FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Twitter usage in Austria

A behavioral analysis about the Twitter usage in Austria and the surrounding border areas

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Armin Müllner, BSc
Matrikelnummer 0728113

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao.Univ.-Prof. Mag. Dr. Wolfdieter Merkl

Wien, 17. April 2015

(Unterschrift Verfasser) (Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Twitter usage in Austria

A behavioral analysis about the Twitter usage in Austria and the surrounding border areas

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Armin Müllner, BSc
Registration Number 0728113

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao.Univ.-Prof. Mag. Dr. Wolfdieter Merkl

Vienna, 17. April 2015

(Signature of Author) (Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Armin Müllner, BSc
Feldgasse 41, 7552 Stinatz

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum)	(Unterschrift Verfasser)

Danksagung

Hiermit möchte ich mich bei dem Betreuer meiner Masterarbeit, Dieter Merkl, recht herzlich für die Unterstützung bedanken. Speziell die Diskussionen haben mir des Öfteren interessante Aspekte aufgezeigt, die ich in die Arbeit einarbeiten konnte. Auch möchte ich mich für die Geduld bedanken, die es erforderte, mich bei dieser Arbeit zu betreuen.

Des weiteren möchte ich mich bei meiner Familie, speziell meinen Eltern bedanken, die mir ermöglicht haben, mein Studium erfolgreich zu absolvieren. Ein Danke auch an meinen Bruder, der in Zeiten niedriger Motivation des Öfteren motivierend zur Seite stand.

Abschließend bedanke ich mich bei meinen Freunden und Studienkollegen, die die Studienzeit zu einem Erlebnis machten und das eine oder andere Mal auch helfend zur Stelle waren.

Abstract

Social Media is a phenomenon which has increasing user numbers. One interesting platform to investigate is Twitter, as the usage is not that high as the usage of other platforms. But as the numbers are rising in the past, it is interesting how this platform is used within a specified area like Austria.
Although Twitter is a microblogging platform there are only some characteristics of normal blogging valid for Twitter. Moreover the usage of Twitter is a combination of the usage of classic social networks and blogs.
In my thesis I crawled Twitter data from Austria and the surrounding border areas within two timeframes (April 2014, June 2014). To define trends and sentiments for the dataset I used several text mining methods. To get insights into the data, I performed several analyses:
First, I performed an overall analysis on the dataset. I found out, that tweets are clustered around big cities. Another observation is, that most tweets are published in English or German, but overall many different languages are used to publish a tweet. The platforms used for publishing a tweet do not represent the real life mobile market share. Because of these two observations Twitter data is not representative for real world populations as the people using Twitter are a special user group and no reflection of the real world population.
I also analyzed sentiments and trends for the Twitter data. The overall sentiment is positive, but as the sentiment algorithm more likely classifies posts as positive, and the positive sentiment value is quite low, this observation is a little bit imprecise. The trends for the dataset mostly contain location terms but also represent real life events.
To perform regional analysis the data was clustered into eight regions. The regions were defined as quadrants, each of equal size. When defining more smaller regions, there would be some of them with too less data. Each region then was analyzed equally as the overall dataset. One observation of the regional analysis is, that trends and sentiment differ in each region. Only a few trends are present in more regions, but also there no regularity can be observed. The results proof that the usage of Twitter is highly dependent on the region where the tweets are posted.
In the end of the work I conclude the results and summarize the results from this work.

Kurzfassung

Die Nutzung von Social Media steigt stetig an. Eine Plattform, die in der Vergangenheit große Wachstumszahlen hatte, ist Twitter. Speziell, wenn man die Nutzungszahlen mit anderen Plattformen vergleicht, wird dieser Anstieg deutlich. In Österreich gibt es bei den Twitter-Nutzerzahlen noch Wachstumspotenzial. Deshalb ist es sehr interessant, zu analysieren, wie Twitter in Österreich genutzt wird.
Twitter ist eine Microblogging Plattform, deshalb sind die Anwendungsbereiche ein wenig anders als bei klassischem Blogging. Die Nutzung von Twitter ist eher eine Kombination aus der Nutzung eines klassischen Sozialen Netzwerks und eines Blogs.
In meiner Diplomarbeit analysiere ich die Nutzung von Twitter in Österreich. Dafür habe ich Twitter Daten aus Österreich und dem nahen Ausland heruntergeladen. Dies passierte in zwei unterschiedlichen Zeiträumen (April 2014, Juni 2014). Um aus diesen Daten Trends und Stimmungen zu extrahieren, habe ich unterschiedliche Text Mining Algorithmen benutzt. Die Ergebnisse wurden dann in weiterer Folge analysiert.
Als Erstes wurden die gesamten Daten analysiert. Hier wurde ersichtlich, dass Tweets um größere Städte gruppiert sind. Auch konnte ich feststellen, dass die meisten Tweets zwar in Deutsch oder Englisch veröffentlicht wurden, jedoch werden sehr viele weitere Sprachen für das Posten verwendet. Die Plattformen, die benutzt werden um zu twittern, unterscheiden sich sehr stark von der Verteilung der mobilen Plattformen unter Personen. Zum Beispiel werden mehr als die Hälfte aller Tweets mit einem iOS-Client veröffentlicht. Aufgrund dieser Tatsachen sind Twitter-Daten nicht repräsentativ für Gesamtbevölkerungen, weil der Benutzerkreis von Twitter ein eingeschränkter ist.
In weiterer Folge wurden auch die Stimmungen und Trends näher analysiert. Generell ist die Stimmung der Posts eher positiv. Dies liegt jedoch daran, dass der Algorithmus zur Definition der Stimmung Tweets eher positiv bewertet. Deshalb ist dieses Ergebnis ein wenig ungenau. Die Trends innerhalb der Twitter-Daten bestehen zum Großteil aus Orten und Plätzen, es sind aber auch Trends ersichtlich, die reale Vorfälle beschreiben.
Um die regionalen Unterschiede zu analysieren wurden die Daten in acht Regionen geteilt. Dafür wurde ein quadratischer Raster angewendet, der acht gleich große Quadranten definiert. Die Anzahl wurde so gewählt, weil bei einer größeren Anzahl von Quadranten einige zu wenig Daten für die Analyse beinhaltet hätten. Jede Region wurde dann gleich analysiert wie die Gesamtdaten. Eine Beobachtung hier ist, dass Trends und Stimmungen in jeder Region variieren. Einige Trends sind zwar in mehr Regionen ersichtlich. Hier kann jedoch keine Regelmäßigkeit erkannt werden. Diese Ergebnisse zeigen, dass die Nutzung von Twitter sehr stark von der Region abhängt und in jeder Region andere Themen relevant sind.

Das Ende dieser Arbeit bildet ein Fazit dieser Arbeit und eine Zusammenfassung der Ergebnisse.

Contents

CHAPTER 1

Introduction

In the introduction chapter a brief overview of this work is given. First the motivation 1.1 of this work in the context of the identified problem statement 1.2 is given. The aim of the work is stated in the next section 1.3. The methodical approach which the thesis is based on is described afterwards 1.4. In the end of this chapter the structure of the work is given 1.5.

1.1 Motivation

With the emerging of the Web 2.0 some years ago the information overload in the Internet also got more important in todays life. The most common definition of Web 2.0 was done by O'Reilly:

> *Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness network effects to get better the more people use them.* O'Reilly [O'R06]

One appliance of Web 2.0 are social networks. When looking at the Social Networks Twitter[1] and Facebook[2] they are widely used in todays world. Based on studies published on the platform Statista[3] Facebook had 143 million users in the US. Compared to the year 2008 this is an increase of 225 percent. Twitter had in 2012 24 million users which are 605 percent more than in 2008. When looking at the ratio of the users of the platforms for every Twitter user there are 6 Facebook users active.

When looking on the statistics of Facebook usage specifically for Austria[4] there were 2,85 million users in September 2012 and 3,24 million users in July 2014. This is an increase of 13,5

[1] `www.twitter.com` accessed July 21, 2014

[2] `www.facebook.com` accessed July 21, 2014

[3] `http://de.statista.com/statistik/daten/studie/224771/umfrage/anzahl-der-nutzer-von-social-media-in-den-usa-nach-plattform/` accessed July 21, 2014

[4] `http://socialmediaradar.at/` accessed July 21, 2014

percent. Twitter had 95 120 users in September 2012 and 121 476 users in July 2014 which is an increase of 27,7 percent. In Austria there are 27 active Facebook users for every Twitter user. From the Twitter accounts in Austria only about half of them are really active, writing or reading, in the network.

Because of these numbers one can see that Twitter is growing much faster than Facebook. This applies to a very mature market of social media usage as well as to a more emerging market like Austria. In Austria there is also a big increase in the number of Twitter users and there is still room for improvement when looking on the relative comparison of the users for the two networks from the US and Austria. Because of the big increase in usage and the high number of not active users Twitter will be an increasingly important platform for the future in Austria. This is, because Twitter is not yet heavily used but has very big growth numbers compared to the most used platform, Facebook.

Another aspect which is important for the analysis of social media platforms is the openess of the platforms. Although nearly all platforms offer an API for extracting data the amount of data which can be extracted is more limited. On Facebook most of the data is private and therefore not returned by the API. As opposite to this most of the posts on Twitter are public available and therefore returned by the API. Therefore extracting data from Twitter is more representative than extracting data from Facebook as for Facebook only a small sample of data could be extracted. As this work analyzes the usage of Twitter in Austria it is important to analyze a broad dataset of all the information that is posted on a platform.

1.2 Problem Statement

As the usage of Twitter can be very versatile it is very interesting to get a deeper insight into the current usage of this network in Austria. Kwak et al. [KLPM10] covered in their work the question if Twitter is like classical social networks or even comparable to a news media. The result was, that Twitter is also frequently used for posting news and in some cases Twitter is also posting news before traditional media does. Also analytics like the clients or the language used for tweeting are very interesting.

Another very interesting aspect on the Twitter usage are regional differences in the usage behavior. On the one hand it is very interesting if there is a difference in the topics of the tweets coming from agglomerations compared to tweets coming from the countryside. But it is also interesting if there are regional trends legible in a dataset of Austrian tweets. This not only applies to trends of the tweets but also for the sentiment.

There is already some research on the general usage of Twitter conducted, like the work from Huberman et al [HRW08], Java et al [JSFT07] and Mislove et al [MLA+11]. Nevertheless it would be very interesting to see results specific for Austria.

1.3 Aim of the Work

The purpose of this thesis is to conduct an extensive analysis of the Twitter usage in Austria following different aspects. The thesis consists of a theoretical and a practical part.

In the theoretical part extensive definitions of social media are given and illuminated. Also Twitter as one appliance of social media is discussed. For the analysis of the tweets from Austria text mining APIs are used to extract trends and sentiment. Therefore also text mining is discussed in the theoretical part of this thesis.

The practical part of the thesis focuses on the extraction and analysis of the tweets from Austria. The implementation contains on the one hand a component which crawls tweets through the Twitter APIs and stores the data in a database. The other part of the implementation uses text mining algorithms to extract trends and sentiment from the tweets to make the data better analyzable. In the analysis part of the work the tweets which were crawled are analyzed in respect to various aspects. Demographic analytics for the tweets are part of the work as well as usage data like the source (platform) the tweet was made with or the language of the post. Also the trends and sentiment extracted are compared and validated. With that results also an evaluation of the used text mining algorithms can be conducted.

1.4 Methodical Approach

The methodical approach of this work is split into 3 different parts:

1. **Literature analysis:** An extensive state of the art analysis concerning social media and it's usage with a special respect to Twitter is conducted.
2. **Experiment:** In the experiment a crawler software was implemented to gain Twitter posts which are located in Austria and the border areas nearby. These posts are analyzed with different text analysis algorithms. The results of the text mining algorithms are validated.
3. **Result analysis:** The results which were generated with the prototype are evaluated and compared. With these results conclusions about the Twitter usage in Austria can be drawn.

1.5 Structure of the Work

This thesis consists of 4 following chapters. The next chapter does a state of the art analysis concerning the field of social media usage. In chapter 3 the practical implementation is explained. Chapter 4 contains the results which were gained with the prototype. In the last chapter the conclusions and potential future work are discussed.

Chapter 2 - State of the art: This chapter contains an extensive state of the art analysis. First the concept of social media and social networks is described. Also Twitter as one specific appliance of a social network is discussed. Because the practical part contains also an analysis based on text mining algorithms also text mining is discussed in this chapter.

Chapter 3 - Twitter Analyzer: In the third chapter the practical implementation is outlined. This is split into two parts. The first component of the algorithm is the data acquisition algorithm which crawls Twitter for posts from Austria and stores them in a database. The second part of the implementation is the integration of the text mining algorithms to extract trends, keywords

and sentiment from the tweets.

Chapter 4 - Results and Evaluation of the Twitter Analyzer: This main part of the work analyzes the data which was acquired by the crawler. First it contains of general analytics of the data set like the clients the tweets were made of, the languages used and also demographic analytics of the publishers. It also contains usage analytics of the Twitter usage in Austria. Also an analysis of the text mining algorithm is conducted and the methods are evaluated and validated.

Chapter 5 - Conclusions and future work: The last chapter of the thesis draws the conclusions of the results gained through this work and also gives an outlook on potential future work in this field.

CHAPTER 2

State of the art

In this chapter the theoretical background of this master thesis is discussed. In the section 2.1 the concept of social networks itself is discussed starting with the emerging of the Internet and the Web 2.0. The next section 2.2 discusses the social network Twitter in detail. In the section 2.3 the keyword and sentiment analysis as well as text mining algorithms are discussed. This chapter is concluded with a short summary.

2.1 Social Media

What is Social Media?

In todays world Social Media is always seen in a close relationship with the Internet. When looking back into the past the concept of Social Media was introduced a long time before computers evolved. In the figure *Social Media Timeline* 2.1 the first appearance of Social Media was the postal service which was introduced 550 BC. Also technological improvements like the telephone and the radio can be seen as early stages of Social Media. Especially the telephone is an example, where technology improved the possibilities to communicate.

With the upcoming trend of the Internet and the email also the speed of innovation in the domain of Social Media enhanced. When the first Usenet systems where created in 1979 it was a big step towards Social Media networks we know today. Usenets gave users the possibility to post articles to specific newsgroups. At the same time also the first board systems where started which were a step towards typical boards we are using nowadays. In the 80s the next new trend came up - the IRCs. These early chat clients gave users the possibility to keep in touch via text messages but it was also possible to start file exchanges with the systems. The most popular appliance of an IRC system was ICQ which was developed in the mid-90s. All these platforms already had several characteristics which apply to social networks.

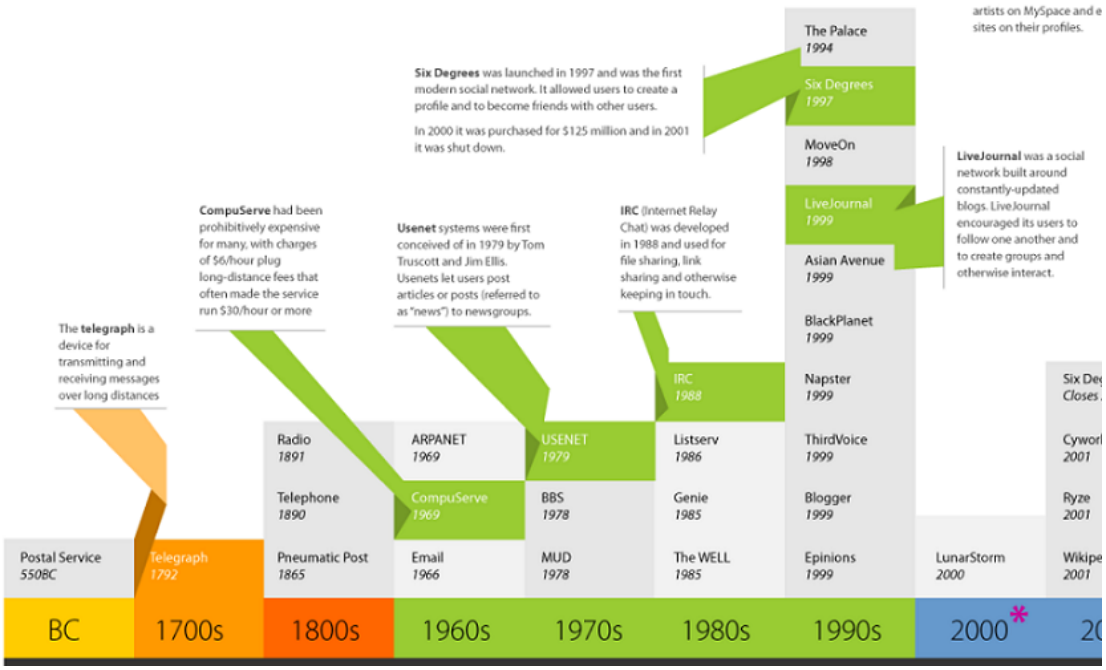

Figure 2.1: Social Media Timeline 1 [Gil10]

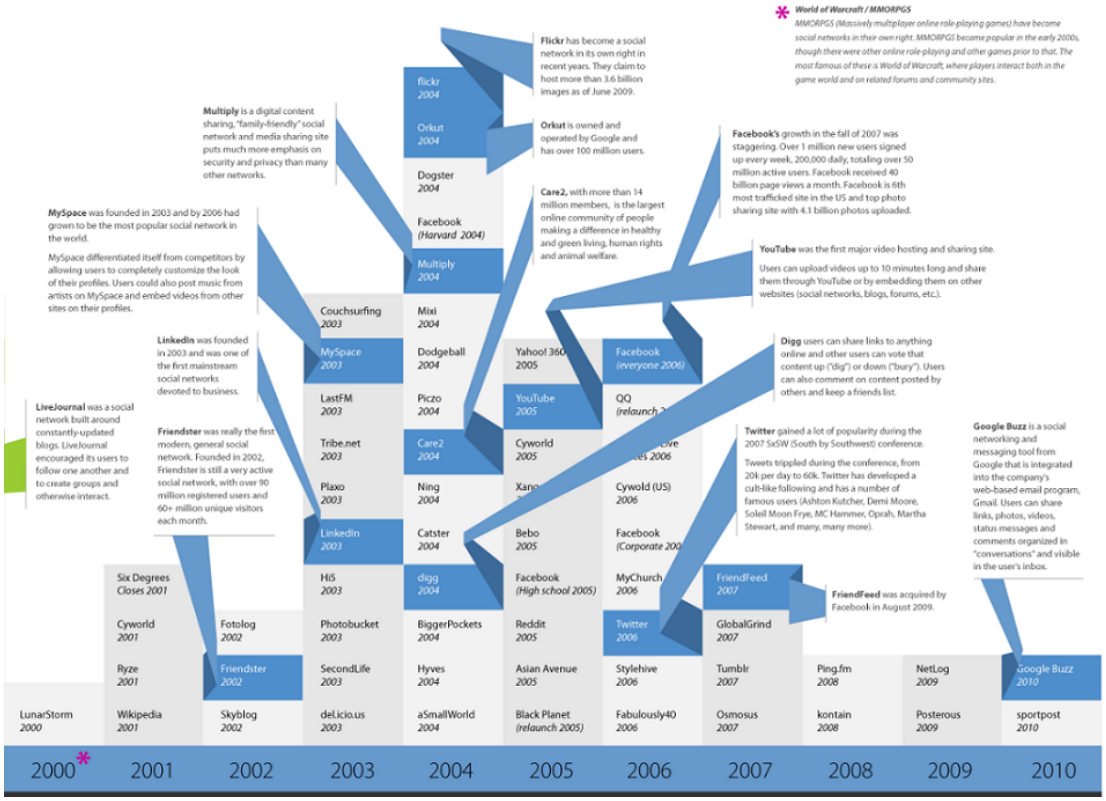

Figure 2.2: Social Media Timeline 2 [Gil10]

Web 2.0 and Social Media

With the emerging of the Web 2.0 some years ago the amount of content in the Internet also emerged, as it got easier to publish content. The most common definition of Web 2.0 was done by O'Reilly:

> *Web 2.0 is the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness network effects to get better the more people use them.* O'Reilly [O'R06]

Within this definition the network effect already is mentioned. Also the improvement of applications relative to the amount of users is mentioned. This network effect is the basic principle behind Social Media. The more people are using a social network, the more people will see a benefit to also use the network. This high impact of the user contribution is nothing new as the Web 1.0 was purely based on the power of user contribution and collective intelligence [O'R06]. But with the upcoming of the Web 2.0 after the dot-com bubble these networking aspects got an even higher importance than ever before.

O'Reilly also defines 5 rules which should apply to Web 2.0 appliances [O'R06]:

1. *Don't treat software as an artifact, but as a process of engagement with your users.* This leads to the fact that for every Web 2.0 appliance the engagement with it's users is crucial.

2. *Open your data and services for re-use by others, and re-use the data and services of others whenever possible*. Also this rule is very valid for Social Media as the re-use of content is very important but also the connection between several other sources and networks is often integrated in social networks.

