the following Sections in detail.

- Recipe Nutrition
- Recipe Healthiness
- Recipe Complexity
- Recipe Presentation
- Recipe Seasonality
- Recipe Innovation
- User Activity and Context

During the process of feature engineering, more than 180 features were created in total. However, many of them were dropped because of minor relevance or temporal independency, at later stages of this thesis.

4.4.2 Recipe Nutrition

The features in this set describe the recipes macro nutritional properties calorific value, fat, protein and carbohydrates. These are among the essential human diet factors and contained in many ingredients. The nutritional factors were crawled from the websites during the studies of Kusmierczyk, Trattner, et al. [2015b] regarding Kochbar.de, respectively Allrecipes.com by Trattner and Elsweiler [2017]. Although the Allrecipes.com dataset would provide more nutritional information of the recipes, Kochbar.de does not. For the means of comparability of the two food websites, only the above mentioned four diet factors will be taken as nutritional metrics. Allrecipes.com's nutritional analysis [*Nutritional Information* 2017] of uploaded recipes is based on ingredient mappings to the ESHA Research's nutrient databases. The nutrients estimation of Kochbar.de's recipes on the other hand is made via the German Nutrient Data Base [8]. This mapping approach is considered to be imperfect as reasearch by Müller et al. [2012] showed.

- Kcal

  The scientific definition of kilocalorie (kcal) is the amount of energy needed to heat one kilogram of water from 0 to 1 C° at sealevel. In terms of nutrition

[8] https://www.blsdb.de/

it is used as the physiological fuel values of food. Alternativley, the term "calorie" is often used in cooking books and food databases, although they are physically not the same (lower case calorie is the gram calorie, upper case means kilogram calorie). Since the adoption of the SI[9] (International System of Units) the joule is the unit of energy in the scientific world. [James L Hargrove, 2007].

- Protein

  Proteins are molecules formed by long chains of amino acids. They are the essential building blocks of living cells. Furthermore, they fullfill specific regulatory functions inside our body e.g. hormones. Proteins are one of the three nutrients that can be used as energy by our body. The others are carbohydrates and fat. The body can produce 4 Calories per gram of protein. [*Medical Definition of Proteins* 2016].

- Carbohydrates

  Chemically defined as a compound of carbon, hydrogen and oxygen. Carbohydrates are one of the principal energy sources of the human body. They come in a variety of forms such as sugars or starches and get mostly broken down to glucose by our digestive system. Carbohydrates deliver 4 Calories per gram to our body. [*Medical Definition of Carbohydrates* 2016]

- Fat

  Compared to proteins and carbohydrates, fat is the nutrient with the highest energy density. It provides up to 9 Calories per gram. The fat found in tissue of animals and humans are compounds chemically defined as fatty acids. [*Medical Definition of Fat* 2016]

4.4.3 Recipe Healthiness

This feature set contains metrics, describing the healthiness of online recipes. A Study conducted by Trattner and Elsweiler [2017] has shown that unhealthy recipes tend to have more user interactions, regarding booksmarks and comments and tend to be rated more positively. This led to the assumption that the healthiness of recipes can be helpful in the means of predicting popularity.

In order to measure the healthiness of online recipes, the same approach as Trattner and Elsweiler [2017] was used. They calculated the recipe healthiness following the nutrient profiling model developed by the World Health Organization (WHO) [Who and Consultation, 2003]. This system for nutrition labeling on food packaging is a international acknowledged standard and is based on the nutritional metadata of foods and beverages.

It was not possible to caluclate the healthiness score in the stated manner, because Allrecipes.com and Kochbar.de do not provide all the nutritional information required by the two guidlines. Therefore only the parameters kcal, protein, carbohydrates and fat were used, since they were available in both datasets. This approach leads to a score that do not represents the actual healthiness of a recipe, but it allows a comparison among the processed recipes. This has a similarity to

[9] http://www.bipm.org/en/publications/si-brochure/

the procedure stated by Howard et al. [2012], as they only used the 7 most important nutritional factors (carbohydrates, fibers, sodium, sugars, proteins, fats and saturated fats) of the human diet.

- WHO Health Score

  This metric is based on the WHO guidelines which defines 15 ranges of essential nutritive substances based on recommended daily intakes. Analog to the approach of Howard et al. [2012] only 4 scales (carbohydraes, proteins, fat and kcal) were used for this feature. The so called "WHO healt score" has a value ranging from 0 to 4, meaning that either none of the nutrients are within the range or all met the recommended quantities.

4.4.4 Recipe Complexity

In order to describe the complexity of a recipe and the effort required to prepare it, 5 features such as preparation time (Minutes), preparation steps or number of ingredients have been defined. These metrics are the main meta data of recipes uploaded to Allrecipes.com or Kochbar.de and are partly required at the uploading process. Like the features in the recipe nutrition set, the complexity metrics have also been utilized crawles of Kusmierczyk, Trattner, et al. [2015b] respectively Trattner and Elsweiler [2017]. The recipe complexity features are:

- **Preparation Time:** The preparation duration in minutes
- **Number of Preparation Steps:** preperation steps needed to prepare
- **Number of Servings:** the number of servings this recipe will produce
- **Number of Ingredients:** the number of ingredients needed for the recipe
- **Number of Catergories:** the number of categories, for example main dish or dessert, a recipe has been assigned to.

4.4.5 Recipe Presentation

The presentation set contains 19 features and is one of the two bigger feature sets. Beside simple text metrics such as number of characters, number of words or number of sentences, also sentiment (as proposed by Parra et al. [2016]) and Shannon entropy of the recipe's title and instructions were measured. Additionally, the readability of the instruction text using the 'Lasbarhetsindex' [Anderson, 1981b] was calculated, since it can be used for English and German text. Furthermore, we followed the approach of San Pedro and Siersdorfer [2009] deriving 8 image features e.g. the brightness, sharpness or colorfulness from the recipe's upload picture. Additionally, we calculated the Shannon entropy of the images. Elsweiler et al. [2017] could show that the recipes images have a influence on the users perception of the recipe.

Image Features: Savakis et al. [2000] have listed several factors that influence the visual appeal of photos to humans. Besides, semantic attribute categories such as "People and Expression”, which are very difficult to compute, they also specified "Objective Measures" such as image quality. In a follow up study, San Pedro and Siersdorfer [2009] successfully used such low level image metrics, in combination with community feedback data, to estimate the attractivness of Flickr images.

All low level image features are measured with the freely available OpenIMAJ Java Framework [10] in version 1.3.5. OpenIMAJ is a collection of tools for analysis multimedia content such as images or video and was developed by the University of Southampton [11].

- Image: Brightness

  The average brightness of an image describes the subjective visual perception of the energy output of a light source [Mike Chaney, 2017]. The brightness of the recipe images was extracted with the AvgBrightness class [12] with the default NTSC weighting scheme and no mask. It uses a standard luminance algorithm

$$Y_{xy} = (0.299 * R_{xy} + 0.587 * G_{xy} + 0.114 * B_{xy}) \tag{4.1}$$
$$avg\_brightness = 1/N \sum_{x,y} Y_{xy}$$

  where Y_{xy} denotes the luminance value and N is the size of the image. R, G and B are the three RGB color space[13] channels of pixel(x,y) [OpenIMAJ Java Library 2017(a),San Pedro and Siersdorfer 2009].

- Image: saturation

  According to the International Commission on Illumination [2017(a)] the image saturation is defined as the "colourfulness of an area judged in proportion to its brightness". It describes the quality of the color effect or vidiness. The OpenImaj Saturation class[14] was utilized for this measurement. In the HSV colorspace [15] the saturation estimation can be calculated via the RGB approximation of

$$S = \sum max(R, G, B) - min(R, G, B) \tag{4.2}$$
$$avg\_saturation = S/N$$

  where N is the sizes of the image and R,G and B are the coordinates of the color in sRGB space. [OpenIMAJ Java Library 2017(a),San Pedro and Siersdorfer 2009].

[10] http://www.openimaj.org

[11] http://www.southampton.ac.uk/

[12] http://openimaj.org/apidocs/org/openimaj/image/feature/global/AvgBrightness.html

[13] https://en.wikipedia.org/wiki/RGB_color_space

[14] http://openimaj.org/apidocs/org/openimaj/image/feature/global/Saturation.html

[15] https://en.wikipedia.org/wiki/HSL_and_HSV

- Image: saturation variation

  This image measure is done with the OpenImaj Saturation Variation class[16]. The algorithm estimates the variation in saturation via the sample standard deviation [17] of all pixel saturations of the image

  $$saturation\_variation = \sqrt{\frac{\sum(S - \overline{S})^2}{n - 1}} \tag{4.3}$$

  where n is the number of pixels and S is the list of saturations [OpenIMAJ Java Library 2017(a),San Pedro and Siersdorfer 2009].

- Image: Colorfulness

  The International Commission on Illumination [2017(b)] has defined colorfulness as an "attribute of a visual perception according to which the perceived colour of an area appears to be more or less chromatic". Colorfulness can be calculated via the individual color distance of the pixels. Therefore, the image needs to be transfered in to sRGB color space using

  $$\begin{aligned} rg &= R - G \qquad (4.4) \\ yb &= 1/2\,(R + G) - B \end{aligned}$$

  and subsequently, colorfulness can be measured with

  $$\begin{aligned} \sigma_{rgyb} &= \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} \qquad (4.5) \\ \mu_{rgyb} &= \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \qquad (4.6) \\ colorfulness &= \sigma_{rgyb} + 0.3 * \mu_{rgyb} \end{aligned}$$

  where R,G and B are the color channels of the pixels and σ is the standard deviation, respectively μ the arithmetic mean. The colorfulness of the recipe images was measured with corresponding class of OpenImaj [OpenIMAJ Java Library 2017(a),San Pedro and Siersdorfer 2009].

- Image: Naturalness

  The concept of naturalness describes the difference (or similarity) between an image and the humans visual perception of the real world, in respect of colorfulness and dynamic range. Altough very subjective, it is an important image quality metric when it comes to color image design [Huang et al., 2006].

  San Pedro and Siersdorfer [2009] describe the naturalness algorithm as follows: First transfer the image color space, if not already, to HSL[18]. Than use

[16] http://openimaj.org/apidocs/org/openimaj/image/feature/global/SaturationVariation.html

[17] https://people.richland.edu/james/lecture/m170/ch03-var.html

[18] https://en.wikipedia.org/wiki/HSL_and_HSV

only pixels within the thresholds $20 \leq L \leq 80$ and $S \geq 0.1$. In the next step, pixels get grouped in to one of the three sets 'Skin','Grass' or 'Sky', based on their H coordinate (hue). In order to calculate the naturalness of each set, the average saturation value of the group (μ_S) is used:

$$N_{Skin} = e^{-0.5(\frac{\mu_S^{Skin}-0.76}{0.52})^2},\ if\ 25 \leqslant \text{hue} \leqslant 70 \tag{4.7}$$
$$N_{Grass} = e^{-0.5(\frac{\mu_S^{Grass}-0.81}{0.53})^2},\ if\ 95 \leqslant \text{hue} \leqslant 135$$
$$N_{Sky} = e^{-0.5(\frac{\mu_S^{Sky}-0.43}{0.22})^2},\ if\ 185 \leqslant \text{hue} \leqslant 260$$

In the final step, the naturalness index can be caluclated using

$$naturalness = \sum_i \omega_i N_i,\ i \in \ \{\text{'Skin','Grass','Sky'}\} \tag{4.8}$$

where ω_i represents the fraction of pixels of the specific group in the whole image. N ranges from 0 (a unnuatural image) to 1 (a natural image). The OpenImaj class 'Naturalness' uses this very algorithm. [OpenIMAJ Java Library 2017(a),San Pedro and Siersdorfer 2009].

- Image: Contrast

  Contrast is the relative difference in brigthness or color of local features in an image. The International Commission on Illumination [2017(c)] defines it as "assessment of the difference in appearance of 2 or more parts of a field seen simultaneously or successively". There are many contrast meassuring algorithms such as Weber[19] or Michelson[20], but for the means of image comparability, the root mean square contrast (RMS-contrast) gets often used.

$$rms\_contrast = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) \tag{4.9}$$

  x is the pixel intensity, $\overline{x}$ represents the arithmetic mean of the pixel intensity and n is the number of pixels in the image. OpenImaj offers the RMSContrast class for this measurement [OpenIMAJ Java Library 2017(c),San Pedro and Siersdorfer 2009].

- Image: RGB Contrast

  The rgb contrast is almost identical to the basic contrast calculation, explained above. However, it is extended to the three-dimensional RGB color space [OpenIMAJ Java Library 2017(b),San Pedro and Siersdorfer 2009].

[19] https://en.wikipedia.org/wiki/Contrast_(vision)#Weber_contrast

[20] https://en.wikipedia.org/wiki/Contrast_(vision)#Michelson_contrast

- Image: Sharpness

  This image measurement is defined as the subjective perception of sharpness or clarity. It is related to the brightness contrast of edges in an image [Photoreview.com, 2017]. The metric was measured with the OpenIMAJ Sharpness class [21] which uses the algorithm described by San Pedro and Siersdorfer 2009: The algorithm utilizes the images Laplacian, divided by the locale average luminance (μ_{xy}) around pixel (x,y):

$$sharpness = \sum_{x,y} \frac{L(x,y)}{\mu_{xy}}, \text{ with } L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{4.10}$$

- Image: Sharpness variation:

  Simmilar to the saturation variation, sharpness variation gets calculated via the standard deviation of all pixel sharpness values.

- Image: Entropy

  In information theory, entropy is known as a measure for randomness or the average information content of a source. The entropy of an image is often used to determine the amount of information that needs to be encoded by a compression algorithm. As an example, a image of moon craters has a very high edge contrast, this leads to a high entropy, meaning it cannot be compressed very well. This indicates that it can be used as a measure of the image's texture [Cornell University 2017,MathWorks 2017]. The Shannon entropy algorithm was used for this image feature: First convert the image to greyscale, where each pixel has only a intensity value. Secondly, count the occurances of each distinct value. Than, following formula was applied:

$$\begin{gathered} p = \frac{\bar{I}}{N} \\ entropy = -\sum_{i=0}^{N} p_i * \frac{log(p_i)}{log(2)} \end{gathered} \tag{4.11}$$

  where $\bar{I}$ are the intensity frequencies, N is the size of the image to process and p_i denotes the probability of an specific intensity.

  Textual Features:

  The processing of the recipes title and instruction text was done by utilizing the Apache OpenNLP [22] natural language toolkit which offers the possibility to tokenize. In terms of lexical analysing this means breaking down texts into scentences, words or smaller meaningful parts [Tutorialspoint, 2017]. However, this functionality is based on machine learning algorithms and models that need to be generated beforehand. Since the dataset of Allrecipes.com was in english and kochbar.de is a german website, two different models are

[21] http://openimaj.org/apidocs/org/openimaj/image/feature/global/Sharpness.html

[22] https://opennlp.apache.org/

necessary. One advantage of Apache OpenNLP is that it already has free and suitable models[23] at disposal.

The detection and counting of sentences, words and characters is very straight forward. After loading and defining the detection models, one just passes the text to the specific class and methods: SentenceDetectorME[24] for sentences and TokenizerME [25] for detecting words. The number characters was measured via the string length of the text, including whitespaces.

- Number of Sentences of the recipes instruction text
- Number of Characters of recipe instruction and title
- Number of Words of recipe instruction and title
- Readability Score of recipe instruction and title

  In order to measure the readabiliy of the recipe texts, we followed the approach of Anderson [1981a] who showed that the "Lasbarhetsindex" workes for european languages, including english, french and also german. Besides the advantage that Lix can be used for Allrecipes.com and Kochbar.de it is very easy to compute. It bypasses difficulties of other readability algorithms, which need to calculate the number of sylables or polyslabic beforehand. It is mainly based on simple text measures:

$$\begin{aligned} RLW &= \frac{n_{lw}}{n_w} \\ ASL &= \frac{n_w}{n_s} \\ LIX &= 100 \times RLW + ASL \end{aligned} \tag{4.12}$$

  where n_{lw} is the number of long words (words > 6 characters), n_w is the word count and n_s is the scentences count.

- Entropy for recipe instruction and title

  Altough this feature measures the entropy of text, it is calculated in the same way as the entropy of the recipe images, described by formula (4.11) . However, there are differences regarding the frequency data. This entropy calulation is based on the characters distribution of the whole text processed.

- Sentiment for recipe instruction and title

  Every text, regardless of written by a human or a "bot" (a recent development [26]), has an intention. Some of them are just of technical nature, but others try to transmit or provoke emotions. This is especially true for advertisment or social media content, where popularity plays an important role. The sentiment analysis of text offers the possibility to approximate the users intended

[23] opennlp.sourceforge.net/models-1.6

[24] http://opennlp.sourceforge.net/api/opennlp/tools/sentdetect/SentenceDetectorME.html

[25] https://opennlp.apache.org/documentation/1.6.0/apidocs/opennlp-tools/index.html?opennlp/tools/tokenize/TokenizerME.html

[26] https://en.wikipedia.org/wiki/Social_bot

manipulation. There are serveral algorithms and toolkits for polarity extraction such as the Natural Language Toolkit [27] (NLTK, available for many programming languages) or Google's cloud natural language API [28] (based on a REST[29] service) and most of them work with statistical classification methods. This feature metric follows the approach of Parra et al. [2016], who used SentiStrength[30] to analyse emotions in Twitter posts about computer science conferences. Similar to the Apache OpenNLP toolkit, which was used for tokenizing the recipe texts, SentiStrength has already trained classification models for different languages. According to studies of Berrios et al. [2015], showing that human process positive and negative sentiment at the same time, SentiStrength provides two contrary measures: On the one hand the negative sentiment strength and on the other hand the positive one. They range from -1 to -5 (-1 not negative, -5 very negative) and from 1 to 5 (1 not positive, 5 very positive). Kucuktunc et al. [2012] stated two metrics that can be derived from the sentiment strengths:

$$\begin{aligned} sentimentality &= \varphi^{+}(t) - \varphi^{-}(t) - 2 \\ attitude &= \varphi^{+}(t) + \varphi^{-}(t) \end{aligned} \tag{4.13}$$

where φ^{+} is the postitive sentiment strength and φ^{-} the negative one. The sentiment feature of the recipe instruction and title was extracted with the attitude function, which gives predominant sentiment of the text. However, sentimentality indicates how sentimental a text is, compared to neutral [Parra et al., 2016]. For the sentiment measurement of recipe instruction and title only the attitude was used.

4.4.6 Recipe Seasonality

Food preference and recipe choices have traditionally been seasonal dependend. This has not only to do with the availabilty of ingredients such as fresh vegetables in summer or pumpkins in fall, but also with cultural influences. For instance in central europe, people bake a lot of the traditional christmas cookies. Another example would be the traditional Thanksgiving turkey in north america. In their study of user Web usage logs from search engines (Google, Bing, Yahoo!), West et al. [2013] clearly found temporal food consumtion patterns. They could show that ingredient prevalence changes of the year (their data showed prevalence for turkey around octorober utill december). Furthermore, variations in nutritional intake (less chalories in summer, more in winter) were observed. Kusmierczyk, Trattner, et al. [2015b] found similar temporal patterns in one of the same datasets (Kochbar.de) this master thesis bases on. The features of this set are:

- Upload Month
- Day of Month

[27] http://www.nltk.org/

[28] https://cloud.google.com/natural-language/

[29] https://de.wikipedia.org/wiki/Representational_State_Transfer

[30] http://sentistrength.wlv.ac.uk/

- Day of Week
- Within Season

The recipes seasonality (Within season) was calculated using the parsed ingredients. Therefore, the seasonal distribution of every ingredient is necessary. Almost every recipe has an upload date and ingredients mapped to it. By assigning the upload date to the related ingredients, a yearly distribution can be calculated over all recipes in the body. A monthly granularity was chosen for this metric:

Repeat for each ingredient:

1. Count the occurences of this ingredient for each month over all recipes.
2. Use the occurences list to calculate the density function with an univariate kernel density estimator such as the KDEUnivariate[31] method of StatsModels.
3. Evaluate the point density[32] of the estimated density function for all twelve months.

Then calculate for each recipe

$$within\_season = \frac{1}{N}\sum_{i=0}^{N}\rho_i^m \tag{4.14}$$

where N is the amount of ingredients of the recipe and ρ_i^m is the seasonalty probability of ingredient i at the recipe upload month m. The seasonality value ranges from 0 (non of the ingredients ate in season at upload date) to 1 (all ingredients are in season).

4.4.7 Recipe Popularity and Appreciation

The popularity measurements of the online recipes are based on the two available factors: ratings and comments. Kusmierczyk, Trattner, et al. [2015b] showed in their research that the users interests in recipes shift over time. They found that for Kochbar.de the rating probability decay is following a power law and the median lifetime of a recipe in the main course categorie is under 60 days. This means that the temporal development of a recipes popularity in terms of ratings and comments, needs to be taken into account. The considered time spans for this metrics are from the time of upload until the first day, week, month and year after.

Besides ratings and comments, a very similar favor/bookmarking paradigm exists for both food community websites. Users can mark a recipes for later, respectively show their interest by "liking" one. However, only the Allrecipes.com dataset features the date of bookmarking. For the sake of comparability of the two websites, the number of bookmarks of a recipe is excluded from the feature list.

[31] http://www.statsmodels.org/stable/generated/statsmodels.nonparametric.kde.KDEUnivariate.html

[32] http://www.statsmodels.org/stable/generated/statsmodels.nonparametric.kde.KDEUnivariate.evaluate.html#statsmodels.nonparametric.kde.KDEUnivariate.evaluate

- Number of Comments received within first day
- Number of Comments received within first week
- Number of Comments received within first month
- Number of Comments received within first year
- Number of Ratings received within first day
- Number of Ratings received within first week
- Number of Ratings received within first month
- Number of Ratings received within first year

The appreciation of the recipes is meassured with the average rating provided and the average comments sentiment. The comment sentiment, needed for the calculation of the "Average Comment Sentiment" feature is extracted in the very same way as the sentiment for the recipe title or instructions. Also, the attitude score is used only.

- Average Rating

$$avg\_rating = \frac{1}{N}\sum_{i=0}^{N} r_i \tag{4.15}$$

N is the number of ratings of a recipe and r is the value of a rating .

- Average Comment Sentiment

$$avg\_sentiment = \frac{1}{N}\sum_{i=0}^{N} s_i \tag{4.16}$$

N denotes the amount of comments of a recipe and s is the extracted sentiment of a comment.

4.4.8 Recipe Innovation

Kusmierczyk, Trattner, et al. [2015a] pointed out that innovation is a driving factor of chefs and restaurants success. In their research about online food innovation in Kochbar.de they found that the number of known ingredients stays almost the same, but the innovation it self is slowly growing in time. This would support the assumption that innovation is as important in the real world as it is in the virtual, since online cooks also face a competitive preassure.

For understanding innovation or creativity, Kerne et al. [2014] stated 4 important factors that can be measured: novelty, flexibility, fluency and quality. Novelty describes the uniqueness or the similarity of an idea (recipe) compared to other ideas. Flexibility is the measure of the variety of ideas or alternative interpretations. In terms of food: how variable is the chefs cooking or recipe vocabulary?

Fluency adresses the amount of ideas and is reasoned with the Darwininan theories. This means, the more ideas a person has, the more likely it is to have a creative one. Quality can be seen as the measure of fitness for purpose and is based on the rating of externals. This feature set tries to measure the novelty of the recipes, using their ingredients combination. Following features were calculated:

- Recipe Innovation Jaccard

  This feature follows the approach of Kusmierczyk, Trattner, et al. [2015a]. They defined an "Innovation Factor" for recipes based on the similarity of their ingredients combinations. This similarity uses the Jaccard's index[33], which calculates the similarity of two recipes r and r':

  $$similarity(r, r') = Jaccard\_Index(r, r') = \frac{|\{i : i \in r \wedge i \in r'\}|}{|\{i : i \in r \vee i \in r'\}|} \tag{4.17}$$

  The innovation function then measures the distance of a specific recipe to the most similar one of all before uploaded recipes in the body.

  $$innovation\_factor(r) = 1 - \underset{r' \prec r}{max}\ similarity(r, r') \tag{4.18}$$

  Operator $\prec$ shows the temporal precedence of the recipes (r') compared to upload date of recipe r. Parameter i denotes the ingredients of the recipes.

- Average Recipe Innovation Jaccard

  The average innovation factor is an adaption of the recipe innocation jaccard desribed above. It is also based on the Jaccard's Index of recipe ingredient sets. However, this metric uses the average distance and not the maximum.

  $$avg\_innovation\_factor(r) = 1 - \underset{r' \prec r}{mean}\ similarity(r, r') \tag{4.19}$$

  The parameters of this equation are described in formula 4.17.

- Recipe Innovation IDF

  This feature is adapted from Kerne et al. [2014], which stated a metric for measuring novelty in ideation processes. Their novelty calculation is based on the the inverse document frequency [34] (IDF) algorithm, which is often used in the information retrieval (e.g. search engines). It uses a list of all answers and an inverted index showing which creators had the same answer. To measure the novelty of recipes, this list contains all ingredients and the inverse index points to all recipes that use them. For the recipe innocation

[33] https://en.wikipedia.org/wiki/Jaccard_index

[34] https://en.wikipedia.org/wiki/Tf%E2%80%93idf

IDF feature, this approach was slightly changed, so that only recipes that were previously uploaded (compared to the recipe of interest) are used.

First, a function that returns all ingredients of a specific recipe gets defined:

$$F(r) = Ingredients(r) \tag{4.20}$$

The before mentioned ingredients list with an inverted index, pointing to the recipes their used in, defines as follows:

$$occurrences(i, R) = \{r | r \in R \land i \in F(r)\} \tag{4.21}$$

where R is the set of before uploaded recipes (including the recipe of interest) in the body and i represents an ingredient. The number of occurences of an ingredient can be seen as a measure for usualness. In order to compute the uniqueness of an ingredient, the inverse of the occurences amount is used:

$$ingredient\_novelty(i, R) = \frac{1}{||occurrences(i, R)||} \tag{4.22}$$

The value of ingredient novelty ranges from 1/n (the ingredient was only used in one recipe) to 1 (the ingredient was used in all recipes). Finally, the novelty for a recipe is calculated as the sum of all its ingredient novelties, normalized by the amount of all ingredients used in the recipe:

$$recipe\_novelty(r, R) = \frac{\sum_{i \in I} ingredient_n ovelty(i, R)}{||F(r)||} \tag{4.23}$$

The resulting recipe novelty (Recipe Innovation IDF) is a rational number between 0 and 1. The theoretical maximum novelty of a recipe is 1.

Further uniquness or innovation metrics utilize the frequency ranks of recipe parameters. Rankings were calcuated by counting the occurences of ingredients, categories and title words in the complete recipes body and ranking them descending, according to the number of occurences. Since most of the recipes have more than one ingredient, category or title word, the average of the specific ranks were calculated. This way, it was possible to derive a single value per parameter type and recipe:

- Ingredients rank

  This metric denotes the average rank of the ingredients used in a recipe:

$$avg\_ingredients\_rank = \frac{1}{N} \cdot \sum_{n=1}^{N} rank(I_n) \tag{4.24}$$

where N is the amount of all ingredients used in this recipe and the funktion $rank()$ returns the frequency rank of the ingredient I_n.

The next two features "categories rank" and "title words rank" are calculated in the same way. However, they are based on the ranks of the categories and title words of a recipe.

- Categories rank

$$avg\_categories\_rank = \frac{1}{N} \cdot \sum_{n=1}^{N} rank(C_n) \tag{4.25}$$

- Title Words rank

$$avg\_title\_words\_rank = \frac{1}{N} \cdot \sum_{n=1}^{N} rank(TW_n) \tag{4.26}$$

In contrast to the mean rank features, the next metrics utilize the median. This function is a more robust measure, since it not as sensitive to skewness (e.g. outliers) of distributions as the mean. Nevertheless, these features use the same ranking functions, as the ones described before:

- Ingredients rank (median)

$$median\_ingredients\_rank = median([rank(I_1), ..., rank(I_n)]) \tag{4.27}$$

- Categories rank (median)

$$median\_categories\_rank = median([rank(C_1), ..., rank(C_n)]) \tag{4.28}$$

- Title Words rank (median)

$$median\_title\_words\_rank = median([rank(TW_1), ..., rank(TW_n)]) \tag{4.29}$$

4.4.9 User Activity and Context

Contribution is a way to gain popularity in social networks. People, or users that contribute a lot of content and meta data to the network, can gain reputation. Such individuals are often called "influencers", meaning that they can spread information fast and with a high reach. The mechanism behind this is, in terms of marketing, known as the "word-of-mouth diffusion"[35]. However, as Bakshy et al. [2011] found in their studies about influencers and diffusion on Twitter that there is probably not this small number of influencers, disproportionately influencing

[35] https://en.wikipedia.org/wiki/Word-of-mouth_marketing

high numbers of others. They showed that “ordinary influencers” can often be more cost effective in the means of diffusion and marketing. Considering this, the influence measurement of all users could potentially help to solve the task of popularity prediction. This assumption is strenghtend by the study of Khosla et al. [2014] who successfully used a set of "social cues" to predict the popularity of Flickr images. They also point out the importance of social context respectively the phenomenon of social influence in the means of popularity of content. Users are more likely to adopt behaviours or content of friends. The metrics they used are for example the mean of number of views of all images of a user, the number of photos uploaded by the user or the duration of membership. Following these activity and influence measuring approches the features of this set were derived.

The next three features capure the users activity in prior to the upload date of a specific recipe of him. This activities include the number of uploaded recipes, the number of written comments and the sum of all comments, his recipes received.

- Recipes Uploaded until Upload
- Comments Written until Upload
- Comments until Upload

In order to measure the broadness and categorical reach of the users recipes, the next four features are used. The idea behind this metric is that a cook’s recipe will be searched and found more easily by others if he applies more distinct ingredients and categories to all of his recipes.

- Number of Distinct Ingredients used
- Average number of Ingredients used (per recipe)
- Number of Distinct Categories used
- Average number of Categories used (per recipe)

The temporal dimension of the users activity is capured by twelve features. This is necessary because a user that is registered for five years, but uploaded only ten recipes, is not as active as a user who is registered for two years and published ten recipes too. The defined activity timespan for this measurment is ranging from the sign up date until the date of his last productive activity (recipe upload, comment written or rating provided). The number of years, months, weeks and days in this timespan is used to calculate the relation between the users activity and his active time:

- Uploaded Recipes per day
- Uploaded Recipes per week
- Uploaded Recipes per month
- Uploaded Recipes per year

- Number of Ratings provided per day
- Number of Ratings provided per week
- Number of Ratings provided per month
- Number of Ratings provided per year
- Number of Comments provided per day
- Number of Comments provided per week
- Number of Comments provided per month
- Number of Comments provided per year

As a last one, a binary feature captures whether a user is located in the origin country of the food websites. This means that if a Allrecipe.com user is living in the USA, the value will be 1 otherwise 0. The same is applied to Kochbar.de and germany located users.

- Cook Living in Germany/USA

4.4.10 Top-20 Categories

This feature set contains the top 20 categories that occure in both datasets. First, all categories of both datasets were ranked by the number of recipes assigned. Secondly, the matching categories with the highest rank were manually mapped together. This was necessary since the predominant language in Allrecipes.com (english) and Kochbar.de (german) is different. This feature set is only used for the predictive modelling based on all recipes. Table 4.1 shows all used categories and their number of recipes.

4.5 Comparative Statistical Analysis

One of the goals of this thesis phase was to identify differences, or simmilarities, between both data stets. This comparitive analysis was based on the predictive features, created in the feature engineering and popularity analysis phase. The outcome of this is especially interessting and relevant for the answer of RQ1 and RQ4.

As a first step, all engineered features were compared. Therefore, statistical standard metrics like mean, median, minimum or maximum of each feature and both datasets got computed. In order to show that there are significancy differences, statistical tests were conducted. Because of the variing feature distributions it was necessary to utilize different tests that can cope with all present distribution assumptions. First, the bigger datasets needed random subsampling to match the size of the smaller ones, as unequal sample sizes can cause bias in the statistical tests. Next, a Brown–Forsythe test for finding statistically equal variance was utilized. The Brown–Forsythe test is used for group comparison based on median

Table 4.1: The top twenty overlapping categories of Allrecipes.com and Kochbar.de

Category	#Recipes (Allrecipes.com)	#Recipes (Kochbar.de)
Main dishes	11194	81241
Desserts	11526	32098
Appetizers and snacks	3978	24547
Side dishes	7549	14279
Soup stews and chili	3618	19434
Cakes	1450	35761
Breakfast and brunch	1550	9199
Salad	3076	20365
Pasta and noodles	1571	13269
Roast	442	16418
Casseroles	1289	23569
Low calorie	442	2347
Healthy	109	21763
Veggie	129	29243
Stir fry	303	5715
Asia style	41	9686
Pizza	203	7236
Deep fried	102	456
Italy and italian style	84	7303
Meat and poultry	2632	37925

absolute deviations (MAD). Compared to the Leven's test, which uses the mean, it is more robust against outliers. The p-value for this test got defined as 0.05, which is a standard value for this type of test. In the case of equal variance of the two feature distributions ($p < 0.05$) a Mann-Whitney U test was performed. When the test rejected the equal variance hypothesis, a two sample Kolmogorov–Smirnov (KS) test was used. A Mann–Whitney U test should not be used when two sample distributions have unequal variances, because this can lead to faulty significance results. KS however, does not make assumptions about the distribution of the sample data. An other feature of the KS test is its robustness: It compares the empirical distribution functions of the two samples. Notwithstanding, it is considered to be not as accurate or precise as the Mann-Whitney U test. In order to complement the statistical significancy test, the effect size of all feature distributions pairs was calcuated.

The comparison between the two datasets and each category distribution was done with a Mann–Whitney U test. Furthermore, the top twenty overlapping categories within each dataset and each popularity/appreciation metric were tested. Therefore, the Kruskal–Wallis test (also named One-way ANOVA on ranks), a Mann–Whitney U test (also called Wilcoxon rank-sum test) for more than two groups got utilized. The exact same analysis was also done for recipe ingredients and recipe title words.

4.6 Predictive Modelling

To reveal the extend to which the popularity of a online recipe can be predicted and which differences are between the two datasets, a predictive modelling approach was followed.

A statistical analysis was performed to find out, how and which features can be used for the purpos of popularity prediction. A common approach is to do a correlation analysis, which exposes possible feature correlations in general and between preditors (features) and outcome variables (popularity proxies). The correlation analysis of this experiment utilizes Spearman's rank correlation coefficient, since it assesses linear and non linear relationships between two variables and can cope with continous and descrete variables. To predict whether a recipe gets popular or not, a classifcation task rather than a regression task, as used by Rokicki, Herder, and Trattner [2017], was chosen. Therefore, binary classificator variables need to be created. These depict to which class a recipe belongs, so the classification models can be validated with test data. As such we calculated medias of the popularity metrics, following the approach of Shulman et al. [2016]. Recipes below the median are considered as negative and above as positive example.

The dataset preparation for the classification experiment includes multiple steps. First, upload date filtering was done to prevent right censoring. Secondly, potential missing data were imputed with the R library Hmisc[36]. Since the evaluation of the classification experiments was performed with balanced datasets, imbalanced classes needed to be dealt with. In order to do that, the major class was subsampled with random boostraping. Next, Information Gain (IG) was used to prune features and to reveal their discrimantive power prior to the classification experiment. The top ten features found by the IG were selected for further investigation.

