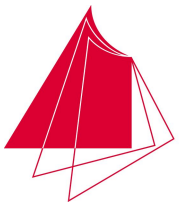

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Mining High Quality Insights in Social Media Data using Machine Learning Methods

Trend Detection and Analysis on Twitter

Scientific Report

Course of Studies: Information Technology

University of Applied Sciences Karlsruhe

by

Lukas Masuch

Henning Muszynski

Benjamin Raethlein

Due Date:	30. January 2015
Student (Id):	Lukas Masuch (50669)
Student (Id):	Henning Muszynski (50170)
Student (Id):	Benjamin Raethlein (50169)
Academic Supervisor:	Prof. Dr. Norbert Link
Academic Supervisor:	Dr. Ingo Schwab

Contents

List of Figures

List of Listings

1 Introduction

1.1 Motivation

The immense rise of social media is one of the driving forces behind the current Big Data trend. With 2.5 billion gigabytes of data created every day and 90% of the worldwide data created in the last two years [IBM12], Big Data has become a top priority for research organizations and companies. The combination of Big Data and powerful analytical technologies makes it possible to gain highly valuable insights that otherwise might not be accessible.

The popularity of social media services, including social networks, micro-blogging tools, wikis, and photo and video-sharing applications has increased exponentially in the last few years [Cam+13a]. Social media allows individuals and organizations to capture and understand the thoughts, opinions, ideas, conversations and feelings of millions of people. As social media services continue to proliferate, the amount of unstructured social data keeps growing.

Emerging Big Data Analysis and advanced Natural Language Processing technologies make it possible to collect and analyze a massive amount of data and enables a fundamentally new approach for the study of society and human beings.

When hurricane Sandy hit the US Eastcoast on October 29 2012, government agencies and individuals turned to social media services "to communicate with the public like never before" [Coh13]. Hurricane Sandy "marked a shift in the use of social media in disasters" [Sec13, p. 6] and attracted many data researchers to monitor and analyze this event [Kum+11; Car+14]. Besides the analysis of natural disasters, big social data analysis has been shown to be useful for many other use cases: The FBI utilizes advanced data analytic technologies to predict crimes and terrorist attacks based on publicly available social data [WGB12]. Several research projects leveraged those technologies on big social data to predict the spread of diseases [Gin+09; Goo14]. Moreover, social media analysis has been proven to predict political sentiment and forecast election winners [BS11]. These successful results of mainly research-based projects helped to open up new business opportunities. Companies already use social media monitoring and analysis techniques to predict the stock market in real time [BMZ11; Alc13]. Further, an increasing number of companies utilize these technologies to analyze the customer satisfaction and research

the public opinion about products and their company itself [Cam+13b]. In addition, newspaper publishers use big social data analysis to mine the public interest and predict how popular their stories might become. In general, the detection, analysis and prediction of trends in an early stage makes it possible to identify big headlines before they go viral. Therefore, detection is very valuable in different areas such as stock market, brand awareness, political discussions and elections and the success of media.

Big social data analysis has grown into a serious business over the past several years and nowadays includes disciplines such as social media analytics, sentiment analysis, social network analysis, trend discovery and opinion mining.

1.2 Objectives

The goal of this project is the early detection and analysis of arising trends on Twitter. The desired outcome is a concept of a data pipeline and a full implementation of several components to collect, analyze and visualize Twitter data. In addition, this system shall be able to detect trends in an early stage to predict the spreading of future trends on Twitter based on the curves and features of trends in the past. Therefore, we want to explore different metrics and dimensions, such as retweets, hashtag and user mention occurrences, user groups and emotions. Furthermore, this paper explains various related implementations, recommends appropriate technologies and common methods to conduct a big social data analysis on Twitter trends. Our implementation will be presented based on the case study of Twitter data related to several major events happened between 15$^{\text{th}}$ of December 2014 and 15$^{\text{th}}$ of January 2015. Additionally, we compare our results with data from Google Trends to check for similarities and the validity of detecting trends on Twitter based on a small statistical sample.

1.3 Overview

This paper is structured into five chapters. The remaining part of the paper is organized as following.