3. *Don't think of applications that reside on either client or server, but build applications that reside in the space between devices.* In todays world with the multiplicity of devices the boundaries between server and client are not relevant any more.

4. *Remember that in a network environment, open APIs and standard protocols win, but this doesn't matter that the idea of competitive advantage goes away.* When looking at todays Social Media platforms every platform offers an open API but also most of the platforms have their competitive advantage.

5. *Chief among the future sources of lock in and competitive advantage will be data.* When looking on the last initial public offerings of social networks also this rule applies as in many cases the main asset of the networks is their data.

When looking at the Web 2.0 principles one can say, that Social Media is a typical Web 2.0 appliance. This trend came up in 2006. One year later, in 2007, Social Media already produced one third of the new Web content in the world. Finin et al. [FJK+08] defines the difference between traditional websites and Web 2.0 sites because they are connected to other networked data. This leads to an enrichment of the standard hyperlinks with social networks, comments and other typical characteristics of Social Media. Because of these changes Social Media has become the most important way to publish information and engage with others in the Web today. The usage numbers are still increasing like one can see in statistics like on Statista[1] where the numbers show a 225 percent increase of Facebook users and a 605 percent increase of Twitter users between 2008 and 2012. Because of all that facts it will be very important to learn more about how these systems work.

Finin et al. see the World Wide Web as a directed graph of Web pages with undifferentiated links between pages [FJK+08]. One very interesting extension of these links between pages are blogs as blogs are connected to even more pages than normal Web sites are. Blogs for instance can also be linked to other blogs and blogrolls. Blogs can also contain links to other Web content. The same complex structure also applies to comments to blog posts. So this leads to very complex network structures. Therefore blogs are a very interesting system to investigate how social interactions take place in Social Media. When it comes to this extensive networks an interesting aspect to look on is the influence of users. The determination of influence on Social Media is a very complex process. A basic number like e.g. the number of followers on Twitter is no single point of truth. A publisher can have a big number of followers although he has no influence as he is never retweeted. Also can a publisher with a quite small reach be

[1] `http://de.statista.com/statistik/daten/studie/224771/umfrage/anzahl-der-nutzer-von-social-media-in-den-usa-nach-plattform/` accessed July 21, 2014

very influential when he is retweeted by publishers with a bigger reach. Similar examples can be found for many other appliances of Social Media.

Another aspect when it comes to Social Media is the business purpose of it which is analyzed by Kaplan and Haenlein [KH10]. In essence one can say that customers today do an extensive research in the Web before getting into an interaction with a business. This also leads to the need for businesses to be present in Social Media. Social Media in that context is a subset of applications which are located in the Internet which can be classified as Web 2.0 applications and give users the opportunity to create and exchange user generated content. There are several opportunities of Social Media for businesses. The first one are collaborative projects where businesses can use user generated content for their business purposes. An example for that can be the Wikipedia page of a company which can also be edited by end-users or win raffles where users can upload images with a product which than can be used for marketing purpose. Also internal social collaboration can be crucial for businesses - a common idea platform to vote for new innovative ideas is only one example for such a network. A second appliance can be company blogs which can be used to inform customers about current news about specific products. Social networking sites for instance can be used to give customers a forum to comment on their experiences with a business. Also advertising in virtual social games and worlds can be a very powerful way to focus on specific target audiences. Kaplan and Haenlein [KH10] also define ten advices for companies for using Social Media. These in some extend are also valid for private users.

1. **Choose carefully:** It is very important to choose the right Social Media platform for the right purpose. Because the amount of social platforms is still emerging one cannot be active in every platform so it is very important to really figure out the purpose of ones Social Media usage ambition and then the right platform should be chosen.

2. **Pick the application, or make your own:** When the purpose of ones Social Media ambition is clearly defined it is important to make the decision whether to use an existing platform or build an own social network application.

3. **Ensure activity alignment:** It is also very important to have all Social Media activities aligned to have one holistic Social Media approach.

4. **Media plan integration:** For businesses it is also very important to integrate the Social Media activities into their traditional marketing approach.

5. **Access for all:** When a business starts with a social strategy it is also important that everyone can access the Social Media platforms.

6. **Be active:** When using Social Media it is very important also to be active on the social network to stay relevant.

7. **Be interesting:** Only being active is not enough to stay relevant - it is also very important to be interesting for the target audience which should be addressed by the Social Media efforts.

8. **Be humble:** It is very important to not act as the inventor of Social Media when communicating on it.

9. **Be unprofessional:** When using a Social Media platform it is important to be authentic and represent the real world.

10. **Be honest:** The most important aspect is to always be honest when using Social Media as the network will figure out which information is right or wrong.

User Generated Content

The concept which lay behind the World Wide Web was to enable normal users to generate content. The idea behind Usenet which was developed by Tom Truscott and Jim Ellis in 1979 was to enable other users to post their opinions on several topics. This platform was not the first appliance of Social Media but the most relevant. Also the term "weblog" was introduced at the same time. Many years later there are platforms present which enable users to generate and publish any content they want. There are networks for image, video and music sharing as well as blogs, virtual worlds and classical social networks to stay in contact. Each platform with the purpose to generate content.

Kaplan and Haenlein [KH10] define User Generated Content (UGC) as the sum of all ways in which people make use of Social Media. Usually the term is used for all form of content which is published in the Web by end-users. The Organization for Economic Co-Operation and Development [OEC07] defines three criteria which must apply to content to be classified as user generated content:

1. The content must be published on a public available website or a social network with a certain reach

2. The content must show a certain amount of creative effort

3. The content must be created outside of professional routines and practices

When looking at these conditions the first one excludes all posts which are directed to be only available to a very small amount of people. Examples therefore would be emails and instant messages. The second condition specifies that copying content is not a form of user generated content. The last condition states that content created with a commercial background does not classify as user generated content.

User generated content is one shape of self-presentation. Self-presentation was discussed by Goffman [Gof59] in 1959, in a time where Social Media did not exist at all. His thesis is that self-presentation is the intentional and tangible component of identity. It is also contextual and based on a specific setting. So one can say that the self-presentation is a very important part of the identity of a person. The self-presentation somehow is that part of the identity which is tangible first. Because this self-presentation is also dependent on a specific setting the behavior of users differ on each social platform they use. If a person is very good in clearly and precisely making a point but is on the other hand very uncreative when it comes to pictures, this person

will more likely have a better self-presentation on blogs or platforms like Twitter than on image platforms like Instagram or Flickr. It is also more likely that the person will be more relevant in the blogs than in the image communities.

Schau and Gilly [SG03] discovered, that users also use personal web pages to explore and display other selves than their real ones. This also is valid for the present Social Media platforms nowadays. On the one hand it is very easy to present only the positive sides of an identity in Social Media. As content on Social Media is available to a wide audience users will not be that open with negative sides of their identity because that can harm their self-presentation. On the other hand it is very easy to use Social Media as a playground for discovering new viewpoints of ones identity as the social web is very agile.

The OECD defined also several drivers of user generated content [OEC07]:

1. **Technological Drivers:** A very important driver for user generated content is the availability of broadband Internet nearly everywhere. With the rise of smart-phones this driver got even more important as everyone nowadays is always online and has the ability to create content on the fly. Another aspect is that technology made it much easier to produce content. When looking on the process of creating a personal website in the past it was a very complex process to do so. Today it is very easy to start with a first blog after minutes without the need of any specific technical knowledge. Lastly also the rise of several Social Media platforms supported the user generated content creation in several ways.

2. **Social Drivers:** Of course there are also several social drivers which increased the amount of user generated content. The shift to younger age groups, the so-called "digital natives", a generation which was born with the Social Media technologies, increased the content creation. Also the increasingly importance of collaboration within the daily life boosted Social Media. As the younger age groups adopted these new technologies very quick there is also the need for the older generation to also adopt these trends very quickly.

3. **Economic Drivers:** The technological drivers also have a very strong tie to the economic drivers. Nowadays it is very cheap to own a smart-phone or a notebook with Internet access and so there is no entry barrier to content creation. The cost decrease applies to hardware and software costs as well as to Internet costs. As it is also important for businesses to adopt Social Media there is also a commercial interest in user generated content. Lastly it is also a fact, that it is much easier to start with a Social Media platform as there are several new financing models like venture capital to boost startups to reach more and more users.

4. **Institutional and Legal Drivers:** New content licensing forms like the Creative Commons license[2] made it also easier to share and re-use content.

They also divided user generated content into different types like text, photos and images, music and audio, video and film and lastly content posted on products or other areas of interest.

[2] `https://creativecommons.org/licenses/` accessed 22.07.2014

Each of the types has different characteristics as well as there are different motivations for posting the different types. The text generation is a quite mature technique as it was possible since the time the first books started. The creation of images, audio and video got a real boost with the technological drivers for user generated content.

Within the same study the OECD also classified several different platforms for content generation [OEC07]:

1. **Blogs:** a blog is a webpage which presents several articles in a chronological order. These articles can be created by one or more publishers which are called bloggers. Typically it is also possible to comment on articles to start a conversation about the article. Normally one can also subscribe to blogs to easily get updates.

2. **Wikis:** a wiki allows users to add, remove and edit content about a topic of interest. Wikis also have a collaborative aspect as it is normal, that more users contribute to the same article.

3. **Group-based aggregation and social bookmarking:** these platforms allow to construct a link collection to articles and media. There it is also possible to vote for collections to make them more relevant.

4. **Podcasting:** Podcasts offer the possibility to subscribe to short audio clips which deal with a specific topic of interest.

5. **Social Networking Sites:** Social Networking Sites like Facebook allow users to get in contact with friends, post and comment on content, post instant messages and a lot more features. Somehow they can also be seen as a combination of other platforms to one more powerful platform.

6. **Virtual World Content:** Virtual World Content presents content in a 3D digital environment. The most famous appliance of these platforms was Second Life.

Schwartz HA, Eichstaedt JC, Kern ML, et al [SEK+13] found another interesting aspect concerning user generated content. Besides logical aspects like publishers who live in high elevations more likely talk about mountains they also found out that neurotic people more often use words like “sick of“ and “depressed“ in their posts which is also in tie with other research. They also found new aspects like the difference in using the possessive phrase “my“ by males and females when talking about their partner. In essence one can say that males use possessive language more often than females do.

Statistics about Social Media

When talking about Social Media it is also important to get a feeling about the numbers which lay behind these platforms and the usage of them. In a research conducted by PewResearch[3]

[3] `http://www.pewInternet.org/fact-sheets/social-networking-fact-sheet/` accessed on 23.07.2014

they found out, that 74% of all online adult users use social networking sites whereby this splits up to 72% men and 76% woman. When looking on the research concerning age groups it shows that from 18 to 29 89% use social networks whereas this number continuously decreases for older age groups. From 30-49 82 % and for 50-64 65% use social networking sites. In the age group of 65+ anyway 49% use them. The study also analyses the usage among educational level and yearly income where no significant differences can be seen.

Who uses social networking sites

% of internet users within each group who use social networking sites

	All internet users	74%
a	Men	72
b	Women	76
a	18-29	89[cd]
b	30-49	82[cd]
c	50-64	65[d]
d	65+	49
a	High school grad or less	72
b	Some college	78
c	College+	73
a	Less than $30,000/yr	79
b	$30,000-$49,999	73
c	$50,000-$74,999	70
d	$75,000+	78

Figure 2.3: Social Network usage by PewResearch

Twitter users

Among online adults, the % who use Twitter

	All internet users	19%
a	Men	22[b]
b	Women	15
a	18-29	35[bcd]
b	30-49	20[cd]
c	50-64	11[d]
d	65+	5
a	High school grad or less	15
b	Some college	20
c	College+	21
a	Less than $30,000/yr	23[c]
b	$30,000-$49,999	15
c	$50,000-$74,999	13
d	$75,000+	21

Figure 2.4: Twitter usage by PewResearch

When looking at the statistics specifically for Twitter a quite similar picture shows up despite the fact, that Twitter is not so widespread with a usage of 19% of all the adult Internet users. Within Twitter there is a slight difference between male and female users. From the worlds population 22% males and only 15% females use Twitter. When looking on the different age groups it also shows a significant change when looking at older users. This increase for Twitter is even bigger than for all social networks as 35% of the group 18 to 29 use Twitter. For the other groups the numbers are as follows: 30 to 49 20%, 50 to 64 11% and 65+ 5%. So Twitter is even more popular in younger age groups. For the educational level and the yearly income the same as for all social networking sites applies.

Lastly they also evaluated the social network usage on mobile phones. There, 40% of the mobile phone users use social networks on their devices with a split of 39% male and 41% female. For this group of users the age differentials are even bigger as for the previous analytics. The age group usage is as follows: 18 to 29 67%, 30 to 49 50%, 50 to 64 18% and 65+ 5%. This is surely connected to the fact that smartphones are more likely used in the younger age groups. Interesting in this numbers is that the mobile social network usage also increases with the yearly income as well as with the level education.

Social Networking on Mobile Phones

% of cell phone owners who use a social networking site on their phone

	All cell phone owners (n=1,954)	40%
a	Men (n=895)	39
b	Women (n=1,059)	41
	Age	
a	18-29 (n=340)	67[bcd]
b	30-49 (n=562)	50[cd]
c	50-64 (n=587)	18[d]
d	65+ (n=429)	5
	Race/ethnicity	
a	White, Non-Hispanic (n=1,404)	36
b	Black, Non-Hispanic (n=234)	48[a]
c	Hispanic (n=180)	49[a]
	Annual household income	
a	Less than $30,000/yr (n=447)	38
b	$30,000-$49,999 (n=316)	40
c	$50,000-$74,999 (n=272)	48[a]
d	$75,000+ (n=538)	45[a]
	Education level	
a	No high school diploma (n=156)	33
b	High school grad (n=542)	37
c	Some College (n=490)	42[a]
d	College + (n=752)	43[ab]

Figure 2.5: Mobile Social Network usage by PewResearch

Digital Insights also compiled a great overview about the Social Media usage in 2014 compiled from various sources[4]. They have split the statistics per social network. So for Facebook they found out that there are 1.28 billion active users per month which is about 18% of the global population. There are also more than 1 billion active mobile users so nearly 79% of the Facebook users also use it mobile. A very interesting fact concerning the usage of Facebook is that 75 % of the engagement on a post happens within the first 5 hours. So the conclusion is to be relevant over time posts need to be done regularly. Google+ in contrast also has more than 500 million active users per month but the average time spent on the platform is only 7 minutes. So there is not that extensive use of the platform like on Facebook. When looking on Twitter there are 255 million active users per month and a total of more than 1 billion total users. These users post more than 500 million tweets per day, on average 2 tweets per person per day. Twitter is also used by 78% of the users via mobile which is nearly the same number as for Facebook. One challenge Twitter is facing is the massive number of inactive users as 44% of the Twitter users have never tweeted a post and 391 million Twitter accounts have no followers. Compared to the worlds population only 3.5% of them are using Twitter. Also the number of blog users is quite impressive as there are more than 6.7 million people who actively blog via blogging sites and more than 12 million people who blog over social networking sites. The content which is provided in blogs is also heavily used as 77% of all Internet users read blogs. Also from a com-

[4] `http://blog.digitalinsights.in/social-media-users-2014-stats-numbers/05205287.html` accessed 23.07.2014

mercial perspective it is very important for businesses to use blogs as B2B marketeers who use blogs generate 67% more leads. All these numbers lead to the fact that 23% of the time spent in the Internet is spent on blogs and social networks.

When looking at the usage numbers for Austria provided by the Social Media Radar Austria[5] there are some differences in the usage of social networking sites. On Facebook there are 3.24 million active users which split up into 49% female and 51% male users. 86% of the users are younger than 49 years old. When comparing it to the population of Austria this leads to a number of 38% active on Facebook. Twitter has a total of 121.476 users in Austria where 67.816 are actively participating in the community by writing or reading tweets. This is a total of 1.5% of the Austrians is on Twitter whereas less than 1% is really actively reading or writing tweets on the platform. When looking at this numbers one can see that Austria is a quite mature market for social networks as the Facebook usage is higher than the average. Nevertheless is the usage of Twitter quite low compared to the average. When the Twitter usage numbers continue growing, the platform will become increasingly important in the social media landscape.

Definition of Social Media

Although O'Reilly's [O'R06] definition of Web 2.0 is quite a good match to give also a definition of Social Media, there are several other definitions that describe the concept of Social Media in a better way. A very simple but quite precise definition of Social Media was done by Safko who defined it as follows:

> *Social Media is the media we use to be social* [SB09]

This is a very basic definition which seems very banal. To give a clearer overview over the term Social Media it is therefore necessary to specify the word media as well as the term social. The term media refers to the toolset which is used for the communication. In the terms we use Social Media nowadays these tools are technology based. In essence one can say that the term media is described by the technology which connects people. The phrase social is defined by the connection between humans. The purpose for this connection is mainly the sharing of thoughts and experiences. So in a more general and accurate definition Social Media can be seen as the technological toolset which connects people tho share their thoughts and experience.

Another very popular definition of the term Social Media was done by Kaplan and Haenlein [KH10].

> *Social Media is a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content.*

So basically this definition connects the definition of Web 2.0 with the concept of user generated content. This definition defines the word media even more precise as an Internet-based application. This definition also specifies the term social in a more detailed way. Social Media is not

[5] `http://socialmediaradar.at/` accessed 23.07.2014

only a concept to exchange content, but also to generate the content which should be exchanged.

The following section gives different classification concepts of Social Media to provide a deeper understanding of this term.

Classification of Social Media

One first classification of Social Media can be done based on the degree of self presentation users want express with the usage of a Social Media. This self presentation can have several reasons. One can be to make a very positive impression for others. Others want to use Social Media to make an exact virtual copy of one. The self presentation is dependent on the self disclosure. Another dimension for classification can be the media richness of a Social Media appliance, because different platforms offer different possibilities in the richness of the media which can be published. This two dimensions lead to the following classification matrix which was introduced by Kaplan and Haenlein [KH10].

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Figure 2.6: Classification Matrix for Social Media

This classification matrix leads to the following six different types of Social Media. These types of Social Media are also related to the classification of user generated content platforms which was done by the OECD [OEC07]. The six different types after Kaplan and Haenlein [KH10] are as follows:

1. **Collaborative projects:** There are two different types of collaborative projects. Wikis, which are also mentioned in the study of the OECD, have the purpose to provide users the possibility to create and manage text-based content. In contrast is social bookmarking a concept which allows a group-based selection of Internet content. The OECD defines social bookmarking as a seperate type.

2. **Blogs:** Blogs represent a very early form of Social Media appliance and are somehow an equivalent to personal websites. The difference is that blogs are managed by a simple administrative user interface which was not possible within the first personal websites. Blogs are also mentioned in the paper of the OECD.

3. **Content communities:** Content communities are platforms which allow the users to publish a specific type of content. Popular appliances are Flickr for images, YouTube for videos or Slideshare for slides. Within the content communities the content has priority over other personal information of the publisher. These platforms are not mentioned in the study of the OECD.

4. **Social networking sites:** The main purpose of social networking sites like Facebook is to stay in contact with friends and to share content only with friends. Social networking sites are also the most used Social Media when looking on the usage numbers. Also the OECD mentions social networking sites as appliance for user generated content platforms.

5. **Virtual game worlds:** These virtual world have the purpose to provide fictional three-dimensional environments for users where they can interact with avatars. Avatars are the virtual copy of the user. The purpose of these platforms remains the same than in classical single-player games - to entertain the user in a game-based approach.

6. **Virtual social worlds:** In contrast to virtual game worlds the purpose of virtual social worlds is to give the users a virtual three-dimensional room for their interactions with other people. Because these two types have quite common characteristics the OECD summarizes them into on appliance of user generated content platforms. Kaplan and Haenlein [KH09] also published a paper about virtual social worlds where they made a differentiation of virtual social worlds compared to other types of Social Media. One difference is that the interaction in virtual worlds happens in real time which is not valid for the other types of Social Media except the virtual game worlds.

A different way to classify Social Media is based on the ties which refer to the network. These ties can also help to differentiate Social Media platforms. Borgatti et al [BMBL09] defined four basic types of ties which can be studied in social network analysis.

Similarities can be based on memberships or attribute like groups in social networks which share the same interests. The similarities could also be based on other information like location. They are not fully seen as social ties but influence social networks. Also content specific platforms like Flickr are based on similarities as they connect users with similar interests.

Another type of tie is the **social relation** which exists among the users of a network. Many networks are based on relations as the network is used to maintain and build relations. A special appliance of a network which was build to maintain specific relations in a social network was StudiVZ which had the aim, to connect students from similar lectures. Most social networks also support affective social networks - the like-functionality in Facebook is one famous example for this.

The next type of tie is the **interaction** which is also a big motivation for social networks. Basically each social network is based on and targeted at interactions. But there are also some special appliances like help communities which are based on the interaction type of helping someone.

The last form of tie is the **flow**. Flows can be tangible or intangible things that are transmitted with the interaction. Whereas Twitter's purpose is more or less to support an information flow,

the purpose of most of the blogs is to support the flow of beliefs. Slideshare can be seen as a platform which supports the flow of resources.