The popularity prediction experiment follows an before publication approach, meaning that no activity after the upload of a recipe will be taken into account. Such an approach is considered difficult by the litrature, for example Shulman et al. [2016]. For possible applications such as supporting a user while uploading a recipe with information that can boost the popularity of his recipe, an cold start approach is more practical. However, the classification experiment was conducted with 3 different classifiers and by using the statistical computing engine R[37]. Besides a Random Forests (RF) classifier, Generalized Linear Models (GLM) and Naive Bayes (NB) were employed. Shulman et al. [2016] stated that RF and NB are successfully applyed in similar 'before puplication' popularity classification studies. A five fold-cross validation was chosen for the evaluation protocol. Furthermore, the variable importance for each feature set was reported. The used R packages for this experiments were: rminer, e1071, caret, Metrics, elasticnet, doMC, ggplot2, MASS, readr, pscl, PerformanceAnalytics, lmtest, stargazer, BaylorEdPsych, caret, VIM, FSelector, rJava, CORElearn, RWeka, and Hmisc.

[36] https://www.rdocumentation.org/packages/Hmisc/versions/4.0-3

[37] https://www.r-project.org/

Chapter 5

Results

“ There are three types of lies – lies, damn lies, and statistics. ”

[Benjamin Disraeli, British politician, 1804–1881.]

This Chapter reports in detail the results of the studies conducted. The first Section explains the popularity analysis results and presents the popularity patterns found. The subsequent statistical analysis compares the created recipe features of both datasets against each other and identifies significant differences. This investigations are followed by a correlation analysis which findes connections between features. The Chapter concludes with an predictive modelling approach that shows to which extent the features can predict the popularity of online recipes, show differences between food cultures and reveal possible similarities in both analysed communities.

5.1 Popularity Analysis (RQ 1)

The following Section describes the results of the popularity analysis of the two datasets. The focus lies on finding popularity patterns of recipes, categories and ingredients.

Former studies have exploited various proxies for measuring the popularity of online content. The classical metric is the number of views, but it is often hidden and not accessible. Modern Web 2.0 community sites such as Allrecipes.com and Kochbar.de provide other indicators, which can be used for this purpose. This study uses the number of comments, ratings and bookmarkings of recipes as proxies for popularity, since both community platforms feature this paradigms. Other metrics were not contained in the present datasets. As already mentioned in Chapter 3, the dataset of Allrecipes.com does not provide separate data for ratings and comments. In a former version of Allrecipes.com’s website, users had to provide both, even when they just wanted to give a rating. For comparison reasons, this thesis always reports the number of ratings and the number of comments for Allrecipes.com. However, they are the same. The second dimension of popularity is ‘time’. The popularity of online content is always time dependend. It evolves over time and either gains or loses popularity. Important to note is that Kochbar.de’s website has a bookmarking like (favorites) feature. Nevertheless,

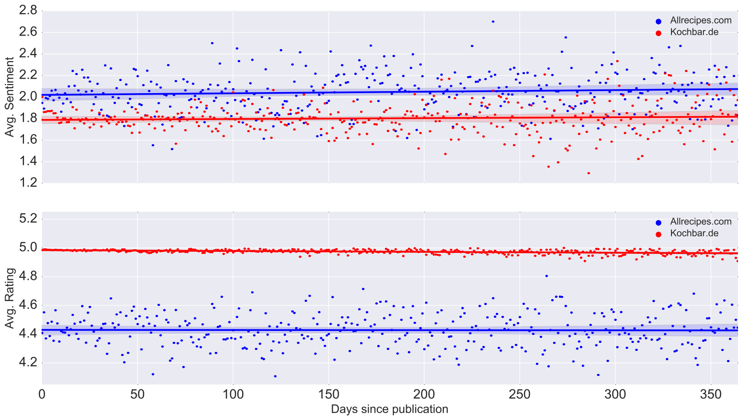
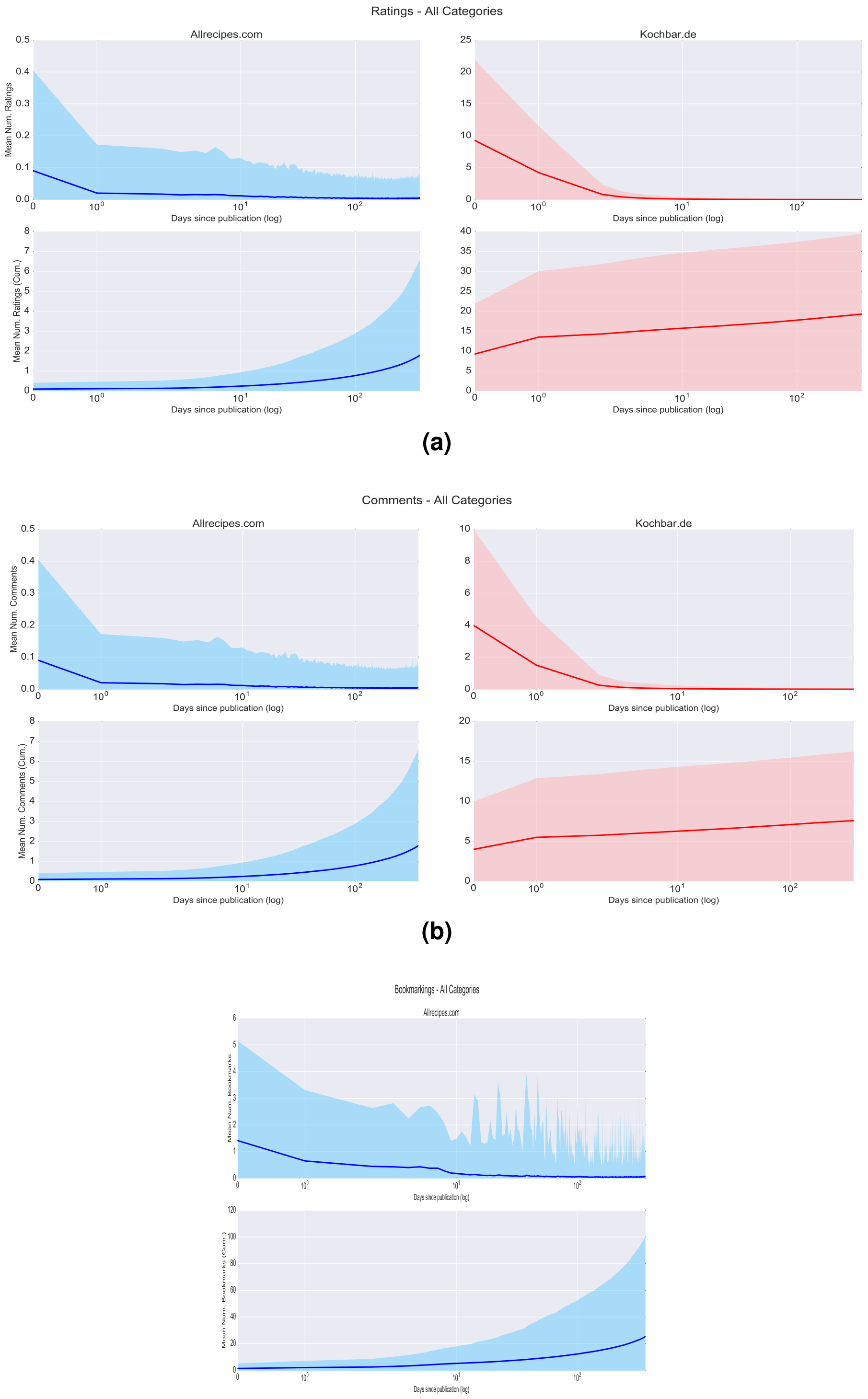


Figure 5.1: This figure illustrate the mean appreciation of recipes measured by the mean sentiment and mean rating of recipes. The lines represent the linear regression of the observations and the lighter colored hulls show the confidence interval of the regression.

this dataset does not contain time data of bookmarkings. Besides popularity, it is possible to measure the appreciation of content. Appreciation is an expression of quality and cannot be quantified by the number of ratings or comments. For that reason, two additional metrics were introduced. First, the mean rating of a recipe and secondly, the mean sentiment of all comments of a recipe.

The Figures 5.1 and 5.2 provide a better understanding of the popularity and appreciataion development over time. All four subfigures have in common that their x-axes measure the number of days since a recipe was uploaded (delta time) until 365 days after. The plots show either the mean or the cumulative mean of the metrics as a continous line and the standard error of the mean as a light colored hull. Figure 5.2a presents the mean number of ratings of all recipes. As one would expect, their is a visual decay in both datasets. The recipes in Allrecipes.com get the most ratings until one day. Within this time period, it rapidly falls from 0.1 ratings on average to about 0.02. After the first day, this development levels more or less out until day eight, where it again starts to fall. Beneath that plot, there is the cumualtive representation of this development. It shows that the mean number of ratings stays under 0.5 until day ten, where it starts to climb to about 2 at day 365. This goes hand in hand with a exponential increase of the standard error starting at about day one. The non-cumulative plot of Kochbar.de has a similar development. However, Kochbar.de's recipes get much more ratings on average. The value falls within the first day from 9 to 4 ratings. Compared to Allrecipes.com, this is a very steep slope. After day one it has a slower rate of falling. At about day seven, it reaches 0.5 and from there on it continuously converges to zero. The cumulative plot on the other hand has a very different appearance. The mean rating starts at 9 and climbs within one day to about 14. After that, it linearly grows until 19 at day 365. The standard error has a almost identical development. Nevertheless, the error is always high, compared to Allrecipes.com. Subfigure 5.2b shows the development of the mean number of comments. When comparing Kochbar.de's plots, ratings and comments seem to have very much the



(a)

(b)

(c)

Figure 5.2: Figure (a) shows the mean number of ratings a recipe receives whereas (b) depicts the mean number of comments of a recipes. Figure (c) presents the mean number of bookmarks of Allrecipes.com recipes. All three figures show a timespan of 360 days. The error of the mean is illustrated by the lighter colored hulls around the lines.

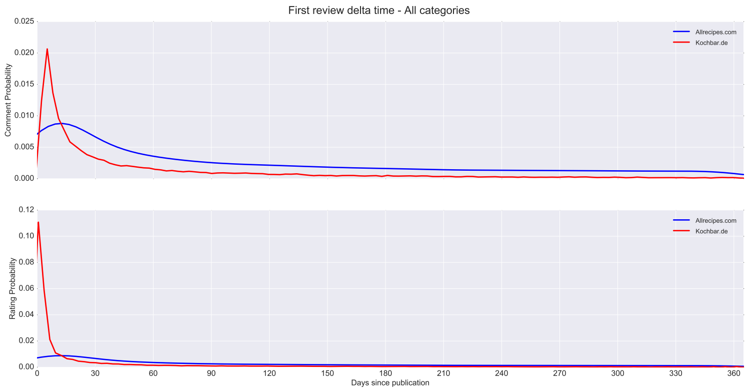


Figure 5.3: The recipes possibility of obtaining the first comment or rating after publication over the course of one year.

same trend. However, the mean number of comments value is lower at all times. It starts at 4 and falls within the first day to 1.5, from there on it again starts to fall to 0.2 at day 7. The same is true for the cumulative plot, as it follows the same trend than the cumulative ratings plot. The value starts at about 4 and grows till 6 after one day. From then on it linearly grows till 7.5 within one year. Also, the error continuously increases over time. Subfigure 5.2c illustrates the bookmarking development of Allrecipes.com. Compared to Allrecipes.com's ratings and comments plots, it has a more than ten times higher value at day zero. Within the first day, it declines from 1.4 to 0.7. From than on, it is falling with a slower rate until day seven, where it reaches about 0.5, afterwards it starts again falling faster. The standard error follows pretty much this development. However, at about day five it begins to fluctuate heavily without any obvious patterns. The cumulative plot underneath, illustrates a more steady development. The mean and error value are slowly increasing. An interesting oberservation is that the users of Allrecipes.com are bookmarking recipes far more, than providing ratings or comments. The number of bookmarks is almost comparable with Kochbar.de's comments. However, the users of Kochbar.de are more active in general. This becomes even more obvious, when comparing the numbers of users who provided ratings or comments in both datasets. In Allrecipes.com, $125,762$ users provided ratings/comments, wheareas in Kochbar.de only $19,444$ provided ratings and $21,951$ provided comments. Subfigure 5.1 exposes the appreciation development over time. In both datasets, mean sentiment as well as the mean rating do only change slightly over time. The standard errors of both mean sentiment plots are comparable. This is in contrast to the mean rating plot, which shows that the users of Kochbar.de seem to provide only good ratings (4 or 5 stars). Allrecipes.com's users also provide good ratings, but not as much as the ones in Kochbar.de. However, it can be seen that the appreciation metrics of both datasets lack variability.

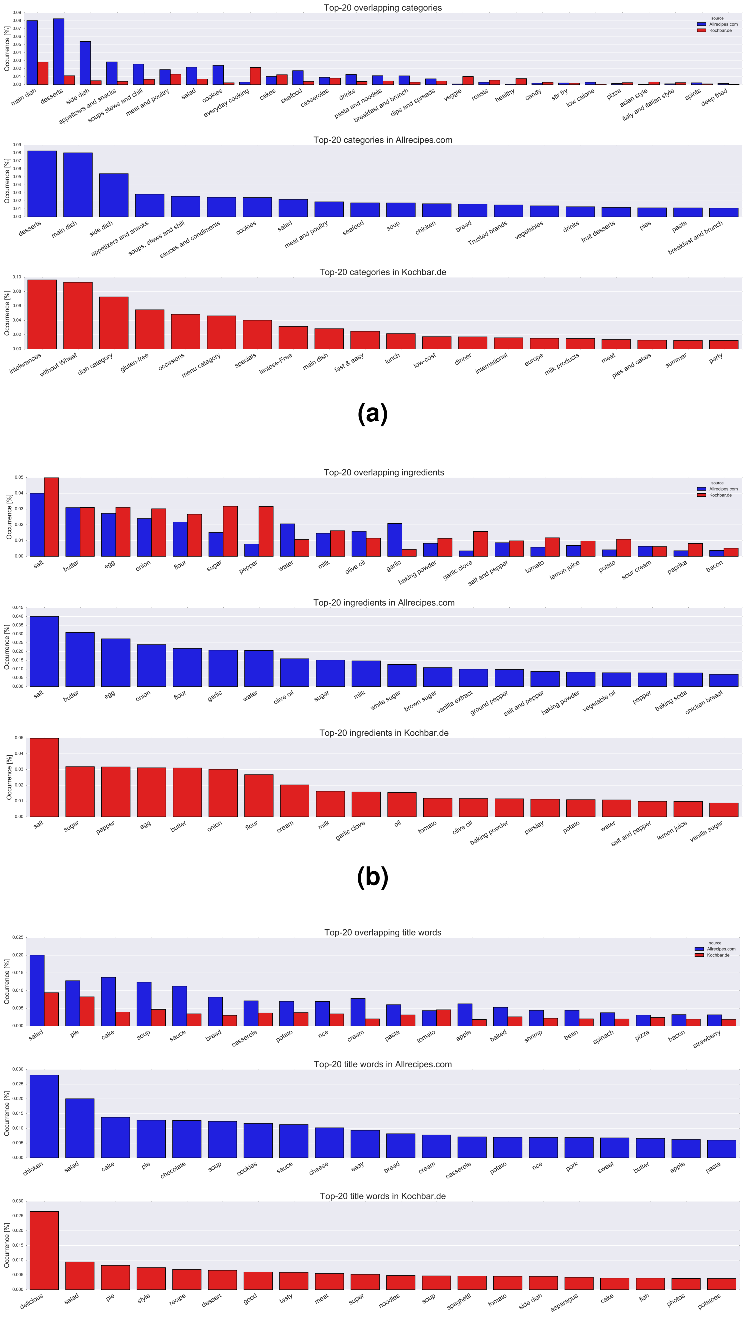
In order to complement the analysis of ratings, comments and bookmarks, Figure 5.3 reveals the probability of a recipe getting the first comment or rating, in relation to the days since it was uploaded. It supports the assumption that Kochbar.de recipes are more likely to get comments or ratings than the ones of Allrecipes.com. And yet, this effect only lasts about two weeks. After this time, recipes uploaded to Allrecipes.com have the higher possibility to gain attention.

The cause of this is not clear and should be further investigated. A possible explanation could be that Allrecipes.com features a recommendation system that brings up older recipes as well. This ensures that as many recipes as possible get ratings or comments and may helps increase user acceptance and satisfaction.

When speaking about recipe popularity, the intuition would suggest that there are recipe categories or ingredients, which are more popular than others. There are also cultural stereotypes, for example that people from the US eat a lot of meat or that germans like sausages and sauerkraut. If these claims are true, such aspects could be very helpful in predicting popularity.

Figures 5.4 shows the top 20 most occuring categories, ingredients and title words of both datasets. The entities of Kochbar.de got translated from german to english, so that they can be compared to the ones of Allrecipes.com. The upper part of the figures shows the overlapping comprisons. These entities have been mapped manually, because some are semantically the same, but are not named equally in Allrecipes.com and Kochbar.de. Subfigure 5.4a illustrates the recipe category occurrences within the datasets. Important to note is that recipes can be assigned to multiple categories. Almost 10% of the recipes in Kochbar.de are labeled as consumable for people with intolerances, followed by recipes without wheat category. This is in contrast to Allrecipes.com, where ‘desserts’ is the most popular category. Second one is ‘main dishes’ with about 8%. The datasets do not show many similarities in the means of category popularity. The highest ranked overlapping category ‘main dish’, does occure in Kochbar.de only at position nine. Subfigure 5.4b gives insightes in to the popularity distributions of the ingredients. This plot does however, show analogies between the datasets. Salt, butter, eggs, onions, flour and sugar are heavily used in both datasets. The last Subfigure 5.4c, shows the popularity of the words used in recipe titles. It is obvious that the users in Kochbar.de do use more adjectives such as ‘delicious’, ‘good’ or ‘tasty’, whereas Allrecipes.com’s title words are more of neutral nature. However, the occurring words are again very similar. Most of the top 20 overlapping title words or ingredients can be found also in the non-overlapping top 20 of the datasets.

The bar-plots 5.4 show that there are entities, which are used more frequently than others. Yet another interessting question to answer is, how the categories, ingredients and title words relate to the popularity proxies and whether comparable patterns can be observed. Table 5.1 illustrates the analysis of the recipe categories. Like the bar-plots of Figure 5.4, this Table shows the top 20 most occuring categories of each dataset and the top 20 overlapping ones. The statistics are calculated over all recipes within the specified category. The top five categories within each metric are denoted with an arrow up (↑) and the bottom five with an arrow down (↓). However, the five most used categories, are not within the top five of any metrics. Only Allrecipes.com’s ‘main dish’ category is under the top five in number of ratings, comments and mean sentiment. The statistics of Allrecipes.com show that there is no direct relation between the number of comments/ratings and average rating/sentiment. Only ‘main dish’ and ‘breakfast and brunch’ are within the top five of the popularity and appreciation metrics. The most popular category is ‘chicken’ (0.28), best rated is ‘trusted brands: recipes and tips’ (4.38) and the highest sentiment has ‘salad’ $(1,95)$. This shows that there is not the one popular or appreciated category. The relative change of the mean ratings/comments value is 67%, whereas the mean rating and sentiment only change 4.5% and 10.7%.



Top-20 overlapping categories

Top-20 categories in Allrecipes.com

Top-20 categories in Kochbar.de

(a)

Top-20 overlapping ingredients

Top-20 ingredients in Allrecipes.com

Top-20 ingredients in Kochbar.de

(b)

Top-20 overlapping title words

Top-20 title words in Allrecipes.com

Top-20 title words in Kochbar.de

(c)

Figure 5.4: Figure (a) shows the number of occurrences of recipe categories, (b) presents the ingredients occurences and (c) illustrates the title words occurences in the whole recipes body. All figures show the occurences numbers for Allrecipes.com and Kochbar.de separatly, as well as the occurences numbers for the overlapping entities (present in both datasets).

This observation goes hand in hand with Figure 5.1. Kochbar.de shows a interesting pattern. Altough ratings and comments are not coupled toghether, they seem to correlate. This may not be an suprising observation, but it indicates that users

are likely to rate and comment in order to express their opinion on a recipe. Mean number of ratings and mean rating seem to correlate too. Futhermore, the number of comments is comparable to the values of Allrecipes.com. The bottom five categories in terms of number of ratings, are also the bottom five of mean rating. The category with the most ratings and comments is 'lunch' (0.84 and 0.29), the one with the highest rating and sentiment is 'meat' (4.96 and 1.85). Relative change of number of ratings is 46% and 48% for number of comments. Mean rating does only change 0.2%, wheareas mean sentiment has a relative change of 6%. The bottom part of Table 5.1 compares the overlapping categories of both datasets. Besides the low relative change of mean rating and mean sentiment, no obvious similarities between Allrecipes.com and Kochbar.de can be found. Only the 'main dish' and 'drinks' categroy seem to have the same tendencies. This observation is underpinned by statistical tests. For almost all categories,
the Mann–Whitney U test rejects the null hypothsis with $p < 0.001$ (donted with a $\star$) that the underlying distributions of the popularity/appreciation metrics of both datasets are equal. Only 'healthy', 'pasta' and 'casseroles' are above the significancy level of $p < 0.1$. Additionally, the categorical distribution tendencies within the datasets and each metric were tested. The Kruskal-Wallis test strongly rejected the null hypothesis for all metrics of both datasets.

The ingredients analysis, reported by Table 5.2, was done in the same fashion as for the categories. Three of the five ingredients with the highest mean rating/comment values are under the five most used ingredients in Allrecipes.com. The most popular ingredient is 'chicken breast' with a value of 0.23, followed by 'backing soda' (0.2) and 'salt' (0.18). The number of comments/ratings do not seem to correlate with the appreciation metrics. The ingredients with the highest mean ratings are 'olive oil' (4.34), 'ground pepper' (4.33) and 'garlic' (4.32). Regarding mean sentiment, the three best are again 'olive oil' (1.94), 'garlic' (1.91) and 'pepper' (1.90). The relative change of the popularity metrics is $46, 8\%$, 3% for mean rating and 0.05% for mean sentiment. Like for the categories, Kochbar.de's ingredient popularity metrics seem to be correlated. The top and bottom five ingredients measured with number of ratings and comments, match exactly. An other interesting observation is that the number of comments is roughly about a third of the number of ratings. The three top ingredients are 'water' (1.01) and 'vanilla sugar' (1.00) with over one rating per recipe on average and 'pepper' (0.96). The most appreciated ingredients in terms of mean rating are 'vanilla sugar' (4.97), 'pepper' (4.97) and 'garlic glove' (4.97). For mean sentiment they are 'cream' (1.89), 'tomato' (1.88) and 'oil' (1.87). The relative change of number of ratings is 27.7%, 28.5% for number of comments, 0.02% for mean rating and $8, 4\%$ for mean sentiment.

The overlapping ingredients, have again no obvious mutual popularity or appreciation pattern. Almost non of the ingredients which are in the Allrecipes.com's top five of the popularity metrics, is it also in the top five of Kochbar.de. Only 'paprika' seem to be popular in both datasets. This is again backed by the Mann–Whitney U which strongly rejects ($p < 0.001$) the null hypothesis for all ingredients and all metrics. Even within the metrics of Allrecipes.com, the distriutions seem to differ. The conducted Kruskal-Wallis test, rejects the null hypothesis for all metrics and both datasets with $p < 0.001$.

The last analysis of this Section covers the popularity of the words used in

Table 5.1: Recipe Category Popularity. $^{\circ}p < 0.1,^{\diamond} p < 0.01,^{\star} p < 0.001$

	Allrecipes.com					Kochbar.de			
Category	Mean Num. Ratings	Mean Num. Comments	Mean Rating	Mean Sentiment	Category	Mean Num. Ratings	Mean Num. Comments	Mean Rating	Mean Sentiment
desserts	0.13 ↓	0.13 ↓	4.26 ↓	1.83 ↓	intolerances	0.47 ↓	0.16 ↓	4.95 ↓	1.78
main dish	0.21 ↑	0.21 ↑	4.27	1.89 ↑	without wheat	0.47 ↓	0.16 ↓	4.95 ↓	1.78
side dish	0.17	0.17	4.32	1.85	dish category	0.71	0.23	4.96	1.79
appetizers and snacks	0.14 ↓	0.14 ↓	4.32	1.82 ↓	gluten-free	0.45 ↓	0.15 ↓	4.95 ↓	1.78
soups, stews and chili	0.17	0.17	4.32	1.85	occasions	0.74	0.24	4.96	1.77
sauces and condiments	0.16	0.16	4.35 ↑	1.83 ↓	menu category	0.74	0.25	4.96 ↑	1.82 ↑
cookies	0.15	0.15	4.20 ↓	1.81 ↓	specials	0.73	0.24	4.96	1.77 ↓
salad	0.13 ↓	0.13 ↓	4.36 ↑	1.95 ↑	lactose-free	0.47 ↓	0.15 ↓	4.95 ↓	1.74 ↓
meat and poultry	0.23 ↑	0.23 ↑	4.23 ↓	1.87	main dish	0.80 ↑	0.28 ↑	4.96 ↑	1.85 ↑
seafood	0.16	0.16	4.31	1.90 ↑	fast & easy	0.76 ↑	0.25 ↑	4.96	1.79
soup	0.17	0.17	4.32	1.85	lunch	0.84 ↑	0.29 ↑	4.96 ↑	1.82 ↑
chicken	0.28 ↑	0.28 ↑	4.20 ↓	1.87	low-cost	0.78 ↑	0.26 ↑	4.96	1.78
bread	0.23 ↑	0.23 ↑	4.28	1.86	dinner	0.81 ↑	0.27 ↑	4.96 ↑	1.79
trusted brands: recipes and tips	0.15	0.15	4.38 ↑	1.93 ↑	international	0.58	0.18	4.96	1.79
vegetables	0.18	0.18	4.30	1.88	europe	0.57 ↓	0.18 ↓	4.95 ↓	1.79
drinks	0.09 ↓	0.09 ↓	4.35 ↑	1.74 ↓	milk products	0.58	0.20	4.96	1.81 ↑
fruit desserts	0.16	0.16	4.29	1.88	meat	0.70	0.25	4.96 ↑	1.85 ↑
pies	0.10 ↓	0.10 ↓	4.33 ↑	1.88	pies and cakes	0.76	0.24	4.95	1.75 ↓
pasta	0.20	0.20	4.18 ↓	1.84	summer	0.73	0.24	4.96	1.77 ↓
breakfast and brunch	0.24 ↑	0.24 ↑	4.31	1.89 ↑	party	0.65	0.21	4.95	1.74 ↓
Overlapping									
main dish	0.21 ↑ ★	0.21 ↑ ★	4.27★	1.89 ↑ ★		0.80 ↑	0.28 ↑	4.96 ↑	1.85 ↑
desserts	0.13 ↓ ★	0.13 ↓ ★	4.26★	1.83★		0.69	0.22	4.96	1.78
side dish	0.17★	0.17★	4.32★	1.85★		1.07 ↑	0.34 ↑	4.97 ↑	1.79
appetizers and snacks	0.14★	0.14★	4.32★	1.82★		0.68	0.22	4.96	1.75
soups, stews and chili	0.17★	0.17★	4.32 ↑ ★	1.85★		0.65	0.22	4.95 ↓	1.80
meat and poultry	0.23 ↑ ★	0.23 ↑ ★	4.23★	1.87★		0.70	0.25	4.96 ↑	1.85 ↑
salad	0.13 ↓ ★	0.13 ↓ ★	4.36 ↑ ★	1.95 ↑ ★		0.66	0.22	4.95 ↓	1.80
cookies	0.15★	0.15★	4.20★	1.81 ↓ ★		0.59 ↓	0.16 ↓	4.96	1.68 ↓
everyday cooking	0.04 ↓ ★	0.04 ↓ ★	4.09 ↓ ○	1.68 ↓ ★		0.84 ↑	0.29 ↑	4.96 ↑	1.82
cakes	0.14★	0.14★	4.29★	1.85★		0.76	0.24	4.95	1.75
seafood	0.16★	0.16★	4.31★	1.90 ↑ ★		0.64	0.23	4.96	1.81
casseroles	0.20★	0.20○	4.17 ↓ ★	1.82		0.63 ↓	0.22	4.96	1.90 ↑
drinks	0.09 ↓ ★	0.09 ↓ ★	4.35 ↑ ★	1.74 ↓ ★		0.43 ↓	0.12 ↓	4.94 ↓	1.55 ↓
pasta	0.20★	0.20○	4.18★	1.84		0.64	0.22	4.95 ↓	1.89 ↑
breakfast and brunch	0.24 ↑ ★	0.24 ↑ ★	4.31★	1.89 ↑ ★		1.07 ↑	0.34 ↑	4.95 ↓	1.68 ↓
dips and spreads	0.13 ↓ ★	0.13 ↓ ★	4.35 ↑ ★	1.85★		0.68	0.20 ↓	4.96	1.72 ↓
veggie	0.17★	0.17★	4.18 ↓ ★	1.68 ↓ ★		0.84 ↑	0.27	4.96	1.73
roasts	0.25 ↑ ★	0.25 ↑ ★	4.41 ↑ ★	2.02 ↑ ★		0.76	0.28 ↑	4.97 ↑	1.83 ↑
healthy	0.28 ↑ ★	0.28 ↑	4.07 ↓ ★	1.86		0.57 ↓	0.19 ↓	4.96	1.78
candy	0.15★	0.15★	4.08 ↓ ★	1.56 ↓		0.61 ↓	0.18 ↓	4.96	1.73 ↓
p-Value	0.0	0.0	5.04e-105	1.06e-33		0.0	0.0	2.75e-131	0.0

recipes titles. For the most part, the kind of words used do not differ much between Allrecipes.com and Kochbar.de. For example, dish names or ingredients are very typical title words. The users in Kochbar.de however, also seem to advertise their recipes with adjectives such as ‘tasty’ or ‘good’, whereas the users of Allrecipes.com often use the word ‘easy’. Similar to the analysis of categories and ingredients, the popularity metrics do not seem to correlate with the appreciation metrics. Only Kochbar.de shows some analogy between number of ratings, number of comments and mean ratings. The most popular title words of Allrecipes.com when, meassured by number of ratings/comments, are ‘chicken’ (0.22), ‘pork’ (0.20) and ‘bread’ (0.20). These metrics have a relative change of 50%. The most popular measured by mean ratings are ‘salad’ (4.35), ‘sweet’ (4.34) and ‘pie’ (4.33). In terms of mean sentiment, ‘salad’ (1.95), ‘bread’ (1.93) and ‘sweet’ (1.92) are representing the top three. The relative changes of mean ratings and sentiment are 3.6% and 7.9%, respectively. Kochbar.de’s top three measured by number of ratings are ‘dessert’ (1.12), ‘side dish’ (1.11) and ‘salad’ (1.02). For number of comments they are ‘side dish’ (0.41), ‘dessert’ (0.39) and ‘salad’ (0.36). The most appreciated title words in terms of mean rating are ‘pho-

Table 5.2: Recipe Ingredients Popularity. $^{\circ}p < 0.1,^{\diamond} p < 0.01,^{\star} p < 0.001$

	Allrecipes.com					Kochbar.de			
Ingredient	Mean Num. Ratings	Mean Num. Comments	Mean Rating	Mean Sentiment	Ingredient	Mean Num. Ratings	Mean Num. Comments	Mean Rating	Mean Sentiment
salt	0.18 ↑	0.18 ↑	4.30	1.88	salt	0.84	0.30	4.97	1.85
butter	0.18 ↑	0.18 ↑	4.31	1.89	sugar	0.91 ↑	0.32 ↑	4.97	1.82 ↓
egg	0.17	0.17	4.27	1.85 ↓	pepper	0.96 ↑	0.34 ↑	4.97 ↑	1.87 ↑
onion	0.17	0.17	4.30	1.87	egg	0.86	0.31	4.96 ↓	1.83 ↓
flour	0.18 ↑	0.18 ↑	4.26 ↓	1.85	butter	0.79	0.28	4.97	1.84
garlic	0.18	0.18	4.32 ↑	1.91 ↑	onion	0.85	0.30	4.96	1.86 ↑
water	0.14 ↓	0.14 ↓	4.27	1.83 ↓	flour	0.87	0.31	4.97	1.84
olive oil	0.16 ↓	0.16 ↓	4.34 ↑	1.94 ↑	cream	0.81	0.29	4.96	1.89 ↑
sugar	0.11 ↓	0.11 ↓	4.31 ↑	1.88	milk	0.78 ↓	0.27 ↓	4.96 ↓	1.83
milk	0.17	0.17	4.22 ↓	1.83 ↓	garlic clove	0.83	0.29	4.97 ↑	1.86
white sugar	0.17	0.17	4.30	1.86	oil	0.73 ↓	0.25 ↓	4.96	1.87 ↑
brown sugar	0.18	0.18	4.29	1.89	tomato	0.82	0.28	4.96 ↓	1.88 ↑
vanilla extract	0.16 ↓	0.16 ↓	4.26	1.84 ↓	olive oil	0.73 ↓	0.25 ↓	4.97	1.85
ground pepper	0.18	0.18	4.33 ↑	1.90 ↑	baking powder	0.77 ↓	0.26 ↓	4.96 ↓	1.75 ↓
salt and pepper	0.17	0.17	4.26 ↓	1.87	parsley	0.79	0.28	4.97 ↑	1.85
baking powder	0.18	0.18	4.21 ↓	1.84 ↓	potato	0.89 ↑	0.32 ↑	4.97 ↑	1.85
vegetable oil	0.18	0.18	4.28	1.87	water	1.01 ↑	0.35 ↑	4.96 ↓	1.81 ↓
pepper	0.14 ↓	0.14 ↓	4.31 ↑	1.91 ↑	salt and pepper	0.78 ↓	0.25 ↓	4.97	1.84
baking soda	0.20 ↑	0.20 ↑	4.28	1.88	lemon juice	0.79	0.27	4.97	1.84
chicken breast	0.23 ↑	0.23 ↑	4.22 ↓	1.90 ↑	vanilla sugar	1.00 ↑	0.33 ↑	4.97 ↑	1.73 ↓
Overlapping									
salt	0.18 ↑ ⋆	0.18 ↑ ⋆	4.30⋆	1.88⋆		0.69	0.23	4.96	1.80
butter	0.18 ↑ ⋆	0.18 ↑ ⋆	4.31⋆	1.89⋆		0.65	0.21	4.95	1.78
egg	0.17⋆	0.17⋆	4.27⋆	1.85⋆		0.67	0.22	4.95 ↓	1.77
onion	0.17⋆	0.17⋆	4.30⋆	1.87⋆		0.68	0.23	4.95	1.84 ↑
flour	0.18 ↑ ⋆	0.18 ↑ ⋆	4.26 ↓ ⋆	1.85⋆		0.66	0.21	4.95 ↓	1.76 ↓
sugar	0.11 ↓ ⋆	0.11 ↓ ⋆	4.31⋆	1.88⋆		0.66	0.21	4.95 ↓	1.75 ↓
pepper	0.14 ↓ ⋆	0.14 ↓ ⋆	4.31⋆	1.91 ↑ ⋆		0.77 ↑	0.27 ↑	4.96 ↑	1.84 ↑
water	0.14⋆	0.14⋆	4.27⋆	1.83 ↓ ⋆		0.78 ↑	0.25 ↑	4.95	1.72 ↓
milk	0.17⋆	0.17⋆	4.22 ↓ ⋆	1.83 ↓ ⋆		0.63 ↓	0.20 ↓	4.95 ↓	1.76 ↓
olive oil	0.16⋆	0.16⋆	4.34 ↑ ⋆	1.94 ↑ ⋆		0.58 ↓	0.19 ↓	4.96	1.81
garlic	0.18⋆	0.18⋆	4.32 ↑ ⋆	1.91⋆		0.83 ↑	0.29 ↑	4.96 ↑	1.79
baking powder	0.18⋆	0.18⋆	4.21 ↓ ⋆	1.84 ↓ ⋆		0.67	0.20	4.95 ↓	1.73 ↓
garlic clove	0.10 ↓ ⋆	0.10 ↓ ⋆	4.36 ↑ ⋆	1.92 ↑ ⋆		0.66	0.22	4.96 ↑	1.83
salt and pepper	0.17⋆	0.17⋆	4.26 ↓ ⋆	1.87⋆		0.59 ↓	0.19 ↓	4.95	1.83
tomato	0.14 ↓ ⋆	0.14 ↓ ⋆	4.32 ↑ ⋆	1.93 ↑ ⋆		0.65	0.22	4.95	1.84 ↑
lemon juice	0.14 ↓ ⋆	0.14 ↓ ⋆	4.30⋆	1.89⋆		0.60 ↓	0.19 ↓	4.96	1.78
potato	0.15⋆	0.15⋆	4.25 ↓ ⋆	1.84 ↓ ⋆		0.72 ↑	0.25 ↑	4.96	1.83
sour cream	0.19 ↑ ⋆	0.19 ↑ ⋆	4.29⋆	1.81 ↓ ⋆		0.60 ↓	0.20 ↓	4.95	1.84
paprika	0.22 ↑ ⋆	0.22 ↑ ⋆	4.29⋆	1.87⋆		0.90 ↑	0.32 ↑	4.96 ↑	1.85 ↑
bacon	0.17⋆	0.17⋆	4.39 ↑ ⋆	1.93 ↑ ⋆		0.65	0.23	4.96 ↑	1.84 ↑
p-Value	2.34e-178	2.34e-178	5.14e-68	5.02e-48		0.0	0.0	2.19e-209	0.0

tos' (4.98), 'side dish' (4.97) and 'potato' (4.97) and for mean sentiment they are 'delicious' (1.89), 'cake' (1.89) and 'tomato' (1.88). Again, the number of comments is about a third of the number of ratings. Also the relative change values are comparable with former results: Number of ratings 48%, number of comments 53%, mean rating 0.6% and mean sentiment 8%.