Chapter 2 provides the theoretical background of this paper. Used terms such as Big Data, Social Media and several data processing techniques are described in this chapter.

The architecture to collect Twitter data and several techniques to analyze and visualize those data are presented in **chapter 3**. Furthermore, we mention some related research and technologies.

In **chapter 4** we analyze three detected Twitter trends with several analysis techniques and explain the gained insights from them.

Chapter 5 provides a conclusion and an overview about possible future work.

2 Theoretical Background

2.1 Big Data

The term *Big Data* describes an enormous amount of data, which cannot be stored, managed or analyzed with conventional database tools [Com11]. Big Data can include different types of data, such as enterprise, machine-generated, sensor or social data [Ora13, p. 3].

In the last few years, the analysis of Big Data became an essential aspect for many companies. Big social data analysis enables these companies to get more information about their customers' sentiment, satisfaction or opinion by collecting and analyzing data from social media services [Ora13].

2.2 Social Media

The term *Social Media* belongs to web applications such as "social networks, blogs, multimedia content sharing sites and wikis" [GS13]. Social networks such as Facebook, Twitter or Google+ are used by an increasing number of people. In September 2013, 73% of online participants used at least one social networking site, of those 71% were active on Facebook and about 19% on Twitter [Cen14]. Twitter, a popular microblogging service with over 284 million active monthly users[1], allows anyone to instantly post 140-characters text messages. Thereby, up to 500 million public Tweets are generated per day in more than 35 languages about nearly any imaginable topic[1].

Social media applications enable people to connect with others as well as to publish content such as their interests, opinions, knowledge and ideas. During the past several years, user-generated-content has become more and more popular, which means that the users participating more in content creation, rather than just content consumption [Agi+08, p. 1]. This behavior leads to an continuously increasing amount of unstructured social data and makes it impossible for humans to read through and analyze this immense amount of unstructured data. Therefore, advanced data mining and analysis techniques are necessary.

[1] `http://about.twitter.com/company` [Online; accessed 12-01-2015]

The traditional approach to gain insights into society, human beings and social relations required "questioning a large number of people about their feelings" [Fla+12, p. 1]. In contrast, social media applications can provide those valuable information about the public "due to the fact that people use them to express their feelings" [Fla+12, p. 1]. For instance, Twitter attracted many professionals to collect and analyze Tweets to gain valuable insights on anything from stock market to natural disasters by offering free API's to access this huge amount of unstructured data.

2.3 Data Analysis

Machine learning describes methods that enable computer systems to automatically learn from empirical data [Dom12; Ins11]. Machine learning methods usually focus on the prediction and classification of information, based on training data that contains truthful information. A wide variety of applications exists for machine learning on big social data such as natural language processing, topic detection, text classification and sentiment analysis.

Data mining is the process of finding valuable insights from large datasets. Therefore, data mining techniques try to extract meaningful patterns and associations in datasets by utilizing artificial intelligence, machine learning or statistical methods [HKP12]. In general, data mining is used as a synonym for the process of discovering knowledge from mostly unstructured data [HKP12, pp. 6 sqq.].

Natural Language Processing (NLP) is "a set of techniques [..] to analyze human language" [Ins11, p. 29]. Those techniques are often based on machine learning or statistical methods that enable computer systems to derive meaning from natural language. Therefore, NLP methods need to analyze and understand the syntax, semantics and the context of a sentence [LM11]. Common application areas of NLP include stemming (described in chapter 3.3.1), named entity recognition[2] and sentiment analysis.

[2] Method to recognize well-known entities (e.g. person, location) in text.

2.4 Trend Detection

Trend detection methods are used to detect emerging topics or trends by using Data Analysis methods. The keyword frequency approach is a popular method to discover trends in a big amount of unstructured text data [Kim+13]. In the majority of cases, the input data is preprocessed to remove meaningless characters and words, as well as to prevent duplicated terms. After preprocessing the input data, the remaining words are ordered by their frequency of occurrence and the top k words stand out as trending keywords [Kim+13, pp. 213 sq.].