Characteristics of Social Media

The previous section discussed Social Media in a more formal way. Social Media also shows many informal characteristics which can't be used for classification but help to get a better understanding of the concepts of the different Social Media platforms.

One characteristic defined by Figueiredo et al [FPB+13] is that Social Media typically includes one main object which is stored in a media type like text, image, audio or video and a variety of other information which is linked to the main object. These links are also called the associated features of the object. Figueiredo also defined several types of features:

- A **textual feature** is descriptive information which is connected to an object. In most cases this information is created and linked by the content creator. Examples are descriptive tags to specify an image or comments to an image. Figueiredo et al [FPB+13] also defined two types of textual features. Collaborative features can be added by any user. Restrictive features can only be added by the content creator. When looking at YouTube the title and the tags of a video are restrictive features as they only can be added by the publisher and the comments for a video are collaborative features as anyone can post comments to a video.

- A **content feature** is information which can be derived from the object. An example for a content feature is the frequency of an audio file.

- A **social feature** is the social context in which the object was created. This is for example the information about who created a video on YouTube.

A broadly used concept within Social Media is the usage of hashtags. The Oxford Dictionaries[6] define the term hashtag as follows:

> *A word or phrase preceded by a hash sign (#), used on Social Media sites such as Twitter to identify messages on a specific topic.*

Another common concept in Social Media is the process of tagging. Enriching data like images or audio files with meta-data is called tagging. One common definition of tagging is done by Wang, Ni, Hua and Chua [WNHC12] who refer to multimedia tagging as the process of assigning a set of keywords to multimedia data to describe their content on semantic or syntactic levels. The concepts of hashtags and tagging are quite similar. Hashtags therefore can be seen as one appliance of tagging in systems which do not support tagging per se.

Most of the Social Media platforms offer the concept of groups. Mislove et al [MMG+07] discovered, that the usage of groups has a significant variance between several social networks.

[6] `http://www.oxforddictionaries.com/definition/english/hashtag` accessed 12.08.2014

They compared the platforms Flickr, LiveJournal[7], YouTube and Orkut[8]. They found that groups are not widely used in Flickr (21%), Orkut (13%) and YouTube (8%). Only on LiveJournal groups are used very often (61%). This leads to the conclusion that the concept of groups is implemented in many platforms but the usage is highly dependent on the purpose of the network.

A very interesting characteristic, which can be discovered in social networks is the connection chain between two randomly chosen individuals in a social network. Milgram [Mil67] stated in his publication *The Small World Problem* that in average each American is connected to any other American over only six hops. Therefore he made an experiment where he gave packages to 60 individuals which should be send to one single person. They were only allowed to send the package directly to the recipient when they directly knew him. Otherwise they should send it to any person they think could know the recipient. From the 60 packages 3 found the way to the recipient with an average path length of six hops. In other executions of the small world experiment from 296 packages 64 reached the target. The other packages did not find the way to the target. Because of the low number of received packages the results from Milgram are very controversial. Leskovec and Horvitz [LH08] validated the experiment of Milgram. They used a dataset of 30 billion conversations among 240 million people done with the Microsoft Messenger instant-messaging system. With that data a communication graph with 180 million nodes was created. With that setting they found out that the average path length among Messenger users is 6.6 hops which is related to the findings of Milgram. Also Ugander et al [UKBM11] proved the thesis of Milgram in a large scale setup.

Schneider et al [SFKW09] investigated social networks from a network perspective. In their work they also found several characteristics of Social Media. One observation is that users tend to stay within the same activities on the networks. So a user who uses a network for messaging tends to stay on the messaging part of the network. Although most of the networks offer a variety of features not all features are used together. Another interesting fact they found is, that although users tend to stay on social networks for a longer time they do not continuously interact with the network. The last observation they made is, that although most social networks offer the functionality to create profiles, this feature is not always the most heavily used feature. In their work they analyzed the networks Facebook, Hi5[9], LinkedIn and StudiVZ. The profile feature was the most popular feature within LinkedIn and StudiVZ, but overall sharing and uploading photos are the most heavily used feature.

A more extensive description of the social network Twitter follows in the section 2.2.

Criticism on Social Media

Although Social Media offers several advantages to the users there are also some problems with the usage of Social Media. Because Social Media platforms are typically free of charge the privacy is one issue which should be critically analyzed as many platforms use the content

[7] `http://www.livejournal.com/` accessed 12.08.2014
[8] `http://www.orkut.com/` accessed 12.08.2014
[9] `http://www.hi5.com/` accessed 12.08.2014

provided by the users for another purpose, like advertising. Another problem is cyberbullying which is a very widespread issue. The partial anonymity and the fact that the communication is not face to face lower the inhibition level to bully another participant in the network.

Privacy

Most of the Social Media platforms reuse the content which is provided by the users to take advantage out of it. As the platforms are free of charge the operators of social networks need other sources of revenue. One very common source is advertising. Advertising is present at nearly all of the platforms. To target the advertisements to the right audience the social networks use personal information like the interests of the users to get better advertising results. The platforms not only store personal data but also use behavioral data for their purposes. Because of these ulterior motives the privacy policies of social networks are very complex and many users do not question them when signing. The process of storing and re-using personal and behavioral data is also called surveillance. Fuchs [Fuc11] defined in his work that surveillance either could be neutral or negative. He also states several points why this surveillance is problematic.

- The terms of use and privacy policies of social networks are very complex and hard to understand. The information of how Facebook uses information they gather from a person contains 2316 words which lead to a seven page document. The information concerning advertisments contains 861 words (four pages) and the section concerning the privacy settings of a user contains 2030 words (five pages). This is only a small selection on the privacy statements of Facebook[10]. The privacy policy of Twitter has 2246 words[11].

- Unequal Internet skills lead to the fact that unexperienced users more likely do not opt-out of several privacy settings or do not activate several privacy settings.

- Social networks do not integrate users in the definition of their privacy policies and so these policies more likely prefer the social network over the user.

- Another big problem is the commercialization of the Internet and therefore every social network wants to benefit from the content they maintain.

- As social networks are more useful the more user they have, there are nowadays only a few relevant, big and global networks which have monopolies and therefore are very powerful in things they do.

- The creators of the content which is used by social networks are mostly unpaid for their efforts.

- A very big problem is the lack of transparency about what is really stored about a user by the surveillance.

[10] `https://www.facebook.com/about/privacy` accessed 12.08.2014
[11] `https://twitter.com/privacy` accessed 12.08.2014

An interesting fact which was discovered by Spiliotopoulos [SO13] is, that social network users from the USA are less concerned about privacy then other users. This also leads to the conclusion that therefore social networks are even more heavily used in the USA compared to e.g. Austria. Another study was done by Blasbalg et al [BCF12] who evaluated privacy concerns among new students. They found out that the students are in general concerned about privacy. Many of them also say, that they have done enough to keep their information private. When researching for their information in many cases it showed that still some information is available public.

Cyberbullying

Another problem with Social Media is cyberbullying. Cyberbullying describes when people use the Internet to harm other people. Heirman and Walrave [WH11] discovered in their studies that 12.1% of the adolescents they asked already cyberbullied someone they know online or offline. One third of the people already experienced cyberbullying. This problem is also more present in higher technical developed countries. One problem which makes cyberbullying even more problematic than offline bullying is, that within cyberbullying more people can participate in bullying. When for instance a person publishes unfavorable pictures of another person the pictures can be seen by many other users of the network. The person which is bullied in that case does not even have the possibility to remove the pictures as they were published by someone else.

Another aspect which makes cyberbullying very harmful is that the information stored in social networks is persisted over a long time and the deletion is not possible in most cases. Public available information is indexed by many other platforms and the social network persists the data.

2.2 Twitter

Twitter Overview

As Twitter is the main topic of this thesis, this section gives an overview about the features of Twitter. Twitter is a Social Media platform which is hard to classify based on the classification scheme defined by Kaplan and Haenlein [KH10]. The platform declares itself as a microblogging service. Therefore Twitter shows several characteristics that can also be observed within blogs. The Twitter user has the ability to post his statements on Twitter, which are called *Tweets*, as he can do within his blog. The difference is, that Twitter limits the post size to 140 characters. Within blogs people can subscribe to blog posts. In Twitter there is a similar mechanism called *following* a publisher to subscribe to his posts. Tweets within Twitter can be *retweeted* which is an equivalent to sharing a blog post.

Twitter on the other hand also shows characteristics of a social networking site as each publisher on the platform also has a profile where he publishes personal information. The platform also supports so called *mentions* of other users to direct a tweet to the other user. This collaboration aspect of Twitter is also discussed in the following sections. Because Twitter also offers mechanisms of favoring Tweets or analysis like *trending topics* Twitter can't be classified only as

blogging platform or a social networking site, but as a mixture of both with some collaborative traits.

Characteristics and Demographics

When looking at Twitter, a very interesting question to discuss is, why is this platform used. This question was discussed by Huberman et al [HRW08]. When looking at Twitter, in comparison to other social networks the links between users are directed. When a user is following another user, the reciprocal relationship does not necessarily also has to be defined.

Also the tweets by a user can be directed to another user by including a @ character together with the username of another user in his post. Indirected posts does not have such a mention in it and are directed to anyone but also directed posts can be read by everyone. Huberman et al [HRW08] found out that 25.4% of all posts on Twitter are directed. Because Twitter does not support a friend-relationship like other networks do, they also defined in their work, that a *friend* on Twitter is a person to whom another user has directed at least 2 of his posts. With this extra relationship they found out that the number of posts initially increases as the number of followers also increases but at a certain point this increase is not measurable anymore. When comparing the number of posts with the number of friends, the dataset showed that the number of posts increases at all time with the number of friends. When building a relationship between the number of friends a user has, and the number of followers and followees a user has, it shows, that the number of friends is very small compared to the number of followers and followees.

This leads to the conclusion that the links between users mainly are meaningless from an interaction point of view. Huberman et al [HRW08] came to the conclusion that the driver of the usage of twitter is a hidden network of connections within the set of followers.

Kwak, Lee, Park and Moon [KLPM10] compared trending topics they crawled from Twitter with trends they received from other media. A good estimator of the activities of the real world are the search terms on Google. In their work they compared the 40 top search topics per day and the 40 trending topics per day. They found out that only 3.6% out of 3479 unique trending topics on Twitter exist in 4597 unique search topics on Google. Nevertheless they found out that more than 85% of the trending topics are headline or persistent news. An interesting aspect is that the topics on Google are not present that long than the trends on Twitter. They also compared the trending topics to the CNN Headline News. In more than half of the time the CNN News were ahead in reporting but there were some news that were present in Twitter before. A very famous example for this is the plane landing in the Hudson River in 2009 where the breaking news of this crash were on Twitter 15 minutes before other news media reported about the accident[12]. The initial tweet of the crash was done by a person who saw the crash and also a person on the ferry to pick up the stranded people tweeted a picture of the crash.

Another work which illuminates the user's intentions for using Twitter was done by Java et al [JSFT07]. The biggest group of users on Twitter are *daily chatters* which talk about the daily

[12] `http://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html` access 18.08.2014

routine and the things they are currently doing. The second biggest user intention is the need for conversation. A conversation is defined by a post which mentions another user. They found out that almost 21% of the users use conversations and 12.5% of all posts are conversations. In the work of Huberman et al [HRW08] the number of conversational posts was twice as big as this number. Another user intent is *sharing information* as approximately 13% of all posts contain some URL in it. The last intention is *reporting news* or comments about current real world events. Java et al [JSFT07] also defined 3 main categories of users:

1. *Information source:* Users which fit into this category are hubs with a large number of followers in Twitter. These user post regular updates. Some of them also post infrequent updates - they in many cases are then posting very valuable information and the amount of followers is high because of this fact.

2. *Friends:* Most relationships on Twitter fall into this category. This can be seen as a subcategory of the following relationship.

3. *Information seeker:* In this group are users which do not post many times but follow many other users.

Another interesting aspect to study are the demographics of Twitter users. Mislove et al [MLA+11] analyzed in their work a dataset of Twitter users which represent over 1% of the overall population of the United States of America. They compare the Twitter population and the US population based on three different categories - geography, gender and ethnicity. Concerning geography over 75% of the users listed a location in their profile but only 8.8% of these locations were also convertible to latitude and longitude coordinates. One finding concerning the geographic distribution is that densely populated countries are overrepresented on Twitter compared to sparsely populated countries. When plotting the users against a map all major cities of the US are clearly visible. When looking at the gender of the Twitter users they found that 71.8% of Twitter users had a male name. Concerning the ethnicity they found that Twitter users represent a highly non-random sample of the overall ethnicity distribution.

Kaplan and Haenlein [KH11] found out that on the day Michael Jackson died 22.6% of all tweets included the term *Michael Jackson*. In opposite Google was not able to handle all the search requests for Michael Jacksons death and classified them as automated requests without a response. They further defined 3 characteristics which are responsible for the success of Twitter. The first reason is *ambient awareness*. They stated that in many cases the information in one tweet on its own is not valuable but only a combination of more tweets can paint a picture of a user's activities. This concept is called ambient awareness. In essence one can say that because of this ambient awareness the social presence on a platform like Twitter increases. The second reason is the *unique type of communication* which is allowed by Twitter. The posts on Twitter are retweeted in many cases which leads to the fact that a message can be transmitted over various networks in a very quick way. The third reason is *virtual exhibitionism and voyeurism*.

Communication on Twitter

As discussed in the previous section Huberman et al [HRW08] found out that 25.4% of all messages posted on Twitter are directed towards one or more other users. This leads to the assumption, that microblogging in contrast to classical blogging also has a collaborative aspect within communication. Whereas normal blogs are used to publish content of any kind, microblogs are not only used to publish content but also to start a conversation with other users. So Twitter can also be used for collaboration.

Honeycutt and Herring [HH09] studied in their paper the conversational and collaborative aspects of Twitter. With the usage of the @ sign to direct messages to specific users a form of addressivity is achieved. With that addressivity also cross-turn coherence could be reached. Coherence can be described as a topic-focused, continuous user-to-user interaction. Because the @ sign is not only used for addressivity they also defined several categories for the usage of the character.

- **Addressivity:** A message is directed to another user. An example would be *@ArminWolf - what do you think of the current situation*.
- **Reference:** A message references another user but is not directed to him like *I would be very happy with the job @ArminWolf has.*
- **Emoticon:** The @ sign could be part of an emoticon.
- **Email:** A tweet could also contain a mail address.
- **Locational:** The @ character can be used to express a location where the user is like *I am happy here @ TU Vienna*.
- **Non-locational:** A message can also contain a @ sign which simply is a placeholder for the word 'at'. An example would be *I am happy to finish my studies @ TU Vienna*.

Honeycutt and Herring [HH09] also analyzed the languages of the tweets they crawled. Of course is the language of the tweets dependent on the time the posts were crawled as the activity of users on Twitter decreases over night. The most tweets were in English followed by Japanese and Spanish. Over all languages the usage of the @ sign remains equal with about 30% of the posts containing a @ sign. This is a significantly higher number than the 12.5% observed by Java et al [JSFT07] and the 25.4% observed by Huberman et al [HRW08]. As 2 years lay between these analysis this leads to the assumption, that the interactive usage of Twitter increases. Another fact they found was that in more than 90% of the uses of the @ sign the intent was to direct a message to another user and in about 5% to reference another user. From all the directed messages 31.2% of the messages received a response within the same hour which leads to the assumption that Twitter is heavily used for conversations. They also identified conversations between more than 2 users but the larger the number of participants is it is also possible to loose the thread or to cross several threads. Therefore Twitter is not the best platform for collaboration within larger groups. For that use case the groups should be split to smaller groups or into dyads. One development towards collaboration in Twitter is the possibility to protect tweets. Protected

tweets can only be read by approved followers.

When it comes to conversations on Twitter another interesting aspect to look at is the practice of retweeting. This topic was analyzed by Boyd et al [BGL10]. Retweeting forwards a message to a different user network than the original post was intended. This invites users to participate in conversations which originally were not directed to them. So retweeting on the one hand can be seen as a form of information diffusion and on the other hand as a possibility to participate in a diffuse participation. In an early version of Twitter the functionality to retweet a post was not present and so a user had to manually copy a tweet to retweet it. Within this copy in many cases also the username of the original publisher was included and therefore retweeting was both information diffusion and conversation. With the new functionality of the retweet-button within Twitter the original post is tweeted again in the network of the new publisher. It is also not possible to modify the original post. From that standpoint the functionality removed the conversational aspect of retweeting.

Twitter Constraints

Compared to other common Social Media platforms Twitter is limited due to some restrictions. Also the APIs for crawling Twitter data have several constraints which developers have to comply to. Another common problem within Social Media platforms is censorship. Also Twitter has to deal with several censorship problems.

Restrictions

Twitter follows several restrictions which limit the functionality of the platform compared to other Social Media providers. The most important restriction is the limitation of the post length to 140 characters. This limit leads to changes in the communication style of the user posting content on Twitter. Another restriction is, that Twitter only supports one type of relationship, the follower-followee relation. Followers may know each other in person but they need not. Also family relationships are not supported within the platform. This leads to the challenge that directed communication, which is achieved by the @ sign within a tweet, can have several different backgrounds which are hard to analyze. Twitter also offers the possibility to protect tweets, so that tweets only can be read by approved followers of the publisher. Another restriction of the Twitter search[13] is, that the search results are pre-filtered. Protected tweets are excluded from the search results on the web page of Twitter as well as when searching data through the APIs Twitter provides.

Twitter also has several limits concerning posts and followers[14]. One limit of Twitter is that an user can only post 2400 updates in a 24-hour period. The limitation applies to the user so it is indifferent which client is used for posting the update. When this limit is exceeded the user is not able to post any update until the time period passes. For direct messages the limit is only

[13] `https://support.twitter.com/groups/53-discover/topics/215-search/articles/42646-twitter-search-rules-and-restrictions` accessed 20.08.2014

[14] `https://support.twitter.com/articles/15364-twitter-limits-api-updates-and-following` accessed 20.08.2014

250 updates. Each invocation of a function the Twitter API offers is counted as one request. The last limitation applies to the maximum number of followers. A user can follow 2000 other users without any problem. After reaching this boundary a ratio of people a user follows and people who follow the user is calculated. The exact set ratio is not available public. There is also a daily restriction for following other users which has a technical reason. This limit is 1000 followers, so in one day a person can only follow 1000 new people. Twitter also limits the API access based on the number of requests. For unauthenticated requests the limit is 150 requests per hour and for authenticated requests 350. These limits apply to the REST API Twitter offers. Twitter offers also a Streaming API which allows to crawl real-time Twitter data.

Censorship Twitter

Censorship is a common problem for Social Media platforms. Also Twitter is facing the problem of censorship in several countries [Wik14]. In China the usage of Twitter is strictly prohibited [BOS12] and the site is blocked. During the *Arab Spring* Twitter was unaccessible in Egypt. The revolution during the Arab Spring was heavily supported by Social Media. Therefore many people believe, that the government was responsible for the blocking of Twitter [Mur11]. Twitter was also blocked during the 2009 Iranian presidential elections because the government was afraid of protest organized over the platform [She09]. Also during the riots in the Russian-Ukrainian-crisis in 2014 several pro-Ukrainian accounts from Russian users were blocked after Russia threatened to ban Twitter entirely [Rie14]. In Turkey Prime Minister Tayyip Erdogan tried to ban Twitter several times. Twitter was also blocked for some days [Rie14]. These are only some examples of censorship of Twitter in the recent past. All this examples show the power a platform like Twitter gives to people to organize themselves in big networks. When looking at the *Arab Spring* Twitter was one very important supporting technique to make the revolution happening[15].

2.3 Text Mining

Keywords and Sentiment

Analyzing a large dataset of tweets manually is a very elaborate task. It is also very hard to do an objective analysis when it comes to trend and sentiment detection. A first indication for a trend can be the extracted keywords from a tweet. The occurrences of the keywords can then be counted and the trends are the keywords which occur the most often. As this is a very straightforward method to detect topics the next section deals with the topic of trend analysis in more detail.

To perform an analysis of the sentiment of a post there are several different methods. Sentiment analysis is a very complex task of text mining as the sentiment is no objective measurement. One can see a positive sentiment in a text whereas another person sees a negative sentiment in the same text. Dealing with sarcasm is another big problem within sentiment detection as sar-

[15] `http://mic.com/articles/10642/twitter-revolution-how-the-arab-spring-was-helped-by-social-media` accessed 02.04.2015

casm often means a negative sentiment which is formulated positive. Also regional differences can lead to different expectations concerning a sentiment rating.

Trend Analysis

A very interesting concept used in Twitter is the detection of trends. Twitter therefore has a functionality called *trending topics* which represents the topics which are discussed very often within a defined region. Wikipedia defines a trending topic as follows: *a word, phrase or topic that is tagged at a greater rate than other tags is said to be a 'trending topic'*.[16]. Aiello et al [APM+13] evaluated 6 different topic detection methods to extract trending topics from a Twitter dataset. To narrow the topic detection down their algorithm needs a predefined set of *seed terms* to filter the dataset down to all posts which contain at least one of the terms, a *timeframe* of interest to detect trends for posts in this period and an *update rate* to determine how often new trends should be detected. To reduce the noise in the posts several pre-processing steps were performed.