The overlapping title words, just like overlapping categories or ingredients, do not show relations between the top or bottom five words for either of the metrics. Only the word 'soup' seem to have a simmilar popularity and appreciation rank. A interessting oberservation is that Mann–Whitney U strongly rejects the null hypothesis only for number of ratings and mean rating metrics. Number of comments and mean sentiment seem to have only a few differing distributions, like for the words 'pasta', 'pie', 'pizza' or 'baked'. The Kruskal-Wallis test strongly rejects for all metrics.

Summary. It has been shown that popularity and appreciation patterns exist and that there are differences between the online recipes of Allrecipes.com and

Table 5.3: Recipe Title Words Popularity. $^\circ p < 0.1, ^\diamond p < 0.01, ^\star p < 0.001$

	Allrecipes.com					Kochbar.de			
Title Word	Mean Num. Ratings	Mean Num. Comments	Mean Rating	Mean Sentiment	Title Word	Mean Num. Ratings	Mean Num. Comments	Mean Rating	Mean Sentiment
chicken	0.22 ↑	0.22 ↑	4.24 ↓	1.89 ↑	delicious	0.71 ↓	0.25 ↓	4.96 ↓	1.89 ↑
salad	0.11 ↓	0.11 ↓	4.35 ↑	1.95 ↑	salad	1.02 ↑	0.36 ↑	4.97	1.82
cake	0.15	0.15	4.30	1.85	pie	0.86	0.29	4.96	1.78 ↓
pie	0.11 ↓	0.11 ↓	4.33 ↑	1.88	style	0.92	0.33	4.97	1.84
chocolate	0.15	0.15	4.25	1.81 ↓	recipe	0.76	0.29	4.96	1.73 ↓
soup	0.15	0.15	4.32 ↑	1.85	dessert	1.12 ↑	0.39 ↑	4.95 ↓	1.82
cookies	0.16	0.16	4.19 ↓	1.81 ↓	good	0.87	0.31	4.96	1.86
sauce	0.14	0.14	4.29	1.87	tasty	0.91	0.32	4.96	1.81
cheese	0.14	0.14	4.29	1.85	meat	0.87	0.31	4.97	1.87
easy	0.18 ↑	0.18 ↑	4.25	1.82 ↓	super	0.58 ↓	0.19 ↓	4.95 ↓	1.83
bread	0.20 ↑	0.20 ↑	4.33 ↑	1.93 ↑	noodles	0.77	0.27	4.96	1.88 ↑
cream	0.15	0.15	4.30	1.86	soup	0.73 ↓	0.25 ↓	4.95 ↓	1.79 ↓
casserole	0.19 ↑	0.19 ↑	4.19 ↓	1.80 ↓	spaghetti	0.74 ↓	0.25 ↓	4.96	1.85
potato	0.17	0.17	4.29	1.85	tomato	0.93 ↑	0.32	4.97 ↑	1.88 ↑
rice	0.14 ↓	0.14 ↓	4.18 ↓	1.82 ↓	side dish	1.11 ↑	0.41 ↑	4.97 ↑	1.80
pork	0.20 ↑	0.20 ↑	4.29	1.88	asparagus	0.91	0.34 ↑	4.97 ↑	1.79 ↓
sweet	0.15	0.15	4.34 ↑	1.92 ↑	cake	0.77	0.26 ↓	4.96 ↓	1.89 ↑
butter	0.18	0.18	4.29	1.86	fish	0.74 ↓	0.26	4.96	1.82
apple	0.12 ↓	0.12 ↓	4.26	1.90 ↑	photos	1.04 ↑	0.36 ↑	4.98 ↑	1.80 ↓
pasta	0.14 ↓	0.14 ↓	4.21 ↓	1.88	potato	0.92	0.33	4.97 ↑	1.88 ↑
Overlapping									
salad	0.09 ↓ ★	0.09 ↓ ◦	4.25★	1.92 ↑ ◦		0.65	0.22	4.95	1.80
cake	0.20★	0.20	4.23★	1.88⋄		0.60	0.19 ↓	4.95	1.73 ↓
pie	0.19★	0.19★	4.30★	1.89★		0.72	0.23	4.95 ↓	1.76 ↓
soup	0.07 ↓ ⋄	0.07 ↓	4.07 ↓ ⋄	0.98 ↓ ◦		0.54 ↓	0.18 ↓	4.95 ↓	1.79 ↓
sauce	0.15 ↓ ★	0.15 ↓ ◦	4.25★	1.85◦		0.62	0.21	4.96	1.80
bread	0.21★	0.21	4.32 ↑ ★	1.96 ↑ ⋄		0.69	0.22	4.95 ↓	1.68 ↓
cream	0.22 ↑ ★	0.22 ↑	4.18 ↓ ★	1.82		0.68	0.23	4.95	1.87 ↑
casserole	0.20★	0.20	4.06 ↓ ★	1.74 ↓ ★		0.60 ↓	0.20 ↓	4.95 ↓	1.93 ↑
potato	0.20★	0.20⋄	4.21★	1.76 ↓		0.78 ↑	0.28 ↑	4.96	1.86
rice	0.16 ↓ ★	0.16 ↓	4.14 ↓ ★	1.80 ↓		0.66	0.21	4.95	1.86
apple	0.16★	0.16◦	4.16 ↓ ★	1.90⋄		0.74 ↑	0.25 ↑	4.95	1.82
pasta	0.17★	0.17★	4.20★	1.90		0.80 ↑	0.28 ↑	4.96 ↑	1.88 ↑
baked	0.35 ↑ ★	0.35 ↑	4.26★	1.91★		0.85 ↑	0.28 ↑	4.96 ↑	1.83
beans	0.17★	0.17	4.23★	1.69 ↓		0.74 ↑	0.25 ↑	4.96 ↑	1.84
shrimp	0.24 ↑ ★	0.24 ↑	4.33 ↑ ★	1.89⋄		0.59 ↓	0.20	4.95	1.82
tomato	0.19★	0.19	4.32★	1.93 ↑		0.71	0.24	4.96 ↑	1.83
spinach	0.26 ↑ ★	0.26 ↑	4.24★	1.91◦		0.55 ↓	0.19 ↓	4.96	1.90 ↑
bacon	0.22 ↑ ★	0.22 ↑	4.35 ↑ ★	1.86◦		0.68	0.24	4.96 ↑	1.87 ↑
strawberry	0.08 ↓ ★	0.08 ↓ ◦	4.58 ↑	2.25 ↑ ◦		0.68	0.23	4.95	1.83
pizza	0.17★	0.17	4.40 ↑ ★	2.09 ↑ ★		0.59 ↓	0.20 ↓	4.94 ↓	1.68 ↓
p-Value	1.68e-05	1.68e-05	6.40e-22	2.49e-10		0.0	3.39e-266	1.37e-85	3.71e-181

Kochbar.de. The time dependend analysis of the proxy metrics, revealed that the number of comments, ratings or bookmarks which a recipe receives, drops fast after only a few days. An interesting observation is that although Allrecipes.com has far more registered users, the recipes get less comments and ratings on average then the ones of Kochbar.de. The appreciation analysis showed that recipes of both datasets have predominantly high mean ratings and mean sentiment. Furthermore, this metrics do not vary much and are stable over time. Another essential point is the popularity of categories, ingredients and title words used in recipes. Generally speaking, entities that are often used in recipes, are not necessarily going hand in hand with appreciation or popularity.

5.2 Comparative Statistical Analysis (RQ 2)

The successful studies of other researchers, as well as the results of the previous Section have led to the development of several features. These features are deemed to capture the popularity and appreciation of online recipes. The results of this Section are presented with a extensive Table comparing each feature of both datasets, based on statistical standard metrics. The differences found are underpinned with statistical tests and effect sizes.

Table 5.4 shows the results of the comparative statistical analysis over all relevant recipes of all categories. All features are assigned to one of seven semantically fitting feature sets. The analysis features five different statistical metrics: mean, median, standard deviation, minimum and maximum value. Furthermore, both feature populations are tested for equally distributed samples following the approach described in the methodology Chapter (4).

Set: Recipe Nutrition

The first feature set groups features that describe the nutritional values of recipes. Recipes of Kochbar.de contain more Kcal (244.19 v 204.91 (AR)) and fat (15.99 v 10.58 (AR)) on average, but Allrecipes.com's recipes contain more protein (6.19 v 5.63 (KB)) and carbohydrates (21.84 v 19.25 (KB)). All four features differ significantly ($p < 0.001$) but with small effect sizes ($r = 0.10 - 0.16$).

Set: Recipe Healthiness

Recipe healthiness is the second feature set. Allrecipes.com's recipes are slightly, but significantly ($p < 0.001$, $r = 0.08$) more healthy, based on the WHO score, than the ones of Kochbar.de (3.50 v 3.38 (KB)).

Set: Recipe Complexity

In terms of complexity, Allrecipes.com has a higher prepartion time (73.01 v 36.52 (KB)) and higher number of servings (11.48 v 5.27 (KB)). On the other hand, Kochbar.de has more preparation steps (3.51 v3.27 (AR)), number of ingrdients (10.00 v 9.02 (AR)) and categories (9.25 v 4.39 (AR)). All are significantly different ($p < 0.001$), but the effect sizes of num. steps ($r = 0.02$) and num. ingredients are very low ($r = 0.00$).

Set: Recipe Presentation

The most of the presentation features are upload image related. Altough these features all differ significantly ($p < 0.001$), they only vary little and have low effect sizes ($r = 0.01 - 0.19$). Image sharpness and contrast have the same mean value for both datasets (0.18 *and* 0.05), but the median of Kochbar.com's image contrast is lower (0.04). Only Kochbar.de's image sharpness variation has a higher value (0.30 v 0.29 (AR)). All other image features are higher on average for Allrecipes.com. The next features describe the recipe instructions. Besides the number of words (95.87 v 94.00 (KB)), all other features have a higher mean value. An possible explanation is that the german language needs more words to describe recipes, than the english. When considering that Allrecipes.com's recipes have one ingredient on average more, this is even more intressting. Furthermore, the Kochbar.de's title sentiment is higher. The users also seem to write the description

Table 5.4: Differences between Allrecipes.com and Kochbar.de online food communities based on the created features (all categories).

	Allrecipes.com (n=58424)					Kochbar.de (n=309360)						
Feature	*M*	*Md*	*SD*	*Min*	*Max*	*M*	*Md*	*SD*	*Min*	*Max*	*p*	*r*
Set: Recipe Nutrition												
Kcal (per 100g)	204.91	174.00	123.81	0.00	877.00	244.19	212.00	167.59	1.00	3434.00	<0.001	0.10
Protein (per 100g)	6.19	5.32	4.25	0.00	46.64	5.63	4.00	5.26	0.00	85.00	<0.001	0.13
Carbohydrates (per 100g)	21.84	15.09	18.19	0.00	270.52	19.25	10.30	21.04	0.00	99.50	<0.001	0.16
Fat (per 100g)	10.58	8.26	9.35	0.00	99.07	15.99	10.80	17.67	0.00	100.00	<0.001	0.10
Set: Recipe Healthiness												
WHO Health Score	3.50	3.00	0.74	0.00	6.00	3.38	3.00	0.66	3.00	6.00	<0.001	0.08
Set: Recipe Complexity												
Preparation Time (Min.)	73.01	45.00	256.05	2.00	42635.00	36.52	30.00	151.81	1.00	29440.00	<0.001	0.34
Num. Preperation Steps	3.27	3.00	1.66	1.00	27.00	3.51	3.00	2.68	1.00	60.00	<0.001	0.02
Num. Servings	11.48	8.00	13.78	1.00	832.00	5.27	4.00	56.17	1.00	10250.00	<0.001	0.49
Num. Ingredients	9.02	9.00	3.74	1.00	35.00	10.00	10.00	3.80	1.00	72.00	<0.001	0.00
Num. Catergories	4.39	5.00	1.18	2.00	6.00	9.25	7.00	5.69	1.00	34.00	<0.001	0.51
Set: Recipe Presentation												
Image: Sharpness	0.18	0.16	0.10	0.00	1.22	0.18	0.16	0.12	0.00	1.55	<0.001	0.01
Image: Sharpness Variation	0.29	0.26	0.16	0.00	1.79	0.30	0.26	0.19	0.00	2.44	<0.001	0.01
Image: Contrast	0.05	0.05	0.02	0.00	0.19	0.05	0.04	0.02	0.00	0.18	<0.001	0.11
Image: Contrast RGB	0.17	0.16	0.07	0.00	0.56	0.16	0.15	0.07	0.00	0.55	<0.001	0.09
Image: Saturation	0.25	0.24	0.11	0.00	0.91	0.21	0.20	0.11	0.00	0.86	<0.001	0.19
Image: Saturation Variation	0.15	0.14	0.05	0.00	0.41	0.14	0.14	0.06	0.00	0.41	<0.001	0.07
Image: Brightness	0.51	0.50	0.11	0.00	0.97	0.48	0.48	0.11	0.00	0.98	<0.001	0.00
Image: Colorfulness	0.23	0.23	0.08	0.00	0.71	0.22	0.21	0.08	0.00	0.71	<0.001	0.12
Image: Entropy	7.38	7.51	0.56	0.00	7.97	7.35	7.42	0.36	0.00	7.97	<0.001	0.12
Image: Naturalness	0.81	0.82	0.14	0.00	1.00	0.77	0.77	0.14	0.00	1.00	<0.001	0.14
Instruction: Num. Chars	569.53	512.00	322.40	5.00	3900.00	661.89	586.00	395.86	1.00	8178.00	<0.001	0.12
Instruction: Num. Words	95.87	86.00	55.55	1.00	699.00	94.00	83.00	59.01	1.00	1940.00	<0.001	0.03
Instruction: Num. Sentences	8.64	8.00	4.64	1.00	52.00	9.05	8.00	5.72	1.00	117.00	<0.001	0.01
Instruction: Readability Score	30.95	30.57	6.23	1.00	81.19	47.74	45.18	15.12	1.00	349.23	<0.001	0.76
Instruction: Entropy	4.36	4.36	0.10	1.50	4.86	4.49	4.50	0.12	1.95	5.02	<0.001	0.57
Instruction: Sentiment	0.22	0.00	0.91	-3.00	4.00	0.36	0.00	1.14	-4.00	4.00	<0.001	0.06
Title: Num. Characters	24.89	23.00	9.70	3.00	115.00	44.10	37.00	29.10	1.00	316.00	<0.001	0.38
Title: Num. Words	4.02	4.00	1.74	1.00	26.00	6.72	5.00	5.21	1.00	61.00	<0.001	0.27
Title: Readbility Score	39.06	36.33	27.00	1.00	105.00	55.14	52.00	25.33	1.00	107.00	<0.001	0.29
Title: Entropy	3.70	3.73	0.33	1.50	4.74	3.94	4.02	0.42	0.00	5.17	<0.001	0.35
Title: Sentiment	0.03	0.00	0.34	-3.00	4.00	0.36	0.00	0.95	-4.00	4.00	<0.001	0.13
Set: Recipe Seasonality												
Upload Month	6.41	6.00	3.40	1.00	12.00	6.15	6.00	3.50	1.00	12.00	<0.001	0.04
Day of Month	15.72	16.00	8.71	1.00	31.00	15.69	16.00	8.72	1.00	31.00	<0.1	0.00
Day of Week	4.18	4.00	1.98	1.00	7.00	3.87	4.00	2.01	1.00	7.00	<0.001	0.08
Within Season	0.08	0.08	0.01	0.04	0.31	0.08	0.08	0.01	0.04	0.20	<0.001	0.08
Set: Recipe Popularity and Appreciation												
Avg. Rating	4.29	4.43	0.66	1.00	5.00	4.95	5.00	0.18	1.00	5.00	<0.001	0.67
Avg. Comment Sentiment	1.86	2.00	0.81	-4.00	4.00	1.78	1.80	0.90	-4.00	4.00	<0.001	0.09
Num. Comments received within day	0.11	0.00	0.35	0.00	9.00	4.02	2.00	6.02	0.00	66.00	<0.001	0.61
Num. Comments received within week	0.21	0.00	0.60	0.00	45.00	6.17	3.00	7.92	0.00	104.00	<0.001	0.73
Num. Comments received within month	0.42	0.00	1.20	0.00	87.00	6.74	4.00	8.28	0.00	93.00	<0.001	0.73
Num. Comments received within year	1.77	1.00	4.70	0.00	304.00	7.57	4.00	8.60	0.00	149.00	<0.001	0.60
Num. Ratings received within day	0.11	0.00	0.35	0.00	9.00	9.45	4.00	12.57	0.00	116.00	<0.001	0.72
Num. Ratings received within week	0.21	0.00	0.60	0.00	45.00	15.40	8.00	18.51	0.00	169.00	<0.001	0.77
Num. Ratings received within month	0.42	0.00	1.20	0.00	87.00	16.73	9.00	19.35	0.00	181.00	<0.001	0.78
Num. Ratings received within year	1.77	1.00	4.70	0.00	304.00	19.22	12.00	19.91	0.00	175.00	<0.001	0.74
Set: Recipe Innovation												
Ingredients rank	292.36	229.36	239.98	3.00	2934.00	156.98	132.62	109.84	2.00	1281.00	<0.001	0.33
Categories rank	77.52	52.25	79.06	1.00	925.00	16.46	15.38	9.78	1.50	160.00	<0.001	0.50
Title Words rank	511.42	265.00	701.02	1.00	7509.00	1224.44	564.00	1899.78	1.00	21614.00	<0.001	0.24
Recipe Innovation Jaccard	0.50	0.55	0.18	0.00	1.00	0.47	0.50	0.16	0.00	1.00	<0.001	0.13
Avg. Recipe Innovation Jaccard	0.96	0.96	0.03	0.82	1.00	0.94	0.94	0.03	0.84	1.00	<0.001	0.31
Recipe Innovation IDF	0.02	0.01	0.04	0.00	1.00	0.00	0.00	0.02	0.00	1.00	<0.001	0.48
Set: User Activity & Context												
Recipes Uploaded until Upload	25.47	0.00	94.30	0.00	1824.00	247.13	68.00	507.81	0.00	5841.00	<0.001	0.61
Comments Written until Upload	10.66	0.00	67.48	0.00	2144.00	1179.88	56.00	3250.44	0.00	37136.00	<0.001	0.61
Comments until Upload	94.13	0.00	544.81	0.00	6367.00	1903.25	174.00	4516.74	0.00	52839.00	<0.001	0.66
Num. Distinct Ingredients used	133.44	44.00	232.34	2.00	1443.00	406.01	346.00	309.03	3.00	1344.00	<0.001	0.56
Num. Ingredients used (per recipe)	9.37	9.31	2.55	1.50	28.50	10.01	10.00	1.76	3.00	25.00	<0.001	0.16
Num. Distinct Categories used	28.01	4.00	76.05	1.00	325.00	68.79	70.00	35.20	2.00	165.00	<0.001	0.69
Num. Categories used (per recipe)	2.71	2.89	0.64	1.00	4.00	9.25	7.78	4.45	2.00	25.13	<0.001	0.85
Uploaded Recipes per day	0.05	0.00	0.15	0.00	1.00	0.32	0.13	0.59	0.00	11.33	<0.001	0.68
Uploaded Recipes per week	0.32	0.02	1.02	0.00	7.01	2.25	0.94	3.84	0.00	63.00	<0.001	0.68
Uploaded Recipes per month	1.41	0.10	4.43	0.01	30.48	9.71	3.94	16.92	0.01	270.00	<0.001	0.68
Uploaded Recipes per year	16.93	1.20	53.17	0.06	365.90	116.62	47.76	200.91	0.17	3285.00	<0.001	0.68
Num. Ratings provided per day	0.02	0.00	0.06	0.00	1.93	5.00	1.06	9.29	0.00	95.28	<0.001	0.64
Num. Ratings provided per week	0.13	0.02	0.44	0.00	13.52	34.38	6.76	63.83	0.00	666.98	<0.001	0.65
Num. Ratings provided per month	0.56	0.11	1.88	0.00	57.93	148.66	29.49	277.07	0.00	2858.49	<0.001	0.64
Num. Ratings provided per year	6.60	1.27	22.34	0.00	687.46	1784.82	356.47	3394.78	0.00	33920.80	<0.001	0.65
Num. Comments provided per day	0.02	0.00	0.06	0.00	1.93	1.83	0.31	4.00	0.00	50.43	<0.001	0.68
Num. Comments provided per week	0.16	0.04	0.48	0.00	13.50	12.74	2.33	27.42	0.00	350.88	<0.001	0.66
Num. Comments provided per month	0.69	0.17	2.09	0.01	58.57	53.12	8.58	118.77	0.01	1535.51	<0.001	0.64
Num. Comments provided per year	8.26	2.03	25.07	0.07	704.03	656.81	101.59	1503.71	0.16	18295.90	<0.001	0.64
Cook Living in Germany/USA	0.87	1.00	0.34	0.00	1.00	0.92	1.00	0.27	0.00	1.00	<0.001	0.04

of the cooking process in a more emotionally loaded manner (0.36 *v* 0.26 (*AR*)). Next, the title text features: They draw a simmilary picture as the instructions. All Kochbar.de title features have higher values than the ones of Allrecipes.com. The sentiment of the recipe title is again higher for Kochbar.de (0.36 *v* 0.03 (*AR*)). This observation is underpinned by the title word occurences analysis of the previous Section, which showed the popularity of adverting adjectives for Kochbar.de. The signficancy test rejects the null hypothesis for all text features of this set. However, when considering the effect sizes, some instrution features such as Num. Words ($r = 0.03$), Num. Scentences ($r = 0.01$) and Sentiment ($r = 0.06$) have very small values.

Set: Recipe Seasonality

The seasonality features try to capture the seasonal character of recipes and espacially their ingredients. Upload month (6.41 *v* 6.15 (*KB*)) and day of month (15.72 *v* 15.69 (*KB*)) show that the recipe uploads are relatively even distributed for both datasets. The day of week feature indicates that the upload behaviour of Allrecipes.com's user, might be sligthly shifted towards weekends
(4.18 *v* 3.87 (*KB*)). The within season feature has in both datasets the same mean and median value (0.08). Almost all features are denoted significantly differnt by the statsitical test. Day of month has only a p value smaller 0.1. However, the effect sizes are again small ($r = 0.04 - 0.08$). Day of month has even a very low effect size ($r = 0.00$).

Set: Recipe Popularity and Appreciation

This feature set directly represents the popluarity and appreciation metrics. Only the two appreciation metrics have similar values. Avg. Rating is higher for Kochbar.de (4.95 *v* 4.29 (*AR*)) and avg. comment sentiment is higher for Allrecipes.com (1.86 *v* 1.78 (*KB*)). This trend was already observed in the previous Section. Both are significant ($p < 0.001$), but the sentiment metric has a low effect size ($r = 0.09$). All other features of this set, are higher for the Kochbar.de dataset. They are significant ($p < 0.001$) with high effect sizes ($0.60 - 0.78$). It seems that Allrecipes.com's users do not comment as much as the ones from Kochbar.de, but when they do, they use a more positive language.

Set: Recipe Innovation

The recipes innovation characteristcs are captured by this set. Allrecipes.com's mean ingredients rank (292.36 *v* 156.98 (*KB*)) and categories rank
(77.52 *v* 16.46 (*KB*)) have higher values. For the title words rank, it is the other way round (1224.44 *v* 511.42 (*AR*)). However, these metrics are not objective. The normed jaccard and IDF features draw a more precise picture of the recipe innovation within the datasets. Allrecipes.com has significantly ($p < 0.001$) higher Jaccard, mean Jaccard and IDF values. The effect sizes are for all features between small and medium ($r = 0.13 - 0.50$).

Set: User Activity and Context

The user activity is also considered to be decisive for the popularity of content, published in a social network. The users of Kochbar.de have significantly ($p < 0.001$) higher values for all activity features of this set. Moreover, almost all effect sizes are medium ($r = 0.56 - 0.85$), only num. ingredients used has a small value ($r = 0.16$). The last feature of this set, cook living in Germany/USA, represents the context of the user in terms of geographical location. Significantly more ($p < 0.001$) recipe uploading user of Kochbar.de, have stated that they are living in Germany, than Allrecipes.com user said that they are living in the USA (0.92 v 0.87 (AR)). However, the effect size of this feature is very low ($r = 0.04$).

Summary. To summarise, the comparative statistical analysis of this Section has shown that almost all feature distributions differ significantly ($p < 0.001$). The finding of the popularity analysis phase that Kochbar.de's users are far more active, has been confirmed. Users of Kochbar.de seem to describe the recipe preparation in more detail. However, this could just be the difference between english and german language. In terms of low level image features, both datasets contain comparable images. An other interessting observation is that the users of Allrecipes.com and Kochbar.de seem to be about equally innovative. Also the recipe uploads are almost even distributed over the year in both datasets. Other important recipe factors are the nutritional values. The findings do not really support the famous stereotype that americans eat more fat and kcals, for this datasets. The recipes of Kochbar.de contain more fat and more kcals on average. Allrecipes.com's recipes have more carbohydrates and protein.

Further results. This analysis is based on all relevant recipes of all categories. Since the popularity analysis phase showed differences in terms of categories popularity, it was interesting to compare features based on recipes of specific categories. Further comparison tables for the three main categories 'appetizers and snacks' (Table A.4), 'main dishes' (Table A.2) and 'desserts' (Table A.3), can be found in the Appendix. In brief, the comparison between all four feature tables did not show major differences. As one would expect, the nutritional values do vary noticeable. For example, desserts have in general higher values for kcal, protein, carbohydrates and fat. Interessting is that all these values are significantly ($p < 0.001$) higher for Allrecipes.com. The health scores are pretty much the same for 'main dishes' and 'appetizers and snacks', with only little or no deviation. In terms of dessert innovation, Kochbar.de seems to be ahead of Allrecipes.com, for example Jaccard's metric (0.83 v 0.42 (AR), $p < 0.001$). However, it is the other way round for all other categories. Also the higher text complexity (title and instrutions) of Kochbar.de could be confirmed. In terms of user activity, there are minor differences between the categories. Allrecipes.com's users upload 'main dishes' the most, but provide more ratings and comments for recipes in the 'appetizers and snacks' catgory. The user of Kochbar.de prefere to upload 'desserts', but do rate and comment more 'main dishes' than others.

Allrecipes.com and Kochbar.de are based on social networks. The importance of the users in such a network is known to be an indicator for the popularity of his uploaded content. However, the networks of both websites are not based on the same paradigms. Allrecipes.com has an following (directed graph) approach, whereas Kochbar.de features an friendship (undirected graph) approach. This means, structural network metrics such as shown in Table A.1 of the Appendix, are not relevant for comparative research, because their emergence is not comparable.

5.3 Predictive Modeling

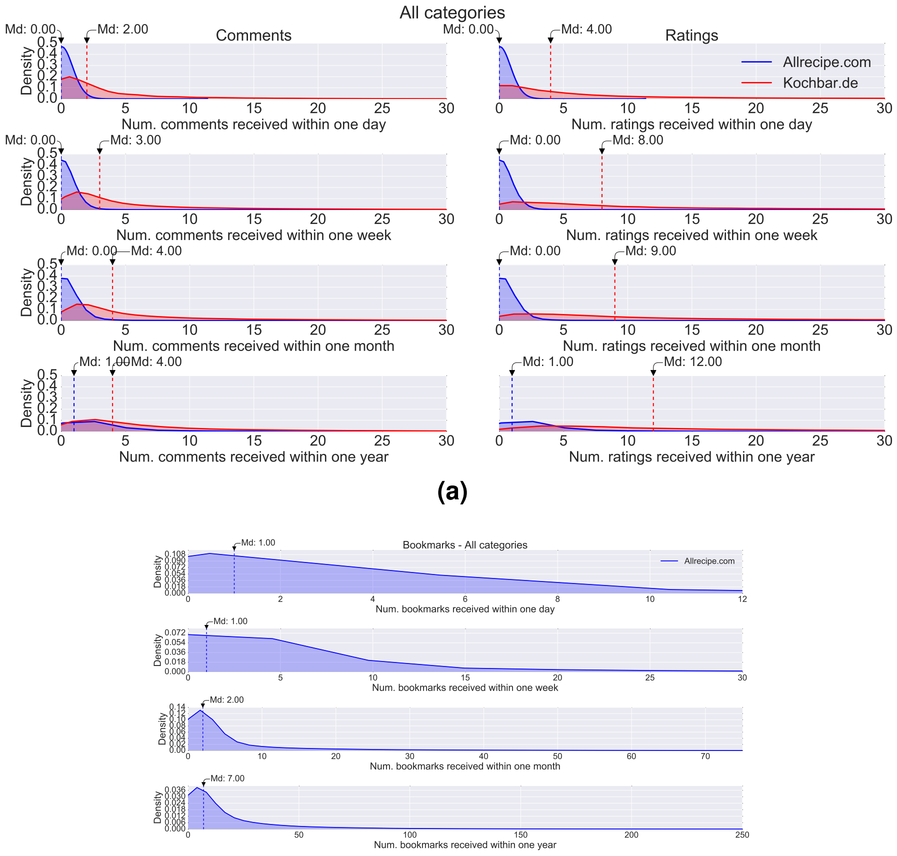
The investigations so far, showed significant patterns in terms of online recipe popularity. This Section depicts to what extent these signals can be exploited, in order to answer the research questions. The chosen classification approach is based on former studies in this domain and related research in the field of popularity prediction of online content.

5.3.1 Dependend and Independend Variables

In terms of machine learning, classification is a wide spread technique for identifing the set of categories, to which observations belong. An obvious example for two such categories is, whether a recipe will be popular or not. This is called binary classification. The possible categories are determined by dependend variables (outcome variables), special characteristics of recipes. Features that determine to which categories a recipe will be assigned, are called independend variables. The designated outcome variables of this experiments are the popularity proxies. It was observed that the appreciation metrics lack variability. A recipe which gets ratings and comments, has most likely a very high average rating (4.4 for AR and 5.0 for KB) and a average comment sentiment (1.8 for Allrecipes.com and 2.0 for Kochbar.de). All average values have a very low standard error of the mean. This attributes led to the decision, to drop the appreciation prediction and concentrate on popularity.

Allrecipes.com's recipes do not often obtain comments or ratings, although Allrecipes.com has far more registered users. Because of this, the number of bookmarks was introduced as third popularity indicator. However, the dataset of Allrecipes.com is the only one that features date information for bookmarkings. So Kochbar.de was excluded from this sub-experiment.

This predictive task follows the approach of Shulman et al. [2016]. Therefore, medias of the popularity metrics needed to be calculated. Figures 5.5 show the density distributions of recipes from all categories, receiving a certain amount of popularity within one day, week, month or year. The left part of Figure 5.5a illustrates the received comments distributions. The median of Allrecipes.com is 0.00 for the time periods of one day, week and month. Only comments received within one year has a value of 1.00. Kochbar.de on the other hand, has medias above or equal to 2.00 for all periods. The ratings part of the figure, draws the same picture. Here, Kochbar.de's recipes receive even more. Below, Figure 5.5b shows the density distribution of the number of bookmarks received. In comparison to the other popularity metrics of Allrecipes.com, recipes get more bookmarks (1.00 or more) within every time period. It is also interessting to note that all distritbutions are heavy tailed, meaning that only a few recipes are very popular. This is a common observation, in terms of online content popularity. The figures A.3



(a)

(b)

Figure 5.5: Density distributions of the popularity metrics number of comments, ratings and bookmarks

present the density distribution plots for all popularity metrics and all recipe categories. The median values show only minor variations for the different categories. Unfortunatly, Allrecipes.com's medians of ratings and comments do not change at all. When considering bookmarks, only the within one year medians do change (3.00−11.00). The sub-categories of Kochbar.de have often higher values. For example 'main dishes' has a median of 12.00 for ratings received within one month, whereas all recipes only have 9.00. However, the median values of the distributions are decisive for the classification outcome. Recipes below the median are considered as negative and above as positive examples of popular recipes. A concrete example: A recipe in the 'main dishes' category of Allrecipes.com, with a ratings received within one month median of 9.00 will be assigned to the negative examples. A recipe with a value of 12.00 and above, will be assigned to the postivie examples.