Many social networks use hashtags to categorize social content, represent a topic or event and help users to discover certain content. A hashtag consists of "a sequence of non-whitespace characters preceded by the hash character" [TR12, p. 644; ZWL13, p. 1427] (e. g. *#NewYear* or *#AirAsia*). Hashtags are well suited for trend detection by measuring the number of uses in a certain time interval [ZWL13, p. 1427]. Using Trend Detection methods on social media content is an effective way to discover frequently used keywords and to show emerging topics in real-time [Kim+13; KML13].

A prominent example for large scale detection and prediction of trends based on huge amount of social data is *Google Flu Trends*. Seasonal influenza, responsible for millions of illnesses and up to 500 thousand deaths per year, is known as a major health issue all over the world. An early detection of epidemics would reduce the significant effect of the pandemic and seasonal influenza. The project Google Flu Trends aims to monitor flu cases in real time based on various keywords with a strong correlation to the appearance and spread of seasonal flu to predict flu trends by analyzing social datasets [Gin+09; Web14; Tec14, p. 1]. Figure 1 visualizes the predicted flu activity in the United States in 2014.

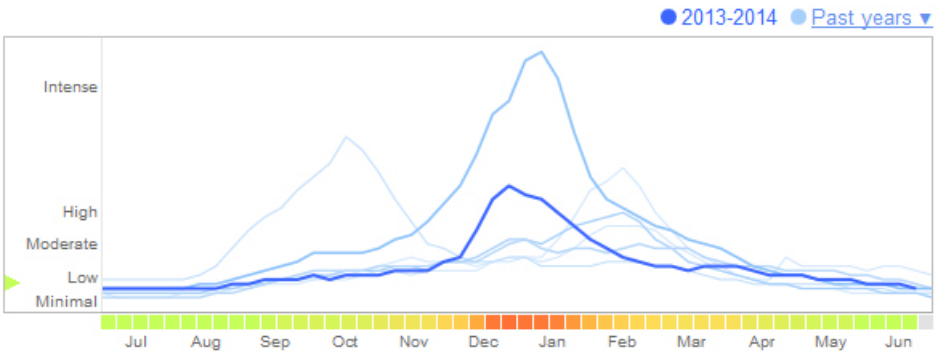

Figure 1: Flu activity in the United States [Goo14]

3 Trend Analysis on Twitter: Concept

3.1 Related Work

The analysis of microblogging data has been shown to provide new and not otherwise attainable information and it is, therefore, an important resource for big social data analysis. There are various tools to collect, analyze and visualize certain aspects of Twitter data. The MapD tweetmap[3] enables users to analyze nearly 350 million historical geolocated Tweets from January 2011 to September 2013 in milliseconds and visualize the results on a map. Sentiment Viz is a web application that allows tracking certain keywords to analyze the sentiment of corresponding tweets in real-time and visualizing the results using different techniques [HR13]. The Arizona State University developed the TweetTracker and TweetXplorer tools to track, analyze, visualize and understand the activity on Twitter. TweetTracker is "capable of monitoring and analyzing location and keyword specific Tweets with near-real-time trending, data reduction, historical review, and integrated data mining tools" [Kum+11, p. 1], whereas TweetXplorer provides a comprehensive set of effective visualization techniques [Mor+13]. Furthermore, other tools are specialized in specific use cases such as the weather sentiment prediction application[4] for analyzing the sentiment about the weather at a specific location, and trendsmap[5] for visualizing upcoming localized trends on a map. Furthermore, the company *Dataminr* scans Twitter for relevant messages characterized by "the right combination of language, context and location" to detect "breaking- and money-making-news" [Alc13].

Moreover, Naaman et al. used Twitter to "identify important dimensions according to which trends can be categorized, as well as the key distinguishing features of trends that can be derived from their associated messages" [NBG11]. They performed their analysis on a previously collected dataset of 48 million tweets. In contrast, our projects aims to achieve results on a smaller dataset and based on real-time data instead of historical data. Zubiaga et al. focused on the classification problem

[3] `http://mapd.csail.mit.edu/tweetmap-desktop` [Online; accessed 12-01-2015]
[4] `http://www.sproutloop.com/prediction_demo` [Online; accessed 12-01-2015]
[5] `http://trendsmap.com` [Online; accessed 12-01-2015]

by "introducing a typology of trending topics, and providing a method to immediately classify trending topics as soon as they appear on the homepage of Twitter" [Zub+11].