- *Tokenization* was done by removing stop words, punctuation and mentions and compressing character repetitions.
- *Stemming* was done to reduce inflected words to their root to reduce the number of possible words for the dataset.
- *Aggregation* was used to aggregate several posts together to one bigger document because topic detection methods suffer from short documents to be analyzed.

The six topic detection methods they used are as follows:

1. *Latent Dirichlet Allocation (LDA):* One method for topic extraction are probabilistic topic models. LDA is a very common topic model. In LDA every document which should be analyzed is a set of terms which are the variables in the model. The distribution of topics within a document and the term composition per topic are unknown and have to be estimated through Bayesian inference. One required input of this method is the number of topics which are expected.
2. *Document-Pivot Topic Detection (Doc-P):* This topic detection method uses a document-pivot approach. The principles of this algorithm are quite similar than the steps the researches used for the aggregation of the posts. The first step is to cluster posts based on their similarity. Clusters with a small item count are filtered out and for the remaining clusters a score which represents the possibility to contain trends, is calculated. The top clusters are returned by the algorithm.
3. *Graph-based Feature-Pivot Topic Detection (GFeat-P):* Within this extraction method a feature-pivot method is used. The first step of this algorithm is a selection of top terms based on the ratio of likelihood. For each term a node in the graph is created. The next step is the linking of the nodes based on the similarity of the terms representing a node.

[16] `http://en.wikipedia.org/wiki/Twitter#Trending_topics` accessed 19.08.2014

The similarity is calculated based on a similarity measure. Choosing the right measure is crucial for the success of this algorithm. After the linking step an algorithm for detecting communities within nodes is applied. Afterwards every community represents a topic. This algorithm also detects node hubs which in the last step of the detection method are linked to the communities if they exceed a specified treshold.

4. *Frequent Pattern Matching (FPM):* This method uses the simultaneous cooccurrence of more than two terms to determine the similarities of posts. This is done with frequent itemset mining. All sets of terms which appear in a set of posts of an itemset are counted. The number of appliances of an itemset is called its support. Each itemset which has a support above a defined treshold is called pattern.

5. *Soft Frequent Pattern Matching (SFPM):* This matching algorithm combines the GFeat-P-method and the FPM-method. The pattern detection method is similar than in the FPM-method but the algorithm is less strict because it does not require all terms in an itemset cooccur frequently.

6. *BNgram:* All the previous mentioned algorithms use unigram terms for the topic detection. Unigrams are single words. This algorithm uses n-grams for the topic extraction. This especially makes sense for Twitter as many messages on Twitter are simple retweets and so there are posts which are completely equal. This method also introduces a new feature selection method. A very interesting thing to look at is the changing frequency of a term over time. All detected n-grams are then ranked and clustered together. The clustering stops when the similarity between the nearest unmerged clusters is below a predefined treshold. All the remaining clusters are then ranked and the highest clusters are the extracted trends.

When looking on the results of the algorithms it shows that the GFeat-P algorithm is the algorithm with the least accurate results followed by the LDA and the Doc-P algorithm. The pattern matching algorithm FPM and SFPM lead to even better results and the BNgram outperforms the other algorithms. This leads to the conclusion that the extraction and aggregation of the keywords from the tweets is the most importing step in detecting trending topics.

Bernhardus and Kalita [BK13] also evaluated methodologies for trend detection in a Twitter dataset. They defined that trends in Social Media are a combination of *chatter* and *spikes*. *Chatter* is characterized by ongoing discussions over time which mostly are initiated by the users. *Spikes* are defined by short-term discussions which mainly are responses to recent real world events. They used several algorithms to define the selection criteria for terms out of Twitter datasets. The results of the selection algorithms were treated as keywords.

1. *Frequency:* The simplest algorithm is to count the frequencies of each term in the dataset. The most frequent terms within tweets are words with no information as 'the', 'and', 'or'. Therefore a stop word algorithm has to be applied which filters out such terms.

2. *TF-IDF:* This algorithm defines a document's relevance based on a composite of the query's term frequency and the inverse document frequency. The term frequency is the

number of documents in which a term occurs in a document set. The inverse document frequency normalizes and dampens the term frequency. Put simply, the weight of a document is higher if the number of occurrences of a word in a document is higher or if the number of documents containing the word is lower.

3. *Normalised term frequency:* This algorithm only uses the term frequency which is normalized with a scaling factor. With the normalization formula each term has a trending score assigned.

4. *Entropy:* In this algorithm the entropy of a term is calculated. The entropy is calculated with a collection of all tweets which contain the term.

For the trend detection they used 2 different methods. In the first method they calculated values for the precision, recall and the F-measure scores comparing their selected keywords with Twitters trending topics. The scores where calculated with true positives, items which were keywords and trending topics, and false positives, items which were keywords but no trending topics. In the second experiment they used recall and relevancy scores. The first method only evaluated exact matches of the keywords and the Twitter topics. The second algorithm used the relevancy score to also use partially matching keywords and trending topics as true positives. With the first method they used, they did not meet their success criteria, therefore this method is defined as no good trend detection algorithm. The second method therefore met the success criteria.

Also Mathioudakis and Koudas [MK10] implemented a system for real-time trend detection within tweets crawled from the Twitter stream. For the trend detection they use *'bursty' keywords* which are defined as keywords which appear in tweets at an unusual high rate. These keywords are then grouped into trends based on their co-occurrences. In essence they define *a trend as a set of bursty keywords that occur frequently together in tweets*.

One aspect interesting to evaluate when it comes to trend is the question how a topic can rise to a trending topic. This question is analyzed by Asur et al [AHSW11]. One finding of their studies was that factors like the number of followers and the activity of a user do not contribute to the formation and propagation of a trend. The key for trends is more the resonance of the content. When a trending topic is highlighted on a very busy website it is more likely that this trend also comes to Twitter as users will tweet about this trend. Within trending topics it is also interesting to observe that many of the trending topics disappear very quickly but some stay present over a longer time. One reason for topics to persist over time is a high number of users posting about the topic. Trends are created by two different types of Twitter users. *Sources* define the trends with their posts. Asur et al [AHSW11] found that there are only a few sources responsible for many different trending topics. One interesting fact they found is that the sources tweet on a regularly base but there is only a weak correlation between the post frequency and the trending topics. The number of followers of the sources was nearly uncorrelated to the trending topics. Therefore the other group of trend creators, the so-called *propagators* are responsible for making a topic to a trending topic. They found that the correlation between the number of retweets of a topic and the duration of the trend is very high.

Sentiment Analysis

Sentiment analysis within short statements is a very tricky part as there are several challenges within this task. The concept of sentiment per se is not a completely objective one as different people see different sentiments in the same message. Sarcasm is also a big problem for sentiment analysis as within sarcasm a negative sentiment is expressed by a message with a positive sentiment. Only when thinking about what the publisher wants to express the real sentiment is revealed. Another problem in sentiment analysis is that the context of the message has to be considered within the sentiment analysis. One good example for this is the word 'hot' within a post. When a person posts about motorbike which looks hot the message is considered to have a positive sentiment. When he posts about a hot exhaust of his motorbike the sentiment is negative. Also can different user groups have a different form to express themselves. In a gamer board messages are positive which in another context would be negative.

Stieglitz and Dang-Xuan [SDX12] analyzed the sentiment of tweets which have a political content. They also examined how the sentiment is influencing the number of retweets these posts have. For their work they analyzed over 64000 german speaking tweets which contain either the political parties of Germany or the prime candidates of the parties. After analyzing the sentiment of the posts they also looked at the influence this sentiment had on the retweet rate. They found that in their dataset a tweet is retweeted 0.43 times and that each tweet contains 0.19 words with a positive and 0.33 words with a negative sentiment. Within the 10 most retweeted users the sentiment occurrences are higher with 0.29 words with a positive and 0.39 with a negative sentiment. This leads to the assumption that the sentiment impacts the retweet-rate of a user significantly.

In the most cases sentiment is divided into two categories, negative and positive. Choudhury et at [CGCH13] studied a very special type of sentiment within tweets. They tried to use tweets to predict a depression of the user who posted the messages. Their major findings were that Social Media can be a very useful signal to detect an onset of a depression. The key indicators for that are a decrease in the social activity, clustered egonetworks, increased relational and medical concerns and also a negative sentiment in the posts. To detect the signals of depression they used a survey to figure out Twitter users with a depression and compare their tweets with a control set of other users. With their findings they also developed a model to predict depression within a set of tweets of a user.

Sentiment is a very interesting aspect to analyze within Social Media. The detection of sentiment though is a very challenging task. Singh et al [SPUW13] introduced a feature-based heuristic for aspect-level sentiment with a dataset of movie reviews. They based their computational method on a public library called SentiWordNet[17] which is a lexical resource for opinion mining. In this algorithm each term of the library has 3 numerical scores assigned. These scores describe the objective, positive and negative polarity of the term. These scores are then combined by eight ternary classifiers. Singh et al [SPUW13] then performed two sentiment classifications - a document-level and a aspect-level sentiment classification. The *document-level sentiment classification* combines all movie reviews of one movie to a single review document and rates

[17] `http://sentiwordnet.isti.cnr.it/` accessed 21.08.2014

this single document. Therefore the reviews were pre-processed to extract terms like adjectives, adverbs and verbs. These terms got a part-of-speech tag assigned and then the sentiment of the terms and tags was rated and the overall sentiment was aggregated. The assignment of the part-of-speech tags is crucial for the success of the rating. A good suggestion is to use adjectives as they are good markers of opinions. The adjective should be combined with adverbs as they further define the opinion (AAC - adjective adverb combination). For example marks the sentence 'this is a good car' a positive sentiment and the sentence 'this is a very good car' an even more positive sentiment. In their algorithm the adverbs are weighted with a scaling factor of 0.35 so that the adjectives get the higher priority over the adverbs. In a second experiment they combined adjectives and adverbs as well as adverbs and verbs (AAAVC - adjective adverb adjective verb combination). In the *aspect-level sentiment classification* they perform the similar tasks but for each review on its own. As dataset they used movie reviews they crawled from a movie rating website. They compared their algorithm with the movie ratings within the movie rating website and with sentiments they calculated with a text mining API, the *AlchemyAPI*[18]. The movie website counted 760 positive reviews, the AAC 678, the AAAVC 688 and the *AlchemyAPI* 634. With the negative reviews the website counted 240, the AAC 98, the AAAVC 99 and the *AlchemyAPI* 140. In relative numbers the website contained 76% positive and 24% negative reviews, the AAC 82% positive and 18% negative, the AAAVC 82.9% positive and 17.1% negative and the *AlchemyAPI* 73.4% positive and 26.6% negative. Their findings were, that the difference between the two methods they used is very small and tends to be biased more to a positive sentiment. A reason could be that the algorithms can't deal with sarcasm which the manual ratings consider. The *AlchemyAPI* appears as a good sentiment rating engine.

Another sentiment classification method was studied by Pak and Paroubek [PP10]. As a first step they crawled Twitter to obtain messages which were clustered in positive, negative and objective posts. For the separation of positive and negative posts they used emoticons within the messages. An objective post is post which does not express an emotion or states a fact. They used this method as Twitter messages mostly contain only one sentence and therefore the usage of an emoticon should represent the sentiment for the whole sentence. These posts were used to train a classification algorithm. For the objective dataset they crawled tweets from popular newspapers and magazines. From the positive and negative tweets they extracted n-grams as binary features. As each post had a sentiment assigned the n-grams can be mapped to the sentiments. They compared their algorithm with other sentiment classifiers and found very accurate sentiment results within their classifier. The higher the sample size was defined, the better the results were.

Automated Text Mining

As analyzing text with respect to keyword extraction, trend detection and sentiment analysis is a very time-consuming and biased task, several Text Mining APIs were developed to perform automated text analysis tasks. Most of the APIs provide also free versions which are very restrictive in the number of allowed transaction per day or month. The *AlchemyAPI* which is used in this work for text analysis purpose also offers an academic version which allows enough

[18] `http://www.alchemyapi.com/` accessed 21.08.2014

transactions a day to analyze also big datasets. The algorithms for trend detection presented by Aiello et al [APM+13] were also released for public usage to perform trend detection. The next 2 sections give short overviews about the *AlchemyAPI* and other APIs for text mining.

AlchemyAPI

The *AlchemyAPI* was launched in 2009. The framework contains of two parts, *AlchemyLanguage* which contains the text mining features and *AlchemyVision* which contains the image recognition features. The solution performs over 3 billion API calls per month. The functionality is available as a Software-as-a-Service offering where users pay for transaction contingents per month. They offer also a free version which allows 1000 transactions per day and an academic version with 30000 transactions per day. The algorithms consist of a very broad feature-set. The text mining methods contain of the following functionality which can be used to analyze text samples, web pages or HTML code:

- **Entity Extraction:** Entities like persons, places and organizations are detected and rated with a relevance score.
- **Keyword Extraction:** Keywords which mark important topics are extracted and rated with a relevance score.
- **Sentiment Analysis:** Sentiments are detected for entities, keywords as well as for entire documents.
- **Concept Tagging:** Concepts are detected within documents and rated with a relevance score. A concept can be seen as a subject area the document is about.
- **Relation Extraction:** Relations like subject, action and object within sentences are detected.
- **Taxonomy Detection:** Texts are categorized into hierarchical taxonomies.
- **Author Extraction:** Information about the author of articles can be extracted.
- **Language Detection:** The language of posts is detected.
- **Text Extraction:** Plain text can be extracted from webpages and undesired content like links or advertisements are removed.
- **Microformat Parsing:** Microformats of webpages can be extracted.
- **Feed Detection:** RSS feeds within webpages can be detected and the links are returned.

For this work the keyword extraction, the sentiment analysis and the language detection is used. The keyword extraction works for the languages English, German, French, Italian, Portuguese, Russian, Spanish and Swedish and the sentiment analysis works for English, French and German.

The *AlchemyAPI* is also widely used in research. Singh et al [SPUW13] used the *AlchemyAPI* to validate 2 new sentiment detection methods for movie reviews. They found that the *AlchemyAPI* sentiment ratings are very valid compared to movie ratings within a movie rating website. Chen, Benedikt and Kharlamov [CBK12] introduced a system for structured querying of annotated documents named QUASAR. Therefore they also implemented a document annotator which used the *AlchemyAPI* for entity and sentiment extraction. Quercia et al [QAC12] made a quantitative analysis of the L-LDA algorithm for topic detection which is a variation of the Latent Dirichlet Allocation. The LDA algorithm returns topics that are numbered distributions over words. The L-LDA is an extension of this algorithm which associates a document with easily-interpretable topics. They used the *AlchemyAPI* to pre-classify their dataset. They also used two other applications, namely OpenCalais[19] and TextWise SemanticHacker[20] to pre-classify the dataset but the per-classification done by the *AlchemyAPI* was more accurate than the results from the other alogrithms. Another work was conducted by Saif et al [SHA11]. They used the *AlchemyAPI*, Zemanta[21] and OpenCalais to extract entities from Tweets and they found out, that the *AlchemyAPI* performed better than the other algorithms in extracting entities out of tweets. OpenCalais uses several sources to retrieve base text data to perform the analysis, namely DBpedia, Wikipedia, Freebase Reuters.com, Geonames, Shopping.com, LinkedMDB and IMDB. Some of the sources used by *AlchemyAPI* are DBpedia, USCensus, Geonames, Freebase, UMBEL, OpenCyc, YAGO, MusicBrainz, CIA Factbook and Crunchbase. The other algorithms do not disclose their base text sources.

Other Frameworks

There are also several other text mining frameworks. A problem with many of them is the availability, as most of them only offer transaction limited free licenses. Zemanata, which contained a free API in the past shut down their publicly available API. The *AlchemyAPI* also offers analytics in more languages than other frameworks do and therefore it is better suited for a dataset of tweets. The topic detection algorithms introduced by Aiello et al [APM+13] were released within a project called SocialSensor[22]. Trend detection is only one part of this project which also contains parts like sentiment analysis and influencer detection. Also Yahoo! offers a term extraction algorithm for keyword detection[23]. Another framework which is available for free is OpenCalais. The *AlchemyAPI* outperformed other algorithms like OpenCalais, Zemanta and TextWise what the research of Quercia et al [QAC12] and Saif et al [SHA11] showed.

2.4 Summary

When looking on the concept of Social Media it becomes very clear that the term Social Media is highly connected to the terms Web 2.0 and user generated content. Social Media is a specialized appliance of the other concepts with various different shapes based on the purpose of the

[19] `http://www.opencalais.com/` accessed 21.08.2014
[20] `http://www.textwise.com/` accessed 21.08.2014
[21] `http://www.zemanta.com/` accessed 21.08.2014
[22] `http://www.socialsensor.eu/` accessed 21.08.2014
[23] `https://developer.yahoo.com/contentanalysis/` accessed 21.08.2014

networks. Social networks like Facebook and Twitter are heavily used all over the world. Also the usage numbers in Austria are rapidly increasing. A very interesting aspect of Social Media is that they supported the findings of Milgram [Mil67] concerning the small world problem. Social Media is also broadly criticized because they have complex privacy terms and the data is re-used in many cases.

When looking especially at the platform Twitter there are several usage drivers. In many cases Twitter also acts as a news media and is the first source to publish information about real world happenings. In more than one third of all posts on Twitter the platform is also used for conversations [HH09]. Twitter follows several restrictions like the post size limit of 140 characters. Another common problem within Social Media platforms is censorship which is done by several countries to prohibit the usage of platforms like Twitter.

The detection of trends in a network like Twitter is a very important task as trending topics give a good indication about real world happenings. The detection of trends is a very tricky task and there are several different algorithms which help in trend detection. The persistence of trends in Twitter is not dependent on the frequency a user who defines trends tweets or the number of followers he has but on how these topics are propagated via retweets.

Text mining is a very complex task therefore several frameworks were developed. The *AlchemyAPI* uses very accurate algorithms for keywords extraction and sentiment rating on tweets.

CHAPTER 3

Twitter Analyzer

In this chapter the practical implementation of this work is discussed. The section 3.1 describes the database model for the crawled Twitter data as well as the implementation of the Twitter crawler. In the section 3.2 the algorithms for keyword detection and sentiment analysis are described. Also the general usage of the trend detection algorithms and the AlchemyAPI is discussed and the restrictions are outlined. Lastly a short summary about the chapter is done.

3.1 Data Acquisition

Twitter offers an API which can be used to perform operations within the database of Twitter. One of these operations returns tweets based on a set of querying parameters. In the first part of the implementation the API was used to collect tweets which were posted within Austria and the regions nearby the Austrian border. To avoid a bias in the analysis of the dataset the data was crawled in two timeframes - the first part between the 16.04.2014 and the 29.04.2014 and the second between the 09.06.2014 and the 18.06.2014.

Implementation

Twitter offers 2 different types of APIs to perform operations on Twitter data. The REST API [1] provides a simple user interface for most of the functionality of Twitter. For querying Twitter, this API offers a GET-request which searches the database. The request needs a search query as input to specify terms after which the database should be searched. As an optional parameter also the latitude and longitude coordinates combined with a search radius can be provided to narrow the search down to specific locations. As for this work all tweets within a region needed to be crawled, the REST API was not applicable because of the needed search query. Another problem of the REST API is the existing rate limit which limits the requests and results which can be processed via the API. A second possibility to crawl data within Twitter is provided by

[1] `https://dev.twitter.com/docs/api/1.1` accessed 01.09.2014

the Streaming API[2]. This API allows to perform real-time searches on different types of streams. The API returns data like tweets as they happen. The first stream which can be queried is the public stream which contains all the public data flowing through Twitter. The second stream is a user stream where all data of a single user's view of Twitter is available. The last stream available is the site stream which represent a multi-user version of user streams. Within the public stream a search for all posts within a location can be started. For getting access to the APIs of Twitter one has to register for a developer account to obtain an API key which is needed for the authentication of an user. The authentication is handled over the OAuth protocol[3]. The current version of the standard is 2.0 and was proposed by Dick Hardt[4].

The implementation was done in Java[5]. For Java there are several libraries for the Twitter API available. In this work twitter4j[6] was used. This library offers an implementation of the REST API as well as of the Streaming API. Also the authentication with the OAuth-protocol is handled within the library. Figure 3.1 shows the Twitter OAuth authentication flow.