5.3.2 Feature correlation analysis (RQ 3)

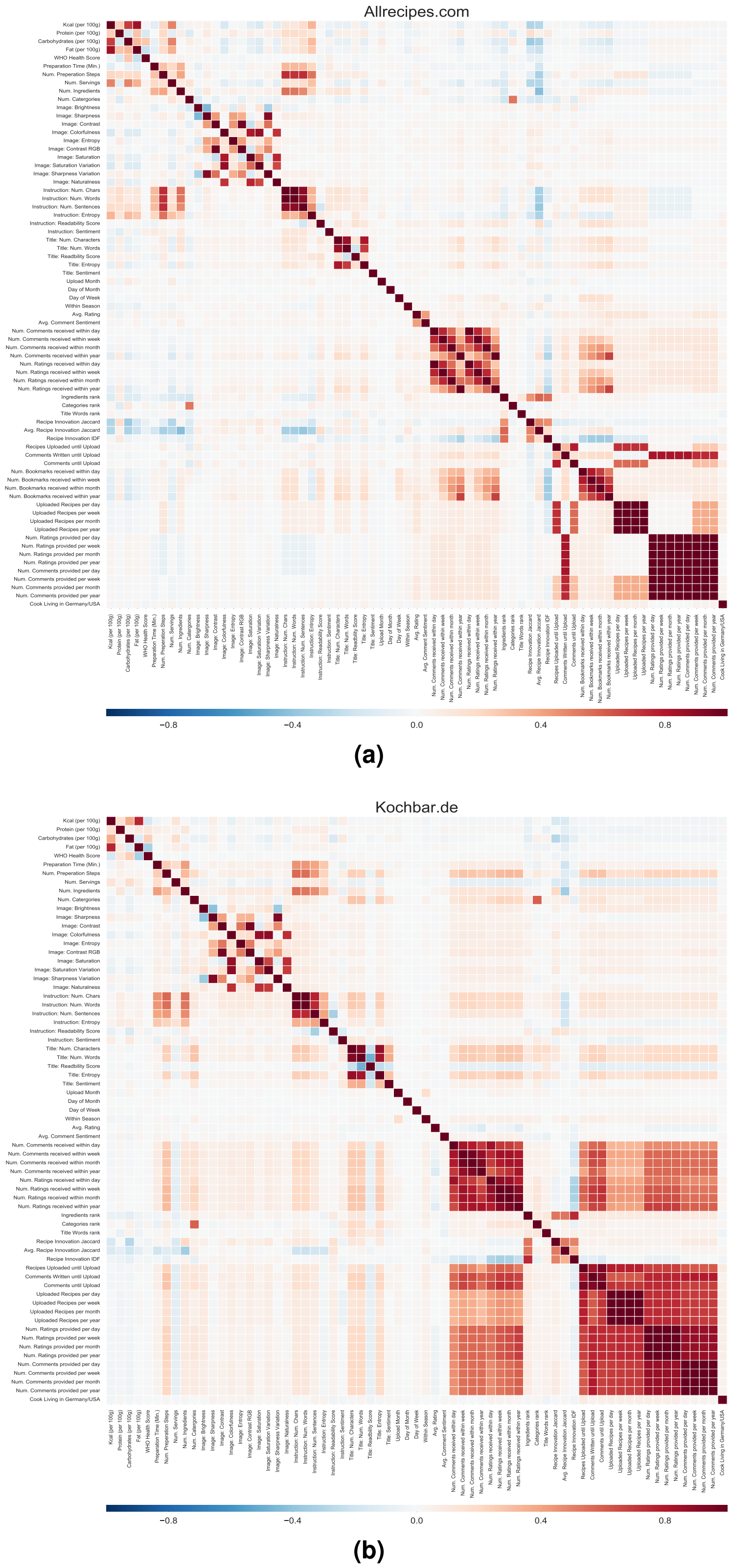
The Previous Sections revealed the presents of popularity patterns regarding various recipe parameters and characteristics. Also, it has been shown that the two datasets do differ in terms of the developed features. The next step is to find out, how and which features can be used for the purpos of popularity prediction. A common approach is to do a correlation analysis, which exposes possible feature correlations. The Figures 5.6 show the outcome of the correlation analysis for Allrecipes.com (5.6a) and Kochbar.de (5.6b), based on all recipe categories. When comparing both matrices, it is obvious that Kochbar.de has more high correlations, positive an negative, than Allrecipes.com. Especially when considering the correlations of the user activity features (number of recipes uploaded, comments written or ratings provided) with the outcome variables (number of comments/ratings received within timespan). This features show high positive correlation. The only features that have simmilar correlation patterns for Allrecipes.com, are the number of bookmarks within outcome variables. This leads to the assumption that the previous user activity may has an impact on the popularity of recipes uploaded in the future. An other interesting observation is that the outcome variables of both datasets do negatively correlate with the recipe innovation IDF feature. Novel or innovative recipes seem to receive less ratings,comments or bookmarks. Also, the innovation factor (Jaccard's index) does negatively correlate with most of the nutritional values for both datasets. This could be a hint that novel recipes follow the healthiness trend. Furthermore, the number of categories and recipe title/instruction complexity, do positively correlate with Kochbar.de's outcome variables. The most of the other higher correlations are more or less trivial. For example, the number of preparation steps does positively correlate with the size of the instruction text, or the image features do correlate among each other. As already mentioned, this analysis was done for all recipes of all categories.

Further results. Like in the comparative statistical analysis phase, the other recipe categories were also analysed in this correlation analysis and results can be found in the Appendix. The Figures A.4 and A.6 show the heatmaps of the other categories for both datasets. Compared to results of all recipes, no major differences could be found neither for Allrecipes.com, nor for Kochbar.de. All matrices within each dataset show pretty much the same correlation patterns. This leads to the assumption that the found popularity indicators apply generally.

5.3.3 Popularity prediction experiment (RQ 4)

According to the results of the temporal popularity analysis, recipes are most likely to get comments or ratings within the first ten to fifteen days. The users interesst for Kochbar.de's recipes drops after about seven to ten days (with a radical drop after seven days) and after ten to fifteen days for Allrecipes.com. Also the probability of getting ratings/comments is dropping fast. It could be argued that the reputation process of a recipe is more or less completed after one month. This led to the decision, to predict the popularity of recipes within one week and within one month.

Multiple data preparation steps, previous to the classification expriment, needed to be done. First, recipes that were uploaded within one week or within one month prior to the crawl dates, were dropped. This way, right censoring can be avoided. The binary classification was performed with balanced datasets. Therefore, the major class was subsampled with random boostraping. A major problem of the Allrecipes.com dataset was that an relevant amount of recipes had no upload user



(a)

(b)

Figure 5.6: Figure (a) shows the feature correlation heatmap for the Allrecipes.com dataset, whereas (b) highlights it for the Kochbar.de dataset. The correlation analysis is based on Spearman's rank correlation coefficient of all categories.

information attached. Since recipes with missing values were dropped, his led to small class sizes. For example, the class sizes for the 'appetizers and snacks' category experiments were under 300, after balancing. In order to tackle that problem, missing values got imputed with a simple mean imputation.

In the next step the Information Gain (IG) of the classification data was calculated. Table 5.5 shows the results of the IG calculation for the dependend variable comments within one week/month. The first thing to notice is that the popularity of Allrecipes.com's recipes are dependend on the upload image, whereas the user activity has more relevance for Kochbar.de. This importance of the users activity for Kochbar.de, was already indicated by the correlation analysis. However, user activity also seems to play a role for Allrecipes.com. A feature of that set, was ranked number nine for the 'within one month' time period. The most important feature of Allrecipes.com is the innovation factor based on IDF, which was also previously indicated by the correlation heatmap. Interesting to note is that the IG values are higher for Kochbar.de ($.0102 - .0171\ v\ .0563 - .1767\ (KB)$). Table 5.6 shows the results of the IG calculation for the outcome variable 'ratings within one week/month'. The Allrecipes.com results are the same for comments and ratings, since they are coupled together in the dataset. Although this is not true for Kochbar.de, the ratings based IG results are almost identical to the comments IG ones. Again, the user activity feature set is most important, followed by the image features. However, the num. ratings within week/month analysis has higher IG values ($0563 - .1767\ (KBcomments) v .0736 - .2143\ (KBratings)$). Table 5.7 shows the IG calculation of 'num. bookmarks within one week/month' of Allrecipes.com. The results include 4 different feature sets, for each of the time periods. Compared to the ratings/comments IG results of Allrecipes.com, not even one image feature is included. The recipe innovation factor IDF remains relevant with rank 1 of 'within one week' and rank 4 of 'within one month', respectively. However, the user activity gained importance in both time periods. For 'within one month' features of that set occupy ranks one to three and five to seven. Also worth noting is that the upload weekday seems to play a role in terms of bookmark popularity. The IG analysis of the three overlapping categories can be found in the appendix (Tables A.5, A.6 and A.7). The results for Kochbar.de do not show major differences. All in all, the same patterns can be observed. The only differences, are regarding Allrecipes.com. The IG algorithm was not always able to report ten relevant features. For example, the 'within one week' period of 'Appetizers and Snacks' has only two features that are considered relevant by the algorithm. One possible reason for that issue is probably the small number of recipes in that categories, since that is the only difference between the specific categories and all recipes. Futhermore, the IG of bookmarks shows a higher relevancy of user activity features for the specific recipe categories.

The actual binary classification experiment was conducted with three different classifieres: Random Forest (RF), Naive Bayes (NB) and Generalized Linear Models (GLM). RF and NB showed good results in comparable previous studies, for example at predicting the future popularity of online news articles. The used GLM is a binomial logistic regression, which is considered as a simpler classifier. Hence, GLM can be seen as a baseline for the results of RF and NB. The evaluation protocol for this experiment is 5 fold cross validation. This way the whole

Table 5.5: Information Gain of the number of comments within one week/-month dependend variable

	Allrecipes.com				Kochbar.de			
	Within one Week		Within one Month		Within one Week		Within one Month	
Rank	IG	Feature	IG	Feature	IG	Feature	IG	Feature
All Categories								
1	.0122	INN:Recipe Innovation IDF	.0171	INN:Recipe Innovation IDF	.1767	ACT:Comments Written until Upload	.1646	ACT:Comments Written until Upload
2	.0118	IMG:Image Entropy	.0161	IMG:Image Entropy	.1681	ACT:Comments until Upload	.1563	ACT:Num. Ratings provided per month
3	.0114	IMG:Image Brightness	.0153	IMG:Image Saturation Variation	.1652	ACT:Num. Ratings provided per month	.1525	ACT:Comments Written until Upload
4	.0112	IMG:Image Contrast RGB	.0153	IMG:Image Colorfulness	.1544	ACT:Comments Written until Upload	.1385	ACT:Comments until Upload
5	.0112	IMG:Image Saturation Variation	.0153	IMG:Image Saturation	.1520	ACT:Uploaded Recipes per month	.1339	ACT:Uploaded Recipes per month
6	.0111	IMG:Image Colorfulness	.0152	IMG:Image Naturalness	.0942	ACT:Recipes Uploaded until Upload	.0730	ACT:Recipes Uploaded until Upload
7	.0111	IMG:Image Sharpness	.0152	IMG:Image Contrast RGB	.0597	IMG:Image Contrast RGB	.0566	IMG:Image Contrast RGB
8	.0108	IMG:Image Saturation	.0152	IMG:Image Contrast	.0594	IMG:Image Saturation Variation	.0564	IMG:Image Saturation Variation
9	.0108	IMG:Image Contrast	.0151	ACT:Comments Written until Upload	.0593	IMG:Image Sharpness variation	.0564	IMG:Image Sharpness variation
10	.0102	IMG:Image Naturalness	.0151	IMG:Image Sharpness	.0592	IMG:Image Contrast	.0563	IMG:Image Contrast

Table 5.6: Information Gain of the number of ratings within one week/-month dependend variable

	Allrecipes.com				Kochbar.de			
	Within one Week		Within one Month		Within one Week		Within one Month	
Rank	IG	Feature	IG	Feature	IG	Feature	IG	Feature
All Categories								
1	.0122	INN:Recipe Innovation IDF	.0171	INN:Recipe Innovation IDF	.2143	ACT:Comments until Upload	.2009	ACT:Comments until Upload
2	.0118	IMG:Image Entropy	.0161	IMG:Image Entropy	.1921	ACT:Comments Written until Upload	.1924	ACT:Num. Ratings provided per month
3	.0114	IMG:Image Brightness	.0153	IMG:Image Saturation Variation	.1805	ACT:Num. Ratings provided per month	.1884	ACT:Comments Written until Upload
4	.0112	IMG:Image Contrast RGB	.0153	IMG:Image Colorfulness	.17860	ACT:Comments Written until Upload	.1848	ACT:Comments Written until Upload
5	.0112	IMG:Image Saturation Variation	.0153	IMG:Image Saturation	.1684	ACT:Uploaded Recipes per month	.1726	ACT:Uploaded Recipes per month
6	.0111	IMG:Image Colorfulness	.0152	IMG:Image Naturalness	.1269	ACT:Recipes Uploaded until Upload	.1128	ACT:Recipes Uploaded until Upload
7	.0111	IMG:Image Sharpness	.0152	IMG:Image Contrast RGB	.0738	IMG:Image Sharpness variation	.0763	IMG:Image Sharpness variation
8	.0108	IMG:Image Saturation	.0152	IMG:Image Contrast	.0737	IMG:Image Contrast RGB	.0762	IMG:Image Contrast RGB
9	.0108	IMG:Image Contrast	.0151	ACT:Comments Written until Upload	.0737	IMG:Image Saturation Variation	.0761	IMG:Image Saturation Variation
10	.0102	IMG:Image Naturalness	.0151	IMG:Image Sharpness	.0736	IMG:Image Contrast	.0761	IMG:Image Contrast

Table 5.7: Information Gain of the number of bookmarks within one week/month dependend variable

	Allrecipes.com			
	Within one Week		Within one Month	
Rank	IG	Feature	IG	Feature
All Categories				
1	.0320	INN:Recipe Innovation IDF	.0523	ACT:Comments Written until Upload
2	.0153	SEA:Day of Week	.0477	ACT:Recipes Uploaded until Upload
3	.0108	INN:Recipe Innovatin Jaccard	.0465	ACT:Comments until Upload
4	.0103	PRE:Instruction Readability Score	.0458	INN:Recipe Innovation IDF
5	.0100	PRE:Instructions Num. Chars	.0289	ACT:Uploaded Recipes per month
6	.0097	PRE:Instructions Num. Words	.0270	ACT:Num. Ratings provided per month
7	.0085	ACT:Recipes Uploaded until Upload	.0212	ACT:Cook Living in Germany/USA
8	.0049	INN:Mean Recipe Innovatin Jaccard	.0196	INN:Mean Recipe Innovatin Jaccard
9	.0047	ACT:Uploaded Recipes per month	.0146	SEA:Day of Week
10	.0043	ACT:Comments Written until Upload	.0130	COM:Preparation Time (Min.)

dataset can be used for training and testing, which is advantageous for potentially small class sizes. Since this is a classification on balanced datasets, accuracy is the only metric used to report how correctly the models identify the conditions. For each dataset, each time period and each feature set, models of all three classifieres were calculated. This way it was possible to find the feature sets with the best classification performances. Furthermore, the differences between the various models of each dataset were caluclated with statistical significancy tests. The experiment it self, was done by utilizing the statistical computing engine R.

Table 5.8 summarizes the results of the comments based classification tasks. The best feature set for each classifier is marked blue and the second and third best are bolded. The results for the top 10 IG features and the results for the all feature sets are separatly reportet (beneath the dashed line). Additionally, the 'all categories' part at the bottom of the Table, shows also results for the 'top-20 categories' features. The results for Allrecipes.com's 'appetizers and snack' and 'main dishes' are quite stable. 'Innovation' worked the best for all classifiers. Also 'Presentation' and 'User activity' seem to have a good predictive power. 'Desserts' and 'All Categories' have slightly different patterns. Their best working classifiers do vary between 'Innovation' and 'User activity'. Overall, the accuarcy is between 50% and 60%, but there are also values under 50%, which is lower than the likelihood of a coin toss. The highest values has the 'within one month' period for the models 'Innovation' (60.90% (*NB*)) and 'User activity' (60.32% (*RF*)). Overall performance of the top 10 IG features, is also good. In the most cases, it would have been under the best three feature sets. Also noteworthy is that the GLM classifier often has higher accuracy values than RF or NB. Kochbar.de's results show very stable patterns. 'User activity' is the best working feature set for all categories, classifiers and both time periods. 'Complexity' and 'Presentation' are the second and third best. This results align with the results of the IG and the correlation experiments. Interesting to not is that somehow the 'within one week' time period has higher accuracy, compared to the 'within one month' period. It is the other way round for Allrecipes.com, where the 'within one month' models have higher accuracy. The reason for that could be the radical popularity drop after seven to ten days in the Kochbar.de dataset. However, Kochbar.de has, compared to Allrecipes.com, generally higher values. They range from 50% ('Healthiness') to over 83% ('User activity'). For the most of the models, the GLM classifier is the one with the lowest performance. However, the differences are not as big, as one would expect.

The results of the ratings based classification experiments are illustrated by Table 5.9. Allrecipes.com's results are again identical to the ones of the comments experiment. To sum it up, Kochbar.de shows quite stable results. Besides the feature sets 'User activity', 'Presentation' and 'Complexity', also 'Innovation' seems to have good predictional power. For example, the 'Innovation' set is always in the top three of 'Desserts' and 'All categories'. The best working feature set however, is again 'User Activity'. Like at the comments based experiment, the 'within one week' time period seems to have better working models in general. The RF classifier even shows accuracy levels over 89%, for example with 'Main Dishes' and 'All categories'. Also for this classification experiment, the differences between the three classifier is not that big. Overall however, it can be said that GLM is probably the one with the lowest predictive performance.

Table 5.8: The classification accuracy for number of comments within week and within month (All categories). The highest accuracy values are marked blue whereas the second and third highest are bolded. The 'User Activity and Context' feature set of Allrecipes.com contains only 6 instead of 7 features, because number of ratings and comments are equivalent.

	Allrecipes.com						Kochbar.de						
	Within one Week			Within one Month			Within one Week			Within one Month			
Feature Set	Rand.For.	GLM	NB	Rand.For	GLM	NB	Rand.For.	GLM	NB	Rand.For	GLM	NB	Num.Feat.
Appetizers and Snacks	*Class size: 553*			*Class size: 862*			*Class size: 5193*			*Class size: 5584*			
Recipe Nutrition	49.01%	53.62%	**54.16%**	49.83%	52.78%	50.81%	51.89%	51.55%	52.49%	51.58%	51.71%	52.09%	4
Recipe Healthiness	49.01%	47.56%	50.00%	51.63%	51.45%	51.39%	50.10%	50.17%	50.66%	50.38%	50.51%	50.66%	1
Recipe Complexity	54.16%	52.72%	53.07%	52.26%	51.97%	48.73%	**57.53%**	**59.12%**	**56.82%**	**56.69%**	**58.52%**	**55.89%**	5
Recipe Presentation	**55.43%**	**57.14%**	**56.50%**	**56.26%**	**54.41%**	55.40%	**66.04%**	**59.58%**	**67.36%**	**65.78%**	**59.43%**	**66.99%**	21
Recipe Seasonality	52.35%	53.25%	49.81%	54.98%	**54.76%**	**54.18%**	53.04%	49.49%	51.10%	52.55%	50.48%	51.60%	4
Recipe Innovation	57.51%	59.77%	59.04%	57.71%	59.51%	60.90%	55.17%	56.48%	52.98%	53.35%	55.51%	53.26%	6
User Activity and Context	**55.25%**	**54.43%**	49.28%	**57.25%**	50.00%	49.30%	81.34%	75.68%	77.22%	79.60%	74.82%	75.68%	7 (6)
Top-10 Features IG	57.78%	57.41%	56.16%	57.95%	58.41%	51.57%	80.96%	75.77%	77.82%	79.42%	75.18%	76.71%	10
All	59.59%	60.39%	55.97%	60.62%	59.92%	58.59%	79.94%	76.56%	77.31%	78.73%	75.81%	78.41%	68 (67)
Main Dishes	*Class size: 1653*			*Class size: 2772*			*Class size: 35483*			*Class size: 37800*			
Recipe Nutrition	51.90%	**54.43%**	**52.90%**	50.23%	52.04%	51.70%	51.15%	51.06%	51.27%	51.08%	51.19%	51.37%	4
Recipe Healthiness	48.28%	48.73%	47.74%	50.87%	50.83%	51.05%	50.31%	50.34%	50.21%	50.11%	50.22%	49.76%	1
Recipe Complexity	52.18%	53.70%	48.10%	51.82%	49.68%	53.01%	**61.67%**	**62.32%**	**63.57%**	**61.33%**	**61.63%**	**62.56%**	5
Recipe Presentation	**55.70%**	52.89%	**53.71%**	**55.01%**	**56.78%**	**55.77%**	**69.43%**	**62.08%**	**70.16%**	**69.07%**	**61.58%**	**69.86%**	21
Recipe Seasonality	53.44%	53.89%	50.90%	52.85%	**53.72%**	**53.07%**	55.26%	51.87%	52.66%	54.79%	51.47%	52.31%	4
Recipe Innovation	56.69%	61.39%	58.13%	57.47%	58.62%	58.66%	57.30%	59.09%	52.62%	56.09%	57.83%	55.23%	6
User Activity and Context	**55.34%**	**56.24%**	49.55%	**56.82%**	48.85%	49.80%	86.04%	79.67%	80.21%	85.04%	78.40%	78.74%	7 (6)
Top-10 Features IG	56.69%	59.49%	57.32%	59.90%	58.01%	52.22%	86.15%	79.63%	81.12%	84.89%	78.42%	79.72%	10
All	59.22%	58.14%	51.54%	60.23%	59.09%	54.36%	85.50%	80.60%	79.15%	84.23%	79.36%	80.70%	68 (67)
Desserts	*Class size: 1391*			*Class size: 1996*			*Class size: 14256*			*Class size: 14914*			
Recipe Nutrition	51.94%	50.18%	49.57%	51.10%	49.50%	48.02%	50.63%	51.92%	52.28%	51.48%	51.79%	51.92%	4
Recipe Healthiness	52.84%	53.06%	53.05%	51.25%	51.28%	51.38%	50.86%	50.86%	50.86%	50.78%	50.70%	50.66%	1
Recipe Complexity	52.55%	49.50%	50.00%	53.06%	53.31%	53.38%	**58.15%**	57.35%	**56.56%**	**57.51%**	**56.91%**	**55.57%**	5
Recipe Presentation	53.38%	53.74%	**54.75%**	**55.86%**	**58.02%**	**57.11%**	**67.07%**	**60.56%**	**78.60%**	**66.87%**	**60.07%**	**67.69%**	21
Recipe Seasonality	**54.17%**	**55.79%**	**55.43%**	55.46%	**56.06%**	55.71%	53.71%	52.39%	52.94%	53.08%	52.76%	52.52%	4
Recipe Innovation	**54.85%**	55.97%	54.60%	**56.14%**	58.49%	**57.99%**	56.14%	**57.50%**	56.39%	55.56%	56.09%	55.00%	6
User Activity and Context	55.82%	**55.86%**	55.50%	60.32%	56.41%	58.52%	82.60%	76.65%	79.05%	81.32%	75.44%	77.44%	7 (6)
Top-10 Features IG	54.53%	56.36%	55.64%	60.42%	57.84%	61.77%	82.79%	76.74%	80.19%	81.86%	75.55%	78.52%	10
All	56.40%	58.30%	57.98%	61.45%	60.95%	62.12%	82.40%	77.68%	78.60%	80.78%	76.72%	79.49%	68 (67)
All Categories	*Class size: 7948*			*Class size: 12583*			*Class size: 127005*			*Class size: 138568*			
Recipe Nutrition	50.58%	52.13%	51.50%	51.33%	52.91%	52.79%	51.40%	51.28%	52.07%	51.51%	51.21%	52.03%	4
Recipe Healthiness	49.84%	50.26%	49.75%	49.97%	50.27%	49.90%	50.00%	49.88%	50.05%	50.17%	49.90%	50.14%	1
Recipe Complexity	51.79%	52.85%	**53.14%**	53.00%	53.45%	51.22%	**61.29%**	**60.84%**	55.75%	**60.44%**	**60.11%**	**58.54%**	5
Recipe Presentation	**53.49%**	**54.65%**	**55.16%**	**55.12%**	**56.49%**	**56.65%**	**67.68%**	**61.67%**	**68.51%**	**67.17%**	**60.99%**	**66.62%**	21
Recipe Seasonality	52.66%	52.77%	53.03%	53.16%	**54.16%**	53.95%	57.07%	52.45%	52.54%	55.88%	52.47%	52.55%	4
Recipe Innovation	**55.20%**	56.73%	56.50%	**56.47%**	57.96%	58.09%	58.15%	59.51%	**57.71%**	56.42%	57.74%	56.21%	6
User Activity and Context	55.20%	**53.46%**	52.19%	58.12%	53.80%	**54.33%**	83.89%	76.15%	78.14%	82.73%	74.77%	76.51%	7 (6)
Top-10 Features IG	56.42%	54.76%	53.43%	59.37%	56.88%	59.90%	83.92%	76.21%	79.53%	82.46%	74.80%	77.77%	10
Top-20 Categories	52.94%	53.21%	51.84%	54.22%	54.20%	50.02%	57.26%	56.84%	50.23%	56.61%	56.28%	50.20%	20
All	57.77%	58.61%	53.24%	60.78%	60.57%	62.29%	83.30%	77.61%	76.61%	81.87%	76.01%	79.21%	68 (67)

The last classification experiments were based on the bookmarking numbers of Allrecipes.com. Table 5.10 summarizes the results of this prediction task. Altough, the bookmarkings based prediction has generally higher accuracy values than the ratings/comments prediction, very simmilar patterns can be observed for Allrecipes.com. The 'Innovation' feature set is the best for predicting popularity of 'Main Dishes'. The second and third vary between 'Presentation' and 'Seasonality'. Also, 'User activity' shows some predictive power. The best three feature sets of the 'Desserts' category are 'Presentation', 'Innovation' and 'User activity'. However, the best working set does again vary. 'Presentation' works best for the GLM classifier and 'User activity' for RF and NB. When considering 'All categories' the results do vary even more. Here, the best sets are 'Presentation' for GLM and 'Innovation' for RF and NB. The accuracy values range between 50% and over 71%. The Desserts 'Top-10 IG' set reaches even 74.87% with RF. Overall, the values are higher for 'within one month' except for 'Desserts'. It is difficult to say which classifier works best and which is the worst. All three have comparable performance.

All experiments have in common that the 'top 20 categories' feature set shows

Table 5.9: The classification accuracy for number of ratings within week and within month (All categories). The highest accuracy values are marked blue whereas the second and third highest are bolded. The 'User Activity and Context' feature set of Allrecipes.com contains only 6 instead of 7 features, because number of ratings and comments are equivalent.

	Allrecipes.com						Kochbar.de						
	Within one Week			Within one Month			Within one Week			Within one Month			
Feature Set	Rand.For.	GLM	NB	Rand.For	GLM	NB	Rand.For.	GLM	NB	Rand.For	GLM	NB	Num.Feat.
Appetizers and Snacks	*Class size: 553*			*Class size: 862*			*Class size: 5675*			*Class size: 5677*			
Recipe Nutrition	49.01%	53.62%	**54.16%**	49.83%	52.78%	50.81%	51.67%	53.60%	54.04%	51.15%	53.65%	53.12%	4
Recipe Healthiness	49.01%	47.56%	50.00%	51.63%	51.45%	51.39%	51.04%	50.48%	51.04%	50.84%	50.41%	50.99%	1
Recipe Complexity	54.16%	52.72%	53.07%	52.26%	51.97%	48.73%	60.12%	**60.07%**	**57.89%**	59.15%	**59.93%**	**57.89%**	5
Recipe Presentation	**55.43%**	**57.14%**	**56.50%**	**56.26%**	**54.41%**	**55.40%**	**69.47%**	59.79%	**70.53%**	**70.14%**	59.66%	**70.97%**	21
Recipe Seasonality	52.35%	53.25%	49.81%	54.98%	**54.76%**	**54.18%**	58.43%	52.25%	54.35%	57.10%	51.93%	53.53%	4
Recipe Innovation	57.51%	59.77%	59.04%	57.71%	59.51%	60.90%	**63.42%**	**63.36%**	57.71%	**61.86%**	**62.20%**	57.25%	6
User Activity and Context	**55.25%**	**54.43%**	49.28%	**57.25%**	50.00%	49.30%	87.13%	82.11%	79.62%	86.59%	81.48%	78.85%	7 (6)
Top-10 Features IG	57.78%	57.41%	56.16%	57.95%	58.41%	51.57%	86.56%	82.08%	81.62%	86.11%	81.72%	80.93%	10
All	59.59%	60.39%	55.97%	60.62%	59.92%	58.59%	86.04%	83.81%	81.40%	85.54%	83.10%	80.89%	68 (67)
Main Dishes	*Class size: 1653*			*Class size: 2772*			*Class size: 38573*			*Class size: 37789*			
Recipe Nutrition	51.90%	**54.43%**	**52.90%**	50.23%	52.04%	51.70%	51.34%	50.77%	51.02%	50.75%	50.57%	50.87%	4
Recipe Healthiness	48.28%	48.73%	47.74%	50.87%	50.83%	51.05%	49.91%	50.18%	49.90%	49.79%	49.68%	49.94%	1
Recipe Complexity	52.18%	53.70%	48.10%	51.82%	49.68%	53.01%	62.71%	**62.89%**	**64.30%**	62.76%	**62.73%**	**64.44%**	5
Recipe Presentation	**55.70%**	52.89%	**53.71%**	**55.01%**	**56.78%**	**55.77%**	**71.10%**	61.94%	**71.92%**	**71.54%**	61.94%	**72.26%**	21
Recipe Seasonality	**53.44%**	**53.89%**	50.90%	**52.85%**	**53.72%**	**53.07%**	62.68%	53.57%	54.44%	61.59%	53.36%	54.23%	4
Recipe Innovation	56.69%	61.39%	58.13%	57.47%	58.62%	58.66%	**64.62%**	**64.92%**	57.59%	**63.45%**	**64.11%**	57.08%	6
User Activity and Context	55.34%	56.24%	49.55%	56.82%	48.85%	49.80%	89.95%	82.56%	80.66%	89.44%	81.91%	79.79%	7 (6)
Top-10 Features IG	56.69%	59.49%	57.32%	59.90%	58.01%	52.22%	89.32%	82.60%	82.75%	88.88%	81.96%	82.12%	10
All	59.22%	58.14%	51.54%	60.23%	59.09%	54.36%	88.45%	84.51%	80.50%	88.09%	83.91%	80.00%	68 (67)
Desserts	*Class size: 1391*			*Class size: 1996*			*Class size: 15034*			*Class size: 15017*			
Recipe Nutrition	51.94%	50.18%	49.57%	51.10%	49.50%	48.02%	51.26%	51.41%	52.53%	51.69%	50.98%	52.33%	4
Recipe Healthiness	52.84%	53.06%	53.05%	51.25%	51.28%	51.38%	50.92%	51.14%	51.01%	51.28%	51.28%	52.33%	1
Recipe Complexity	52.55%	49.50%	50.00%	53.06%	53.31%	53.38%	60.45%	60.10%	51.85%	60.95%	59.86%	51.93%	5
Recipe Presentation	53.38%	53.74%	**54.75%**	**55.86%**	**58.02%**	**57.11%**	**70.39%**	**61.63%**	**71.16%**	**70.61%**	**61.52%**	**71.57%**	21
Recipe Seasonality	**54.17%**	**55.79%**	**55.43%**	55.46%	56.06%	55.71%	59.50%	54.91%	54.66%	58.52%	54.57%	54.58%	4
Recipe Innovation	**54.85%**	55.97%	54.60%	**56.14%**	58.49%	**57.99%**	**65.43%**	**64.68%**	**59.47%**	**63.76%**	**63.59%**	**59.19%**	6
User Activity and Context	55.82%	**55.86%**	55.50%	60.32%	**56.41%**	58.52%	87.54%	81.74%	80.25%	87.11%	81.18%	79.57%	7 (6)
Top-10 Features IG	54.53%	56.36%	55.64%	60.42%	57.84%	61.77%	86.74%	81.85%	81.89%	86.30%	81.18%	81.45%	10
All	56.40%	58.30%	57.98%	61.45%	60.95%	62.12%	86.71%	83.56%	81.61%	86.24%	83.02%	81.30%	68 (67)
All Categories	*Class size: 7948*			*Class size: 12583*			*Class size: 144428*			*Class size: 142011*			
Recipe Nutrition	50.58%	52.13%	51.50%	51.33%	52.91%	52.79%	51.83%	50.56%	51.74%	51.72%	50.75%	51.71%	4
Recipe Healthiness	49.84%	50.26%	49.75%	49.97%	50.27%	49.90%	49.99%	50.09%	50.04%	49.99%	50.13%	50.08%	1
Recipe Complexity	51.79%	52.85%	**53.14%**	53.00%	53.45%	51.22%	63.99%	**62.56%**	58.57%	63.49%	**62.32%**	59.91%	5
Recipe Presentation	**53.49%**	**54.65%**	**55.16%**	**55.12%**	**56.49%**	**56.65%**	**69.94%**	62.11%	**70.75%**	**70.15%**	61.93%	**70.96%**	21
Recipe Seasonality	52.66%	52.77%	53.03%	53.16%	**54.16%**	53.95%	67.61%	53.35%	53.81%	66.26%	52.76%	53.35%	4
Recipe Innovation	**55.20%**	56.73%	56.50%	**56.47%**	57.96%	58.09%	**68.77%**	**68.01%**	**62.63%**	**67.05%**	**66.74%**	**62.23%**	6
User Activity and Context	55.20%	**53.46%**	52.19%	58.12%	53.80%	**54.33%**	89.24%	81.13%	78.59%	88.62%	80.51%	77.77%	7 (6)
Top-10 Features IG	56.42%	54.76%	53.43%	59.37%	56.88%	59.90%	88.39%	81.17%	80.35%	87.76%	80.52%	79.87%	10
Top-20 Categories	52.94%	53.21%	51.84%	54.22%	54.20%	50.02%	58.77%	58.10%	50.03%	58.61%	57.89%	50.03%	20
All	57.77%	58.61%	53.24%	60.78%	60.57%	62.29%	87.87%	83.57%	78.42%	87.17%	82.67%	78.75%	68 (67)

poor predictive performance. This observation is backed by Tatar et al. [2014]. They stated that categorical features are in general problematic for such tasks.