3.2 Limitations

For this case study, Twitter is used as the only data source. However, other social media sources for additional public social data could easily be integrated into the current data flow. This case study is limited to only collect tweets in English language since NLP in English is more advanced, offers a proper comparison and is simpler to use. In addition, the Twitter Streaming API is restricted to to a small fraction of the total volume of tweets at any given moment[6]. In addition, we restricted our dataset only on geo-tagged tweets from the United States to enable a more advanced geospatial analysis, to concentrate on more unified trends from US and to prevent spam content[7].

3.3 Analysis Methods

3.3.1 Data Preparation

Stop word removal describes the process of removing the most common words out of a text. Words like *to*, *the* or *a* have little influence in any analysis and are most of the time omitted to avoid unnecessary indices and clean the dataset. Normally so called *stop lists* are defined containing all words which should be removed before the analysis [MRS08, p. 27]. However, in some cases it can be dangerous or simply wrong to remove too many stop words or to remove stop words at all. For example when searching for some "well known pieces of verse consist entirely of words that are commonly on stop lists (To be or not to be, Let It Be, I don't want to be, ...)" [MRS08, p. 27].

Stemming is used to bring related terms and words to a common base form. This is often needed when texts are analyzed and words in different forms are used like *am*, *are* or *is* a stemming algorithm would then find the common base form as *be* [MRS08, p. 32]. Different approaches for stemming exist, like just cutting off the ends of words and hoping for a good result. More advanced approaches try to find

[6] `https://dev.twitter.com/faq` [Online; accessed 12-01-2015]
[7] We assume that most spam content on Twitter is not tagged with location information.

the correct base "with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only" [MRS08, p. 32].

A **bag of words model** describes a technique to analyze documents by counting and weighting their words. Each word can be a so called bag. The technique can be further enhanced to include weighting of words, for example stop words should have less weight. Another approach is to use a stemming algorithm on each word in order reduce the amount of bags [MRS08, p. 117].

3.3.2 Sentiment Analysis

Sentiment Analysis is a widely used NLP technique to analyze social media data. Therefore, many companies, such as AlchemyAPI[8], ViralHeat[9] and TextAlytics[10], offer commercial web services to detect sentimental information of any text data by utilizing machine learning techniques. Several open-source machine learning toolkits, e.g. Weka[11] and Mallet[12], offer similar algorithms that can be trained to classify and compute the corresponding sentiment. Further, these libraries are also suited for topic modeling, information extraction and pattern recognition on big social data.

Figure 2 shows the sentiment analysis for the comedy movie "The Interview". The movie features North Korean dictator Kim Jong-un and was detected as a trend on Twitter after hackers attacked Sony Pictures Entertainment and demanded the cancellation of the planned release of the movie. The sentiment analysis shows the three clearly distinguishable sentiments *negative*, *neutral* and *positive*. There are a lot of negative reactions to the movie itself, but not about Sony Pictures putting it online for everyone to watch. About Sony's decision to put the movie online and only show it in selected cinemas was mostly reported with a neutral sentiment.

[8] `http://alchemyapi.com` [Online; accessed 12-01-2015]
[9] `http://viralheat.com` [Online; accessed 12-01-2015]
[10] `http://textalytics.com` [Online; accessed 12-01-2015]
[11] `http://cs.waikato.ac.nz/ml/weka` [Online; accessed 12-01-2015]
[12] `http://mallet.cs.umass.edu` [Online; accessed 12-01-2015]