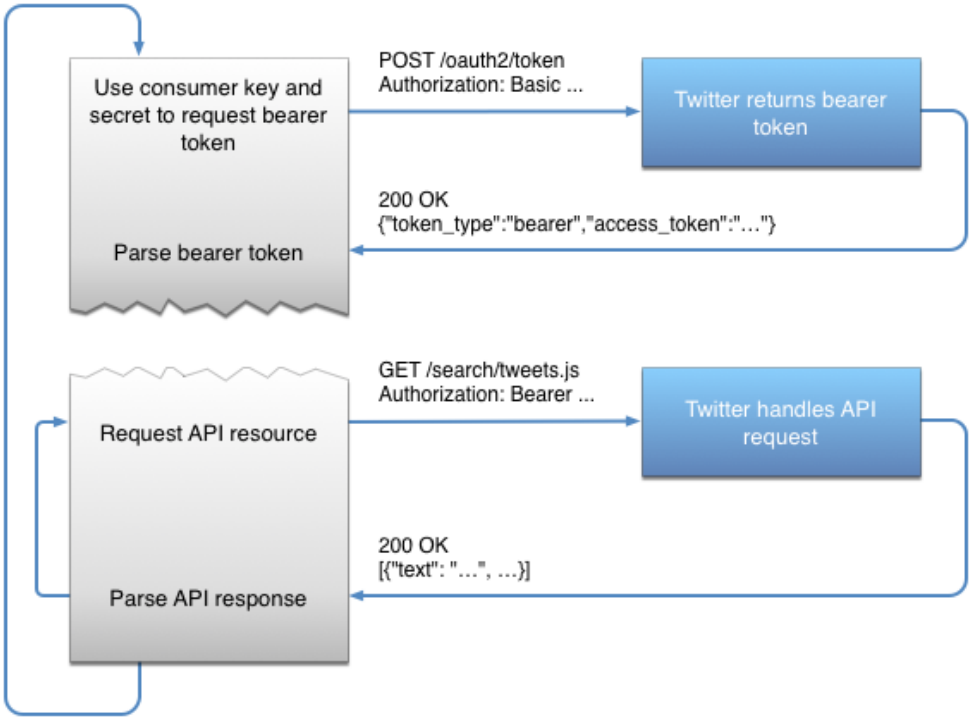

Figure 3.1: Twitter OAuth Authentication Flow

[2] `https://dev.twitter.com/docs/api/streaming`, accessed 01.09.2014
[3] `http://oauth.net/` accessed 01.09.2014
[4] `http://tools.ietf.org/html/rfc6749` accessed 01.09.2014
[5] `https://www.java.com/de/` accessed 01.09.2014
[6] `http://twitter4j.org/en/index.html` accessed 01.09.2014

The public stream of Twitter offers an endpoint which limits the returned posts within one request. The endpoint `POST statuses/filter` needs at least one predicate parameter out of the following three:

1. **follow:** A list of user ID's must be provided for which the API returns all posts.

2. **location:** A set of bounding boxes must be specified. Each bounding box contains location coordinates in a rectangular form. The API returns all tweets posted within these areas.

3. **track:** A list of keywords must be provided and the API returns all posts containing the keywords.

For this work the location parameter was used to obtain all tweets posted in the bounding box between `46.41794, 8.20349` and `49.11300, 17.52155`. This box specifies a rectangle containing Austria as well as the border regions. This box is visualized in figure 3.2 The major cities within this box are Vienna, Bratislava, Maribor, Bolzano, Zurich and Munich. The cities outside of Austria are included because of various reasons. One reason is, that the definition of a bounding box containing all data from Austria also contains cities outside from Austria like Munich. Also for the regional comparisons within different parts from Austria it is interesting to compare with the biggest cities outside of Austria - e.g. for comparing southern Austrian posts with posts in Maribor and Bolzano or western Austrian posts with posts from Zurich.

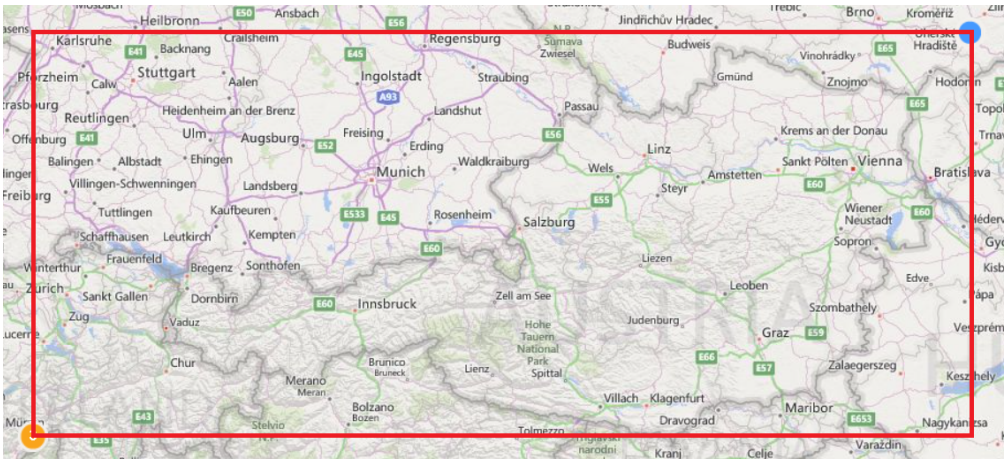

Figure 3.2: Bounding Box visualized on Map

The data was crawled in several crawling runs. To avoid a bias in the data, two separate datasets where created with the application. The first sample was crawled between the 16.04.2014 and the 29.04.2014 and the second one between the 09.06.2014 and the 18.06.2014 whereby the application did not crawl the posts continuously during the time frames. The crawler was launched at intermittent times due to technical limitations of the database service where the

crawled posts where saved. This service, Microsoft Azure[7], closes the database connection after some time. The resulting dataset was stored within a SQL database hosted on Microsoft Azure.

Resultset Datamodel

The dataset which was crawled using the Twitter Streaming API contains basic information for each tweet. In addition to this information also the user data from the publisher of the message is returned. Tweets are also linked to a place where the tweet was published. For all tweets in the resultset also the place information was crawled.

Tweet

The main information for each tweet is the text which was published within the post. This text is the base frame for the analysis of the Twitter data. Each tweet is identified by an explicit id. Also the exact date when the tweet was posted is stored in the dataset.

The location information for each tweet is stored in two different ways. For some tweets the coordinates are stored within the tweet information. These coordinates are reported by the user himself or the client application used to publish the tweet. Tweets can also be associated with a place entity. The place entity is stored separately. Tweets without location information were not crawled by the implementation.

When a tweet is a reply to another tweet the name of the user from the original tweet is also stored. Also the information if a post is a retweet is stored. Twitter also supplies language information for each post. The language is detected by Twitter and stored in the BCP 47[8] format. In the source field the utility which was used to post the tweet is stored in HTML format. Each tweet also contains a flag which indicates possibly sensitive content in the message. This detection only works for tweets containing an URL whereby the content of the webpage referenced by the link is analyzed. Each tweet stores the distinct id of the publishing user.

User

As each tweet is associated with an user also an user object is crawled by the implementation. Each user object is identified by an unique id. The date when the user was created within Twitter is also returned by the API. For a user the total count of favorites, followers and friends is managed. A friend is defined as a person which is followed by the user. Also the number of public lists the user is member of and the number of tweets the user has published is counted.

The language information which was defined as the user's user interface language is stored in the user object. Within Twitter each user can specify her or his location which is also included in the object. Twitter stores two different names for each user. The name contains the name the user defined within her or his profile. The screenname is an unique name which represents a user. This name is also used for mentions within a post.

[7] `https://azure.microsoft.com/de-de/` accessed 09.09.2014
[8] `http://tools.ietf.org/html/bcp47` accessed 11.09.2014

Place

Tweets are also linked to a place entity. Each place is identified by an unique id. The definition of a place is done with a bounding box, so each place contains four latitude and longitude pairs which define the location of a place. In addition also the country containing the place is stored. A place also contains a readable name which can be assigned to the place.

Twitter differentiates several place types which is also stored in the result set. This dataset contains of five different place types, namely city, neighborhood, POI, country and admin where POI stands for *point of interest* and admin represents an administration. Places can also be hierarchically nested. The order for place types is as follows: a country can contain several administrations which can contain several cities which can contain several neighborhoods which can contain several points of interest.

3.2 Text Mining Algorithms

The algorithm which is used by Twitter for trend detection is not disclosed. Therefore the second part of the implementation uses methods besides Twitter to detect trending topics. Also other features like the language or the sentiment of a tweet are detected by these algorithms. The used APIs and libraries also have some limitations which are discussed individually for each algorithm.

Trend Detection

The algorithms which were analyzed in the work of Aiello et al [APM+13] were published as a Java framework[9]. For the detection of topics in the sample dataset the following 3 algorithms were used:

1. Document-Pivot Topic Detection (Doc-P)
2. Latent Dirichlet Allocation (LDA)
3. Soft Frequent Pattern Matching (SFPM)

Implementation

The framework uses an itemset of tweets containing the text, the ID and the language of the post to detect the trending topics within the itemset. The framework also performs the pre-processing steps needed for the trend detection, namely tokenization, stemming and aggregation. Each of the algorithms needs several parameters which need to be set in order to adjust the algorithm.

The only parameter for the **Doc-P** algorithm is the similarity treshold which is needed for the cluster assignment within the algorithm. This measure defines the degree of similarity which is needed to assign a post to an existing cluster of similar posts. These similar posts are then used to identify trending topics. For this work the parameter was set to 0.5 because for short texts like

[9] `http://socialsensor.eu/results/software/87-topic-detection-framework`, accessed 22.10.2014

tweets the similarity of two items is either close to 0 or close to 1. Choosing a parameter close to 1 would only use duplicates or near duplicates, e.g. retweets for trend detection.

The parameters for the **LDA** algorithm are the number of topics which should be extracted, the number of training iterations and the number of keywords per topic. This algorithm mainly depends on the right choice of the number of topics to extract. The number of topics in this work was set to 10 with 300 training iterations and 5 keywords per topic. The amount of topics was chosen to generate a similar amount of topics as the other methods automatically do. The number of training iterations was defined based on experiments, which showed, that more training iterations do not lead to better results. The number of keywords per topic was chosen because each topic also contains short terms like articles. With 5 keywords not only these words are returned.

The **SFPM** algorithm uses the same term selection method and parameters as the GFeat-P algorithm. Additionally there are two sigmoid parameters which define the shape of the sigmoid function used for the expansion of the topics.

Restrictions

One conclusion of Aiello et al [APM+13] is, that the trend detection algorithms work well for datasets dedicated to a specific event and struggle with very noisy datasets like the dataset used in this work. Therefore the quality of the detected trends is not optimal. This is caused by the fact, that similar term patterns are not likely to occur in short messages like tweets. Retweets therefore can bias the topic detection in a massive way. Within this dataset it is very interesting that the amount of tweet duplicates is very small.

Another problem within the dataset is the occurrence of many different languages. The algorithms are dependent on the cooccurrences between words. In different languages these cooccurrences will become rare. Therefore the trend detection algorithms should be used only on datasets in one language.

AlchemyAPI

The AlchemyAPI is a set of different algorithms used to perform text mining purposes. The text mining API contains of the following functionality, which can be used to analyze text samples, web pages or HTML code:

- **Entity Extraction:** Entities like persons, places and organizations are detected and rated with a relevance score.
- **Keyword Extraction:** Keywords which mark important topics are extracted and rated with a relevance score.
- **Sentiment Analysis:** Sentiments are detected for entities, keywords as well as for entire documents.
- **Concept Tagging:** Concepts are detected within documents and rated with a relevance score. A concept can be seen as a subject area the document is about.

- **Relation Extraction:** Relations like subject, action and object within sentences are detected.
- **Taxonomy Detection:** Texts are categorized into hierarchical taxonomies.
- **Author Extraction:** Information about the author of articles can be extracted.
- **Language Detection:** The language of posts is detected.
- **Text Extraction:** Plain text can be extracted from webpages and undesired content like links or advertisements are removed.
- **Microformat Parsing:** Microformats of webpages can be extracted.
- **Feed Detection:** RSS feeds within webpages can be detected and the links are returned.

Implementation

The AlchemyAPI provides a JAVA library to access the functions implemented in the API. Within this work the tweets were posted against the API and the ranked keywords and the sentiment was analyzed. The API also detects the language of the tweet. The extracted keywords are then counted and the words which were present in most of the tweets can therefore be seen as trending topics.

Restrictions

The AlchemyAPI currently only supports a few languages and therefore many tweets in the sample dataset are not processed by the API. The keyword extraction works for the languages English, French, German, Italian, Portuguese, Russian, Spanish and Swedish. As languages like Slovakian, Croatian, Turkish and others are missing not all tweets were used for keyword extraciton.

The sentiment analysis is only supported for the languages English, German and French. Another problem of the AlchemyAPI is, that the text mining functions do not work for text parts which are very short. In the dataset there are also some tweets with too little characters. They are therefore not processed by the API.

3.3 Summary

In the implementation several different parts where developed. The crawling component did crawl Twitter using the Streaming API offered by the service. The query was setup to crawl data within a bounding box surrounding Austria including the near border regions and major cities nearby. The crawling happened between two intermittent time frames, one in April 2014 and one in June 2014. Each tweet crawled by the application contains of metadata concerning the tweet as well as the publishers user information and place information which is associated to the post. For trend detection several different methods where used. For all of the algorithms there are several restrictions which apply to the quality of the dataset.

CHAPTER 4

Results and Evaluation of the Twitter Analyzer

In this chapter the results of the implementation work are analyzed and discussed. The section 4.1 describes the crawled tweets and analyzes the data. In the section 4.2 the comparison of the used APIs are discussed. Also the results of the analysis of the API are validated. The chapter is closed with a short summary.

4.1 Twitter Data

In this work Twitter data was crawled based on a selection of tweets in a bounding box as described in the section 3.1. Twitter provides tweets which were tweeted within the specified boundaries as well as all tweets linked to a place which also has parts within the boundaries. Therefore the dataset also contains tweets which were not within the boundaries but contain a place which partly lies within the boundaries. An example would be tweets from southern Croatia linked to the place Croatia which is partly within the boundaries. Figure 4.1 shows a map where all posts are tweets are plotted onto. The colors indicate the amount of tweets. Red areas mark a high post volume.

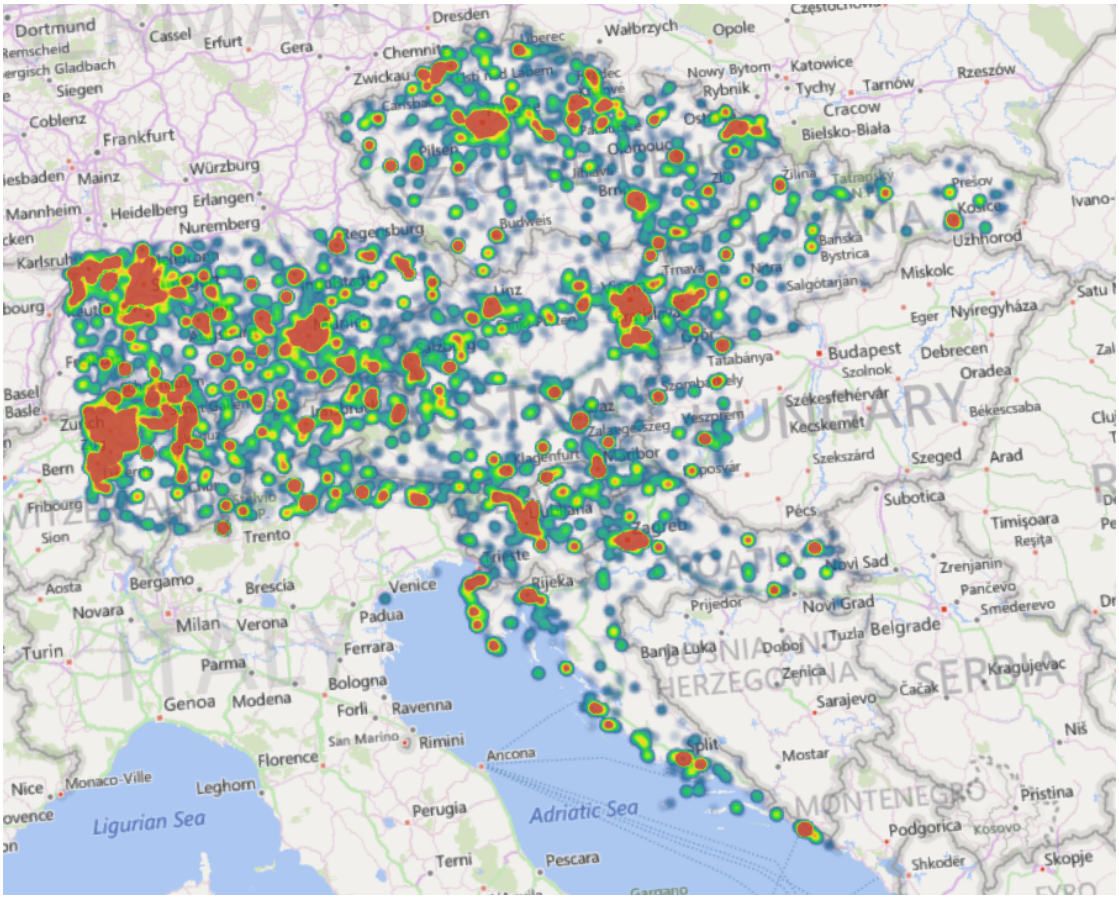

Figure 4.1: Map visualization of all tweets with location information in the post

Data Analysis

Overall 158 220 tweets were collected during the implementation. From this dataset 76 675 tweets (48.46%) were posted with latitude and longitude coordinates which lay inside the specified bounding box. 48293 posts (30.52%) do not contain location information and 33 252 (21.02%) were posted outside of the bounding box but are linked to a place which partly lays inside. The map also shows, that tweets mostly occur in cities like Vienna, Zurich or Munich. Although tweets are clustered within bigger cities, Twitter is no urban-only phenomenon. The number of tweets from the four biggest cities (Vienna, Munich, Stuttgart, Zurich) within the dataset is 28 157 which is 36.72% of the overall post amount. When adding the cities Graz, Linz, Salzburg, Innsbruck, Bozen, Marburg and Bratislava the number of tweets is 34298 which is 44.73% of the overall post amount. So more than the half of the tweets happen in places outside of the biggest cities within the dataset. The figure 4.2 shows a map containing all tweets with latitude and longitude coordinates within the specified bounding box.

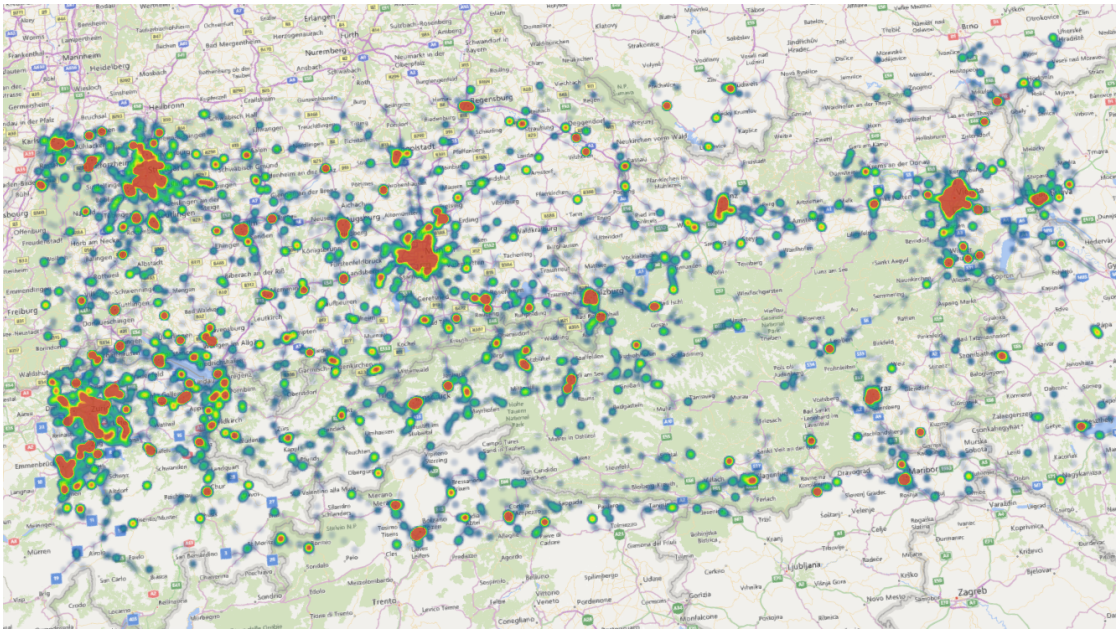

Figure 4.2: Map visualization of all tweets within the specified region

Tweeting Frequency

When comparing the frequency of tweets from the April and June datasets it shows that there was an increase in activity. For the comparison three different periods where defined. One from 00:00 - 10:00, one from 17:00 - 22:00 and one from 22:00 - 24:00. In the first two periods two equal days where compared (Saturday and Sunday). In the third period a Thursday in April was compared with a Friday in June. The numbers in the following table 4.1 outline the increase of tweeting frequency.