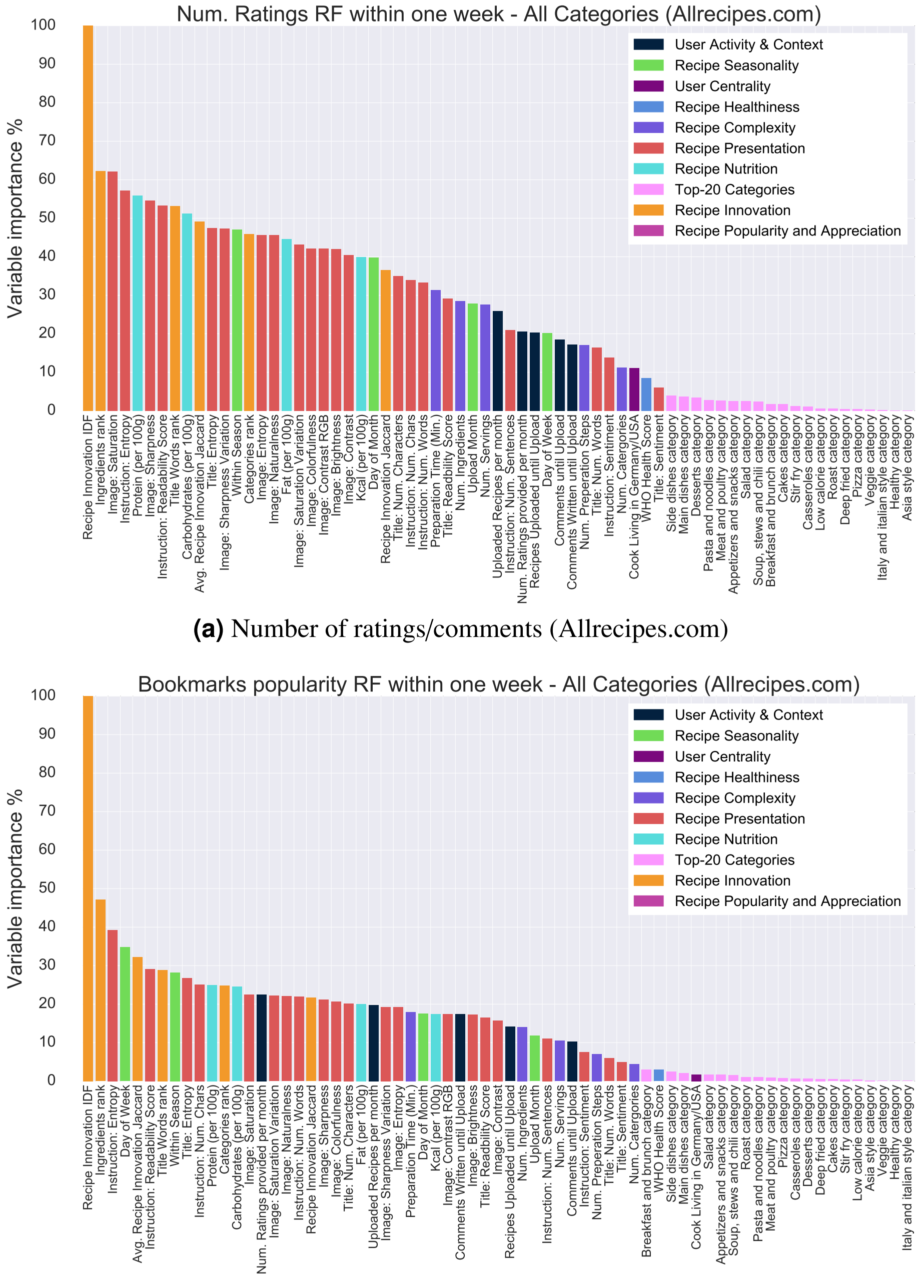
Figures 5.7 and 5.8 illustrate the decisive importance on feature level for the 'all categories' popularity 'within one week' experiment (RF). The most important features do slightly differ from the top-10 identified by the IG experiment. However, the feature sets are the same. In general, it can be said that recipes innovation and presentation play a major role in Allrecipes.com's popularity. All three RF models chose 'Recipe Innovation IDF' (100%) as the most relevant feature. It is either followed by another innovation or a presentation feature. Interesting to note is that nutritional values also have influence on the decision. This has not been a obvious results of IG or the correlation experiments. Kochbar.de's results on the other hand, draw a different picture. 'User activity' features are the most important ones, followed by 'Innovation' and 'Presetation' features. The results also do line up almost exactly with the outcome of the IG experiments. The results of the 'within one month' experiments are summarized by Figures 5.9 and 5.10. Kochbar.de's results, do not show relevant differences to the 'within one week' feature importance. The 'User activity' features are again the most impor-

Table 5.10: The classification accuracy for bookmarks of ratings within week and within month (All categories, only Allrecipes.com). The highest accuracy values are marked blue whereas the second and third highest are bolded.

	Allrecipes.com						
	Within one Week			Within one Month			
Feature Set	Rand.For	GLM	NB	Rand.For	GLM	NB	Num.Feat.
Appetizers and Snacks	*Class size: 1845*			*Class size: 1798*			
Recipe Nutrition	50.11%	52.93%	52.03%	51.84%	54.78%	52.98%	4
Recipe Healthiness	51.63%	51.84%	51.63%	50.50%	50.78%	50.08%	1
Recipe Complexity	54.31%	55.31%	53.44%	54.42%	55.42%	54.12%	5
Recipe Presentation	56.96%	**62.93%**	**56.99%**	57.54%	**64.27%**	**59.46%**	21
Recipe Seasonality	**58.48%**	**59.35%**	**59.27%**	**58.01%**	**58.62%**	**58.70%**	4
Recipe Innovation	68.48%	66.91%	66.72%	69.97%	67.35%	68.13%	6
User Activity and Context	**63.17%**	48.92%	49.81%	**64.04%**	50.11%	49.64%	6
Top-10 Features IG	69.51%	65.88%	55.07%	69.97%	67.30%	54.31%	10
All	69.49%	68.24%	59.73%	71.77%	69.16%	61.04%	67
Main Dishes	*Class size: 5437*			*Class size: 5360*			
Recipe Nutrition	51.22%	51.86%	51.70%	51.63%	52.68%	52.84%	4
Recipe Healthiness	50.62%	50.87%	50.30%	51.62%	51.87%	51.87%	1
Recipe Complexity	53.64%	54.89%	51.75%	55.71%	55.75%	53.42%	5
Recipe Presentation	**60.30%**	**65.63%**	**61.44%**	**61.74%**	**66.73%**	**62.58%**	21
Recipe Seasonality	58.27%	**60.72%**	**60.61%**	59.10$	**60.77%**	**60.49%**	4
Recipe Innovation	68.94%	66.51%	67.45%	69.89%	68.07%	68.40%	6
User Activity and Context	**62.49%**	43.93%	49.74%	**63.58%**	43.16%	49.82%	6
Top-10 Features IG	69.42%	66.48%	55.98%	70.63%	68.00%	58.21%	10
All	69.99%	69.85%	60.51%	71.99%	70.93%	62.59%	67
Desserts	*Class size: 3458*			*Class size: 4552*			
Recipe Nutrition	52.57%	55.81%	55.33$	51.82%	55.37%	55.16%	4
Recipe Healthiness	48.26%	48.99%	49.28%	51.83%	50.30%	50.26%	1
Recipe Complexity	61.15%	55.65%	52.68%	**61.40%**	54.89%	51.45%	5
Recipe Presentation	**62.72%**	68.81%	**60.83%**	61.34%	67.70%	**60.25%**	21
Recipe Seasonality	58.44%	59.07%	59.48%	57.29%	57.65%	57.74%	4
Recipe Innovation	**70.05%**	**66.60%**	**68.44%**	**67.42%**	**65.26%**	**66.51%**	6
User Activity and Context	71.80%	**66.28%**	71.01%	71.54%	**64.07%**	70.50%	6
Top-10 Features IG	74.87%	68.25%	74.28%	73.52%	66.72%	72.29%	10
All	73.57%	73.86%	73.00%	71.31%	71.19%	70.64%	67
All Categories	*Class size: 23245*			*Class size: 22971*			
Recipe Nutrition	52.10	55.20	54.76	52.54%	55.84%	55.62%	4
Recipe Healthiness	51.04	51.08	51.03	51.22%	51.26%	51.20%	1
Recipe Complexity	57.22	54.50	50.45	57.57%	54.52%	51.74%	5
Recipe Presentation	**59.10**	**64.17**	**59.88**	**60.41%**	64.58%	**61.14%**	21
Recipe Seasonality	58.92	58.47	59.02	59.90%	**58.45%**	59.07%	4
Recipe Innovation	67.03	**62.61**	**65.75**	68.56%	**64.46%**	67.04%	6
User Activity and Context	**66.86**	63.70	65.76	**67.19%**	57.20%	**65.42%**	6
Top-10 Features IG	70.58	63.61	68.84	70.92%	64.99%	68.80%	10
Top-20 Categories	57.28	57.19	50.17	56.46%	56.27%	50.08%	20
All	70.24	68.56	67.83	71.79%	69.94%	69.11%	67

tant ones. Allrecipes.com's bookmarks feature importance however, shows a shift towards 'User activity'. 'Recipe Innovation IDF' is still the most decisive feature, but the one second is now 'Recipes Uploaded until Upload'. The feature importance for the three other categories, can be found in the Appendix (A.8 - A.13). In brief, no major differences are observed. The feature importance patterns are more or less stable. Kochbar.de shows a higher tendency towards 'User activity' for the specific categories. The same is true for the bookmarks experiments of Allrecipes.com's 'Desserts' category.

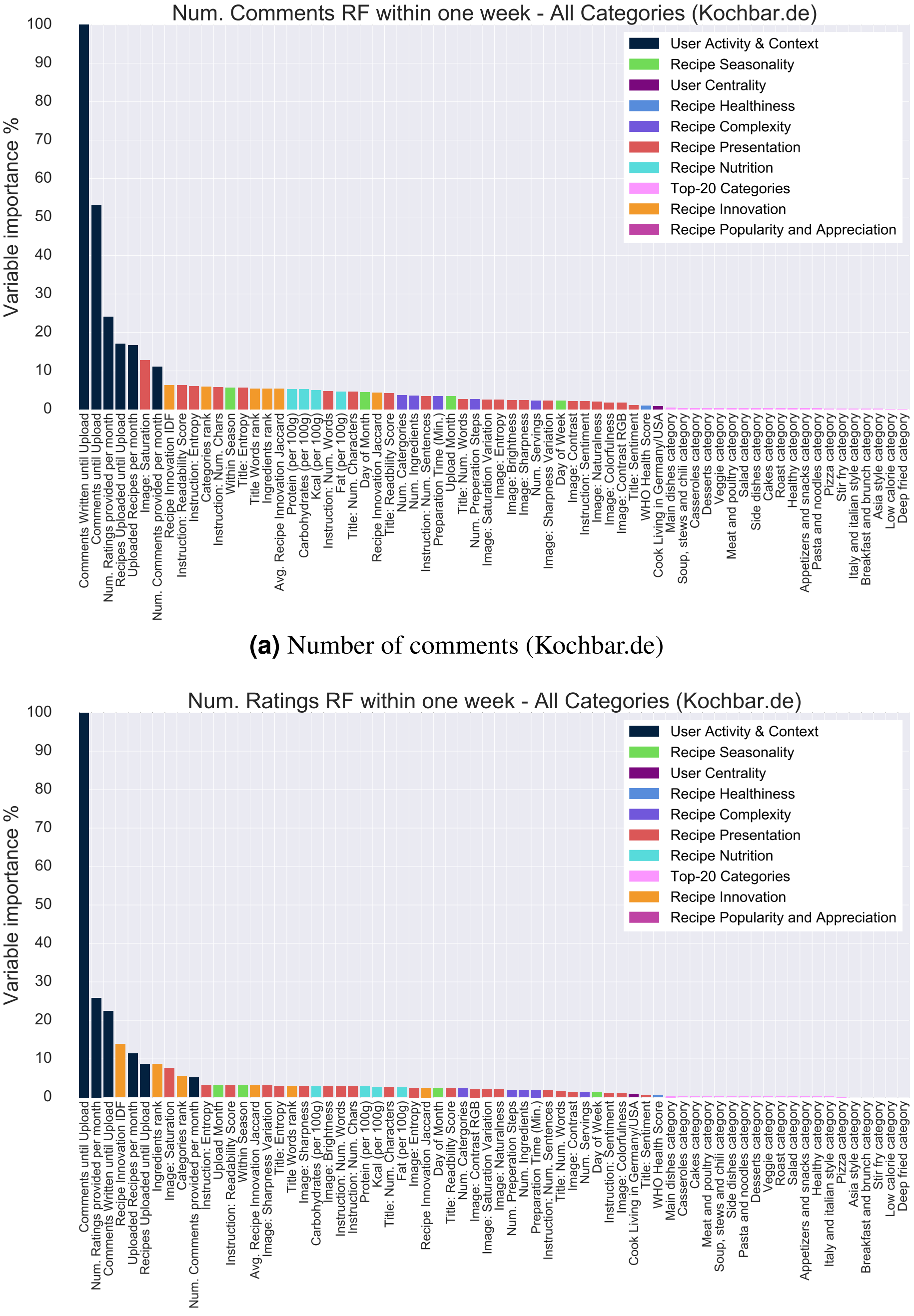
Summary. It has been shown that it is possible to predicit, within a defined time period, whether a recipe gets popular above average or not. The experiments are based on features, exclusively available at upload. Three different classifiers have been used. Random Forest, Naive Bayes and Generalized Linear Models. It cannot be said that one of them generally outperformes the others, but Random Forest seems to have the best predictive performance. The binary popularity classes of online recipes can be predictied with up to 89% accuracy, for certain configurations. The models of Kochbar.de have overall higher accuracy. Furthermore, the revealed patterns suggest that there are generally valid charateristics, such as the upload users former activity that influence the popularity of online recipes. However, there are also other influencing features such as the novelty of a recipe idea or the presentation of it.



(a) Number of ratings/comments (Allrecipes.com)

(b) Number of bookmarks (Allrecipes.com)

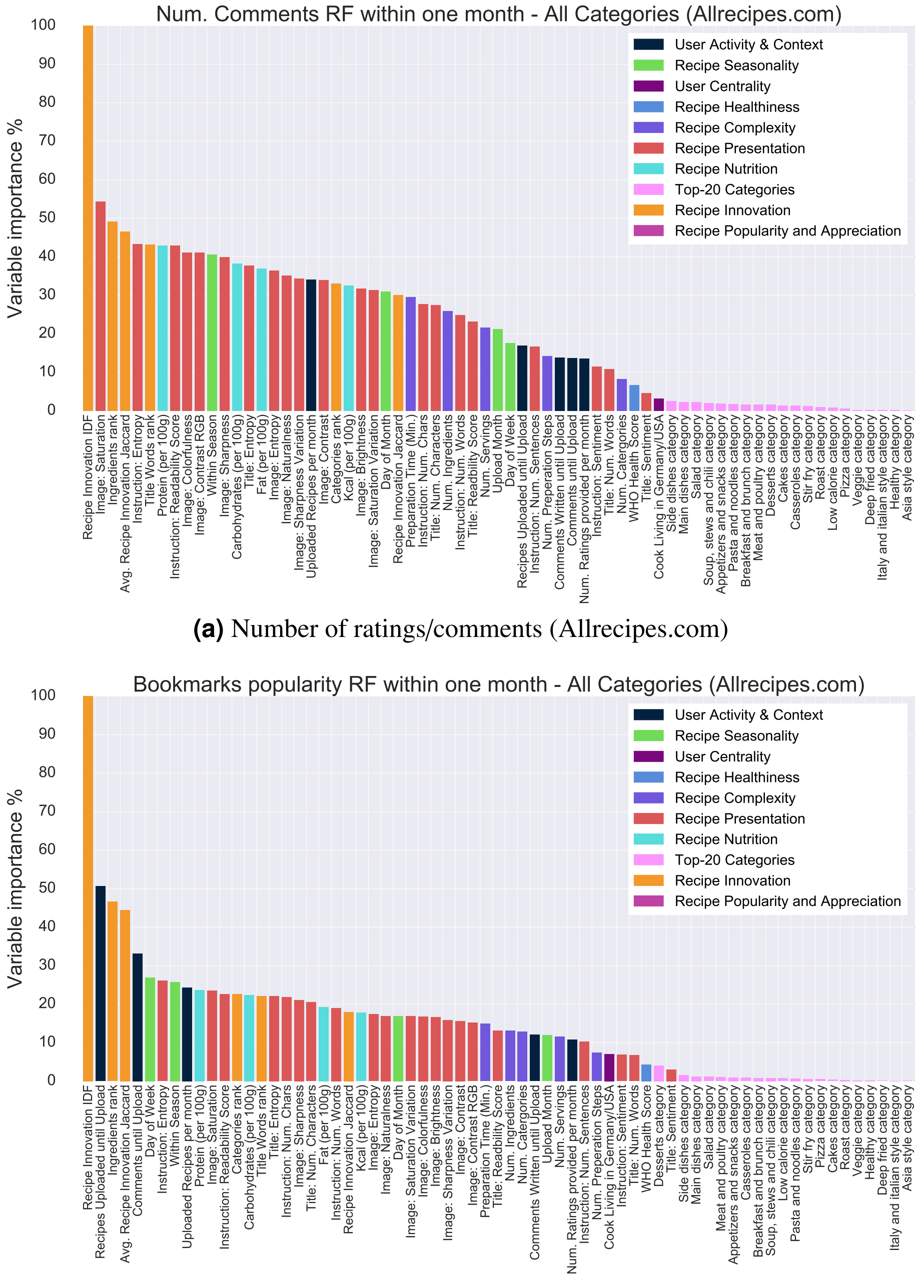
Figure 5.7: Variable importance of the 'within one week' predictions of the Allrecipes.com dataset. Important to note is that number of comments is equivalent to number of ratings for Allrecipes.com.



(a) Number of comments (Kochbar.de)

(b) Number of ratings (Kochbar.de)

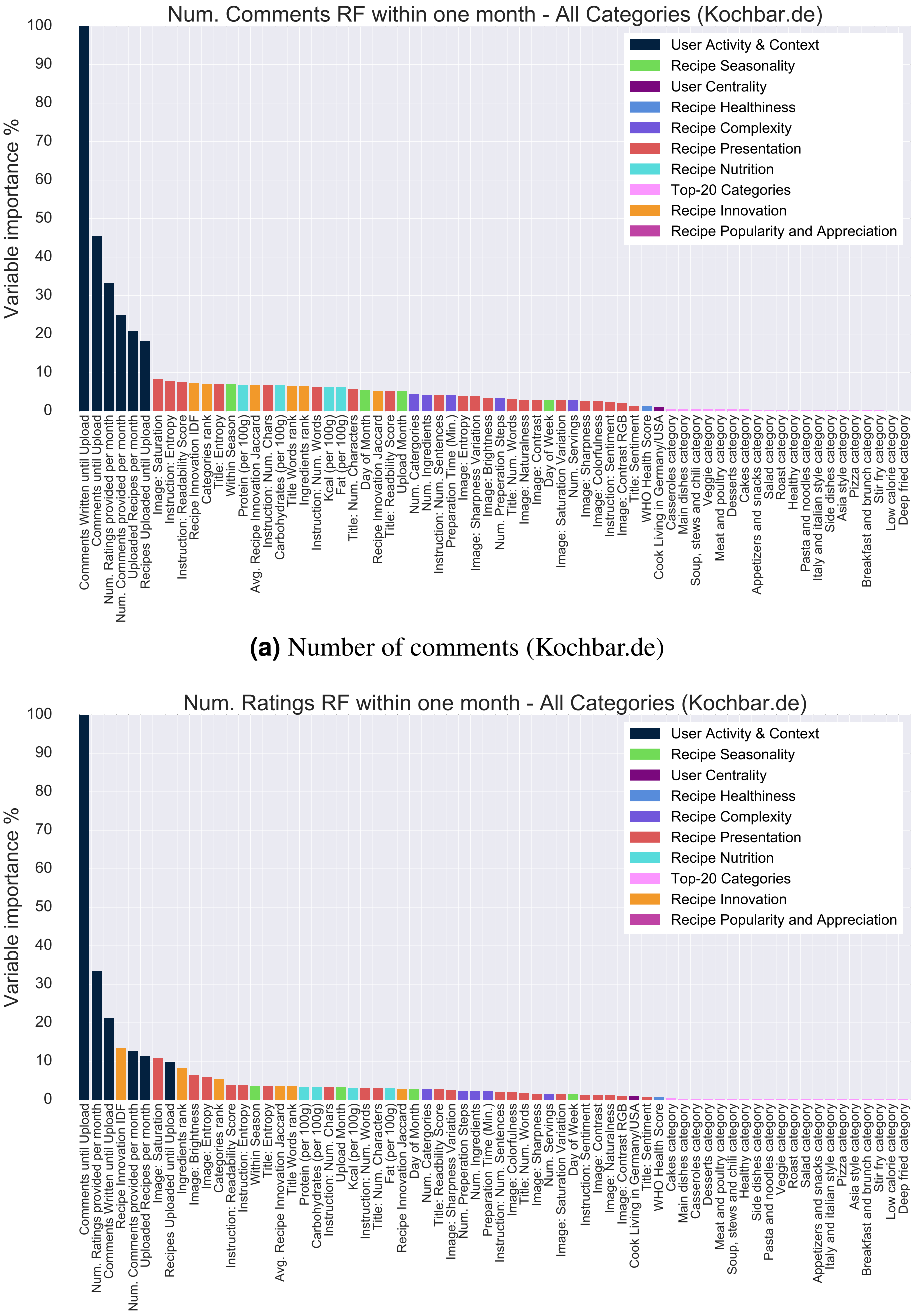
Figure 5.8: Variable importance of the 'within one week' predictions of the Kochbar.de dataset.



(a) Number of ratings/comments (Allrecipes.com)

(b) Number of bookmarks (Allrecipes.com)

Figure 5.9: Variable importance of the ‘within one month’ predictions of the Allrecipes.com dataset. Important to note is that number of comments is equivalent to number of ratings for Allrecipes.com.



(a) Number of comments (Kochbar.de)

(b) Number of ratings (Kochbar.de)

Figure 5.10: Variable importance of the 'within one month' predictions of the Kochbar.de dataset.

Chapter 6

Conclusions and Future Work

“ Sometimes the questions are complicated
and the answers are simple. ”

[Theodor Seuss Geisel, American author, 1904–1991.]

This Chapter concludes the thesis by summarizing findings and discussing possible limitations of the approach. In addition, an outlook on possible improvements and follow up studies is given.

The central theme of this master thesis is to provide better insights into the hidden patterns and workings of online recipe's popularity. Understanding this sociodynamic processes, could help to develop more elaborate health aware recommender systems. Such recommender systems have the potential to solve many of the food related health issues, rapidly growing in modern societies. The approach followed, included the statistical analysis of datasets from two popular food community websites: Allrecipes.com and Kochbar.de. They represent two separate stereotypical food cultures of western civilization. By comparatively analysing both, a more general view on the popularity processes was revealed. This analyses are based on characteristics of the recipes and the underlying social network of the websites. These so called features, are developed on the results of releated studies regarding the popularity of online content or food preferences. In order to reveal the predictive power of the features and to test the statistical findings, a predictive modelling experiments were conducted. The findings of this master thesis can be summarized as follows:

Popularity patterns in the two datasets (RQ 1): It has been shown that, popularity patterns exist and that there are differences between the two recipe communities. A temporal analysis revealed that within the first seven to ten days, Kochbar.de's recipes gain the most attention. Allrecipes.com has such a trend too, but within the first ten to fifteen days. After that periods, the attention drops very quickly. An interesting observation is that although Allrecipes.com has far more registered users, compared to Kochbar.de, the recipes get less comments and ratings on average. Because of this, the number of bookmarks of a recipe was additionally introduced as a popularity metric for Allrecipes.com. An appreciation analysis revealed that recipes get primarily high ratings and have positively written comments. Also, this does not change over time. Furthermore, the popularity of categories, ingredients and title words got analysed. It showed that there are more popular entities than others, regarding the number of occurences. However, entities that are often used in recipes, are not necessarily contributing to the appreciation or popularity of recipes.

Differences and similarities between the communities (RQ 2): In order to answer this research question, a comparative statistical analysis was conducted. It has been shown that almost all feature distributions differ significantly ($p < 0.001$). The observations of the explorative data analysis phase, such as that Kochbar.de's users are more active, have been confirmed. Interesting to note is that the findings do not support the stereotype that americans eat heftier food than europeans. An other finding is that in terms of novel recipes, users of Allrecipes.com and Kochbar.de are about equally innovative. Furthermore, it was observed that the users of Kochbar.de do describe the recipe preparation steps in more detail. Also the titles of Kochbar.de's recipes, are often used in an advertising fashion. Although, the food communities do differ in many ways, they are nevertheless similar. A possibly reason could be that the american cuisine has its origins also in europe.

Potential useful feature correlations (RQ 3): A correlation analysis, based on Spearman's rank correlation coefficient, was used to expose possible feature relations. It has been shown that the Kochbar.de's user activity features have high correlation with the popularity metrics based on comments and ratings. Allrecipes.com does show similar correlation patterns for the bookmarks based metrics. It can be assumed that former user activity may has an important impact on the popularity of recipes uploaded in the future. An other interesting observation is that the popularity metrics of both datasets do negatively correlate with features that meassure the innovation of a recipe. Novel recipes seem to no attract much attention. However, the healthiness of recipes do also negatively correlate with innovation. This maybe indicates that new and innovative recipes, are often also following the trend of healthy recipes. In general, Kochbar.de has more high correlating features, positive and negative, than Allrecipes.com. However, the most correlations, such es correlations between image features, are trivial.

Popularity prediction and relevant features (RQ 4): The predictive modelling experiment showed that the developed features do have good predictive power. These binary classification experiments are based on features, exclusively available at upload. It's goal was to predict, whether a recipe gets popular above average or not, within a defined time period. Therefore, the three different classifiers Random Forest, Naive Bayes and Generalized Linear Models have been used. It was possible to predicit the outcome class with an accuracy of up to 89%, for certain configurations. Overall, the Kochbar.de's models show the highest accuaracy values. The best working classifier seems to be Random Forest. However, neither one of the classifiers generally outperformes the others. The results of the predictive modelling confirmed the assumptions of previous experiments: There are generally valid charateristics, such as the upload users former activity, the presentation or the novelty of a recipes idea that have strong influence on the future popularity of online recipes. However, user activity features, such as written and obtained ratings/comments or the number of recipes uploaded, have a higher impact on Kochbar.de's recipes popularity. Whereas innovation features, such as recipe innovation or ingredients popularity rank and image features, for instance, saturation or image entropy seem to have more influence on the popularity of Allrecipes.com its recipes.

6.1 Limitations and Future Work

This master thesis revealed interessing observations. Due to the different approaches and paradigms of the analysed online recipe community websites, some limitations exist. The results, presented in this thesis, may only be valid for the two available datasets of Allrecipes.com and Kochbar.de. Hence, further investigations are required to be able to determine if and what results are generally applicable. The features used to predict the popularity of recipes, have been chosen according to "best practices" in literature and are expected to provide good results. However, the ingredients text parsing approach is error-prone. The reason for that is the ambigous nature of human language and the typing error encouraging free form text input fields, at the upload page of Allrecips.com and Kochbar.de. An optimal recipe parsing approach could possibly reveal new or different popularity patterns. Furthremore, Allrecipes.com's dataset lacks user data for several recipes, which made imputation of missing values necessary. The reason for that, is not clear. However, a new crawl could maybe solve this problem. A other limitation arises from the different time periods both websites were founded in. Allrecipes.com was one of the first food databases on the Internet and was developed in the early 2000s. Kochbar.de on the other hand, was founded in 2007. This leads to the question, if both datasets are fully comparable. Maybe it is necessary to extract recipes from specific time periods and compare them exclusively. Possible future applications are for example better health aware recommenders or systems that support the user in uploading recipes. Such a system could provide suggestions in how to alter the characteristics of a recipe, with the goal of maximizing its future popularity.

Appendix A

Appendix

This Appendix Chapter holds further feature descriptions and experiment results for both datasets which could not all be presented in the previous Chapters. The first section (i) describes features that capture the users connections in the social network of Allrecipes.com and Kochbar.de. Section two (ii) details further results of the popularity analysis phase and section three (iii) holds more results of the comparative statistical analysis phase, including results of the social network analysis. The last section (iv) shows further results of the predictive modelling experiments.

A.1 Feature Engineering

A.1.1 User Centrality

This feature set was dropped during the course of the experiments. Altough the social network centrality of a user is vital to his reputation, it is only a current snap-shot of its development. Since the datasets did not feature a history of the user network centrality, it cannot be taken into account for this thesis. It would bias the results, because the recipes popularity is clearly time dependend.

The features in this set measure the importance of a user within the food community. Following the assumptions described by feature set ‘user activity and context’, the importance of a user should also be a result of his activities. A very active user, in the means of uploading, rating or commenting, is a more important (or central) node within the user network. In his study about centrality in social networks Freeman [1978] defined three concepts that describe the structural importance of a node: degree, betweenness and closeness. All three were utilized for this feature set.

It is important to note that the two food communities do not use the same network paradigm. Allrecipes.com has a follower based (more like Twitter.com[1]), whereas Kochbar.de uses a friendship based (like Facebook.com[2]) approach. This major difference makes direct comparison difficult. Adding a friend to my network requires the approval of the other user, but following an other user does

[1] https://twitter.com/

[2] https://www.facebook.com/

not. In network theory, such user networks get represented as graphs built from vertices (nodes) and the connecting edges (relations). Furthermore, edges can have associated directions with them. Because of the difference in the underlying approaches of the two websites, Allrecipes.com spans a undirected graph and Kochbar.de a directed one.

The user centrality features were measured utilizing the Stanford Network Analysis Project[3] (SNAP). It is a popular and well documented general purpose network analysis tool and can also be used for graph mining. An alternative library would have been NetworkX [4], which is available for Python[5] too. Nevertheless, NetworkX was not as efficient and fast in calculating the network metrics as SNAP. This observation, however, can not be generalised and was just seen in this specific task and on one specific computer.

- Network Degree

  The degree of a node v is the number of direct neighbors t it has in the network. It is a simple measurment based on the adjacency matrix[6] a_{vt} of a graph. This square matrix indicates whether nodes of a finite graph, are connected or not.

  Generally speaking, nodes of an directed graph can have inbound^{+} and outbound^{-} connections. The degree of a node is the sum of both:

$$\begin{aligned} deg^{+}(v) &= \sum_t a_{vt} \\ deg^{-}(v) &= \sum_t a_{tv} \\ degree(v) &= deg^{+}(v) + deg^{-}(v) \end{aligned} \tag{A.1}$$

  The SNAP methods used for this calculation were GetNodeInDegV[7] and GetNodeOutDegV[8]. The adjacency matrix of an undirected graph is symmetric; hence there is no in or out differentiation and the degree calculation can be simplified:

$$degree(v) = \sum_t a_{vt} \tag{A.2}$$

  Besides the basic degree feature, in- and out-degree were used as additional features:

- Network In-Degree

$$in\_degree(v) = deg^{+}(v) \tag{A.3}$$

[3] http://snap.stanford.edu/

[4] https://networkx.github.io/

[5] https://www.python.org/

[6] https://en.wikipedia.org/wiki/Adjacency_matrix

[7] https://snap.stanford.edu/snappy/doc/reference/GetNodeInDegV.html

[8] https://snap.stanford.edu/snappy/doc/reference/GetNodeOutDegV.html

- Network Out-Degree

$$out\_degree(v) = deg^{-}(v) \tag{A.4}$$

Since Allrecipes.com's user network is based on a undirected graph, it's in- and out-degree equals the general degree.

- Network Degree Centrality

The degree centrality is simplest of the three centrality metrics described by Freeman [1978]. It can be seen as the users local entanglement with the direct neighbors. By utilizing the degree value, it uses the local structure of a node v only and is calculated as follows:

$$degree\_centrality(v) = \frac{degree(v)}{(N-1)} \tag{A.5}$$

where N is the number of nodes in the network. Although the simplicity is an advantage, the metric does not take the whole network and global structure into account. Since the used SNAP method GetDegreeCentr[9] can only calculate the degree centrality of undirected graphs, the Kochbar.de network needs to be converted.

- Network Betweenness Centrality

Betweenness tries to measure the amount of control a node wields over the whole communication within a network. A high betweenness means that a the node stands between other nodes and the information has to pass through that node. Internet service providers (ISP) are good examples of that phenomenon. The SNAP library offers the GetBetweennessCentr[10] method, which is based on the improved betweenness algorithm stated by Brandes [2001]. Simmilar to the standard betweenness formular described by Freeman [1978], involves the pairwise distances and shortest path lenghts between nodes. However, the approach of Brandes [2001] calculates the betweenness metric with $O(n+m)$ space and $O(nm)$ time ($O(nm + n^2 logn)$ for weighted graphs), compared to $\Theta(n^3)$ and $\Theta(n^2)$. The following pseudo code

[9] https://snap.stanford.edu/snappy/doc/reference/GetDegreeCentr.html

[10] https://snap.stanford.edu/snappy/doc/reference/GetBetweennessCentr.html

of the algorithm was taken from Brandes [2001]:

```
$C_B[v] \leftarrow 0, v \in V$;
for $s \in V$ do
    $S \leftarrow$ empty stack;
    $P[w] \leftarrow$ empty list, $w \in V$;
    $\sigma[t] \leftarrow 0, t \in V \; \sigma[s] \leftarrow 1$;
    $d[t] \leftarrow -1, t \in V \; d[s] \leftarrow 0$;
    $Q \leftarrow$ empty queue;
    enqueue $s \rightarrow Q$;
    while $Q$ not empty do
        dequeue $v \leftarrow Q$;
        push $v \rightarrow S$;
        foreach neighbor $w$ of $v$ do
            // w found for the first time?
            if $d[w] < 0$ then
                enqueue $w \rightarrow Q$;
                $d[w] \leftarrow d[v] + 1$;
            end
            // shortest path to w via v?
            if $d[w] = d[v] + 1$ then
                enqueue $w \rightarrow Q$;
                $\sigma[w] \leftarrow \sigma[w] + \sigma[v]$;
                append $v \rightarrow P[w]$;
            end
            $\delta[v] \leftarrow 0, v \in V$ ;
            // S returns vertices in order of non-increasing
             distance from s
            while $S$ not empty do
                pop $w \leftarrow S$;
                for $v \in P[w]$ do
                    $\delta[v] \leftarrow \delta[v] + \frac{\sigma[v]}{\sigma[w]} \cdot (1 + \delta[w])$;
                    if $w \neq s$ then
                        $C_B[w] \leftarrow C_B[w] + \delta[w]$;
                    end
                end
            end
        end
    end
end
```

Algorithm 1: Betweenness centrality in unweighted graphs.

- Network Closeness Centrality

  Freeman [1978] describes closeness as the measurement of how long information would take to advance through the network from a give node v (point centrality). A message commenced in the absolute center of a graph, would by definition spread through the network in minimum time. However, this is only true for connected graphs, since the distance of disconnented nodes can be seen as infinite. If this is not the case, a solution would be calculate the closeness within the largest graph component. Closeness is determined as the reciprocal of the nodes farness, which is the sum of the distances to all other vertices in the network:

$$farness\_centality(v) = \sum_{t \in T} dist(v, t) \tag{A.6}$$
$$closeness\_centrality(v) = \frac{1}{farness\_centrality(v)}$$

where T is the set of nodes without node v. For this measurement the SNAP method GetClosenessCentr[11] was used.

- Network Eigenvector Centrality

  In a typical modern social network, for example Youtube.com[12], not every node is as important as others. Many of the users are just passive consumers, but some create, share and influence. In contrast to the other centrality metrics, the eigenvector is calculated with respect to the linking nodes importance. This metric can be seen as an extension of the degree (especially in-degree) centrality. However, in-degree just counts the number of inbound connection regardless of importance. Eigenvector centrality awards nodes that are linked to them by other prominent nodes. However, this does not mean that a node with an high eigenvector centrality, is necessarily highly linked. Even only a few, but important linkers can have this effect on the node. This approach was first described by Gould [1967]:

$$eigenvector\_centrality(v) = \frac{1}{\lambda} \sum_{t} a_{t,v} \cdot EC(t) \quad , \lambda \neq 0 \tag{A.7}$$

  where a is the adjacency matrix of the graph, v is the node for which this value gets calculated and t are the other nodes. λ is a constant and often choosen as the largest eigenvalue of of the adjacency matrix. $EC()$ is just a short form of the eigenvalue centrality function itself. The SNAP method for this metric is GetEigenVectorCentr[13].