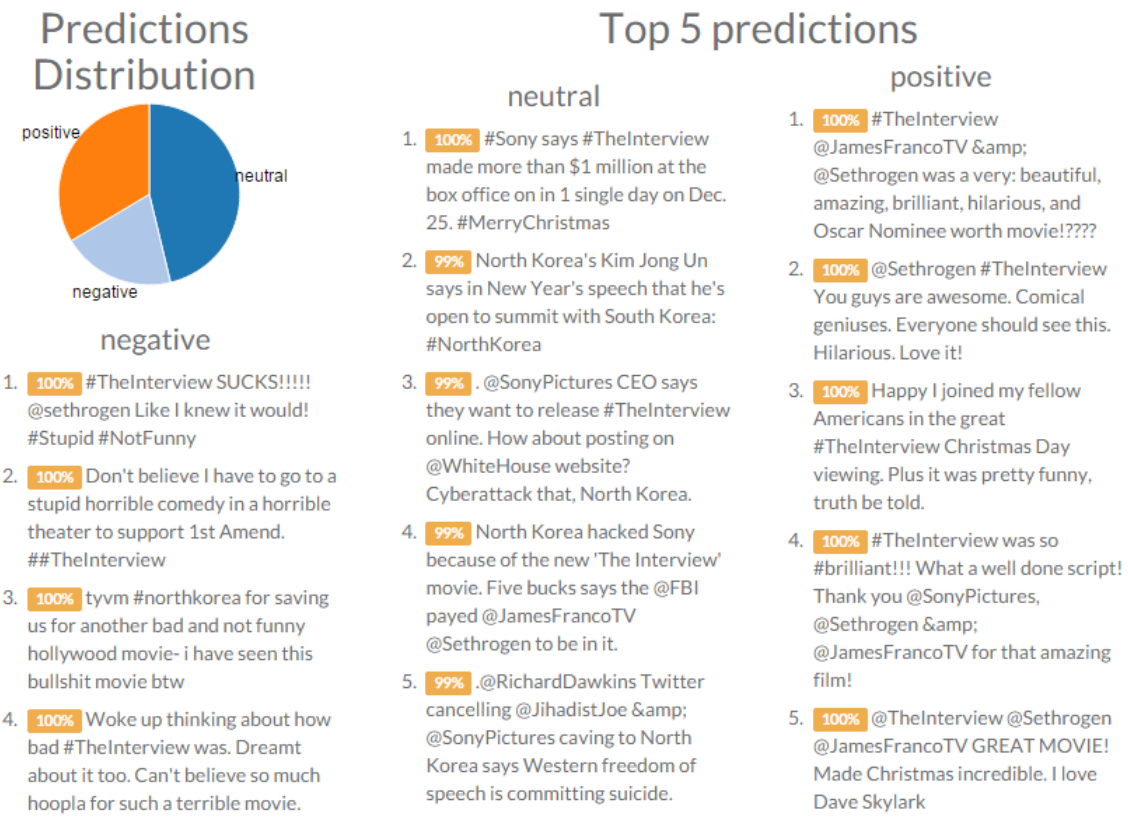

Figure 2: Sentiment - Sony Hack Concerning the Movie “The Interview”

3.3.3 Topic Modeling

Topic modeling is a statistical machine learning model for the automatic discovery of abstract topics occurring in a collection of documents (content entities). Moreover, it allocates the analyzed documents to the discovered topics and clusters the most common words (terms). LDA is a common method of topic modeling introduced by Blei et al. [Ble+03]. The LDA method assumes that each document contains a mixture of topics where each word attributes to one of these topics [Ble+03]. The parameters for executing LDA for topic modeling are a collection of documents, a specified number of topics and specified number of iterations.

When analyzing detected trends on Twitter, we utilize LDA to find all topics related to hashtags. The collected tweets are the documents needed for LDA and the parameters topic count and iteration count are varied depending on the trend. Ramage et al. concluded in their research paper that the 140 characters of a tweet are sufficient as a document for LDA [RDL10].