Number of Tweets in timeframe			
	00:00 - 10:00	17:00 - 22.00	22:00 - 24:00
April 2014	4340	6481	2822
June 2014	5083	8815	3703
Increase	17.12%	36.02%	31.22%

Table 4.1: Tweeting frequency comparison April and June 2014

The values from the dataset in April 2014 can also be compared with a control dataset from March 2015. Whereas the website Social Media Radar Austria[1] shows still a slight increase in the usage of Twitter the number of tweets doesn't show an increase. As sample data the tweets

[1] `http://socialmediaradar.at/twitter` accessed 19.03.2015

from one day in April 2014 (22 289) are compared to the tweets from a similar weekday in March 2015 (22 430) which is an increase of 0.63%. Also the June 2014 dataset can be compared to the March 2015 dataset. Based on that numbers the conclusion can be drawn that the usage of Twitter is already decreasing. The results are shown in the following table 4.2.

Number of Tweets in timeframe			
	00:00 - 10:00	17:00 - 22.00	22:00 - 24:00
June 2014	5083	8815	3703
March 2015	4523	8415	1851
Increase	-11.02%	-4.54%	-50.01%

Table 4.2: Tweeting frequency comparison June 2014 and March 2015

Language and Client Analysis

When looking at the used languages for tweeting overall 53 languages appear in the dataset and 4.34% have an undetermined language. The most used languages are English (27.83%), German (19.15%), Turkish (6.95%), Slovak (6.54%), Slovenian (5.45%), Spanish (4.70%), French (3.47%), Italian (3.16%), Arabic (2.49%) and Russian (2.43%). Figure 4.3 shows a chart displaying the number of tweets per language.

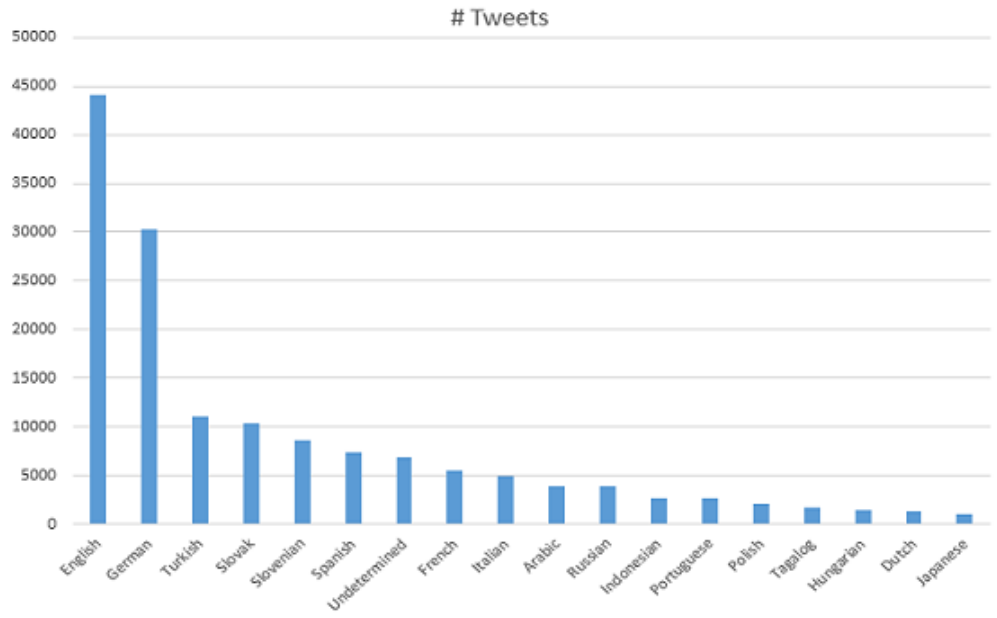

Figure 4.3: Overview languages, Source: Twitter

As English is commonly used within the Internet it is obvious that most of the tweets are in English. The appearance of most of the other languages is also logical as the dataset also contains tweets from Switzerland, Slovakia and Slovenia. Interesting is the high number of Turkish tweets. Compared to the number of Turkish people living in Austria [2], which is 1.35% of the inhabitants of Austria, the number of Turkish tweet is higher. In other sources[3] the number of people with Turkish background lays somewhere between 3% and 4%. These numbers lead to the assumption that Twitter is heavily used by Turkish people living in our region as the share of Turkish tweets is relatively high compared to the tweets with other languages.
The AlchemyAPI also includes a language detection functionality within the keyword extraction process. The results are different to the languages returned by the Twitter API. In most of the cases (85.07%) the language of the tweet was detected as unsupported language. The supported languages are English, German, French, Italian, Portuguese, Russian, Spanish and Swedish. The API detected 10.51% as English posts, 3.43% as German posts and 0.66% as French posts. 0.34% of the tweets had not enough text to detect the language. This already shows the first problem in using the AlchemyAPI to analyze a very noisy dataset like tweets. The language coverage of the API is not sufficient to perform a complete analysis. Also the difference in English and German posts compared to the Twitter data lead to the assumption that the results from the AlchemyAPI are a little bit imprecise. Figure 4.4 shows a chart displaying the number of tweets per language classified by the AlchemyAPI.

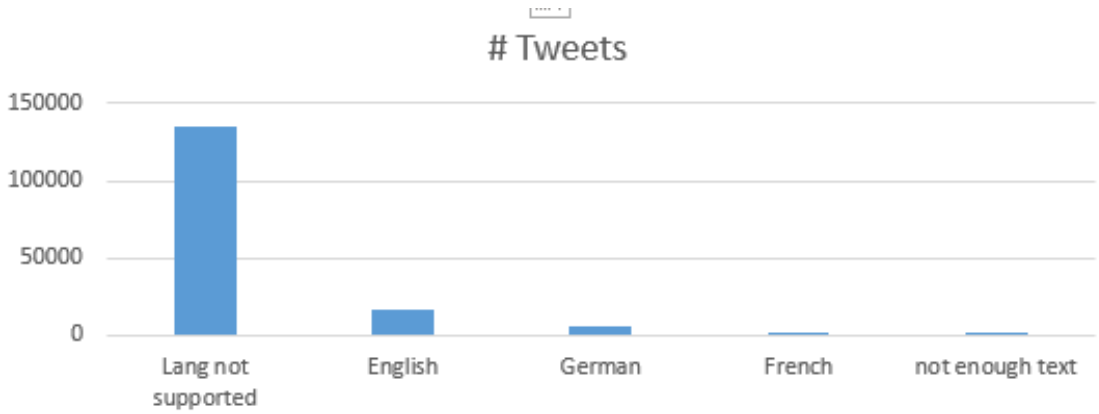

Figure 4.4: Overview languages, Source: AlchemyAPI

Another aspect to look on, when analyzing Twitter data, are the used platforms for posting the tweets. The Twitter Streaming API also returns the client which was used to post the tweet. The top sources are *Twitter for iPhone* (26.95%), *web* (18.44%), *Twitter for Android* (17.43%),

[2] `http://www.statistik.at/web_de/statistiken/bevoelkerung/bevoelkerungsstruktur/bevoelkerung_nach_staatsangehoerigkeit_geburtsland/022498.html`, accessed 07.04.2015

[3] `http://medienservicestelle.at/migration_bewegt/2011/08/17/turkische-community-zahlen-und-daten/`, accessed 07.04.2015

Twitter Web Client (6.41%), *foursquare* (5.60%) and *Instagram* (5.36%). One interesting aspect when looking at the client numbers is, that many tweets are made with mobile platforms, namely more than 70%. So Twitter is mostly used on mobile devices. In figure 4.5 a diagram showing the number of tweets per client is shown.

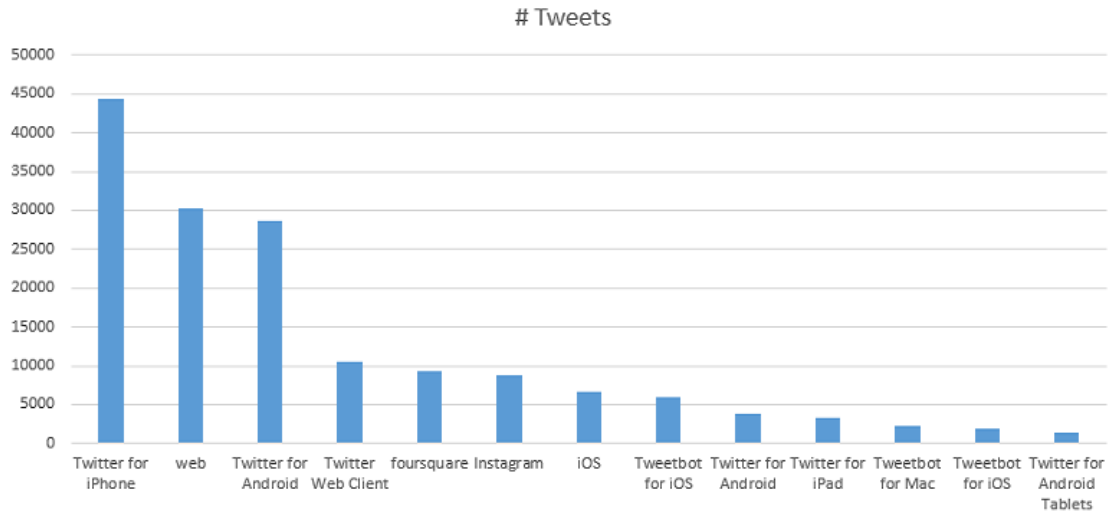

Figure 4.5: Overview clients

When summarizing the amount of tweets made with platforms, which can be clearly identified as part of one of the four main mobile platforms (Android, iOS, BlackBerry, Windows Phone), the dominance of mobile posts also is shown. Overall 37.88% of the tweets were made on iOS devices and 21.07% were made with Android devices. Windows Phone and BlackBerry devices represent only less than 0.50%. The figure 4.6 shows a comparison chart of the tweets per client and the mobile marketshare in Austria.

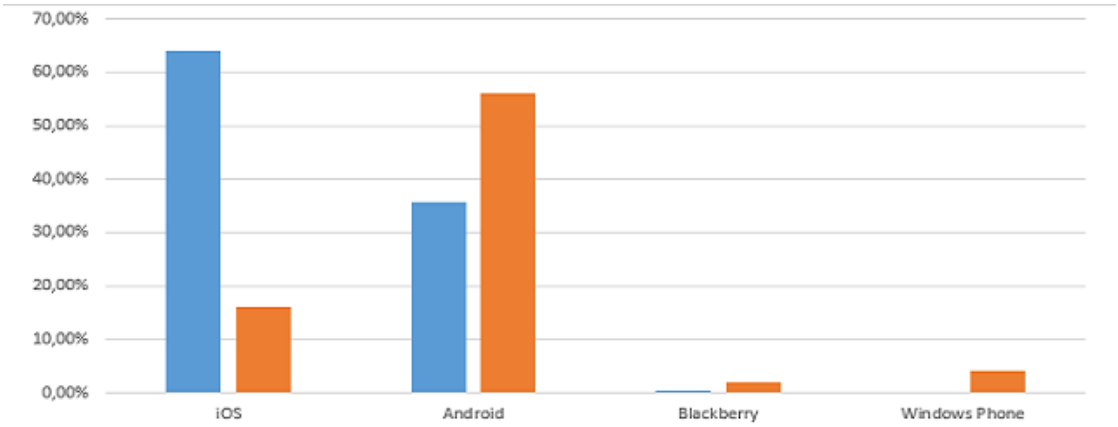

Figure 4.6: Tweets per client vs. Mobile Marketshare in Austria

Compared to the mobile platform market share the Twitter usage numbers are very different. In Austria the most used platform is Android (56%), followed by iOS (16%) and Windows Phone (4%)[4]. From this it follows, that the Austrian users of Twitter are not representative for all Austrian mobile phone users.

Characteristics

Huberman et al [HRW08] analyzed in their work the conversational aspects of Twitter and found out, that 25.4% of all tweets are directed to another user. Directed to another user means, that the tweet contains the @ character directly followed by a text. So Twitter was used for conversations in one fourth of all tweets. Java et al [JSFT07] also analyzed the number of conversations on Twitter. In their work the number of conversations was 12.5%. In Honeycutt and Herring's [HH09] work the number of conversations was 30%. So over the years 2007 to 2009 the number of conversations on Twitter almost doubled. In the dataset from this work the number of conversations is 49.09% so the conversational aspect of Twitter continued to increase. Today half of the posts on Twitter are used with a conversational motivation. Another number, which strengthens the conversational aspects of Twitter is, that 36.05% of all tweets in the dataset are replies to other tweets.
Java et al [JSFT07] also found out, that 13% of posts contain some URL in it and are therefore used for information sharing. Within this dataset 26.46% of the tweets contain an URL in it. So also the information sharing purpose of Twitter increased over time. On the other hand only 36.28% of the tweets were singletons, tweets without an URL or a @ sign. Twitter also classifies tweets if they contain possibly sensitive content. This classification can be done by the person who published the tweet or by users who are viewing the tweet. Twitter defines sensitive content

[4] `http://de.statista.com/statistik/daten/studie/300835/umfrage/genutzte-handy-betriebssysteme-in-oesterreich-nach-geschlecht/`, accessed 07.04.2015

as *nudity, violence or medical procedures*. Within the dataset 0.95% of the posts are marked with a sensitivity flag.

Sentiment and Trend Analysis

The AlchemyAPI also offers sentiment analysis functionality. The sentiment rating currently works for the languages English, French and German. When looking at the overall dataset the sentiment score is 0.1267 on a scale from -1 to 1. So the tweets within the dataset tend to be a little bit more positive than negative. When looking at the German tweets the sentiment is 0.0734, for the French the sentiment is 0.0392 and for the English posts the rating is 0.1548. Also the sentiment for the tweets posted in Vienna is below average with a rating of 0.1075.
Also when looking at the sentiment of the tweets posted with specific clients it shows that the mobile tweets posted with *Twitter for iPhone* (0.1011) and *Twitter for Android* (0.0999) are slightly more negative then the average posts. The tweets posted over the *web* client are quite average (0.1209) whereas the tweets made with *foursquare* (0.2820) and *Instagram* (0.2804) have a much better sentiment then the average posts. Also the tweets with the sensitivity flag have a better sentiment than the average (0.2311).

The AlchemyAPI was also used to extract keywords. For defining trends the extracted keywords are counted and the keywords with the highest count are used as trending topics. The 10 most detected keywords are *co, http, love, time, fuck, Wien, at, München, Adidas and Bayern*. As one can see the keywords also contain parts of URLs like *co, http and at*. Therefore the data which was analyzed with the AlchemyAPI needs a manual clearing process to avoid such topics. These words have been filtered out. The trending topics are shown in the figure 4.7.

Figure 4.7: Tagcloud with cleared trending topics, AlchemyAPI

The other algorithms, which were used for generating trending topics were described in the work of Aiello et al [APM+13]. In detail the *Document-Pivot Topic Detection (Doc-P), Latent Dirichlet Allocation (LDA)* and the *Soft Frequent Pattern Matching (SFPM)* methods were used. The algorithm parameters were described in section 3.2.
The first used algorithm, the Doc-P, did not return any trending topics within the overall dataset. The reasons for that are described in the section 4.2.
The LDA algorithm returned 10 topics with 5 keywords for each topic as specified in the algorithm properties. The topics are *de la le en est*, *praha wien prague hotel stuttgart*, *da je ne se na*, *ich die und der das*, *bu bir ya ne ve*, *de la el en mi*, *love follow happy day easter*, *na se je tak si*,

im xd haha rich st and *nchen bayern good time vienna*. As one can see also within these trends many short term words like German articles are included. Therefore also this method needs a manual clearing process. After this the remaining topics are shown in the figure 4.8. When comparing the cleared results with the results from the AlchemyAPI there are several equal topics like *Praha, Time, Love, Prague* and *Bayern*. Another interesting aspect in the trends generated by the LDA algorithm contain locations like *Prague, Praha, Wien, Vienna, Stuttgart* and *Bayern*. Also interesting is, that two cities appear in the trends in 2 different languages. The trending topics generated with the LDA algorithm are shown in figure 4.8.

Figure 4.8: Tagcloud with cleared trending topics, LDA

The SFPM algorithm returned 21 different topics. The topics are *für café çok wäre niederösterreich schön hemmings würde ostern praha weiß zürich não más mám über württemberg když güzel için münchen*. The topics after the manual clearing are shown in the figure 4.9. Compared with the results from the other trend detection methods the only equal topic is *Praha* which was returned as topic from the LDA algorithm as well. The appearance of topics like *Vienna, Stuttgart* or *Bayern* are obvious as these locations are also situated within the area where data was crawled. In opposite to that, the equal topic within all methods, *Praha*, which is also a location is not within the bounding box of crawled locations. Another interesting observation is, that in the SFPM the topic *Ostern* and in the LDA the topic *Easter* were detected. This is not surprising as the Easter time 2014 was within the crawling time frames. Whereas the AlchemyAPI and the LDA mostly returned topics in English and German the SFPM also returned trends in other languages as well. The trending topics generated with the SFPM algorithm are shown in figure 4.9.

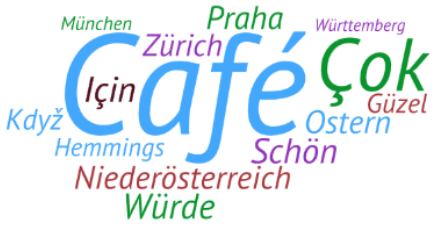

Figure 4.9: Tagcloud with cleared trending topics, SFPM

Regional Analysis

When taking a detailed look on the plotted Twitter data on the map one can see that most of the tweets in Austria are located in the area around Vienna, and the cities Linz, Graz and Salzburg. In Germany the same applies for the regions around Munich and Stuttgart. When looking at the number of tweets per person living in a city no significant difference between the biggest cities in Austria is measurable. The percentages are 0.8566% (Innsbruck), 0.6597% (Munich), 0.6497% (Vienna), 0.5479% (Stuttgart), 0.2967% (Graz) and 0.2673% (Linz), so there are less than 1 post per 100 inhabitants. The figures 4.10 and 4.11 show maps, where the amount of tweets which are clustered around Austrian and German cities is visualized. The colors indicate the amount of tweets.

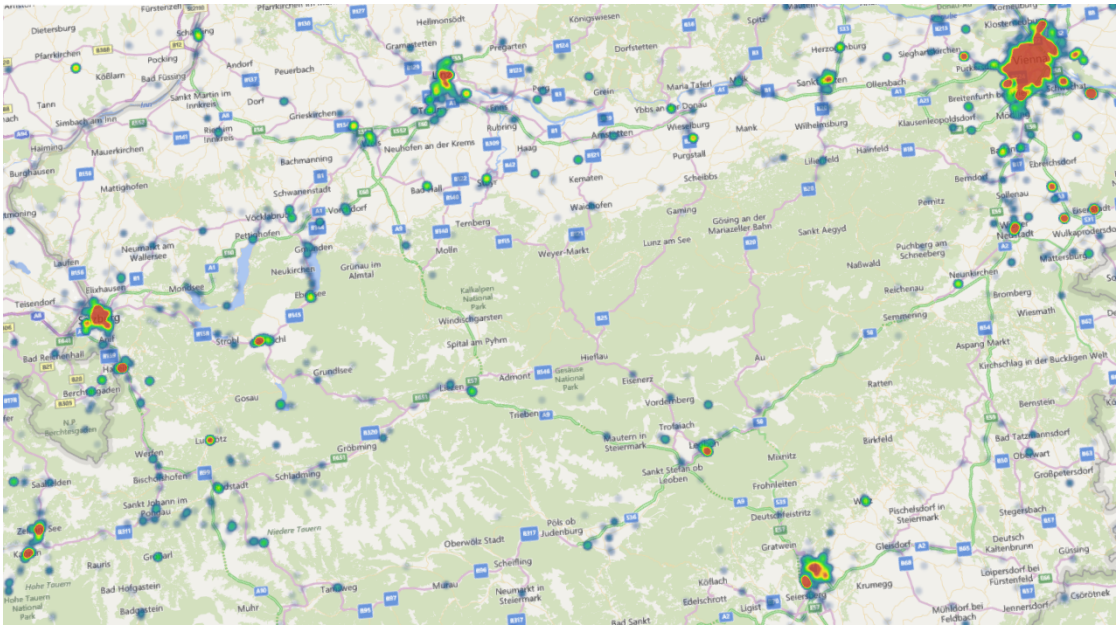

Figure 4.10: Detailed map of tweets concentrated around Austrian cities

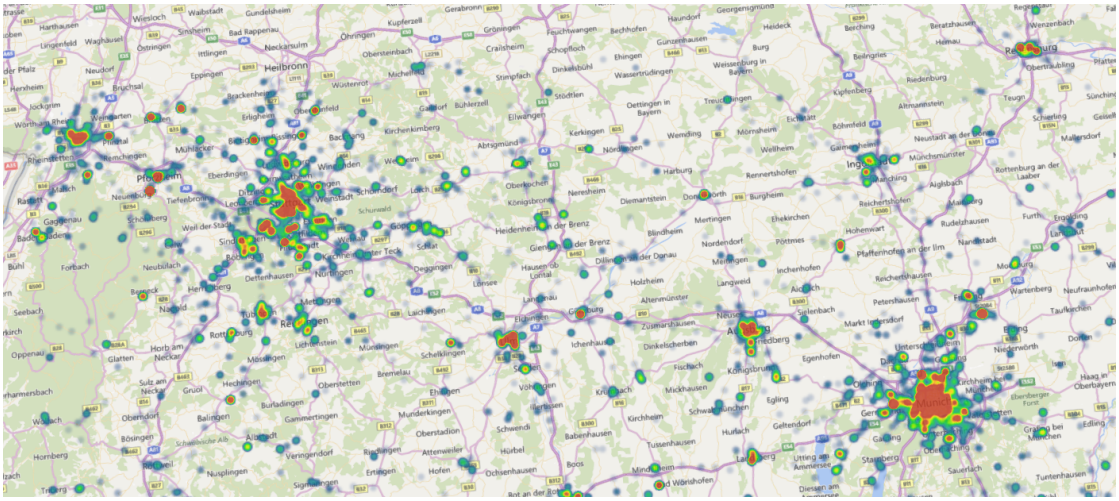

Figure 4.11: Detailed map of tweets concentrated around German cities

For a more comprehensive analysis of regional differences the dataset was divided into eight regional quadrants. Each quadrant is of the same size. The figure 4.12 shows the quadrants visualized on a map.