- PageRank

  This very famous network metric was created by Larry Page and Sergey Brin, the two founders of Google Inc.[14]. The algorithm currently used by Google is highly complex an kept secret. This metric uses the version of their first puplished paper [Brin and Page, 2012]. The algorithm follows the same basic principles than the eigenvector centrality. PageRank was intended to measure the importance of website pages and ranks them according to the popularity of inbound linked pages:

[11] https://snap.stanford.edu/snappy/doc/reference/GetClosenessCentr.html?highlight=closeness

[12] https://www.youtube.com/

[13] https://snap.stanford.edu/snappy/doc/reference/GetEigenVectorCentr.html?

[14] https://www.google.de/

$$PageRank(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{A.8}$$

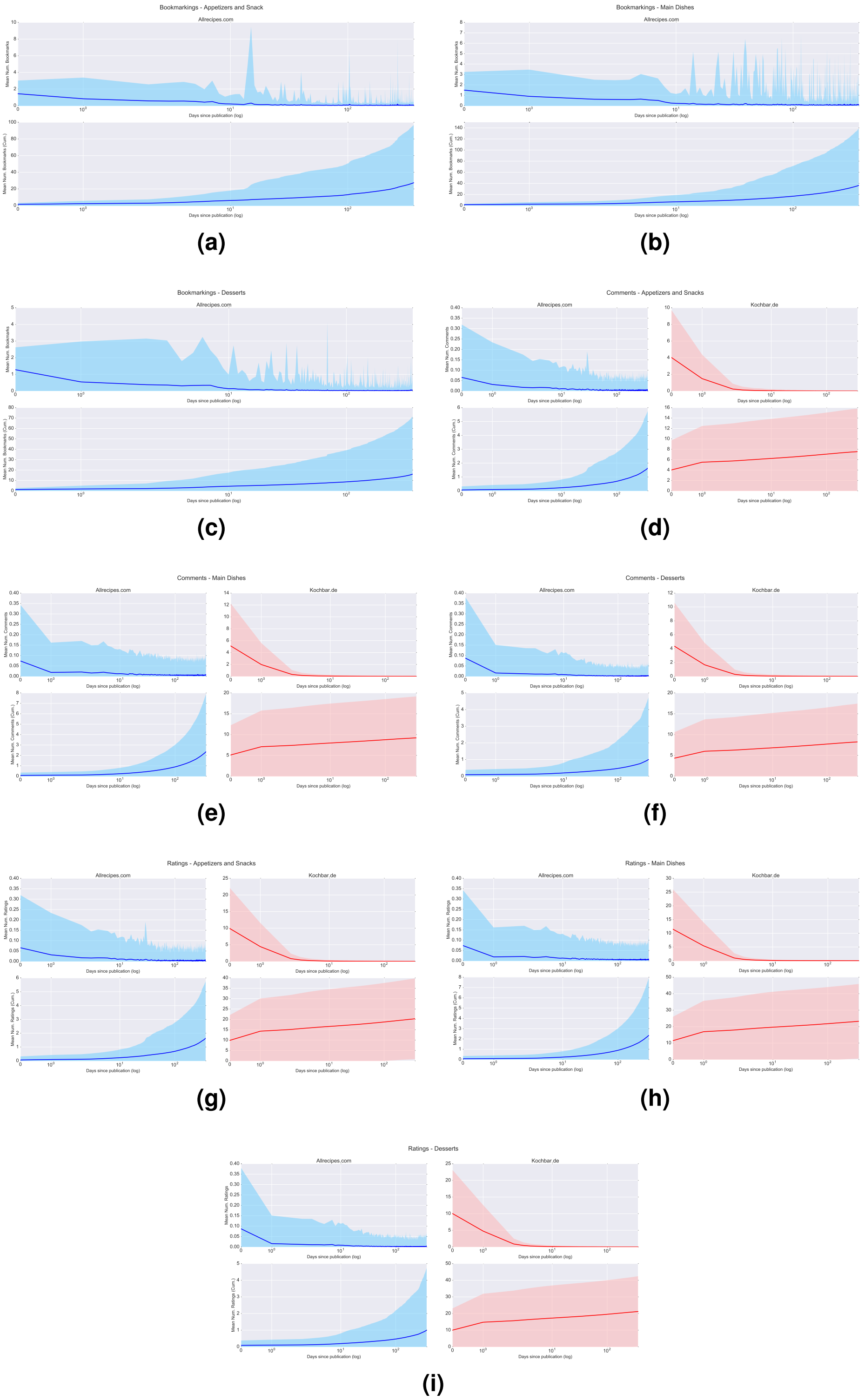
where the set M denotes the web pages that link to page pi. N is the total number of nodes in the graph. L is a function that returns the number of outgoing links and $PR()$ is the recursive called PageRank function it self. The $1 - d$ part of the left side of the summation symbol is considered as the random surfer. This is a model that takes random surfing into account. This means that with probability $1 - d$ (d is the damping factor and generally 0.85) a user jumps to a random webpage from bookmarks or by typing it in the address bar of the browser and does not surf there via a link of a page. The PageRank of Allrecipes.com an Kochbar.de users was measured with the SNAP GetPageRank[15] method.

- Num. Follower

  Compared to the other nodes importance measurements, the number of followers is very simple. It is just the number of followers of a user in Allrecipes.com and the number of friends in Kochbar.de.

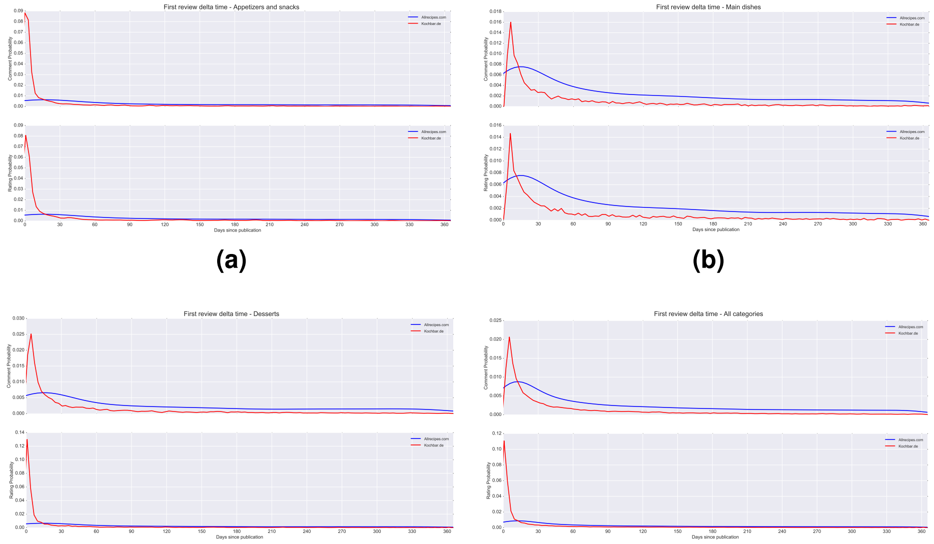
[15] https://snap.stanford.edu/snappy/doc/reference/GetPageRank.html

A.2 Popularity Analysis



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure A.1: Ratings/comments/bookmarks popularity of recipes over time.



(a)

(b)

(c)

(d)

Figure A.2: Days until first comment/rating (delta time).

A.3 Comparative Statistical Analysis

Table A.1: Differences between Allrecipes.com and Kochbar.de online food communities regarding structural network metrics of user.

	Allrecipes.com (n=171665) Follower based					Kochbar.de (n=199749) Frienship based				
Feature	*M*	*Md*	*SD*	*Min*	*Max*	*M*	*Md*	*SD*	*Min*	*Max*
Set: User Centrality										
Network Degree	3.51	2	27.21	1	3825	12.75	2	42.67	1	1429
Network In-Degree	2.15	1	14.15	0	1876	12.75	2	42.67	1	1429
Network Out-Degree	1.36	1	21.49	0	3651	12.75	2	42.67	1	1429
Network Degree Centrality	5.62498e-05	1.854e-05	0.0004	1.854e-05	0.07034	0.0008	0.0001	0.002	6.4292e-05	0.091
Network Betweeness Centrality	0.0002	0	0.007	0	1	0.001	3.53348e-07	0.01	0	1
Network Closeness Centrality	0.07	0.11	0.07	0	0.23	0.24	0.25	0.07	6.4292e-05	0.41
Network Eigenvector Centrality	0.0009	1.183e-06	0.004	0	0.24	0.00264812	0.0003	0.007	0	0.14
Pagerank	2.23006e-05	1.0348e-05	5.32085e-05	5.289e-06	0.006	6.38207e-05	2.22865e-05	0.0001	1.2248e-05	0.006
Num. Follower	0.33	0	10.65	0	3651	20.33	4	53.77	1	1374

Table A.2: Differences between Allrecipes.com and Kochbar.de online food communities based on the created features (Main Dishes).

	Allrecipes.com (n=11194)					Kochbar.de (n=81232)						
Feature	*M*	*Md*	*SD*	*Min*	*Max*	*M*	*Md*	*SD*	*Min*	*Max*	*p*	*r*
Set: Recipe Nutrition												
Kcal (per 100g)	162.24	153.00	65.02	4.00	697.00	210.84	170.00	156.72	1.00	955.00	<0.001	0.09
Protein (per 100g)	9.44	9.08	4.00	0.13	27.62	6.53	4.80	5.46	0.00	36.00	<0.001	0.35
Carbohydrates (per 100g)	11.54	10.23	8.44	0.00	79.74	11.71	5.60	15.09	0.00	94.00	<0.001	0.18
Fat (per 100g)	8.66	7.64	5.43	0.03	53.97	15.54	10.00	17.41	0.00	97.00	<0.001	0.13
Set: Recipe Healthiness												
WHO Health Score	3.35	3.00	0.66	3.00	6.00	3.35	3.00	0.64	3.00	6.00	<0.1	0.00
Set: Recipe Complexity												
Preparation Time (Min.)	86.22	50.00	113.54	3.00	1740.00	41.46	30.00	32.38	1.00	960.00	<0.001	0.32
Num. Preperation Steps	3.75	4.00	1.63	1.00	19.00	4.15	4.00	2.86	1.00	25.00	<0.1	0.01
Num. Servings	6.31	6.00	3.95	1.00	72.00	3.75	4.00	1.74	1.00	55.00	<0.001	0.47
Num. Ingredients	10.00	9.00	4.00	1.00	35.00	11.42	11.00	4.05	1.00	46.00	<0.001	0.00
Num. Catergories	4.31	4.00	0.88	3.00	6.00	12.63	12.00	5.80	2.00	36.00	<0.001	0.77
Set: Recipe Presentation												
Image: Sharpness	0.18	0.17	0.09	0.02	0.81	0.17	0.15	0.10	0.01	1.09	<0.001	0.09
Image: Sharpness Variation	0.29	0.26	0.15	0.00	1.50	0.29	0.26	0.17	0.02	2.00	<0.1	0.02
Image: Contrast	0.05	0.05	0.02	0.00	0.16	0.05	0.04	0.02	0.00	0.16	<0.001	0.00
Image: Contrast RGB	0.17	0.16	0.07	0.00	0.49	0.16	0.15	0.07	0.01	0.46	<0.001	0.00
Image: Saturation	0.27	0.26	0.10	0.02	0.75	0.23	0.21	0.10	0.00	0.78	<0.001	0.22
Image: Saturation Variation	0.15	0.15	0.05	0.00	0.41	0.15	0.15	0.05	0.01	0.40	<0.001	0.03
Image: Brightness	0.49	0.49	0.10	0.12	0.89	0.49	0.49	0.11	0.09	0.89	<0.001	0.04
Image: Colorfulness	0.24	0.24	0.07	0.03	0.54	0.23	0.22	0.08	0.00	0.64	<0.001	0.10
Image: Entropy	7.47	7.53	0.32	3.16	7.96	7.37	7.43	0.33	4.32	7.95	<0.001	0.16
Image: Naturalness	0.83	0.85	0.12	0.00	1.00	0.79	0.80	0.13	0.00	1.00	<0.001	0.14
Instruction: Num. Chars	682.57	616.00	327.94	86.00	2866.00	781.44	695.50	422.56	27.00	4571.00	<0.001	0.12
Instruction: Num. Words	115.99	104.00	56.07	14.00	498.00	110.06	97.00	61.72	9.00	804.00	<0.001	0.00
Instruction: Num. Sentences	9.73	9.00	4.34	1.00	40.00	10.43	10.00	5.96	1.00	64.00	<0.001	0.05
Instruction: Readability Score	30.42	30.45	5.50	8.00	54.49	48.67	45.96	15.87	2.00	295.59	<0.001	0.81
Instruction: Entropy	4.35	4.35	0.08	3.96	4.63	4.51	4.51	0.10	2.85	5.01	<0.001	0.70
Instruction: Sentiment	0.15	0.00	0.99	-3.00	4.00	0.29	0.00	1.13	-4.00	4.00	<0.001	0.07
Title: Num. Characters	25.43	24.00	9.93	3.00	110.00	51.72	45.00	32.16	5.00	347.00	<0.001	0.51
Title: Num. Words	4.13	4.00	1.85	1.00	19.00	7.67	6.00	5.51	1.00	56.00	<0.001	0.37
Title: Readbility Score	39.99	36.33	27.61	1.00	105.00	54.57	52.00	22.59	1.00	105.00	<0.001	0.29
Title: Entropy	3.72	3.76	0.33	1.50	4.54	4.04	4.11	0.37	1.92	5.08	<0.001	0.46
Title: Sentiment	0.03	0.00	0.33	-3.00	4.00	0.38	0.00	0.96	-3.00	4.00	<0.001	0.14
Set: Recipe Seasonality												
Upload Month	5.96	6.00	3.28	1.00	12.00	6.12	6.00	3.45	1.00	12.00	<0.01	0.02
Day of Month	15.71	16.00	8.74	1.00	31.00	15.72	16.00	8.63	1.00	31.00	0.987	0.00
Day of Week	4.23	5.00	2.01	1.00	7.00	3.85	4.00	2.01	1.00	7.00	<0.001	0.00
Within Season	0.08	0.09	0.01	0.04	0.25	0.08	0.08	0.01	0.05	0.15	<0.001	0.09
Set: Recipe Popularity and Appreciation												
Avg. Rating	4.27	4.38	0.58	1.00	5.00	4.97	5.00	0.13	2.50	5.00	<0.001	0.73
Avg. Comment Sentiment	1.89	2.00	0.73	-4.00	4.00	1.85	1.88	0.85	-3.00	4.00	<0.001	0.06
Num. Comments received within day	0.09	0.00	0.30	0.00	5.00	5.13	2.00	7.18	0.00	65.00	<0.001	0.65
Num. Comments received within week	0.20	0.00	0.50	0.00	7.00	7.93	4.00	9.41	0.00	90.00	<0.001	0.77
Num. Comments received within month	0.43	0.00	0.97	0.00	24.00	8.34	5.00	9.47	0.00	87.00	<0.001	0.75
Num. Comments received within year	2.30	1.00	5.43	0.00	212.00	9.11	5.00	9.84	0.00	99.00	<0.001	0.57
Num. Ratings received within day	0.09	0.00	0.30	0.00	5.00	11.86	6.00	14.50	0.00	117.00	<0.001	0.76
Num. Ratings received within week	0.20	0.00	0.50	0.00	7.00	19.16	11.00	20.79	0.00	160.00	<0.001	0.81
Num. Ratings received within month	0.43	0.00	0.97	0.00	24.00	20.65	12.00	21.75	0.00	161.00	<0.001	0.81
Num. Ratings received within year	2.30	1.00	5.43	0.00	212.00	23.06	15.00	22.22	0.00	179.00	<0.001	0.75
Set: Recipe Innovation												
Ingredients rank	295.54	245.05	198.80	3.00	2030.00	154.39	135.11	96.12	3.25	1281.00	<0.001	0.43
Categories rank	45.55	21.50	56.44	2.00	329.67	16.37	15.79	8.20	4.40	68.57	<0.001	0.19
Title Words rank	526.63	264.00	732.14	1.00	7436.00	1131.91	547.50	1699.97	1.00	21582.00	<0.001	0.22
Recipe Innovation Jaccard	0.57	0.60	0.14	0.00	0.92	0.51	0.53	0.13	0.00	0.94	<0.001	0.00
Avg. Recipe Innovation Jaccard	0.96	0.97	0.02	0.87	1.00	0.94	0.94	0.02	0.85	1.00	<0.001	0.50
Recipe Innovation IDF	0.02	0.01	0.05	0.00	0.79	0.00	0.00	0.02	0.00	0.51	<0.001	0.57
Set: User Activity & Context												
Recipes Uploaded until Upload	16.96	0.00	78.71	0.00	1386.00	263.80	92.00	388.27	0.00	5792.00	<0.001	0.70
Comments Written until Upload	9.24	0.00	55.26	0.00	2027.00	1581.71	107.00	3966.74	0.00	38124.00	<0.001	0.66
Comments until Upload	73.36	0.00	514.37	0.00	6310.00	2472.89	300.00	5389.24	0.00	51808.00	<0.001	0.74
Num. Distinct Ingredients used	144.43	46.00	248.89	2.00	1443.00	451.64	406.00	302.19	5.00	1344.00	<0.001	0.59
Num. Ingredients used (per recipe)	9.84	9.75	2.56	2.00	28.50	10.46	10.35	1.88	3.00	24.31	<0.001	0.17
Num. Distinct Categories used	31.98	4.00	84.35	1.00	325.00	78.79	81.00	33.42	4.00	165.00	<0.001	0.00
Num. Categories used (per recipe)	2.44	2.57	0.69	1.00	4.00	11.31	11.19	4.34	2.00	23.70	<0.001	0.86
Uploaded Recipes per day	0.05	0.00	0.16	0.00	1.00	0.33	0.16	0.49	0.00	5.10	<0.001	0.69
Uploaded Recipes per week	0.38	0.02	1.11	0.00	7.01	2.41	1.13	4.08	0.00	63.00	<0.001	0.69
Uploaded Recipes per month	1.64	0.10	4.83	0.01	30.48	10.62	5.18	17.10	0.01	154.87	<0.001	0.70
Uploaded Recipes per year	19.68	1.16	57.98	0.07	365.90	125.69	60.96	210.10	0.22	3285.00	<0.001	0.69
Num. Ratings provided per day	0.02	0.00	0.05	0.00	0.91	6.08	1.71	10.37	0.00	74.09	<0.001	0.70
Num. Ratings provided per week	0.11	0.02	0.35	0.00	6.35	41.33	11.24	71.05	0.00	666.98	<0.001	0.69
Num. Ratings provided per month	0.46	0.10	1.48	0.00	27.23	172.87	43.12	295.20	0.00	2858.49	<0.001	0.70
Num. Ratings provided per year	5.49	1.23	17.55	0.00	323.18	2094.13	568.24	3459.02	0.00	33920.80	<0.001	0.70
Num. Comments provided per day	0.02	0.00	0.05	0.00	0.91	2.25	0.48	4.43	0.00	50.43	<0.001	0.71
Num. Comments provided per week	0.13	0.04	0.38	0.00	6.35	15.85	3.50	29.46	0.00	350.88	<0.001	0.68
Num. Comments provided per month	0.58	0.17	1.65	0.01	27.60	69.83	15.97	132.98	0.01	1535.51	<0.001	0.68
Num. Comments provided per year	6.94	1.99	19.76	0.07	331.35	840.17	194.45	1570.78	0.17	10459.50	<0.001	0.69
Cook Living in Germany/USA	0.88	1.00	0.33	0.00	1.00	0.91	1.00	0.29	0.00	1.00	<0.1	0.03

Table A.3: Differences between Allrecipes.com and Kochbar.de online food communities based on the created features (Desserts).

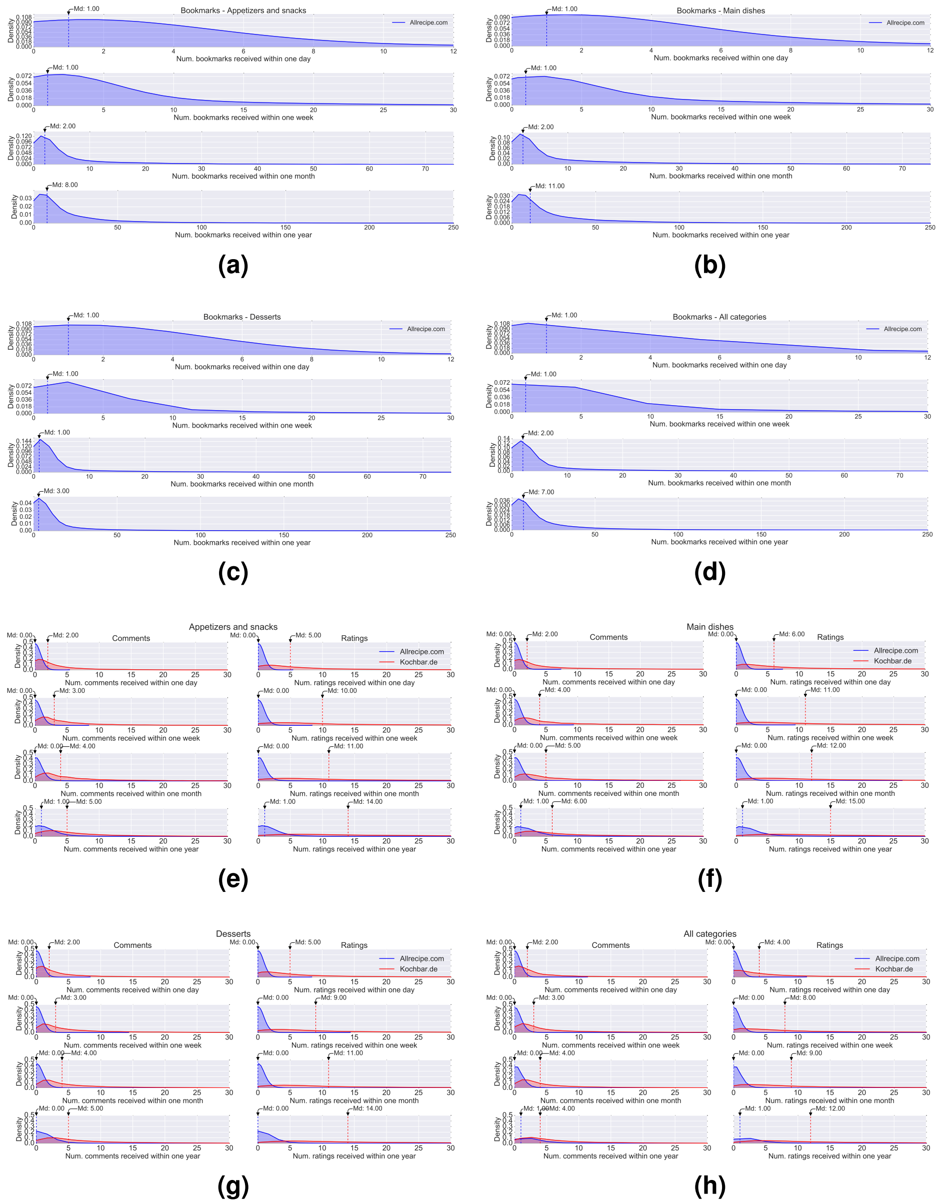
	Allrecipes.com (n=11526)					Kochbar.de (n=32098)						
Feature	*M*	*Md*	*SD*	*Min*	*Max*	*M*	*Md*	*SD*	*Min*	*Max*	*p*	*r*
Set: Recipe Nutrition												
Kcal (per 100g)	328.97	343.00	110.54	7.00	709.00	263.94	257.00	135.14	3.00	1000.00	<0.001	0.27
Protein (per 100g)	4.51	4.30	2.34	0.00	35.54	4.26	3.50	3.49	0.00	79.90	<0.001	0.13
Carbohydrates (per 100g)	43.57	45.47	14.44	1.84	99.05	30.63	25.50	20.99	0.00	100.00	<0.001	0.37
Fat (per 100g)	16.10	15.65	8.67	0.00	74.85	13.31	11.00	12.58	0.00	99.30	<0.001	0.19
Set: Recipe Healthiness												
WHO Health Score	3.56	3.00	0.69	3.00	6.00	3.43	3.00	0.70	3.00	6.00	<0.001	0.10
Set: Recipe Complexity												
Preparation Time (Min.)	59.47	45.00	140.68	3.00	9598.00	32.00	30.00	26.24	1.00	650.00	<0.001	0.42
Num. Preperation Steps	3.77	4.00	1.57	1.00	27.00	3.89	3.00	3.02	1.00	45.00	<0.001	0.09
Num. Servings	19.61	12.50	17.22	1.00	300.00	6.17	4.00	42.14	1.00	4500.00	<0.001	0.68
Num. Ingredients	9.07	9.00	3.56	1.00	30.00	8.82	8.00	3.41	1.00	48.00	<0.001	0.04
Num. Catergories	5.04	5.00	0.66	3.00	6.00	12.04	11.00	6.01	2.00	36.00	<0.001	0.66
Set: Recipe Presentation												
Image: Sharpness	0.19	0.16	0.11	0.01	1.20	0.18	0.15	0.12	0.01	1.55	<0.001	0.06
Image: Sharpness Variation	0.30	0.26	0.17	0.02	1.59	0.31	0.26	0.20	0.02	1.89	0.552	0.00
Image: Contrast	0.05	0.05	0.02	0.00	0.18	0.05	0.05	0.02	0.00	0.16	<0.001	0.07
Image: Contrast RGB	0.17	0.16	0.07	0.01	0.54	0.16	0.15	0.07	0.01	0.58	<0.001	0.06
Image: Saturation	0.22	0.21	0.11	0.02	0.75	0.19	0.17	0.11	0.00	0.86	<0.001	0.16
Image: Saturation Variation	0.13	0.12	0.05	0.01	0.35	0.13	0.12	0.06	0.00	0.39	0.353	0.01
Image: Brightness	0.49	0.49	0.10	0.13	0.93	0.48	0.47	0.11	0.08	0.94	<0.001	0.09
Image: Colorfulness	0.21	0.20	0.08	0.02	0.60	0.20	0.19	0.08	0.02	0.71	<0.001	0.05
Image: Entropy	7.38	7.47	0.39	2.73	7.96	7.33	7.40	0.38	3.90	7.97	<0.001	0.00
Image: Naturalness	0.76	0.77	0.15	0.00	1.00	0.73	0.73	0.15	0.00	1.00	<0.001	0.00
Instruction: Num. Chars	632.80	573.00	318.85	66.00	3686.00	677.53	587.00	409.50	40.00	5658.00	<0.001	0.03
Instruction: Num. Words	106.46	96.00	55.01	11.00	637.00	96.62	83.00	59.52	6.00	972.00	<0.001	0.13
Instruction: Num. Sentences	10.08	9.00	4.49	1.00	52.00	8.96	8.00	5.55	1.00	70.00	<0.001	0.15
Instruction: Readability Score	30.69	30.30	5.25	8.50	66.22	47.07	44.19	15.83	19.38	379.91	<0.001	0.78
Instruction: Entropy	4.41	4.41	0.08	4.06	4.86	4.49	4.49	0.12	3.54	5.28	<0.001	0.39
Instruction: Sentiment	0.24	0.00	0.76	-3.00	3.00	0.68	1.00	1.19	-4.00	4.00	<0.001	0.19
Title: Num. Characters	22.67	21.00	8.39	3.00	115.00	45.97	38.00	30.98	5.00	297.00	<0.001	0.46
Title: Num. Words	3.66	3.00	1.47	1.00	26.00	7.25	6.00	5.63	1.00	61.00	<0.001	0.37
Title: Readbility Score	40.12	36.33	27.97	1.00	105.00	54.62	50.31	24.37	1.00	105.00	<0.001	0.27
Title: Entropy	3.62	3.65	0.32	1.50	4.58	3.94	4.02	0.44	1.79	5.13	<0.001	0.42
Title: Sentiment	0.02	0.00	0.35	-3.00	3.00	0.48	0.00	1.03	-4.00	4.00	<0.001	0.18
Set: Recipe Seasonality												
Upload Month	6.71	7.00	3.39	1.00	12.00	6.13	6.00	3.37	1.00	12.00	<0.001	0.00
Day of Month	15.75	16.00	8.59	1.00	31.00	15.74	16.00	8.79	1.00	31.00	0.950	0.00
Day of Week	4.11	4.00	1.94	1.00	7.00	3.85	4.00	2.03	1.00	7.00	<0.001	0.07
Within Season	0.08	0.08	0.01	0.04	0.31	0.08	0.08	0.01	0.05	0.18	<0.001	0.03
Set: Recipe Popularity and Appreciation												
Avg. Rating	4.26	4.46	0.74	1.00	5.00	4.96	5.00	0.16	1.00	5.00	<0.001	0.67
Avg. Comment Sentiment	1.83	2.00	0.84	-4.00	4.00	1.78	1.80	0.87	-3.00	4.00	<0.001	0.08
Num. Comments received within day	0.10	0.00	0.33	0.00	6.00	4.36	2.00	6.18	0.00	81.00	<0.001	0.63
Num. Comments received within week	0.17	0.00	0.51	0.00	12.00	6.62	4.00	8.05	0.00	81.00	<0.001	0.75
Num. Comments received within month	0.30	0.00	1.01	0.00	41.00	7.26	4.00	8.56	0.00	107.00	<0.001	0.76
Num. Comments received within year	1.02	0.00	3.67	0.00	195.00	8.28	5.00	9.07	0.00	92.00	<0.001	0.72
Num. Ratings received within day	0.10	0.00	0.33	0.00	6.00	10.38	5.00	13.17	0.00	111.00	<0.001	0.74
Num. Ratings received within week	0.17	0.00	0.51	0.00	12.00	16.93	10.00	19.09	0.00	155.00	<0.001	0.80
Num. Ratings received within month	0.30	0.00	1.01	0.00	41.00	18.14	11.00	19.78	0.00	166.00	<0.001	0.81
Num. Ratings received within year	1.02	0.00	3.67	0.00	195.00	21.11	14.00	20.87	0.00	183.00	<0.001	0.81
56.14 *Set: Recipe Innovation*												
Ingredients rank	238.64	161.07	239.65	4.75	2162.60	160.50	133.80	115.29	3.80	1072.00	<0.001	0.12
Categories rank	76.69	56.33	65.80	1.00	402.33	17.67	17.50	7.96	6.67	75.06	<0.001	0.63
Title Words rank	551.65	281.50	759.77	2.00	7482.00	1338.83	647.26	2035.06	1.00	21513.00	<0.001	0.26
Recipe Innovation Jaccard	0.38	0.38	0.18	0.00	1.00	0.42	0.43	0.17	0.00	0.86	<0.001	0.10
Avg. Recipe Innovation Jaccard	0.93	0.93	0.04	0.82	1.00	0.94	0.95	0.03	0.85	1.00	<0.001	0.17
Recipe Innovation IDF	0.02	0.00	0.04	0.00	1.00	0.00	0.00	0.02	0.00	0.78	<0.001	0.46
Set: User Activity & Context												
Recipes Uploaded until Upload	17.04	0.00	67.06	0.00	1352.00	263.99	74.00	536.83	0.00	5814.00	<0.001	0.64
Comments Written until Upload	10.29	0.00	66.26	0.00	1870.00	1315.66	64.00	3552.11	0.00	36511.00	<0.001	0.63
Comments until Upload	40.58	0.00	336.51	0.00	6367.00	2047.48	213.00	4748.46	0.00	51315.00	<0.001	0.70
Num. Distinct Ingredients used	112.14	39.00	194.26	2.00	1443.00	412.70	355.00	311.41	4.00	1344.00	<0.001	0.59
Num. Ingredients used (per recipe)	9.17	9.00	2.59	1.50	25.00	9.84	9.88	1.78	3.75	24.31	<0.001	0.17
Num. Distinct Categories used	21.71	4.00	60.79	1.00	325.00	74.47	77.00	33.69	4.00	144.00	<0.001	0.73
Num. Categories used (per recipe)	2.91	3.00	0.57	1.00	4.00	11.19	11.14	4.59	2.00	27.00	<0.001	0.86
Uploaded Recipes per day	0.03	0.00	0.11	0.00	1.00	0.31	0.13	0.50	0.00	5.10	<0.001	0.71
Uploaded Recipes per week	0.23	0.02	0.80	0.00	7.01	2.15	0.86	3.84	0.01	63.00	<0.001	0.71
Uploaded Recipes per month	1.01	0.09	3.49	0.01	30.48	9.81	3.74	17.22	0.01	240.00	<0.001	0.71
Uploaded Recipes per year	12.15	1.11	41.93	0.07	365.90	126.82	48.23	205.52	0.16	1558.43	<0.001	0.71
Num. Ratings provided per day	0.02	0.00	0.07	0.00	1.93	5.01	1.23	8.64	0.00	95.28	<0.001	0.65
Num. Ratings provided per week	0.14	0.03	0.51	0.00	13.52	34.22	8.61	60.23	0.00	429.69	<0.001	0.64
Num. Ratings provided per month	0.61	0.11	2.21	0.00	57.93	147.34	37.10	266.81	0.00	2858.49	<0.001	0.65
Num. Ratings provided per year	7.24	1.32	26.18	0.00	687.46	1834.13	412.29	3422.65	0.00	33920.80	<0.001	0.66
Num. Comments provided per day	0.02	0.00	0.07	0.00	1.93	2.02	0.32	4.42	0.00	50.43	<0.001	0.68
Num. Comments provided per week	0.17	0.04	0.56	0.00	13.50	13.18	1.91	29.10	0.00	350.88	<0.001	0.63
Num. Comments provided per month	0.75	0.18	2.43	0.01	58.57	52.93	8.47	112.90	0.01	871.86	<0.001	0.65
Num. Comments provided per year	8.95	2.15	29.24	0.07	704.03	674.66	100.56	1490.82	0.16	18295.90	<0.001	0.64
Cook Living in Germany/USA	0.85	1.00	0.36	0.00	1.00	0.88	1.00	0.33	0.00	1.00	0.146	0.03

Table A.4: Differences between Allrecipes.com and Kochbar.de online food communities based on the created features (Appetizers and Snacks).