Listing 3.1 shows the topics that were modeled by using LDA on all tweets related to the comedy “The Interview”. The results are based on nearly 5000 tweets, 5 topics and 1000 iterations. Every found topic describes another aspect of this trend. The

```
1 theinterview (1109) jamesfrancotv (782) sethrogen (573) movie
     (308) interview (204) funny (182) hilarious (148)

2 northkorea (253) sonyhack (140) korea (125) north (117)
     internet (75) sony (72) amp (57)

3 theinterview (695) sonypictures (264) sony (234) movie (147)
     korea (121) north (111) interview (100)

4 theinterview (223) aint (96) hate (76) cuz (58) jealous (52)
     anus (33) peanutbutter (26)

5 theinterview (537) christmas (132) day (89) theaters (67)
     freedom (66) theater (65) showing (57)
```

Listing 3.1: Topic Model for the Sony Hack Concerning the Movie "The Interview"

first topic probably covers recommendations for the movie itself while the second topic is about the hack of Sony Pictures. The third topic seems to be purely informational. The fourth topic references a hilarious scene from the movie which we do not want to spoiler in this paper. The last topic is about Sony's decision to stream the movie online and to show it in selected cinemas on christmas.

3.3.4 Visualization

To understand and interpret the results of this trend analysis, we used a variety of visualization techniques that help to get valuable insights about certain aspects.

Time Series

The time series visualization is used to display the course of an event or a trend. It displays the dates in which the trend has been monitored on the horizontal axis against the count of tweets collected for that topic (based on our sampled twitter dataset) on the vertical axis. The analysis of the time series is particularly usable to detect new trends. Most trending topics will not show up in previous data at all, but as soon as they begin to spread on Twitter they will be clearly visible in the time series graph.

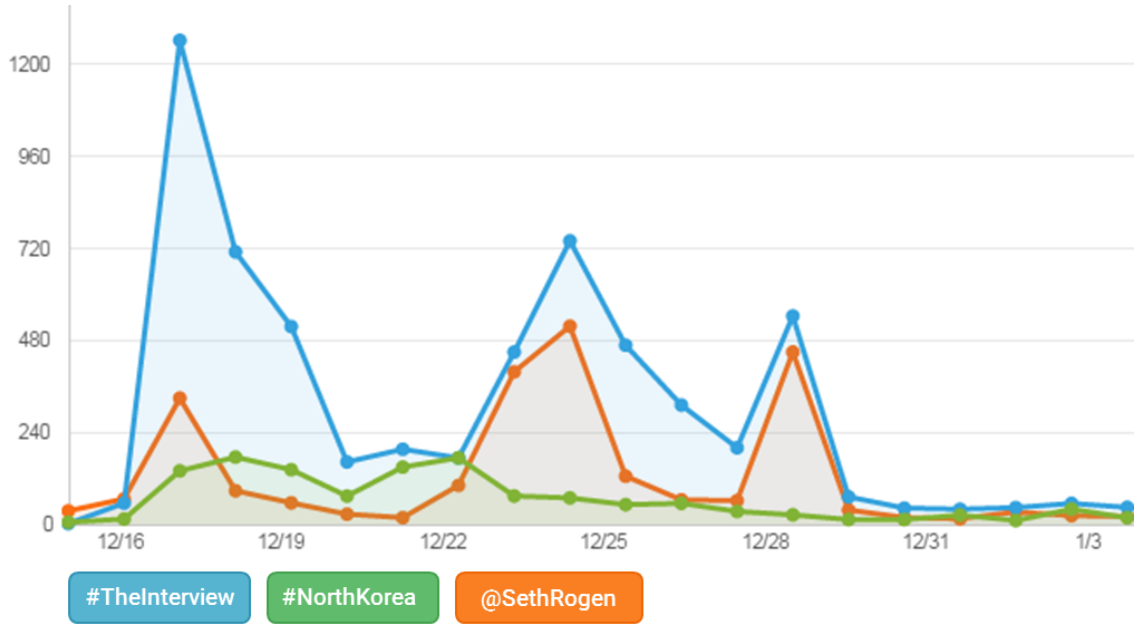

Figure 3: Time Series - Sony Hack Concerning the Movie "The Interview"

Figure 3 shows a time series visualization of the hashtags *#theinterview* and *#northkorea* and the user mention *@SethRogen (main actor)* all belonging to the trend of the comedy movie "The Interview". There is an observable peak for both hashtags and the user mention on 18th of December when Sony announced that they will not screen the movie as a reaction to the hackers' threats. On 24th of December, Sony decided to make the movie available via an online stream. This decision is visible as a peak in the time series visualization as well. The third and last peak in the visualization is based on a live tweeting event from main actor Seth Rogen who tweeted his comments and stories about the movie and the production [Rob14].