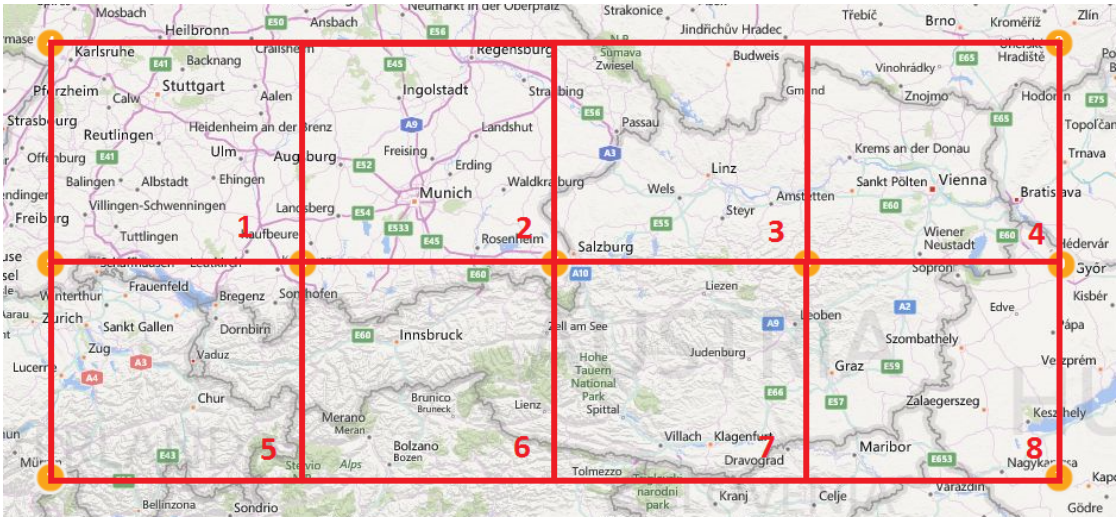

Figure 4.12: Regional quadrant overview

When looking at the amount of tweets per quadrant one can see that, although the number of tweets posted within cities is not significantly higher than those posted outside, the quadrants containing the biggest cities in the dataset also have the highest tweet amounts. This leads to

the assumption that tweets are clustered around cities more tightly than in rural areas. Therefore the tweeting amount is influenced by the location. The numbers for the tweeting frequency per quadrant are shown in figure 4.3.

Quadrant	1	2	3	4	5	6	7	8
No of Tweets	13384	15679	3884	16124	17067	5992	2371	2890
Percent	17.29%	20.28%	5.02%	20.83%	22.05%	7.74%	3.06%	3.73%

Table 4.3: Tweeting frequency per quadrant

In the following subsections the differences in tweeting behavior within the quadrants is discussed.

Language Analytics

As the overall dataset contains tweets in many different languages it is interesting to look also into the languages per quadrant. In most of the quadrants English and German are the dominating languages but in some areas also other languages like Turkish, Italian or Hungarian are used very often. In the quadrant 1 and 2 most tweets are in German (33.82% and 35.99%) and English (21.98% and 22.72%). The third language which is also used very frequently is Turkish (13.55% and 8.77%). The other tweets are distributed over many other languages. The same applies for quadrant 4 (English 29.59%, German 18.38%, Turkish 17.54%). Interesting for this quadrant is, that in opposite to quadrants 1 and 2 which are located mainly in Germany, here the majority of the tweets is in English and the amount of German and Turkish tweets is nearly equal. In this quadrant also Vienna is located. Therefore it is obvious that in Vienna more tweets are posted in Turkish language than in other areas. In quadrant 3 also German (30.43%) and English (28.84%) tweets are dominating, followed by Turkish tweets (10.63%). Interesting here is, that the next frequent language is Spanish with 9.42%. This is the highest appearance of Spanish tweets within the dataset. In quadrant 5 the numbers are quite equal to quadrant 3 (German 29.49%, English 28.61%, Turkish 7.15%, Spanish 7.27%), but the Spanish tweets are the third biggest majority. Interesting here that in the dataset no location was crawled where Spanish is a major language. In quadrant 6 the highest number of tweets is in Italian (28.30%) followed by English (19.81%) and German (17.47%). This is obvious as a big part of the quadrant lies in Italy. Quadrant 7 also is very equal to 1 and 2 (English 25.60%, German 23.75%, Turkish 13.62%). Also 6.92% tweets in Slovenian language are posted within this quadrant. In the quadrant 8 the highest number of tweets is in English (40.62%) followed by Hungarian (14.33%) and German (11.21%). Also here a quite big area of the quadrant lies in Hungary, therefore it is the second biggest part in the dataset. The figure 4.13 displays a chart where the tweet amounts per quadrant grouped by languages is shown.

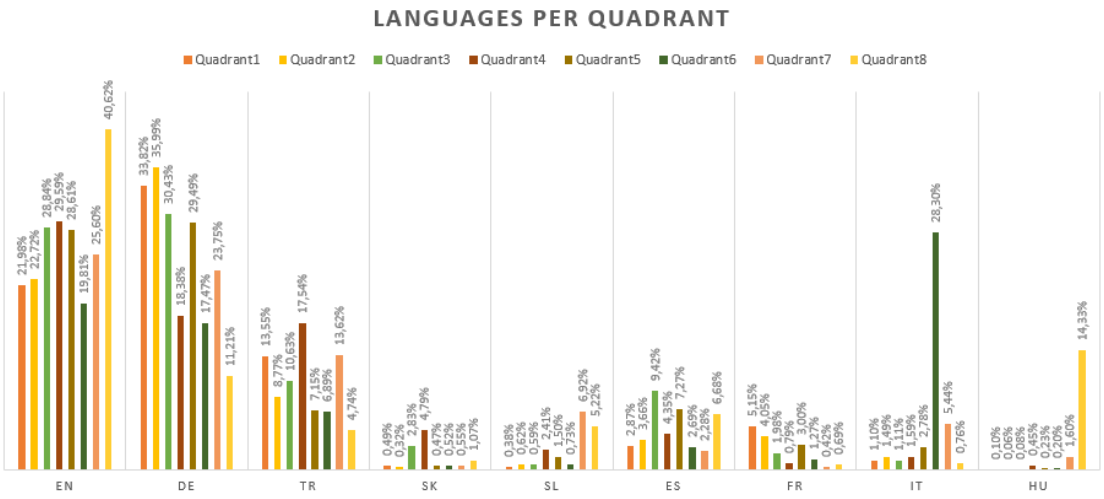

Figure 4.13: language overview per quadrant

Characteristics

When looking at the conversational aspects of Twitter per quadrant one can also see differences per quadrant. For the amount of conversational tweets there are two significant differing values - quadrant 1 has a lower rate of conversations (41.53%) and quadrant 8 has a higher rate of conversations (56.64%). When looking at the information sharing aspect of Twitter in quadrant 2 and 4 there is a higher rate of shared URLs in tweets (37.66% and 36.73%). For the replies there are also some significantly higher values in quadrants 5 and 6 (34.83% and 36.18%). And for the sensitive tweets there are lower values for quadrant 2 and 7 (0.48% and 0.38%). Sensitive tweets are tweets that are marked with a sensitivity flag within Twitter and can contain content like nudity or violence. Table 4.4 displays the percentages of conversational tweets, information sharing tweets, replies and sensitive tweets per quadrant.

Quadrant	1	2	3	4	5	6	7	8
Conversations	41.53%	47.72%	47.76%	46.29%	48.80%	49.30%	44.33%	56.64%
Sharing	28.70%	37.66%	31.33%	36.73%	33.84%	29.54%	30.70%	27.72%
Replies	28.95%	30.45%	33.55%	28.28%	34.83%	36.18%	30.75%	28.86%
Sensitive	0.81%	0.48%	0.90%	1.02%	0.90%	0.90%	0.38%	1.07%

Table 4.4: Tweeting characteristics per quadrant

As replies are also a special type of conversation one can see that in quadrants 5, 6 and 8 more conversations take place on Twitter than in average. In the quadrants 6 and 8 there are no major cities from the dataset located. This means that in 2 out of 4 rural quadrants the

conversational aspects of Twitter are higher developed. The quadrant 1 with less conversational posts than average contains a major city, namely Stuttgart. The quadrants with a higher amount of posts sharing some URL are 2 and 4 which contain the major cities Munich, Vienna and Bratislava. So for 2 out of 4 urban quadrants the information sharing aspects of Twitter are higher developed. The quadrants with the least amount of sensitive posts are quadrants 2 and 7 which is one urban and one rural quadrant. The figure 4.14 shows charts for the amount of tweets classified as conversation, information sharing, replies and sensitive tweets per quadrant. The diagram also shows the median and the standard deviation for each type.

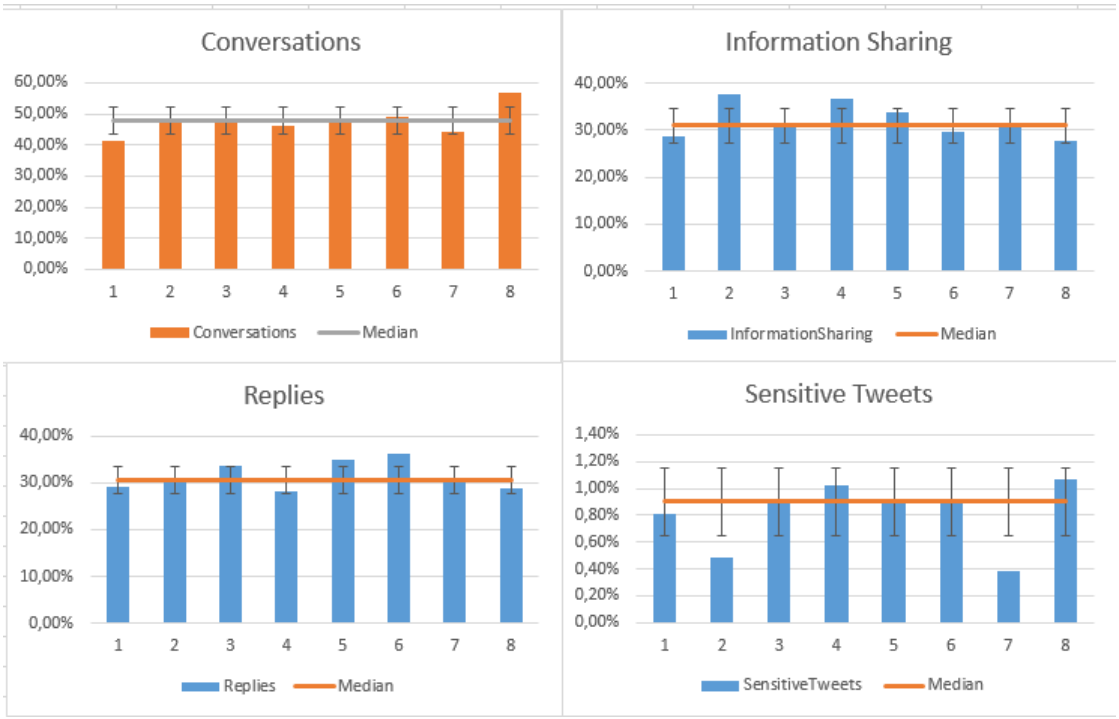

Figure 4.14: Quadrant analysis

Based on these analytics one can draw the conclusion that in rural parts where Twitter is used, the conversational aspects of Twitter are higher developed and in urban parts more URLs are shared within Twitter posts.

Sentiment Analysis

Another regional analysis can be done based on the different sentiments within the quadrants. Also for this analysis the AlchemyAPI was used. As the sentiment analysis only works for the languages English, German and French the results are biased towards these languages. The sentiment values for each quadrant can be found in table 4.5.

Quadrant	1	2	3	4	5	6	7	8
Sentiment	0.0759	0.0753	0.0613	0.1281	0.1835	0.1365	0.1059	0.1059

Table 4.5: Tweeting characteristics per quadrant

Interesting when looking at the sentiments per quadrant is, that only in quadrant 4 the sentiment is nearly equal to the average sentiment. The sentiment in quadrant 6 is higher than average and in quadrant 5 the value is significantly higher than the average. In the quadrants 1, 2, 7 and 8 the posts are more negative then average and in quadrant 3 the sentiment is significantly lower. Nevertheless, the sentiment value in all quadrants is positive, so the tweets are slightly more positive then negative. When looking at the more rural quadrants 3, 6, 7 and 8 the sentiment is lower than average, accept in quadrant 6. So in 3 of 4 rural quadrants the sentiment rating is below average. In the urban quadrants two of them are more negative, one is average and one is more positive than average. The positive quadrant 5 is with respect to other characteristics quite average, therefore it seems that especially in the Swiss region the tweets are more positive. The sentiments per quadrant are shown in figure 4.15. The chart also contain the median and the standard deviation of the sentiment values.

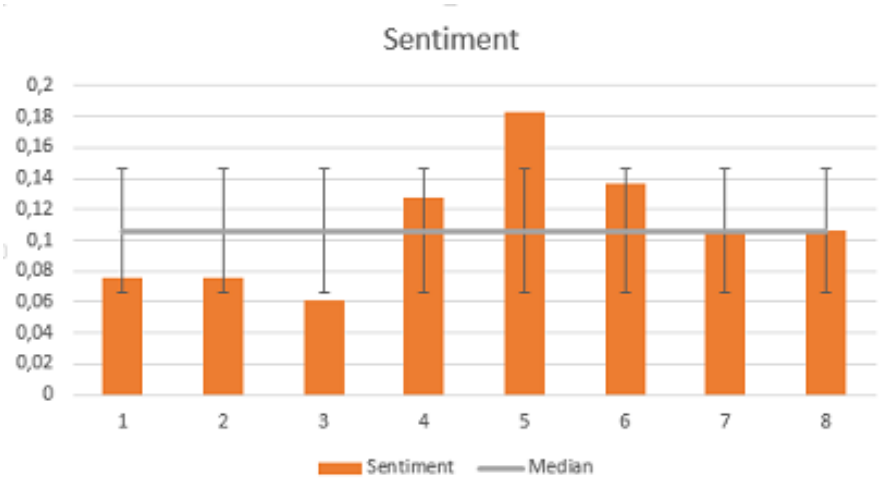

Figure 4.15: Sentiments per quadrant

Trend Analysis

The method to extract the trends for each quadrant was the same than described in section 4.1. Also within the extracted trends from each quadrant a manual data clearing process took place. As for the overall dataset, the Document-Pivot algorithm did not return any trends for the quadrant datasets. The following paragraphs describe the trends for each quadrant.

Trends for Quadrant 1, Figure 4.16 The trends generated by the AlchemyAPI are as follows: *Stuttgart, Baden-Württemberg, rain, Temperature, Humidity, wind, Barometer, messen, km/h* and *Germany*. The LDA algorithm returned the topics *stuttgart, baden, württemberg, love, germany, today, rain, vorhersage, bahnhof* and *ostern*. The results from the SFPM algorithm are *weingarten, württemberg* and *luftfeuchtigkeit*.

Figure 4.16: Tagcloud quadrant 1

When looking at the results of the methods it is obvious that the AlchemyAPI and the LDA algorithm have generated very similar topics. The topics mainly contain of location names like *Stuttgart* and *Württemberg* and information about the weather like *rain* and *Humidity*. Interesting is, that although weather information was detected within each method, the exact terms describing the weather are different within each method.

Trends for Quadrant 2, Figure 4.17 The trends generated by the AlchemyAPI are as follows: *Bayern, München, hPa, munich, km/h, weather, Allianz Arena, f*ck, cloudy* and *MobileCityWalk München*. The LDA algorithm returned the topics *metar, nosig, hPa, munich, germany, happy, world, münchen, flughafen* and *love*. The results from the SFPM algorithm are *münchen, eddm* and *strauß*.

Figure 4.17: Tagcloud quadrant 2

The AlchemyAPI generated topics mostly containing location and weather information as for quadrant 1. The other algorithms also generated locations as trends but also more general terms like *flughafen*. Interesting also that the LDA algorithm generated *Flughafen* as topic whereas the SFPM algorithm generated *strauß* which is a part of the airport name of Munich. The AlchemyAPI also generated the topic *Allianz Arena* which is obvious as in the timeframe there was a soccer match inside this stadium and the participants tweeted about the game.

Trends for Quadrant 3, Figure 4.18 The trends generated by the AlchemyAPI are as follows: *owyposadas, fuck, followback, Salzburg, Austria, PlusReed* and *Adidas*. The LDA algorithm returned the topics *followback, peter, time, krumlov, oberösterreich, linz, world, salzburg, austria* and *good*. The results from the SFPM algorithm are *salzburg, wels, oberösterreich, krumlov* and *linz*.

Figure 4.18: Tagcloud quadrant 3

Although the trends for this quadrant also contain location information, especially the AlchemyAPI also generated other topics like *owyposadas* which is a Twitter user. Because of the low amount of tweets within that quadrant, the AlchemyAPI only generated 8 topics. Although the dataset is smaller than for quadrants 1 and 2 the SFPM algorithm now generated 5 trends.

Trends for Quadrant 4, Figure 4.19 The trends generated by the AlchemyAPI are as follows: *Wien, Vienna, IndyRef, Austria, WeeklyChris, time, Bratislava, love, Adidas* and *photo*. The LDA algorithm returned the topics *wien, vienna, austria, bratislava, good, easter, people, love, foto* and *schönbrunn*. The results from the SFPM algorithm are *schwechat, schönbrunn, niederösterreich* and *pottendorf*.

Figure 4.19: Tagcloud quadrant 4

Also the trends from quadrant 4 contain many location terms within each of the 3 algorithms. The trend *easter* is obvious as the crawling time frame contained the easter time 2014.

Trends for Quadrant 5, Figure 4.20 The trends generated by the AlchemyAPI are as follows: *Adidas, Switzerland, TweetClips, Anton Ewald, Zurich, Allin, Zürich, time, people* and *clothes*. The LDA algorithm returned the topics *switzerland, good, time, zurich, morgen, love, people, originals, follow* and *friend*. The results from the SFPM algorithm are *ewald* and *lszh*.

Figure 4.20: Tagcloud quadrant 5

The trends of this quadrant contain of less location terms than the quadrants 1-4. Although this quadrant is one of the biggest datasets the SFPM algorithm only generated 2 trends. Once again the trend selection of the algorithms is interesting as the AlchemyAPI generated the trend *Adidas* and the LDA algorithm the trend *originals* from the word *Adidas originals*.

Trends for Quadrant 6, Figure 4.21 For quadrant 6 the AlchemyAPI did not generate any topics as no keyword was mentioned more than 10 times within the dataset. The LDA algorithm returned the topics *good, merci, austria, love, hotel, rispondi, neuschwanstein, bolzano, concorso* and *innsbruck*. The results from the SFPM algorithm are *neuschwanstein*.

Figure 4.21: Tagcloud quadrant 6

Within this quadrant an interesting observation is, that the trends also contain Italian terms which is reasonable as Italian is the dominating language within this quadrant.

Trends for Quadrant 7, Figure 4.22 In this quadrant the AlchemyAPI generated only the topic *NashForJanina*, because the tweet amount in this quadrant is very low and therefore the keywords are not used very often. The LDA algorithm returned the topics *hallstatt, photo, night,*

leoben, klagenfurt, kärnten, easter, fenerbahce, sinking ferry and *hotel*. The results from the SFPM algorithm are *hallstatt, 11year, kärnten, leoben, fenerbahce* and *klagenfurt*.

Figure 4.22: Tagcloud quadrant 7

Aslo this quadrant is dominated by trends based on locations. One interesting topic within this dataset is the term *sinking ferry* which represents a real life incident[5]. Within this dataset this is the only appearance of a real world incident as trending topic. So Twitter is not that heavily used for spreading real life incidents like in the work of Kwak, Lee, Park and Moon [KLPM10] stated.