	Allrecipes.com (n=3978)					Kochbar.de (n=11973)						
Feature	*M*	*Md*	*SD*	*Min*	*Max*	*M*	*Md*	*SD*	*Min*	*Max*	*p*	*r*
Set: Recipe Nutrition												
Kcal (per 100g)	226.98	206.00	130.74	0.00	868.00	292.06	277.00	166.66	1.00	1990.00	<0.001	0.21
Protein (per 100g)	7.59	7.20	4.62	0.00	38.41	7.39	6.10	5.70	0.00	36.00	<0.001	0.06
Carbohydrates (per 100g)	14.52	9.55	13.66	0.00	82.35	19.72	9.70	21.46	0.00	92.70	0.249	0.01
Fat (per 100g)	15.86	13.18	12.52	0.00	98.07	20.62	16.00	20.09	0.00	100.00	<0.001	0.07
Set: Recipe Healthiness												
WHO Health Score	3.39	3.00	0.63	0.00	6.00	3.40	3.00	0.67	3.00	6.00	0.529	0.00
Set: Recipe Complexity												
Preparation Time (Min.)	47.36	35.00	61.18	5.00	1480.00	28.94	25.00	23.29	1.00	600.00	<0.001	0.31
Num. Preperation Steps	2.81	3.00	1.61	1.00	19.00	3.78	3.00	2.76	1.00	23.00	<0.001	0.16
Num. Servings	14.39	10.00	13.47	1.00	272.00	5.40	4.00	9.55	1.00	500.00	<0.001	0.60
Num. Ingredients	7.53	7.00	3.25	1.00	27.00	9.39	9.00	3.64	1.00	43.00	<0.001	0.27
Num. Catergories	4.65	5.00	0.77	3.00	6.00	14.42	15.00	6.57	2.00	36.00	<0.001	0.76
Set: Recipe Presentation												
Image: Sharpness	0.18	0.16	0.09	0.00	0.78	0.19	0.16	0.12	0.01	1.16	<0.1	0.03
Image: Sharpness Variation	0.28	0.25	0.15	0.00	1.24	0.31	0.26	0.19	0.02	2.04	<0.01	0.04
Image: Contrast	0.05	0.05	0.02	0.00	0.16	0.05	0.04	0.02	0.00	0.15	<0.001	0.00
Image: Contrast RGB	0.17	0.16	0.07	0.00	0.46	0.16	0.15	0.07	0.01	0.47	<0.001	0.00
Image: Saturation	0.26	0.25	0.11	0.03	0.76	0.22	0.20	0.11	0.02	0.79	<0.001	0.21
Image: Saturation Variation	0.15	0.15	0.05	0.00	0.36	0.14	0.14	0.05	0.02	0.37	<0.001	0.00
Image: Brightness	0.50	0.49	0.10	0.00	0.88	0.48	0.48	0.11	0.09	0.95	<0.001	0.10
Image: Colorfulness	0.24	0.23	0.08	0.00	0.59	0.21	0.21	0.08	0.03	0.60	<0.001	0.00
Image: Entropy	7.44	7.52	0.37	0.00	7.97	7.38	7.44	0.36	4.08	7.97	<0.001	0.00
Image: Naturalness	0.81	0.82	0.13	0.00	1.00	0.77	0.78	0.14	0.14	1.00	<0.001	0.12
Instruction: Num. Chars	459.93	402.50	275.81	23.00	2881.00	651.66	566.50	401.96	58.00	5386.00	<0.001	0.29
Instruction: Num. Words	77.34	67.00	48.16	4.00	490.00	92.36	79.00	59.17	3.00	819.00	<0.001	0.15
Instruction: Num. Sentences	6.76	6.00	3.98	1.00	34.00	8.53	8.00	5.33	1.00	49.00	<0.001	0.17
Instruction: Readability Score	31.79	31.35	6.75	2.00	71.45	47.57	44.73	15.55	16.36	241.25	<0.001	0.73
Instruction: Entropy	4.32	4.32	0.10	3.64	4.61	4.50	4.51	0.11	3.24	4.99	<0.001	0.69
Instruction: Sentiment	0.02	0.00	0.81	-3.00	3.00	0.53	0.00	1.16	-2.00	4.00	<0.001	0.22
Title: Num. Characters	22.58	21.00	8.80	4.00	80.00	49.87	44.00	30.82	5.00	297.00	<0.001	0.55
Title: Num. Words	3.73	3.00	1.61	1.00	16.00	7.75	7.00	5.49	1.00	49.00	<0.001	0.44
Title: Readbility Score	35.19	36.33	27.54	1.00	104.00	53.92	50.00	23.70	1.00	105.00	<0.001	0.35
Title: Entropy	3.64	3.68	0.34	1.92	4.53	4.05	4.13	0.40	2.00	5.17	<0.001	0.53
Title: Sentiment	0.04	0.00	0.38	-3.00	4.00	0.48	0.00	1.06	-3.00	4.00	<0.001	0.17
Set: Recipe Seasonality												
Upload Month	6.14	6.00	3.40	1.00	12.00	6.14	6.00	3.36	1.00	12.00	0.294	0.00
Day of Month	15.98	16.00	8.82	1.00	31.00	15.94	16.00	8.76	1.00	31.00	0.869	0.00
Day of Week	4.24	5.00	1.96	1.00	7.00	3.86	4.00	2.05	1.00	7.00	<0.001	0.09
Within Season	0.08	0.08	0.01	0.05	0.13	0.08	0.08	0.01	0.05	0.17	<0.001	0.06
Set: Recipe Popularity and Appreciation												
Avg. Rating	4.32	4.48	0.64	1.00	5.00	4.96	5.00	0.17	1.00	5.00	<0.001	0.65
Avg. Comment Sentiment	1.82	2.00	0.82	-3.00	4.00	1.74	1.75	0.88	-3.00	4.00	<0.001	0.09
Num. Comments received within day	0.10	0.00	0.33	0.00	3.00	3.98	2.00	5.64	0.00	55.00	<0.001	0.63
Num. Comments received within week	0.20	0.00	0.51	0.00	6.00	6.18	3.00	7.62	0.00	63.00	<0.001	0.75
Num. Comments received within month	0.40	0.00	1.22	0.00	45.00	6.73	4.00	7.79	0.00	64.00	<0.001	0.75
Num. Comments received within year	1.61	1.00	4.01	0.00	104.00	7.77	5.00	8.69	0.00	99.00	<0.001	0.63
Num. Ratings received within day	0.10	0.00	0.33	0.00	3.00	9.92	5.00	12.42	0.00	102.00	<0.001	0.75
Num. Ratings received within week	0.20	0.00	0.51	0.00	6.00	16.44	10.00	18.16	0.00	131.00	<0.001	0.81
Num. Ratings received within month	0.40	0.00	1.22	0.00	45.00	17.38	11.00	18.17	0.00	133.00	<0.001	0.81
Num. Ratings received within year	1.61	1.00	4.01	0.00	104.00	20.34	14.00	19.42	0.00	158.00	<0.001	0.79
Set: Recipe Innovation												
Ingredients rank	317.85	253.47	249.06	3.00	2111.50	168.72	143.20	119.62	4.25	1002.67	<0.001	0.36
Categories rank	69.75	40.50	73.67	4.00	445.50	24.15	24.00	7.27	10.12	63.50	<0.001	0.22
Title Words rank	549.17	295.83	709.90	1.00	7135.00	1230.67	572.92	1862.81	1.00	20105.00	<0.001	0.22
Recipe Innovation Jaccard	0.54	0.57	0.16	0.00	0.91	0.47	0.50	0.17	0.00	0.89	<0.001	0.21
Avg. Recipe Innovation Jaccard	0.97	0.97	0.02	0.86	1.00	0.94	0.94	0.03	0.86	1.00	<0.001	0.49
Recipe Innovation IDF	0.02	0.01	0.06	0.00	0.75	0.00	0.00	0.02	0.00	0.63	<0.001	0.50
Set: User Activity & Context												
Recipes Uploaded until Upload	17.78	0.00	87.60	0.00	1380.00	254.38	73.00	492.72	0.00	5770.00	<0.001	0.68
Comments Written until Upload	11.25	0.00	65.11	0.00	1278.00	995.70	59.50	2855.02	0.00	35695.00	<0.001	0.65
Comments until Upload	67.77	0.00	510.73	0.00	6353.00	1796.65	225.00	4369.31	0.00	52812.00	<0.001	0.73
Num. Distinct Ingredients used	136.47	42.00	250.02	2.00	1443.00	414.43	336.00	310.87	4.00	1344.00	<0.001	0.58
Num. Ingredients used (per recipe)	8.70	8.71	2.35	2.00	21.50	10.02	10.03	1.85	4.00	30.00	<0.001	0.31
Num. Distinct Categories used	28.52	4.00	77.87	1.00	325.00	80.89	82.00	33.47	5.00	165.00	<0.001	0.00
Num. Categories used (per recipe)	2.68	2.67	0.61	1.00	4.00	12.05	12.27	4.70	3.21	25.13	<0.001	0.86
Uploaded Recipes per day	0.05	0.00	0.16	0.00	1.00	0.34	0.14	0.55	0.00	5.10	<0.001	0.70
Uploaded Recipes per week	0.36	0.02	1.13	0.00	7.01	2.44	0.98	4.10	0.00	59.50	<0.001	0.69
Uploaded Recipes per month	1.56	0.10	4.90	0.01	30.48	10.35	3.99	17.69	0.03	240.00	<0.001	0.69
Uploaded Recipes per year	18.77	1.18	58.88	0.06	365.90	123.05	50.27	208.05	0.26	3102.50	<0.001	0.69
Num. Ratings provided per day	0.02	0.00	0.07	0.00	0.82	4.72	1.24	8.45	0.00	95.28	<0.001	0.68
Num. Ratings provided per week	0.15	0.02	0.46	0.00	5.75	34.44	7.89	62.47	0.00	666.98	<0.001	0.66
Num. Ratings provided per month	0.64	0.11	1.98	0.00	24.63	142.29	33.83	243.65	0.00	2858.49	<0.001	0.66
Num. Ratings provided per year	7.56	1.26	23.46	0.00	292.23	1616.67	401.47	2943.21	0.00	33920.80	<0.001	0.67
Num. Comments provided per day	0.02	0.00	0.07	0.00	0.82	1.46	0.25	3.42	0.00	28.66	<0.001	0.69
Num. Comments provided per week	0.18	0.04	0.50	0.00	5.74	11.18	2.11	26.42	0.00	200.42	<0.001	0.67
Num. Comments provided per month	0.77	0.17	2.17	0.01	24.97	53.96	8.26	132.03	0.01	1535.51	<0.001	0.64
Num. Comments provided per year	9.29	2.05	26.05	0.07	299.64	583.23	101.59	1411.09	0.17	18295.90	<0.001	0.65
Cook Living in Germany/USA	0.89	1.00	0.31	0.00	1.00	0.93	1.00	0.25	0.00	1.00	0.115	0.04

A.4 Predictive Modelling

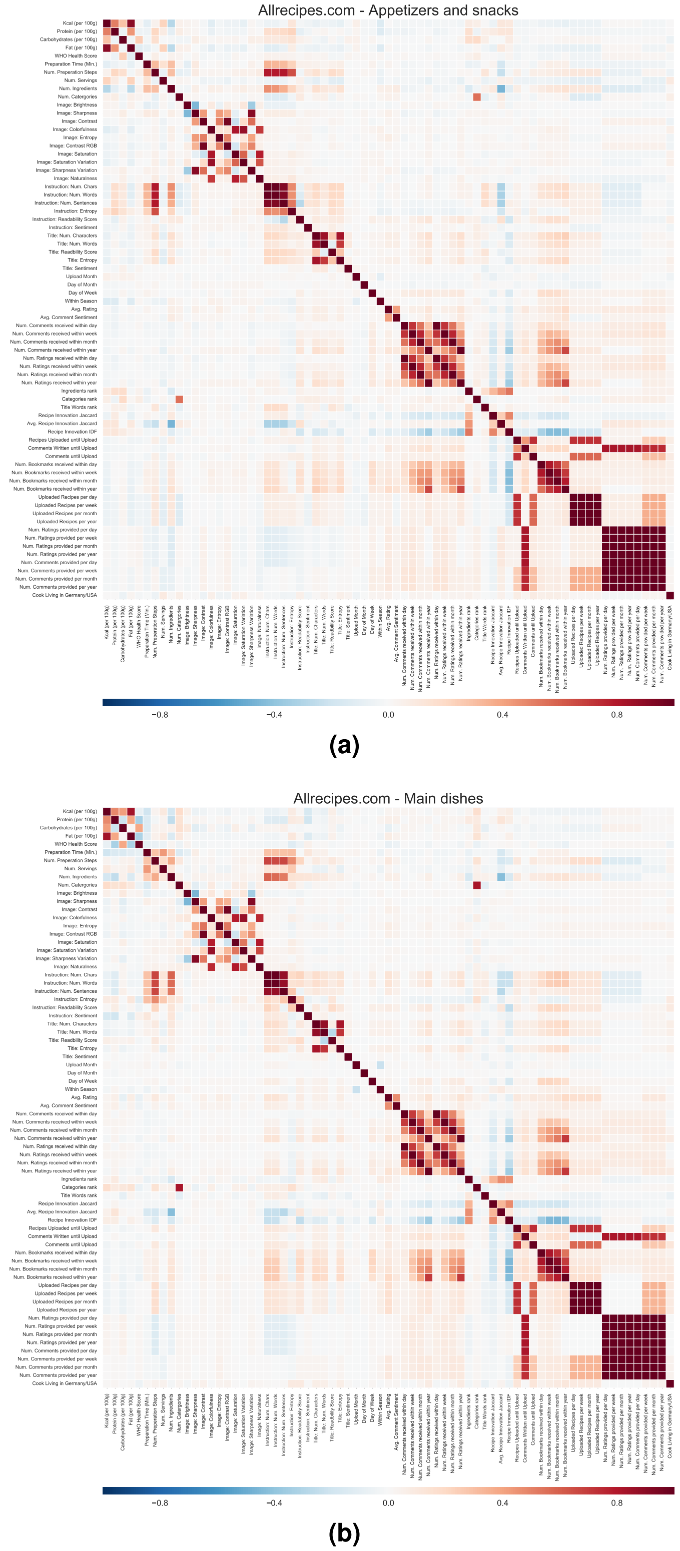
A.4.1 Dependend and Independend Variables Analysis



(a) **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)**

Figure A.3: Density distribution plots of all dependend variable analyses and all recipe categories.

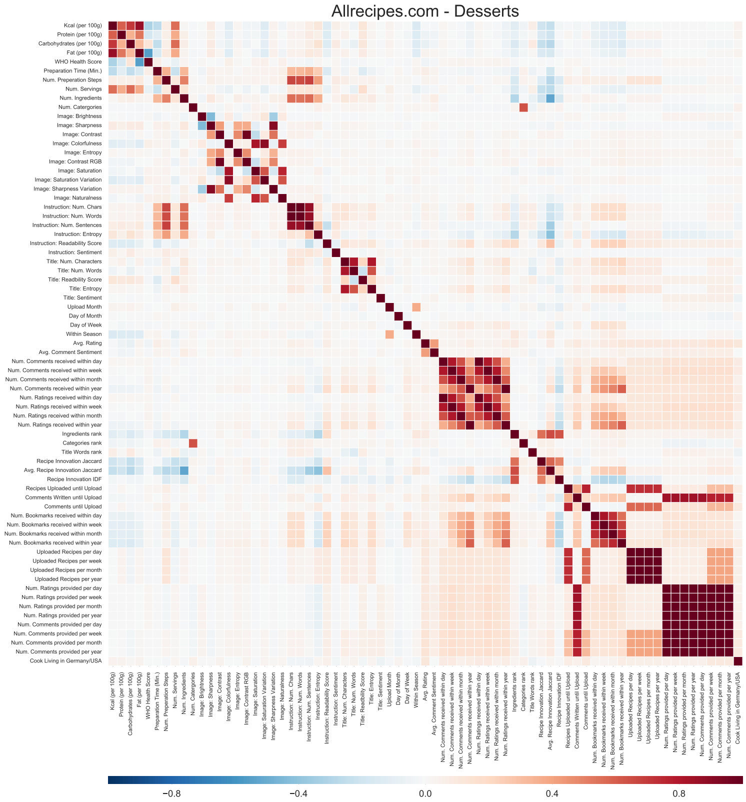
A.4.2 Feature Correlation Analysis



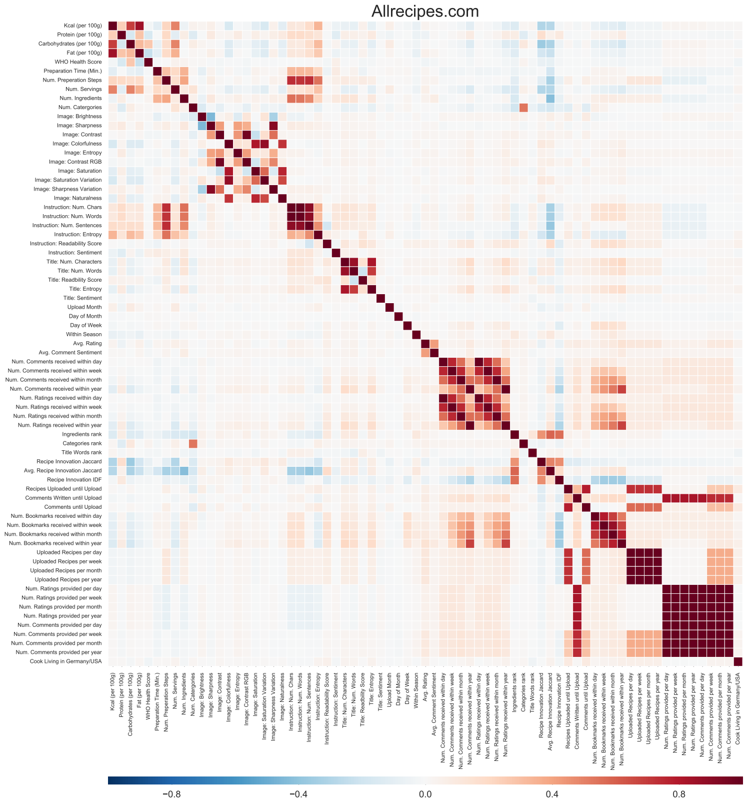
(a)

(b)

Figure A.4: Feature correlation heatmap based on Spearman's rank correlation coefficient of all categories (Allrecipes.com) 1/2.

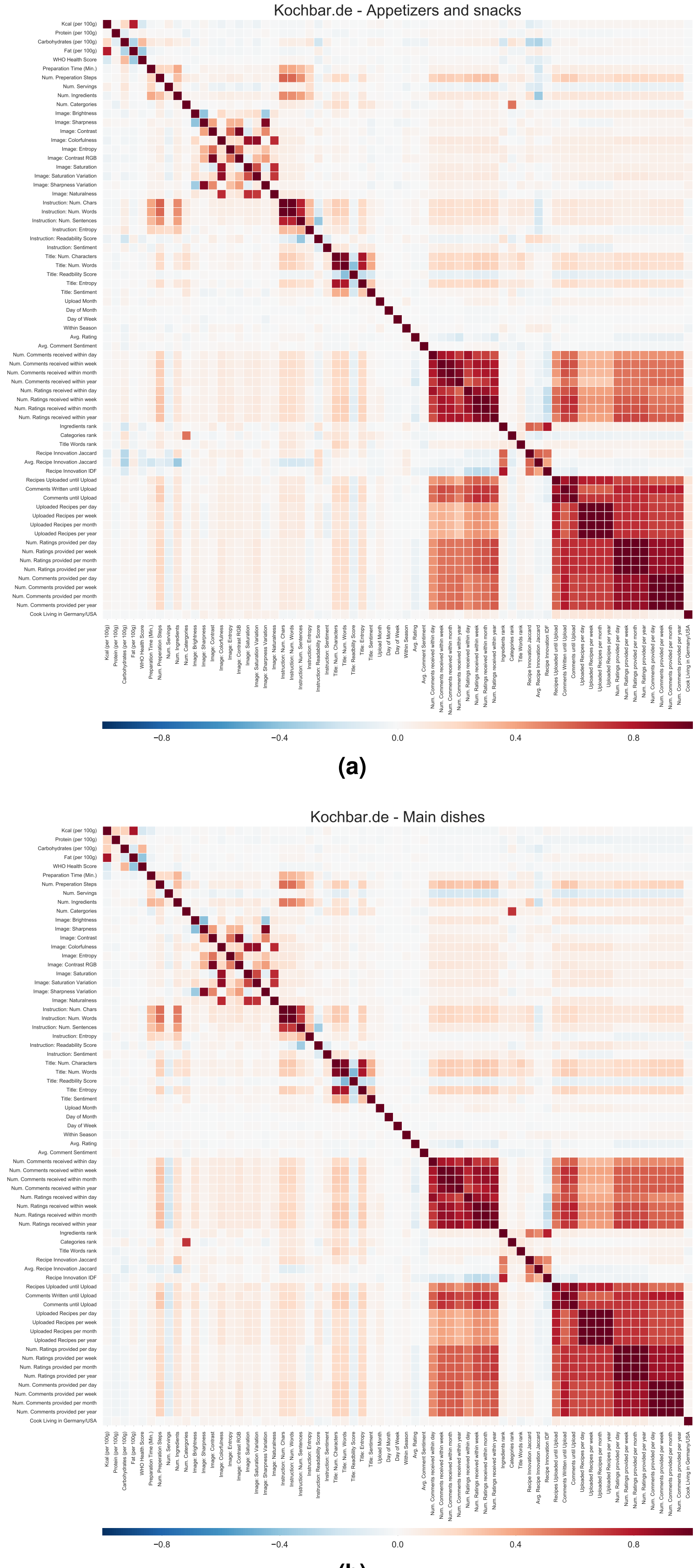


(a)



(b)

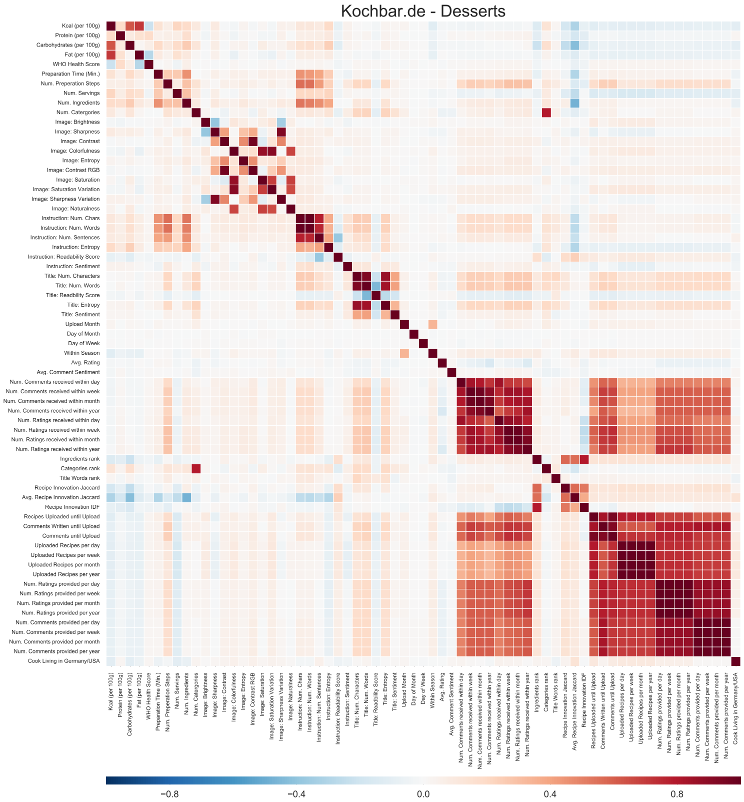
Figure A.5: Feature correlation heatmap based on Spearman's rank correlation coefficient of all categories (Allrecipes.com) 2/2.



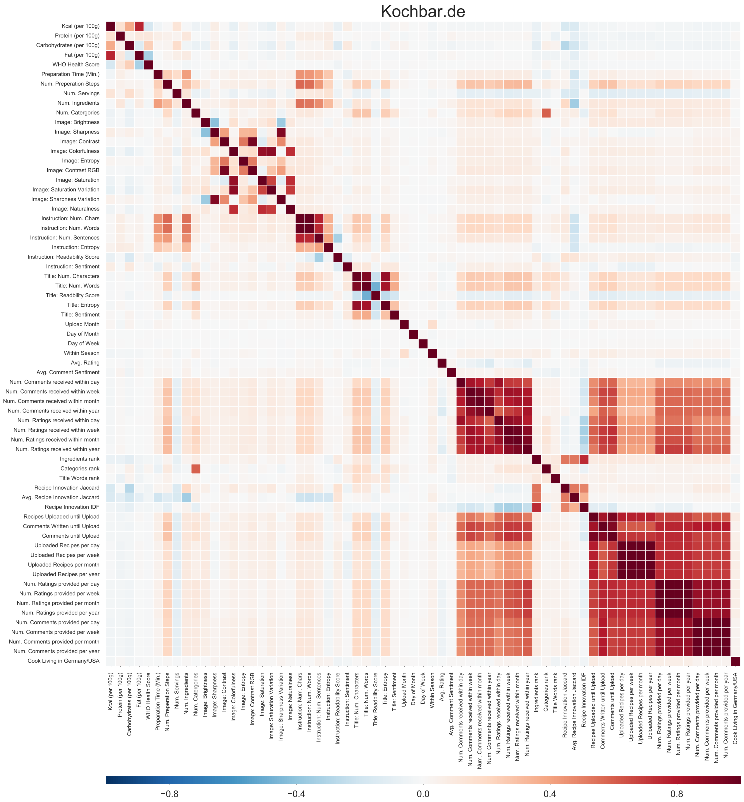
(a)

(b)

Figure A.6: Feature correlation heatmap based on Spearman's rank correlation coefficient of all categories (Kochbar.de) 1/2.



(a)



(b)

Figure A.7: Feature correlation heatmap based on Spearman's rank correlation coefficient of all categories (Kochbar.de) 2/2.

A.4.3 Information Gain Results

Table A.5: Information Gain of the num. comments within one week/-month dependend variable.

	Allrecipes.com				Kochbar.de			
	Within one Week		Within one Month		Within one Week		Within one Month	
Rank	IG	Feature	IG	Feature	IG	Feature	IG	Feature
Appetizers and Snacks								
1	.0265	INN:Recipe Innovation IDF	.0258	ACT:Comments Written until Upload	.1535	ACT:Comments Written until Upload	.1376	ACT:Comments Written until Upload
2	.0187	INN:Recipe Innovatin Jaccard	.0246	INN:Recipe Innovation IDF	.1404	ACT:Comments until Upload	.1207	ACT:Comments until Upload
3	-	-	.0190	ACT:Uploaded Recipes per month	.1161	ACT:Comments Written until Upload	.1183	ACT:Comments Written until Upload
4	-	-	.0175	ACT:Cook Living in Germany/USA	.1138	ACT:Num. Ratings provided per month	.1121	ACT:Num. Ratings provided per month
5	-	-	.0121	INN:Recipe Innovatin Jaccard	.0779	ACT:Recipes Uploaded until Upload	.0637	ACT:Recipes Uploaded until Upload
6	-	-	-	-	.0519	ACT:Uploaded Recipes per month	.0556	ACT:Uploaded Recipes per month
7	-	-	-	-	.0518	IMG:Image Entropy	.0492	IMG:Image Saturation Variation
8	-	-	-	-	.0516	IMG:Image Saturation Variation	.0491	IMG:Image Entropy
9	-	-	-	-	.0515	IMG:Image Brightness	.0489	IMG:Image Contrast RGB
10	-	-	-	-	.0512	IMG:Image Sharpness	.0489	IMG:Image Colorfulness
Main Dishes								
1	.0119	INN:Recipe Innovation IDF	.0168	INN:Recipe Innovation IDF	.2055	ACT:Comments Written until Upload	.1850	ACT:Comments Written until Upload
2	.0089	INN:Recipe Innovatin Jaccard	.0140	ACT:Comments Written until Upload	.1909	ACT:Comments until Upload	.1786	ACT:Num. Ratings provided per month
3	.0070	ACT:Uploaded Recipes per month	.0134	ACT:Comments until Upload	.1816	ACT:Num. Ratings provided per month	.1756	ACT:Comments Written until Upload
4	.0066	ACT:Cook Living in Germany/USA	.0133	ACT:Num. Ratings provided per month	.1788	ACT:Comments Written until Upload	.1637	ACT:Comments until Upload
5	-	-	.0053	INN:Recipe Innovatin Jaccard	.1478	ACT:Uploaded Recipes per month	.1478	ACT:Uploaded Recipes per month
6	-	-	.0042	COM:Preparation Time (Min.)	.0971	ACT:Recipes Uploaded until Upload	.0769	ACT:Recipes Uploaded until Upload
7	-	-	-	-	.0697	IMG:Image Saturation Variation	.0676	IMG:Image Contrast RGB
8	-	-	-	-	.0697	IMG:Image Saturation	.0676	IMG:Image Saturation
9	-	-	-	-	.0694	IMG:Image Contrast RGB	.0674	IMG:Image Saturation Variation
10	-	-	-	-	.0693	IMG:Image Brightness	.0671	IMG:Image Contrast
Desserts								
1	.0180	ACT:Num. Ratings provided per month	.0238	ACT:Comments Written until Upload	.1703	ACT:Comments Written until Upload	.1522	ACT:Comments Written until Upload
2	.0166	ACT:Cook Living in Germany/USA	.0210	ACT:Num. Ratings provided per month	.1604	ACT:Comments until Upload	.1460	ACT:Num. Ratings provided per month
3	.0127	ACT:Uploaded Recipes per month	.0203	ACT:Comments until Upload	.1479	ACT:Comments Written until Upload	.1421	ACT:Comments Written until Upload
4	.0099	ACT:Comments Written until Upload	.0153	ACT:Recipes Uploaded until Upload	.1472	ACT:Num. Ratings provided per month	.1327	ACT:Comments until Upload
5	-	-	.0191	INN:Recipe Innovation IDF	.0905	ACT:Recipes Uploaded until Upload	.0916	ACT:Uploaded Recipes per month
6	-	-	.0157	ACT:Cook Living in Germany/USA	.0872	ACT:Uploaded Recipes per month	.0707	ACT:Recipes Uploaded until Upload
7	-	-	.0153	ACT:Uploaded Recipes per month	.0561	IMG:Image Contrast RGB	.0551	IMG:Image Saturation Variation
8	-	-	.0063	SEA:Day of Week	.0561	IMG:Image Contrast	.0551	IMG:Image Contrast RGB
9	-	-	-	-	.0558	IMG:Image Sharpness variation	.0551	IMG:Image Sharpness variation
10	-	-	-	-	.0558	IMG:Image Saturation Variation	.0550	IMG:Image Contrast
All Categories								
1	.0122	INN:Recipe Innovation IDF	.0171	INN:Recipe Innovation IDF	.1767	ACT:Comments Written until Upload	.1646	ACT:Comments Written until Upload
2	.0118	IMG:Image Entropy	.0161	IMG:Image Entropy	.1681	ACT:Comments until Upload	.1563	ACT:Num. Ratings provided per month
3	.0114	IMG:Image Brightness	.0153	IMG:Image Saturation Variation	.1652	ACT:Num. Ratings provided per month	.1525	ACT:Comments Written until Upload
4	.0112	IMG:Image Contrast RGB	.0153	IMG:Image Colorfulness	.1544	ACT:Comments Written until Upload	.1385	ACT:Comments until Upload
5	.0112	IMG:Image Saturation Variation	.0153	IMG:Image Saturation	.1520	ACT:Uploaded Recipes per month	.1339	ACT:Uploaded Recipes per month
6	.0111	IMG:Image Colorfulness	.0152	IMG:Image Naturalness	.0942	ACT:Recipes Uploaded until Upload	.0730	ACT:Recipes Uploaded until Upload
7	.0111	IMG:Image Sharpness	.0152	IMG:Image Contrast RGB	.0597	IMG:Image Contrast RGB	.0566	IMG:Image Contrast RGB
8	.0108	IMG:Image Saturation	.0152	IMG:Image Contrast	.0594	IMG:Image Saturation Variation	.0564	IMG:Image Saturation Variation
9	.0108	IMG:Image Contrast	.0151	ACT:Comments Written until Upload	.0593	IMG:Image Sharpness variation	.0564	IMG:Image Sharpness variation
10	.0102	IMG:Image Naturalness	.0151	IMG:Image Sharpness	.0592	IMG:Image Contrast	.0563	IMG:Image Contrast

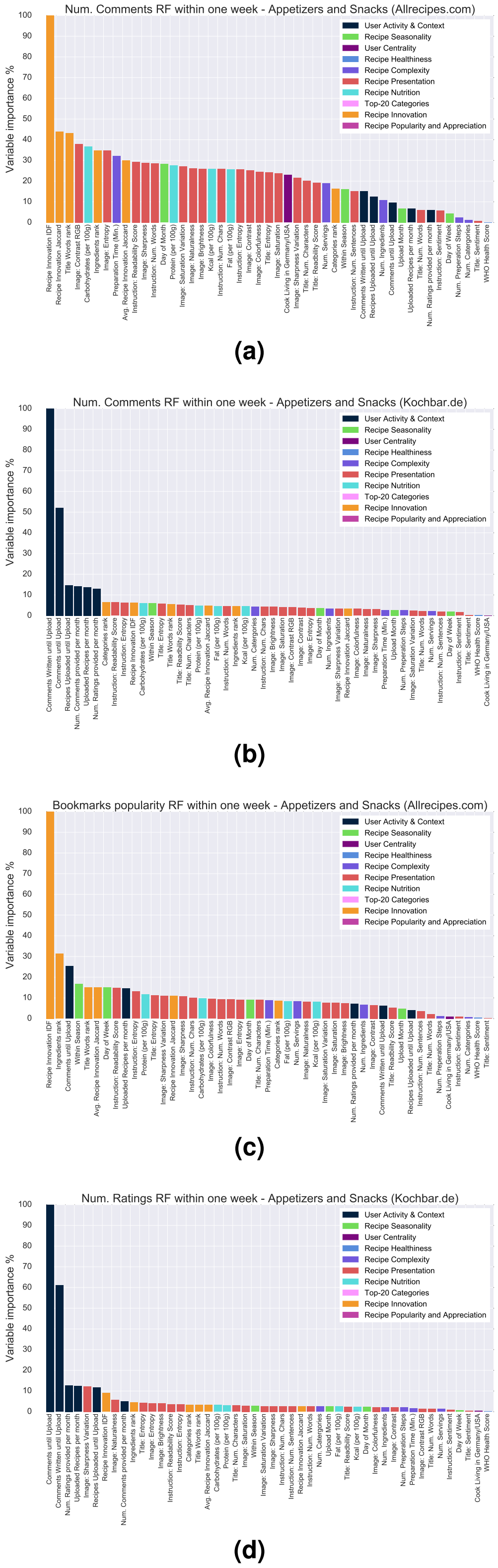
Table A.6: Information Gain of the num. ratings within one week/month dependend variable.