Trends for Quadrant 8, Figure 4.23 For quadrant 6 the AlchemyAPI did not generate any topics as no keyword was mentioned more than 10 times within the dataset. The LDA algorithm returned the topics *ilysm, maribor, disco, world, amazing, graz, steiermark, happy easter, twinkle* and *bra*. The results from the SFPM algorithm are *hemmings, sebi506, graz, twinkle, steiermark* and *maribor*.

Figure 4.23: Tagcloud quadrant 8

[5] http://www.bbc.com/news/world-asia-27045512, accessed 09.04.2015

The topics of this quadrant do not contain that many locations as others do. Also within this dataset the topic *happy easter* appears which is present in other topics as well.

When comparing the results from the different quadrants one observation is, that the trends are mainly dominated by location terms which describe the location where the tweet was published. Recurring topics over more quadrants were the terms *good, love, easter* and *Adidas*. Some of these topics are also trending topics for the overall dataset. From the overall trending topics some of them do not appear in the regional topics, like *Praha*. Nevertheless the trending topics within each quadrant are very different, which leads to the assumption, that trending topics on Twitter are differing based on the region where the post was published.

4.2 Text Mining Validation

The algorithms which were used for topic generation and for sentiment analysis also have some problems, especially with datasets like tweets. These problems are described in section 4.2. The algorithms also need to be validated and interpreted manually, which is done in section 4.2.

Problems of the APIs

The method which was used to detect trending topics with the AlchemyAPI is a very simple one as the keywords which were extracted by the API in the next steps were counted and the most used keywords were defined as trending topics. In contrast to the other used algorithms, which also use near duplicates to define trends, therefore this algorithm is more strict in trend detection and in smaller dataset it is more likely to not return sufficient trends as the number of keywords which are used more often is low. In this methods also keywords like *Adidas* and *Adidas original* are handled separate and therefore possibly not defined as trends, as the number of occurrences is lower for each of the two terms.
Another problem in this method is, that the AlchemyAPI supports keyword extraction only for some languages. As the dataset for this work contains many different languages, the trends are biased towards the supported languages. As most of the tweets still are in English or German the algorithm works for generating trends, but some trends are ignored by this method as one can see when comparing the generated trends from the other algorithms.
Also when it comes to sentiment analysis this language problem is present and therefore not all posts are sentiment rated. Therefore the overall sentiment value is biased as well towards the supported languages. Another problem for the sentiment analysis is, that as tweets are very short texts, it is hard to detect sentiment. Also some tweets mostly contain special characters, here the sentiment detection also struggles.

One problem of the text mining algorithms described by Aiello et al [APM+13] is, that they struggle with the processing of text in different languages as they mostly are based on methods which use co-occurrences of words. As it is not likely that words in different languages are the same this is one limitation of the algorithms. Therefore it is recommended to use these algorithms for datasets with only one language. Because of the high amount of languages also the

resource demand for the trend calculation is very high.
Another remark which applies to the crawled data is, that the dataset contains only few duplicates or near duplicates. As the amount of text per tweet is very limited due to the post size constraint of 140 characters it is not very likely that frequent co-occurrences of terms happen. Normally this is not that relevant for Twitter data tough as many retweets and copied messages are posted. Within this dataset the amount of retweets is quite low, therefore the algorithms struggle with the noisiness of the dataset.
When looking at the LDA algorithm it shows, that this was the method, which returned the expected amount of topics for all datasets which were analyzed with that algorithm. This is because the algorithm defines each tweet as topic and then assigns all terms within the dataset to the tweets. After some training iterations it then defines the most clustered topic as trending topic. This method also works for short texts but in the result often topics like *at, co* were generated.
The DocP algorithm did not return any topics for each of the datasets. This is, because the algorithm clusters the whole tweet and as the dataset did not contain many duplicates like retweets the algorithm did not detect any trending topics.
The SFPM algorithm also returned several topics, but the amount of topics was smaller than with the LDA algorithm. The algorithm uses co-occurrences of any number of terms (possibly more than 2) within tweets. Therefore it is obvious, that the algorithm returns less topics, as co-occurrence of e.g. 3 terms from one tweet in another tweet is less likely than co-occurrence of one term from one tweet.

Validation of the Results

The validation of the text mining algorithm needs to be done manually, as no 100% accurate method is known yet for topic detection and sentiment analysis. When looking at the sentiment values, the automatic rating is a tricky task, as an algorithm always will struggle with issues like sarcasm in posts. For the validation an sample of 10 tweets was extracted, 5 positive and 5 negative rated in the languages English, German and French.

1. Positive (0.363252): *Dank unseren Sponsoren, die die 9. SOMENIKA ermöglichen: @karlshochschule u. @KarlsruheTweets (Stadtmarketing KA) #smcka*

2. Positive (0.209081): *@Claus_Pandi @mehrenhauser Ja. Bring ein Sackerl für sein Gackerl. #ballhausplatz*

3. Positive (0.667051): *Got this from my Kindergarden. :D they are so sweet. #Easter love it. @ Kindergarten http://t.co/kJEUvWGDK4*

4. Positive (0.586015): *@AlexandreLorot France here!This Brige&every mega-structures are always impressive.If U like Breaking Bad stuff and Fun U know what to do :)*

5. Positive (0.091211): *@Alexaandre_hb @_AfterTheDream MDR Claire ! Non mais je ne comprends pas comment on fait pour être aussi con !*

6. Negative (-0.126261): *@_shipperin_ DU HAST DAS NEUE PROFIL DESIGN OMG NEIN DAS BEDEUTET DASS ICH ES AUCH BALD HABEN WERDE*

7. Negative (-0.023338): *Was für ein verrücktes Wetter???*

8. Negative (-0.693377): *So keeping myself busy resulted in a relapse of glandular fever and now I'm just been to the gym. #seriouslydead #reallydead #died*

9. Negative (-0.263729): *Barca's funeral. #CopaDelRey*

10. Negative (-0.351243): *@kevinkeves @mariejully08 si c'est sa et qu'ils immobilisent une vingtaine de flics pour sa..... Pitié... Bientôt yaura le GIGN aussi*

The tweet 1 has a positive sentiment and when looking on the text it can be seen as positive as someone thanks for a sponsoring. Tweet 2 was also rated positive, but no real sentiment is detectable in the tweet. Post 3 is rated very positive which is true as the publisher used a laughing smiley, said that the kindergarden kids are sweet and he loves easter. Also tweet 4 was rated positive and is positive as well which the use of the smiley proofs. Tweet 5 which was rated slightly positive is in real life a quite negative post as the publisher asks another person how he could be so stupid. So overall from 5 positive posts, 3 were really positive, 1 was neutral and 1 was rated wrong as it was negative.
For tweet 6 the negative sentiment is accurate as the person somehow complains about the new profile design which will be applied shortly to it's profile. Tweet 7 was rated slightly negative. When looking on the tweet it can be seen as neutral as the person asks about the crazy weather, but no real sentiment is detectable. As the automatic rating for negative is very small, this can also be seen as a valid rating. Post 8 was rated very negative and is also quite negative as someone states that he got a relapse of glandular fever. Also post 9 is negative in reality as someone makes a statement about a soccer match where Barcelona lost. Therefore the sentiment is accurate again. Tweet 10 tells about immobilization as well as about police and what a pity this is also this negative sentiment rating is accurate. When looking over all these negative posts, the sentiment rating for the negative posts are valid in every sample.
The result for the sentiment rating is, that the automated approach works well for negative tweets but has some problems with positive tweets. Therefore the overall sentiment of 0.1267 must be treated with caution as the algorithms tend to classify more positive than the reality is. This is also validated in the work of Singh et al [SPUW13].

When looking at the overall trends generated by the AlchemyAPI many of them represent real world locations which are accurate as trends, as person often check-in at places and post about places they were. Also other trends like *Bayern* are accurate trend as Bayern Munich played several soccer matches within the crawled data timeframes and people tweeted about this games. Overall the trends generated by the AlchemyAPI tend to be very valid although they mostly represent locations. This also applies for the overall trends genereated by the LDA algorithm. This trend also contains the term *Easter* which is a valid trend, as the Easter time was within the crawling time. Also here, the trends are dominated by locations. Also the trends from the SFPM algorithm contain some locations, but not that much as the other algorithms did. Also

here the term *Ostern* appears, which is the translation for *Easter*. Overall the trends generated by the LDA and SFPM methods seem to be more accurate as they do not mostly contain location information.
When looking at the trends per quadrant some of the observations of the overall dataset, like the location trends, are also observable. In quadrant 1 many trends contain weather information. This is due to the fact, that within that location a meteorological office tweeted weather information per day, which influenced the trends. In quadrant 2 the soccer game influenced the trends with the terms *Bayern* and *Allianz Arena*. Quadrants 3, 4, 6 and 7 are dominated by location information. In quadrant 5 on trend is *Anton Ewald*, who is a Swedish singer. He released a new single in June 2014 which was in the crawling time. Therefore this trend also represents a real world event. The trends for quadrant 8 also contain the term *Easter*.
Overall the trends are biased towards location trends, but also represent some real world trends. So the trends are valid trending topics for Twitter. There are no big similarities in the regional trends, so one can assume, that the tweet topics vary based on the region of the publisher.

4.3 Summary

When looking at the retrieved data an interesting observation is, that the Twitter API also returns tweets which are outside of the specified geo locations. The number of tweets with wrong location information is significantly high. Although the majority of tweets is not published within the biggest cities in the dataset, the tweets are clustered around the big cities. The four quadrants containing the biggest cities contain far more tweets than the other four.
The tweet frequency over time increased for the two datasets in 2014 where Twitter seems to have had a peak in usage. Compared to data from 2015 the amount of tweets already decreased. The dataset contains overall 53 languages which makes trend detection and sentiment analysis very cumbersome. Nevertheless the German and English tweets dominate the dataset. The used platforms to publish tweets differ from the real world platform distribution - therefore a sample of Twitter users is not representative for the real world population.
Twitter is nowadays used more conversational than in the past, so more posts are either replies or directed to another user and not just messages without connection to another user. This is shown by comparison of this work with the work of Huberman et al [HRW08] or Java et al [JSFT07]. The overall sentiment is more positive, but this is biased towards the languages of the tweets as the sentiment rating only works for three languages from the dataset. The sentiment algorithm also tends to rate tweets more positive, as they are in real world. When looking on the trends, three trend detection methods performed well and returned accurate trends. Only the Document-Pivot algorithm did not return any trends. The trends are valid trending topics for Twitter. When looking on regional differences within the dataset one can see, that the tweeting behavior differs between each region. There are some average regions, but also some where other languages were used or where the sentiment is more positive or negative. The conversational aspect is quite equal within each region. The trends per region are also quite different, but there are some region independent trends like *Easter*.

CHAPTER 5

Conclusions and future work

5.1 Conclusions

In my thesis I gave an extensive analysis of the usage of the microblogging platform Twitter in Austria. The analysis shows, that the peak of usage of Twitter in Austria already was reached and is declining again at the moment. Twitter is used heavily for conversational messages. Here a big increase to the past can be seen. Also the sharing of information increased over time. When looking at the languages and platforms people use Twitter with, it shows, that the dataset of Twitter users is not representative for the real worlds population. When looking at the regional differences in sentiment and trends it shows, that for each regions, the topics as well as the sentiment is varying. Another observation is, that the majority of tweets is not posted within the biggest cities. Nevertheless the tweets are clustered around these cities.

For this work I crawled Twitter data based on location information. So every tweet should be crawled which was posted inside a bounding box surrounding Austria and the near border areas. Here it showed, that Twitter returns also tweets which were posted outside of this bounding box. To these tweets then several text mining algorithms were applied. The AlchemyAPI was used to detect sentiments and to identify trends. A big restriction of this API is the lack of support of multiple languages, so the results of this API are limited to few languages. The other algorithms which were used for trend detection are the Document-Pivot Topic Detection (DocP), the Latent Dirichlet Allocation (LDA) and the Soft Frequent Pattern Matching (SFPM). The DocP method did not return any trends, but the LDA and the SFPM algorithms performed very good in trend detection. This data afterwards was analyzed with respect to various dimensions.
I plotted the tweets to a map to see differences in tweeting behavior. After this, one can see, that the tweet heatmaps show clusters where bigger cities are located. But when looking into details the amount of tweets posted in the cities is not bigger than the number posted outside. Nevertheless, the tweets are clearly clustered around the cities. Another aspect I analyzed was the frequency of tweets. Whereas there was an increase measurable in the data April 2014 and June 2014, there is a decrease when comparing with tweet amounts from March 2015.

Another analysis dimension is the languages which were used for posting tweets. Although most of the tweets were posted in English in German there are some interesting observations with respect to the languages. For instance in the region around Vienna the amount of Turkish tweets is nearly the same than for the German ones. Another example would be, that in the region also containing Italian areas, the amount of Italian tweets is the highest, with a distance to English and German tweets. Also when looking at the platforms used for tweeting there are differences to the real life mobile platform market share. Therefore a dataset of Twitter users is not representative for a real world population.
In this work I also then analyzed the data with respect to regional differences. Therefore I clustered the data into 8 quadrants. The results show, that the trends and sentiments are very different for each region. There are only some common topics. In simplified terms, although people in e.g. Tyrol or Upper Austria post about different topics than people in Vienna do, they also post about different topics compared to other regions which are geographically nearer, like Munich. This leads to the assumption that the usage of Twitter is highly dependent on the region the user is when tweeting.

My work gives a first overview about the overall usage of Twitter in the Austrian region and evaluates several text mining methods. It also shows, that the usage of Twitter is not rising anymore and therefore the importance of the platform won't increase in the near future.

5.2 Future Work

This analysis of the Twitter usage is limited to some extend. Therefore there are several improvement areas for the future which are outlined in the following enumeration:

1. **Data Acquisition**: The dataset for this work contains not only data from Austria, but also for the border areas. Therefore the crawling algorithm could be optimized to only crawl data which was posted within Austria. Also the posts without coordinates but with an associated place which lies in Austria can be included in the analysis. Within this work the data was crawled very unsteady. This can help to avoid a bias towards a specific event. Nevertheless on improvement could also be, to crawl data constant over time also to get better comparisons for analyzing changes in tweeting frequency.

2. **Sentiment Analysis**: This work performed the sentiment analysis only with one algorithm. This algorithm is limited to the languages English, German and French. An area of improvement would be, to also use other sentiment algorithms, to on the one hand validate the sentiment results but also to get analysis for other languages.

3. **Trend Detection**: The trend detection struggles with very noisy datasets like I used within this work. Therefore increasing the amount of data would be helpful to make the trend detection more accurate. Also a data pre-processing would be helpful to improve the trend results. Within this pre-processing tweets and terms, that are not relevant for trend detection can be filtered out.

4. **Demographic Analysis**: A part which was not topic of this thesis is a detailed demographic analysis of the users which post tweets. Also this can be included in a future work.

APPENDIX A

Appendix

List of Tables

List of Figures

Bibliography

[AHSW11] Sitaram Asur, Bernardo A. Huberman, Gábor Szabó, and Chunyan Wang. Trends in social media : Persistence and decay. *CoRR*, abs/1102.1402, 2011.

[APM+13] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker, Ioannis Kompatsiaris, and Alejandro Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, 2013.

[BCF12] Jacob Blasbalg, Ryan Cooney, and Steven Fulton. Defining and exposing privacy issues with social media. *J. Comput. Sci. Coll.*, 28(2):6–14, December 2012.

[BGL10] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[BK13] James Benhardus and Jugal Kalita. Streaming trend detection in twitter. *IJWBC*, 9(1):122–139, 2013.

[BMBL09] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

[BOS12] David Bamman, Brendan O'Connor, and Noah Smith. Censorship and deletion practices in chinese social media. *First Monday*, 17(3), 2012.

[CBK12] Luying Chen, Michael Benedikt, and Evgeny Kharlamov. Quasar: querying annotation, structure, and reasoning. In Elke A. Rundensteiner, Volker Markl, Ioana Manolescu, Sihem Amer-Yahia, Felix Naumann, and Ismail Ari, editors, *EDBT*, pages 618–621. ACM, 2012.

[CGCH13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff, editors, *ICWSM*. The AAAI Press, 2013.

[FJK+08] Tim Finin, Anupam Joshi, Pranam Kolari, Akshay Java, Anubhav Kale, and Amit Karandikar. The Information ecology of social media and online communities. *AI Magazine*, 29(3):77–92, September 2008. (PREPRINT).

[FPB+13] Flavio Figueiredo, Henrique Pinto, Fabiano BelémM, Jussara Almeida, Marcos GonçAlves, David Fernandes, and Edleno Moura. Assessing the quality of textual features in social media. *Inf. Process. Manage.*, 49(1):222–247, January 2013.

[Fuc11] Christian Fuchs. New media, Web 2.0 and surveillance. *Sociology Compass*, 5(2):134–147, February 2011.

[Gil10] Curt Gilstrap. Social media timeline. `http://socialmediacertificate.net/2010/12/social-media-timeline/`, December 2010. [Online; accessed 2014-07-22].

[Gof59] Erving Goffman. *The Presentation of Self in Everyday Life*. Anchor, June 1959.

[HH09] Courtenay Honeycutt and Susan C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42). Los Alamitos, CA.*, pages 1–10, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[HRW08] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope, 2008. cite arxiv:0812.1045.

[JSFT07] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[KH09] Andreas M. Kaplan and Michael Haenlein. The fairyland of second life: Virtual social worlds and how to use them. *Business Horizons*, 52(6):563–572, 2009.

[KH10] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1):59–68, January 2010.

[KH11] Andreas M Kaplan and Michael Haenlein. The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2):105–113, 2011.

[KLPM10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[LH08] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 915–924, New York, NY, USA, 2008. ACM.

[Mil67] Stanley Milgram. The small world problem. *Psychology Today*, 67(1):61–67, 1967.

[MK10] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.

[MLA+11] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of twitter users. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.

[MMG+07] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[Mur11] Dan Murphey. Inspired by tunisia, egypt's protests appear unprecedented. `http://www.csmonitor.com/World/Backchannels/2011/0125/Inspired-by-Tunisia-Egypt-s-protests-appear-unprecedented`, 2011. [Online; accessed 20-August-2014].

[OEC07] OECD. Participative web and user-created content. web 2.0, wikis, and social networking. 2007.

[O'R06] Tim O'Reilly. Web 2.0 compact definition: Trying again. `http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html`, December 2006. [Online; accessed 2014-07-22].

[PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[QAC12] Daniele Quercia, Harry Askham, and Jon Crowcroft. Tweetlda: Supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 247–250, New York, NY, USA, 2012. ACM.

[Rie14] Brian Ries. Twitter blocks pro-ukrainian political account for russian users. `http://mashable.com/2014/05/19/twitter-blocks-account-russia/`, 2014. [Online; accessed 20-August-2014].

[SB09] L. Safko and D.K. Brake. *The Social Media Bible: Tactics, Tools, and Strategies for Business Success*. Wiley, 2009.

[SDX12] Stefan Stieglitz and Linh Dang-Xuan. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and

retweet behavior. In *Proceedings of the 2012 45th Hawaii International Conference on System Sciences*, HICSS '12, pages 3500–3509, Washington, DC, USA, 2012. IEEE Computer Society.

[SEK+13] H A Schwartz, J C Eichstaedt, M L Kern, L Dziurzynski, S M Ramones, M Agrawal, A Shah, M Kosinski, D Stillwell, M E Seligman, and L H Ungar. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*, 8(9), 2013.

[SFKW09] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger. Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, pages 35–48, New York, NY, USA, 2009. ACM.

[SG03] Hope J. Schau and Mary C. Gilly. We Are What We Post? Self-Presentation in Personal Web Space. *The Journal of Consumer Research*, 30(3):385–404, 2003.

[SHA11] Hassan Saif, Yulan He, and Harith Alani. Semantic smoothing for twitter sentiment analysis. In *The 10th International Semantic Web Conference (ISWC)*, Bonn, Germany, 2011.

[She09] Ali Sheikholeslami. Iran blocks facebook, twitter sites before elections (update1). `http://www.bloomberg.com/apps/news?pid=newsarchive&sid=anh.uW3gNZp4`, 2009. [Online; accessed 20-August-2014].

[SO13] Tasos Spiliotopoulos and Ian Oakley. Understanding motivations for facebook use: Usage metrics, network structure, and privacy. 2013.

[SPUW13] V.K. Singh, R. Piryani, A Uddin, and P. Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on*, pages 712–717, March 2013.

[UKBM11] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

[WH11] Michel Walrave and Wannes Heirman. Cyberbullying: Predicting victimisation and perpetration. *Children and Society*, 25(1):59–72, 2011.

[Wik14] Wikipedia. Censorship of twitter — wikipedia the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Censorship_of_Twitter&oldid=621549634`, 2014. [Online; accessed 20-August-2014].

[WNHC12] Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*, 44(4):25:1–25:24, September 2012.