	Allrecipes.com				Kochbar.de			
	Within one Week		Within one Month		Within one Week		Within one Month	
Rank	IG	Feature	IG	Feature	IG	Feature	IG	Feature
Appetizers and Snacks								
1	.0265	INN:Recipe Innovation IDF	.0258	ACT:Comments Written until Upload	.2160	ACT:Comments until Upload	.1975	ACT:Comments until Upload
2	.0187	INN:Recipe Innovatin Jaccard	.0246	INN:Recipe Innovation IDF	.1971	ACT:Comments Written until Upload	.1865	ACT:Comments Written until Upload
3	-	-	.0190	ACT:Uploaded Recipes per month	.1431	ACT:Num. Ratings provided per month	.1455	ACT:Num. Ratings provided per month
4	-	-	.0175	ACT:Cook Living in Germany/USA	.1252	ACT:Recipes Uploaded until Upload	.1101	ACT:Comments Written until Upload
5	-	-	.0121	INN:Recipe Innovatin Jaccard	.1145	ACT:Comments Written until Upload	.1070	ACT:Recipes Uploaded until Upload
6	-	-	-	-	.0796	ACT:Uploaded Recipes per month	.0779	IMG:Image Saturation Variation
7	-	-	-	-	.0710	IMG:Image Saturation Variation	.0774	IMG:Image Saturation
8	-	-	-	-	.0710	IMG:Image Sharpness	.0772	IMG:Image Naturalness
9	-	-	-	-	.0710	IMG:Image Entropy	.0772	IMG:Image Sharpness
10	-	-	-	-	.0708	IMG:Image Contrast	.0771	IMG:Image Contrast
Main Dishes								
1	.0119	INN:Recipe Innovation IDF	.0168	INN:Recipe Innovation IDF	.2315	ACT:Comments until Upload	.2142	ACT:Comments until Upload
2	.0089	INN:Recipe Innovatin Jaccard	.0140	ACT:Comments Written until Upload	.2151	ACT:Comments Written until Upload	.2043	ACT:Comments Written until Upload
3	.0070	ACT:Uploaded Recipes per month	.0134	ACT:Comments until Upload	.1736	ACT:Num. Ratings provided per month	.1841	ACT:Num. Ratings provided per month
4	.0066	ACT:Cook Living in Germany/USA	.0133	ACT:Num. Ratings provided per month	.1664	ACT:Comments Written until Upload	.1618	ACT:Comments Written until Upload
5	-	-	.0053	INN:Recipe Innovatin Jaccard	.1474	ACT:Uploaded Recipes per month	.1543	ACT:Uploaded Recipes per month
6	-	-	.0042	COM:Preparation Time (Min.)	.1283	ACT:Recipes Uploaded until Upload	.1118	ACT:Recipes Uploaded until Upload
7	-	-	-	-	.0814	IMG:Image Saturation Variation	.0835	IMG:Image Saturation Variation
8	-	-	-	-	.0813	IMG:Image Saturation	.0834	IMG:Image Naturalness
9	-	-	-	-	.0810	IMG:Image Sharpness variation	.0834	IMG:Image Saturation
10	-	-	-	-	.0810	IMG:Image Contrast	.0830	IMG:Image Contrast
Desserts								
1	.0180	ACT:Num. Ratings provided per month	.0238	ACT:Comments Written until Upload	.2170	ACT:Comments until Upload	.2027	ACT:Comments until Upload
2	.0166	ACT:Cook Living in Germany/USA	.0210	ACT:Num. Ratings provided per month	.1987	ACT:Comments Written until Upload	.1906	ACT:Comments Written until Upload
3	.0127	ACT:Uploaded Recipes per month	.0203	ACT:Comments until Upload	.1616	ACT:Num. Ratings provided per month	.1654	ACT:Num. Ratings provided per month
4	.0099	ACT:Comments Written until Upload	.0153	ACT:Recipes Uploaded until Upload	.1267	ACT:Comments Written until Upload	.1392	ACT:Comments Written until Upload
5	-	-	.0191	INN:Recipe Innovation IDF	.1254	ACT:Recipes Uploaded until Upload	.1110	ACT:Recipes Uploaded until Upload
6	-	-	.0157	ACT:Cook Living in Germany/USA	.1062	ACT:Uploaded Recipes per month	.0994	ACT:Uploaded Recipes per month
7	-	-	.0153	ACT:Uploaded Recipes per month	.0753	IMG:Image Contrast RGB	.0784	IMG:Image Saturation Variation
8	-	-	.0063	SEA:Day of Week	.0751	IMG:Image Contrast	.0782	IMG:Image Contrast RGB
9	-	-	-	-	.0751	IMG:Image Sharpness variation	.0782	IMG:Image Contrast
10	-	-	-	-	.0750	IMG:Image Saturation Variation	.0781	IMG:Image Naturalness
All Categories								
1	.0122	INN:Recipe Innovation IDF	.0171	INN:Recipe Innovation IDF	.2143	ACT:Comments until Upload	.2009	ACT:Comments until Upload
2	.0118	IMG:Image Entropy	.0161	IMG:Image Entropy	.1921	ACT:Comments Written until Upload	.1924	ACT:Num. Ratings provided per month
3	.0114	IMG:Image Brightness	.0153	IMG:Image Saturation Variation	.1805	ACT:Num. Ratings provided per month	.1884	ACT:Comments Written until Upload
4	.0112	IMG:Image Contrast RGB	.0153	IMG:Image Colorfulness	.17860	ACT:Comments Written until Upload	.1848	ACT:Comments Written until Upload
5	.0112	IMG:Image Saturation Variation	.0153	IMG:Image Saturation	.1684	ACT:Uploaded Recipes per month	.1726	ACT:Uploaded Recipes per month
6	.0111	IMG:Image Colorfulness	.0152	IMG:Image Naturalness	.1269	ACT:Recipes Uploaded until Upload	.1128	ACT:Recipes Uploaded until Upload
7	.0111	IMG:Image Sharpness	.0152	IMG:Image Contrast RGB	.0738	IMG:Image Sharpness variation	.0076	IMG:Image Sharpness variation
8	.0108	IMG:Image Saturation	.0152	IMG:Image Contrast	.0737	IMG:Image Contrast RGB	.0762	IMG:Image Contrast RGB
9	.0108	IMG:Image Contrast	.0151	ACT:Comments Written until Upload	.0737	IMG:Image Saturation Variation	.0761	IMG:Image Saturation Variation
10	.0102	IMG:Image Naturalness	.0151	IMG:Image Sharpness	.0736	IMG:Image Contrast	.0761	IMG:Image Contrast

Table A.7: Information Gain of the num. bookmarks within one week/month dependend variable.

	Allrecipes.com			
	Within one Week		Within one Month	
Rank	IG	Feature	IG	Feature
Appetizers and Snacks				
1	.0634	INN:Recipe Innovation IDF	.0624	INN:Recipe Innovation IDF
2	.0451	ACT:Comments Written until Upload	.0562	ACT:Comments Written until Upload
3	.0433	ACT:Comments until Upload	.0557	ACT:Comments until Upload
4	.0415	ACT:Recipes Uploaded until Upload	.0525	ACT:Recipes Uploaded until Upload
5	.0219	ACT:Num. Ratings provided per month	.0308	ACT:Uploaded Recipes per month
6	.0172	ACT:Uploaded Recipes per month	.0307	ACT:Num. Ratings provided per month
7	.0168	ACT:Cook Living in Germany/USA	.0260	ACT:Cook Living in Germany/USA
8	.0148	SEA:Day of Week	.0177	PRE:Title Num. Chars
9	.0116	PRE:Title Num. Chars	.0139	SEA:Day of Week
10	.0105	INN:Recipe Innovatin Jaccard	.0121	title entropy
Main Dishes				
1	.0640	INN:Recipe Innovation IDF	.0707	INN:Recipe Innovation IDF
2	.0466	ACT:Comments Written until Upload	.0524	ACT:Comments Written until Upload
3	.0453	ACT:Comments until Upload	.0502	ACT:Comments until Upload
4	.0442	ACT:Recipes Uploaded until Upload	.0494	ACT:Recipes Uploaded until Upload
5	.0233	ACT:Uploaded Recipes per month	.0281	ACT:Uploaded Recipes per month
6	.0219	SEA:Day of Week	.0252	ACT:Num. Ratings provided per month
7	.0205	ACT:Num. Ratings provided per month	.0205	PRE:Instructions Num. Chars
8	.0155	PRE:Instructions Num. Chars	.0200	SEA:Day of Week
9	.0148	ACT:Cook Living in Germany/USA	.0185	PRE:Instructions Num. Words
10	.0146	PRE:Instructions Num. Words	.0176	ACT:Cook Living in Germany/USA
Desserts				
1	.0945	ACT:Comments Written until Upload	.0772	ACT:Comments Written until Upload
2	.0850	ACT:Recipes Uploaded until Upload	.0671	ACT:Recipes Uploaded until Upload
3	.0844	ACT:Comments until Upload	.0663	ACT:Comments until Upload
4	.0427	ACT:Num. Ratings provided per month	.0402	ACT:Num. Ratings provided per month
5	.0419	ACT:Uploaded Recipes per month	.0401	ACT:Uploaded Recipes per month
6	.0377	INN:Recipe Innovation IDF	.0321	COM:Preparation Time (Min.)
7	.0338	ACT:Cook Living in Germany/USA	.0309	INN:Recipe Innovation IDF
8	.0331	COM:Preparation Time (Min.)	.0300	ACT:Cook Living in Germany/USA
9	.0304	INN:Mean Recipe Innovatin Jaccard	.0270	INN:Mean Recipe Innovatin Jaccard
10	.0204	PRE:Instruction Readability Score	.0187	PRE:Instruction Readability Score
All Categories				
1	.0320	INN:Recipe Innovation IDF	.0523	ACT:Comments Written until Upload
2	.0153	SEA:Day of Week	.0477	ACT:Recipes Uploaded until Upload
3	.0108	INN:Recipe Innovatin Jaccard	.0465	ACT:Comments until Upload
4	.0103	PRE:Instruction Readability Score	.0458	INN:Recipe Innovation IDF
5	.0100	PRE:Instructions Num. Chars	.0289	ACT:Uploaded Recipes per month
6	.0097	PRE:Instructions Num. Words	.0270	ACT:Num. Ratings provided per month
7	.0085	ACT:Recipes Uploaded until Upload	.0212	ACT:Cook Living in Germany/USA
8	.0049	INN:Mean Recipe Innovatin Jaccard	.0196	INN:Mean Recipe Innovatin Jaccard
9	.0047	ACT:Uploaded Recipes per month	.0146	SEA:Day of Week
10	.0043	ACT:Comments Written until Upload	.0130	COM:Preparation Time (Min.)

A.4.4 Classification Variables Importance



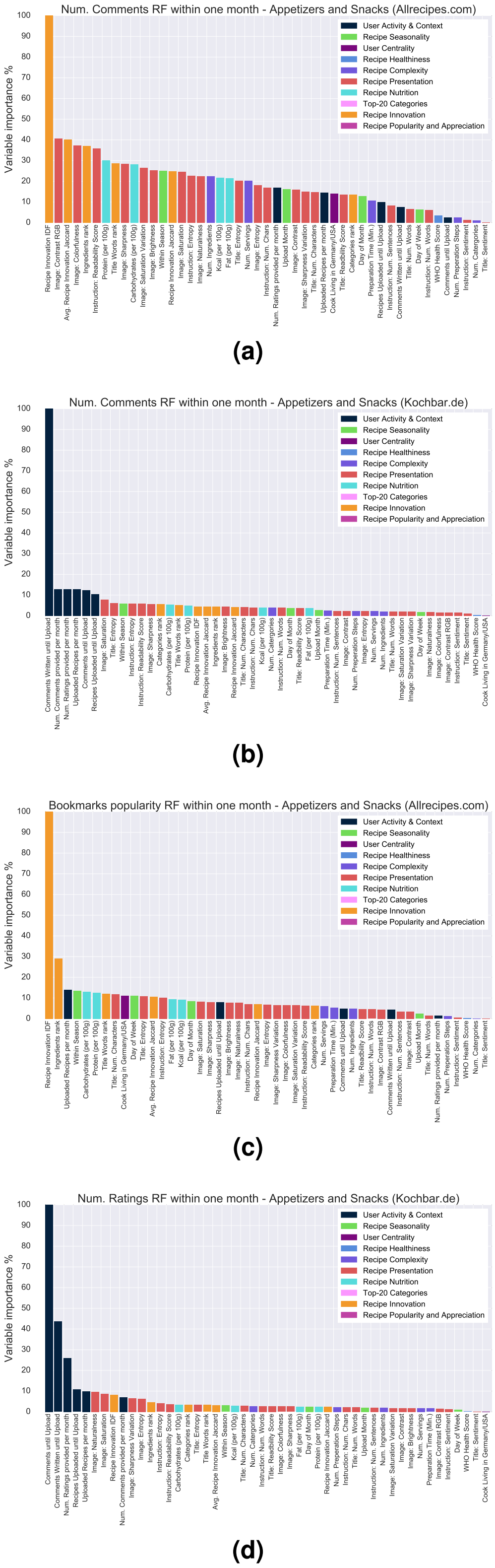
(a)

(b)

(c)

(d)

Figure A.8: Variable importance of within one week predictions (Appetizers and Snacks).



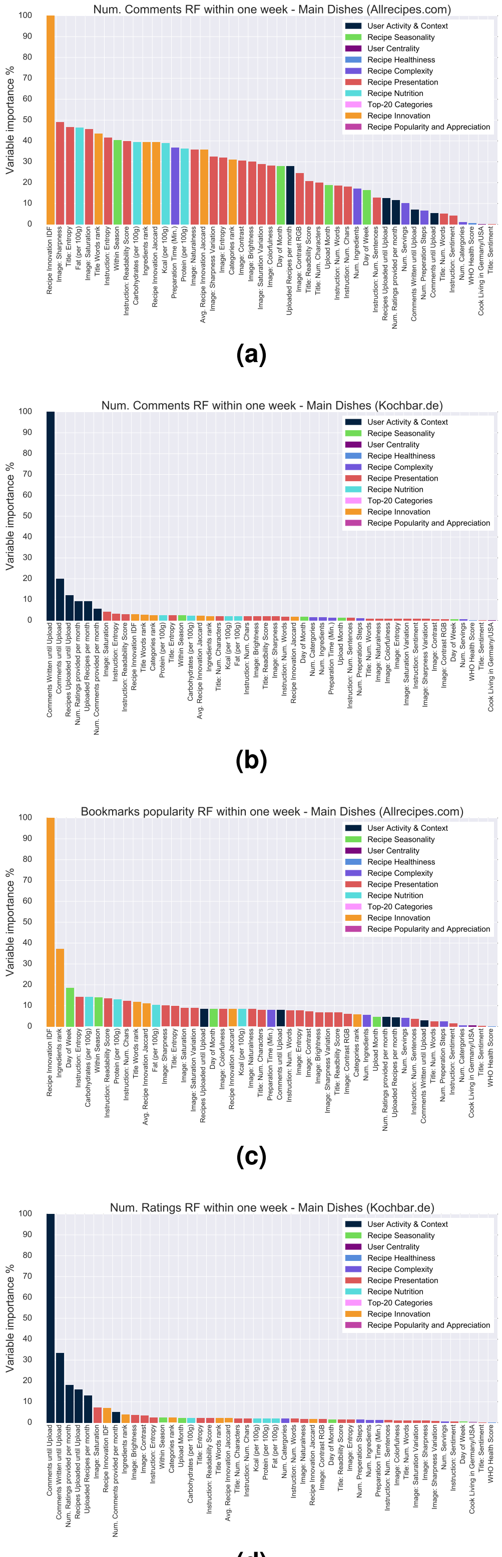
(a)

(b)

(c)

(d)

Figure A.9: Variable importance of within one month predictions (Appetizers and Snacks).



Num. Comments RF within one week - Main Dishes (Allrecipes.com)

(a)

Num. Comments RF within one week - Main Dishes (Kochbar.de)

(b)

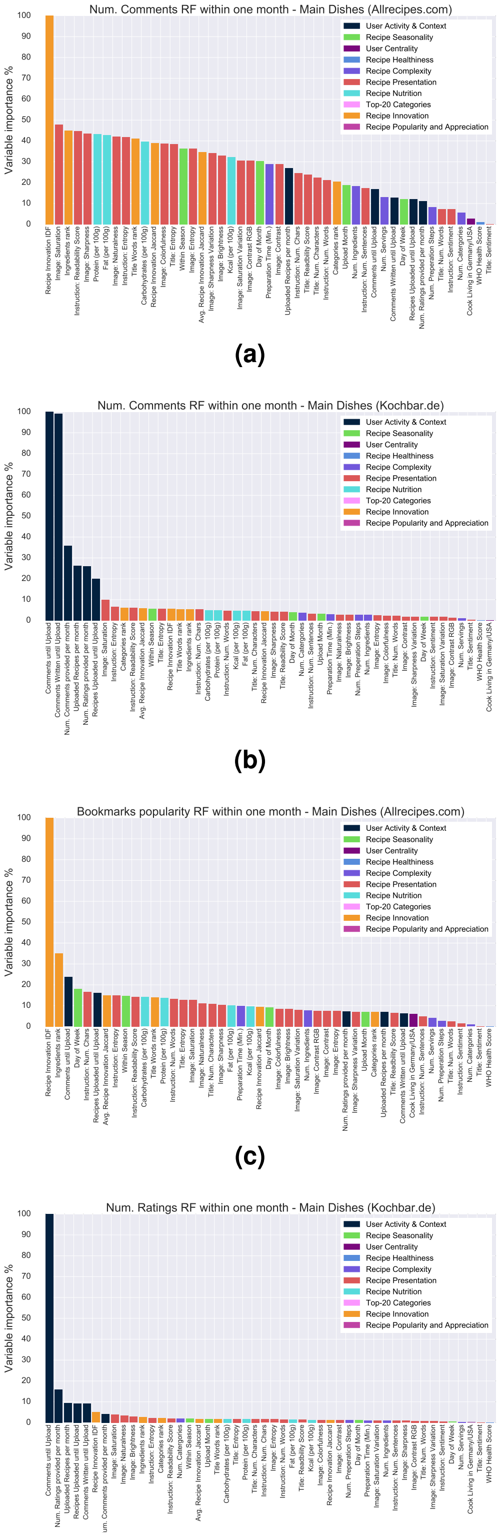
Bookmarks popularity RF within one week - Main Dishes (Allrecipes.com)

(c)

Num. Ratings RF within one week - Main Dishes (Kochbar.de)

(d)

Figure A.10: Variable importance of within one week predictions (Main Dishes).



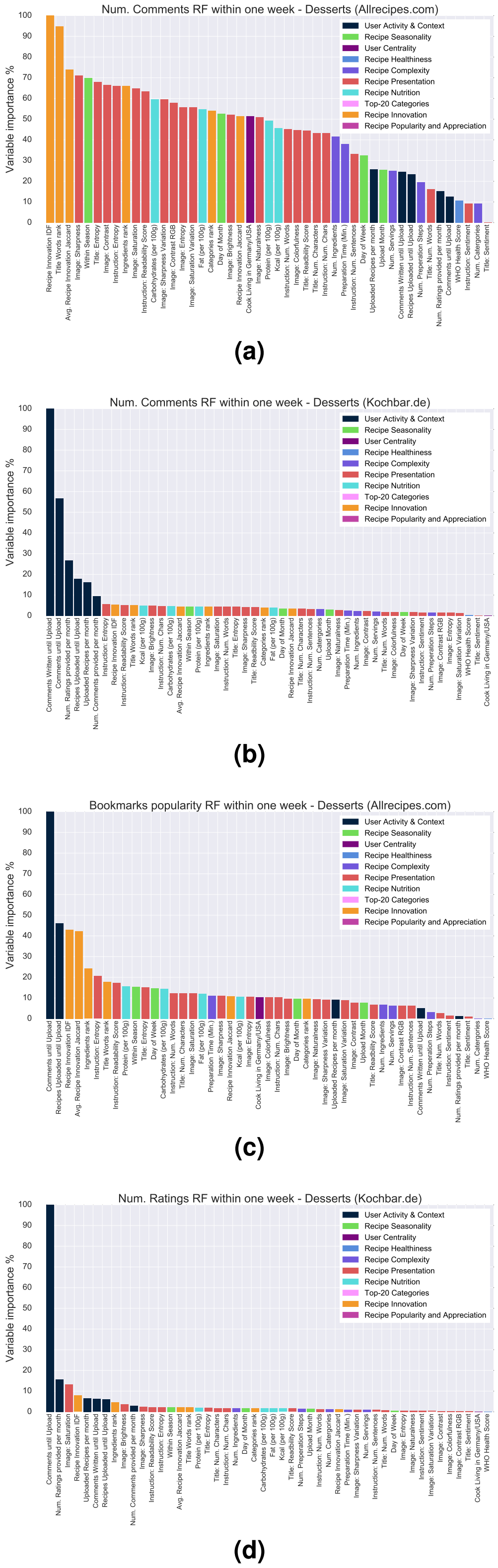
(a)

(b)

(c)

(d)

Figure A.11: Variable importance of within one month predictions (Main Dishes).



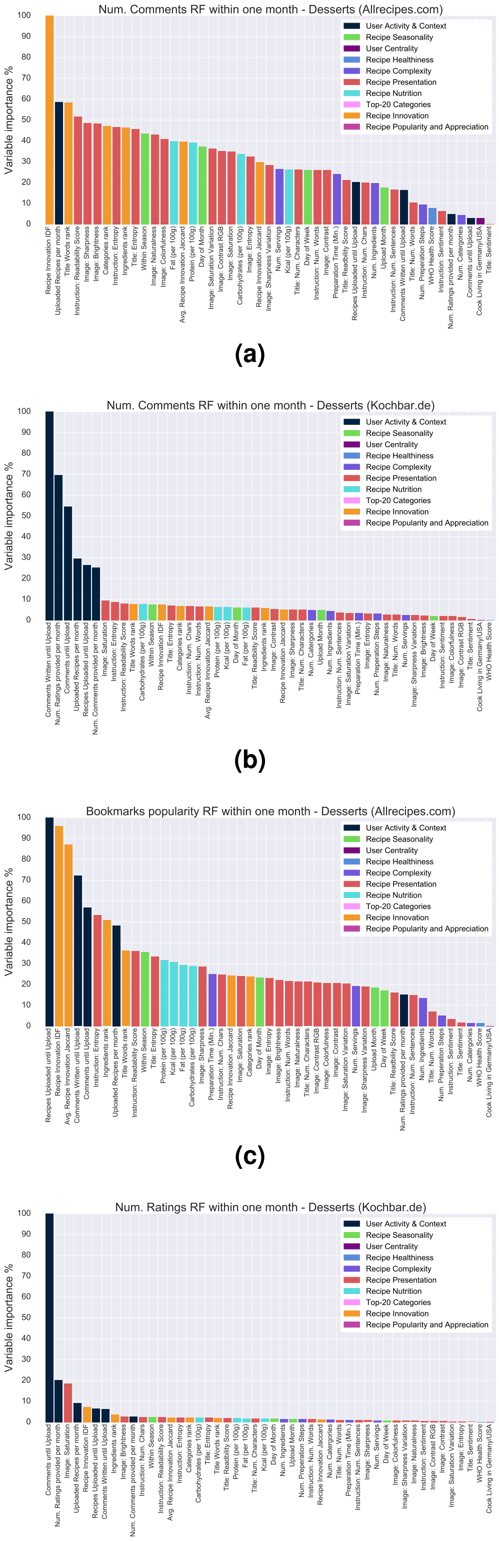
(a)

(b)

(c)

(d)

Figure A.12: Variable importance of within one week predictions (Desserts).



(a)

(b)

(c)

(d)

Figure A.13: Variable importance of within one month predictions (Desserts).

Bibliography

Abbar, Sofiane, Yelena Mejova, and Ingmar Weber [2015]. "You tweet what you eat: Studying food consumption through twitter". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM. 2015, pages 3197–3206 (cited on pages 17–18).

Ahn, Yong-Yeol, Sebastian E Ahnert, James P Bagrow, and Albert-László Barabási [2011]. "Flavor network and the principles of food pairing". *Scientific reports* 1 [2011] (cited on pages 29–30).

Anderson, J [1981a]. "Analysing the Readability of English and Non-English Texts in the Classroom with Lix Paper presented at the Australian Reading Association Conference". *Darwin, August (ED 207 022)* [1981] (cited on page 51).

Anderson, Jonathan [1981b]. "Analysing the Readability of,English and Hon-English Texts in the Classroom with Lix." [Aug 1981] (cited on page 46).

Bakshy, Eytan, Jake M Hofman, Winter A Mason, and Duncan J Watts [2011]. "Everyone's an influencer: quantifying influence on twitter". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pages 65–74 (cited on page 57).

Bandari, Roja, Sitaram Asur, and Bernardo A Huberman [2012]. "The pulse of news in social media: Forecasting popularity." *ICWSM* 12 [2012], pages 26–33 (cited on pages 9–11, 43).

Berrios, Raul, Peter Totterdell, and Stephen Kellett [2015]. "Eliciting mixed emotions: a meta-analysis comparing models, types, and measures". *Frontiers in psychology* 6 [2015] (cited on page 52).

Brandes, Ulrik [2001]. "A faster algorithm for betweenness centrality". *Journal of mathematical sociology* 25.2 [2001], pages 163–177 (cited on pages 99–100).

Brin, Sergey and Lawrence Page [2012]. "Reprint of: The anatomy of a large-scale hypertextual web search engine". *Computer networks* 56.18 [2012], pages 3825–3833 (cited on page 101).

Chunara, Rumi, Lindsay Bouton, John W Ayers, and John S Brownstein [2013]. "Assessing the online social environment for surveillance of obesity prevalence". *PloS one* 8.4 [2013], e61373 (cited on pages 21–22).

Cooper, Jenny [2015]. *Cooking Trends Among Millennials: Welcome to the Digital Kitchen*. Jun 2015. https://www.thinkwithgoogle.com/consumer-insights/cooking-trends-among-millennials/ (visited on 11/07/2017) (cited on page 16).

Cornell University. *Image Entropy*. http://www.astro.cornell.edu/research/projects/compression/entropy.html (visited on 26/05/2017) (cited on page 50).

Culotta, Aron [2013]. "Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages". *Language resources and evaluation* [2013], pages 1–22 (cited on page 18).

De Choudhury, Munmun, Sanket Sharma, and Emre Kiciman [2016]. "Characterizing dietary choices, nutrition, and language in food deserts via social media". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM. 2016, pages 1157–1170 (cited on pages 17, 19–20).

Elsweiler, David, Christoph Trattner, and Morgan Harvey [2017]. "Exploiting Food Choice Biases for Healthier Recipe Recommendation". In: *In Proceedings of ACM SIGIR Conference, Tokyo, Japan*. SIGIR'17. 2017, page 11 (cited on pages 14, 16, 43, 46).

Erica Greene [2015]. *Extracting Structured Data From Recipes Using Conditional Random Fields*. 25 Sep 2015. https://open.blogs.nytimes.com/2015/04/09/extracting-structured-data-from-recipes-using-conditional-random-fields/?_r=1 (visited on 01/06/2016) (cited on page 42).

Famaey, Jeroen, Frédéric Iterbeke, Tim Wauters, and Filip De Turck [2013]. "Towards a predictive cache replacement strategy for multimedia content". *Journal of Network and Computer Applications* 36.1 [2013], pages 219–227 (cited on page 6).

Freeman, Linton C [1978]. "Centrality in social networks conceptual clarification". *Social networks* 1.3 [1978], pages 215–239 (cited on pages 97, 99–100).

Fried, Daniel, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, and Dane Bell [2014]. "Analyzing the language of food on social media". In: *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE. 2014, pages 778–783 (cited on pages 18–19).

Gould, Peter R [1967]. "On the geographical interpretation of eigenvalues". *Transactions of the Institute of British Geographers* [1967], pages 53–86 (cited on page 101).

Hong, Liangjie, Ovidiu Dan, and Brian D Davison [2011]. "Predicting popular messages in twitter". In: *Proceedings of the 20th international conference companion on World wide web*. ACM. 2011, pages 57–58 (cited on page 11).

Howard, Simon, Jean Adams, and Martin White [2012]. "Nutritional content of supermarket ready meals and recipes by television chefs in the United Kingdom: cross sectional study". *BMJ* 345 [2012], e7607 (cited on pages 14, 46).

Huang, Kai-Qi, Qiao Wang, and Zhen-Yang Wu [2006]. “Natural color image enhancement and evaluation algorithm based on human visual system”. *Computer Vision and Image Understanding* 103.1 [2006], pages 52–63 (cited on page 48).

International Commission on Illumination. *17-1136 saturation*. http://eilv.cie.co.at/term/1136 (visited on 26/05/2017) (cited on page 47).

International Commission on Illumination. *17-233 colourfulness*. http://eilv.cie.co.at/term/233 (visited on 26/05/2017) (cited on page 48).

International Commission on Illumination. *17-251 contrast colourfulness*. http://eilv.cie.co.at/term/251 (visited on 26/05/2017) (cited on page 49).

James L Hargrove [2007]. *Does the history of food energy units suggest a solution to Calorie confusion?* 20 Aug 2007. https://nutritionj.biomedcentral.com/articles/10.1186/1475-2891-6-44 (visited on 22/05/2017) (cited on page 45).

Kerne, Andruid, Andrew M Webb, Steven M Smith, Rhema Linder, Nic Lupfer, Yin Qu, Jon Moeller, and Sashikanth Damaraju [2014]. “Using metrics of curation to evaluate information-based ideation”. *ACM Transactions on Computer-Human Interaction (ToCHI)* 21.3 [2014], page 14 (cited on pages 54–55).

Khosla, Aditya, Atish Das Sarma, and Raffay Hamid [2014]. “What makes an image popular?” In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pages 867–876 (cited on page 58).

Kucuktunc, Onur, B Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu [2012]. “A large-scale sentiment analysis for Yahoo! answers”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pages 633–642 (cited on page 52).

Kusmierczyk, Tomasz and Kjetil Nørvåg [2016]. “Online Food Recipe Title Semantics: Combining Nutrient Facts and Topics”. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM. 2016, pages 2013–2016 (cited on pages 28–29).

Kusmierczyk, Tomasz, Christoph Trattner, and Kjetil Nørvåg [2015a]. “Temporal patterns in online food innovation”. In: *Proceedings of the 24th International Conference on World Wide Web*. ACM. 2015, pages 1345–1350 (cited on pages 25–26, 34, 36, 42, 54–55).

Kusmierczyk, Tomasz, Christoph Trattner, and Kjetil Nørvåg [2015b]. “Temporality in online food recipe consumption and production”. In: *Proceedings of the 24th International Conference on World Wide Web*. ACM. 2015, pages 55–56 (cited on pages 44, 46, 52–53).

MathWorks. *Entropy of grayscale image*. https://de.mathworks.com/help/images/ref/entropy.html (visited on 26/05/2017) (cited on page 50).

Medical Definition of Carbohydrates [2016]. 13 May 2016. http://www.medicinenet.com/script/main/art.asp?articlekey=15381 (visited on 22/05/2017) (cited on page 45).

Medical Definition of Fat [2016]. 13 May 2016. http://www.medicinenet.com/script/main/art.asp?articlekey=3394 (visited on 22/05/2017) (cited on page 45).

Medical Definition of Proteins [2016]. 13 May 2016. http://www.medicinenet.com/script/main/art.asp?articlekey=15380 (visited on 22/05/2017) (cited on page 45).

Mike Chaney. *BRIGHTNESS, CONTRAST, SATURATION, AND SHARPNESS*. http://www.steves-digicams.com/knowledge-center/brightness-contrast-saturation-and-sharpness.html#b (visited on 22/08/2017) (cited on page 47).

Müller, Manuel, Morgan Harvey, David Elsweiler, and Stefanie Mika [2012]. "Ingredient matching to determine the nutritional properties of internet-sourced recipes". In: *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on*. IEEE. 2012, pages 73–80 (cited on page 44).

Nutritional Information. http://dish.allrecipes.com/customer-service/nutrition-information/ (visited on 22/05/2017) (cited on page 44).

OpenIMAJ Java Library. *Class AvgBrightness*. http://openimaj.org/apidocs/src-html/org/openimaj/image/feature/global/AvgBrightness.html (visited on 26/05/2017) (cited on pages 47–49).

OpenIMAJ Java Library. *Class RGBRMSContrast*. http://openimaj.org/apidocs/org/openimaj/image/feature/global/RGBRMSContrast.html (visited on 26/05/2017) (cited on page 49).

OpenIMAJ Java Library. *Class RMSContrast*. http://openimaj.org/apidocs/org/openimaj/image/feature/global/RMSContrast.html (visited on 26/05/2017) (cited on page 49).

Parra, Denis, Christoph Trattner, Diego Gómez, Matías Hurtado, Xidao Wen, and Yu-Ru Lin [2016]. "Twitter in academic events: a study of temporal usage, communication, sentimental and topical patterns in 16 computer science conferences". *Computer Communications* 73 [2016], pages 301–314 (cited on pages 46, 52).

Photoreview.com. *Sharpness, Acutance and Resolution*. http://www.photoreview.com.au/tips/shooting/sharpness,-acutance-and-resolution (visited on 22/08/2017) (cited on page 50).

Rokicki, Markus, Eelco Herder, Tomasz Kuśmierczyk, and Christoph Trattner [2016]. "Plate and prejudice: Gender differences in online cooking". In: *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*. ACM. 2016, pages 207–215 (cited on pages 27–28).

Rokicki, Markus, Eelco Herder, and Christoph Trattner [2017]. "How Editorial, Temporal and Social Biases Affect Online Food Popularity and Appreciation." In: *ICWSM*. 2017, pages 192–200 (cited on pages 2, 12–14, 44, 61).

Said, Alan and Alejandro Bellogìn [2014]. “You are What You Eat! Tracking Health Through Recipe Interactions.” In: *RSWeb@ RecSys*. 2014 (cited on pages 30–31).

San Pedro, Jose and Stefan Siersdorfer [2009]. “Ranking and Classifying Attractiveness of Photos in Folksonomies”. In: *Proceedings of the 18th International Conference on World Wide Web*. WWW ’09. Madrid, Spain: ACM, 2009, pages 771–780 (cited on pages 46–50).

Savakis, Andreas E, Stephen P Etz, and Alexander CP Loui [2000]. “Evaluation of image appeal in consumer photography”. In: *Electronic Imaging*. International Society for Optics and Photonics. 2000, pages 111–120 (cited on page 46).

Scheibehenne, Benjamin, Linda Miesler, and Peter M Todd [2007]. “Fast and frugal food choices: Uncovering individual decision heuristics”. *Appetite* 49.3 [2007], pages 578–589 (cited on pages 2, 43).

Shulman, Benjamin, Amit Sharma, and Dan Cosley [2016]. “Predictability of Popularity: Gaps between Prediction and Understanding.” In: *ICWSM*. 2016, pages 348–357 (cited on pages 61, 77).

Tatar, Alexandru, Marcelo Dias de Amorim, Serge Fdida, and Panayotis Antoniadis [2014]. “A survey on predicting the popularity of web content”. *Journal of Internet Services and Applications* 5.1 [2014], page 8 (cited on pages 5–7, 85).

Trattner, Christoph and David Elsweiler [2017]. “Investigating the healthiness of internet-sourced recipes: implications for meal planning and recommender systems”. In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2017, pages 489–498 (cited on pages 1, 44–46).

Trattner, Christoph, David Elsweiler, and Simon Howard [2017]. “Estimating the healthiness of internet recipes: a cross-sectional study”. *Frontiers in public health* 5 [2017] (cited on pages 24–25, 33).

Trattner, Christoph, Denis Parra, and David Elsweiler [2017]. “Monitoring obesity prevalence in the United States through bookmarking activities in online food portals”. *PloS one* 12.6 [2017], e0179144 (cited on page 1).

Tsagkias, Manos, Wouter Weerkamp, and Maarten De Rijke [2009]. “Predicting the volume of comments on online news stories”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pages 1765–1768 (cited on pages 8–11).

Tutorialspoint. *Compiler Design - Lexical Analysis*. `https://www.tutorialspoint.com/compiler_design/compiler_design_lexical_analysis.htm` (visited on 22/08/2017) (cited on page 50).

Wagner, Claudia and Luca Maria Aiello [2015]. “Men eat on Mars, Women on Venus?: An Empirical Study of Food-Images”. In: *Proceedings of the ACM Web Science Conference*. ACM. 2015, page 63 (cited on pages 20–21).

Wagner, Claudia, Philipp Singer, and Markus Strohmaier [2014]. “The nature and evolution of online food preferences”. *EPJ Data Science* 3.1 [2014], page 38 (cited on pages 23–24).

West, Robert, Ryen W White, and Eric Horvitz [2013]. “From cookies to cooks: Insights on dietary patterns via analysis of web usage logs”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pages 1399–1410 (cited on pages 17, 52).

Who, Joint and FAO Expert Consultation [2003]. “Diet, nutrition and the prevention of chronic diseases”. *World Health Organ Tech Rep Ser* 916.i-viii [2003] (cited on page 45).