Word Cloud Visualization

The word cloud visualization highlights the most frequently occurring terms in the current twitter activity related to a trending topic. Thereby, the importance of a term is expressed using its font size. This visualization type is known as an effective summarizing technique and helps to detect the related topics to a trend.

The current implementation uses all tweets related to a trend and transforms them into a word cloud. Therefore, all tweets are fetched from the database and then the frequency of each word in the text is counted using wordfreq.js[13]. Finally, every unique word and the associated frequency is forwarded to wordcloud2.js[14], a JavaScript visualization library, to render the corresponding word cloud.

[13] `http://timdream.org/wordfreq` [Online; accessed 12-01-2015]
[14] `http://timdream.org/wordcloud2.js` [Online; accessed 12-01-2015]

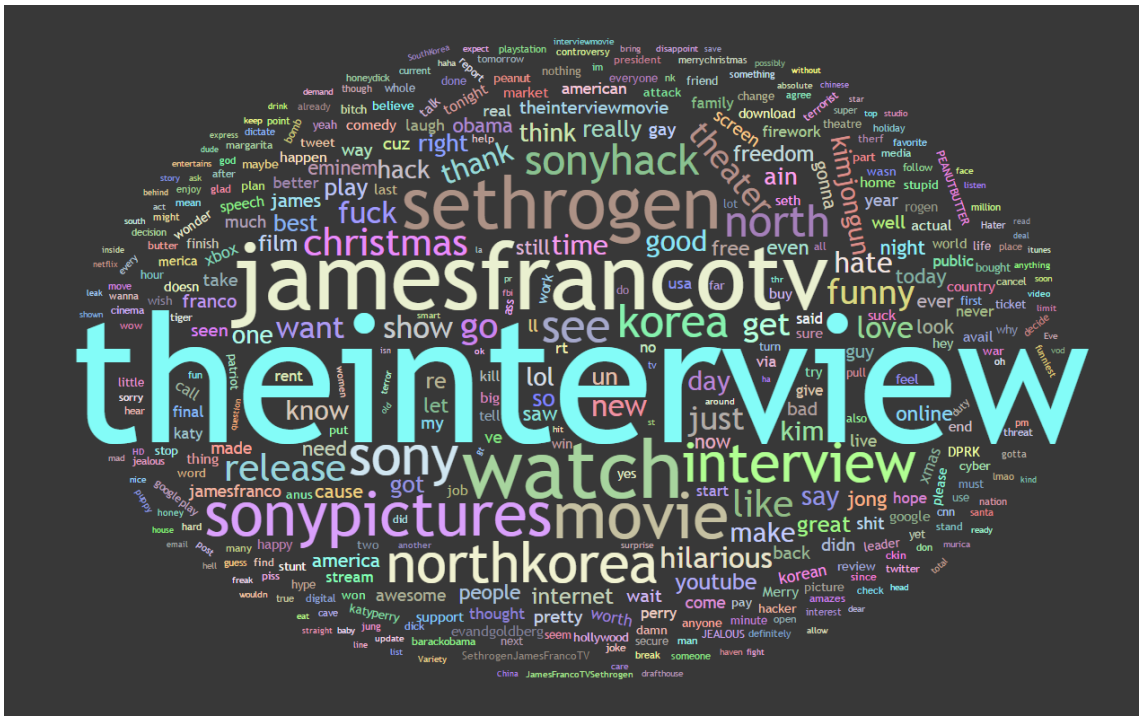

Figure 4: Word Cloud - Sony Hack Concerning the Movie “The Interview”

An example word cloud is visualized in figure 4. It shows the most common words for all tweets related to the comedy movie “The Interview”. The most frequent words were, of course, the movie title itself and the two main actors Seth Rogen and James Franco (his twitter username is *@JamesFrancoTV*). Additionally, Sony Pictures and North Korea are mentioned quite often since they are strongly related to this trend. This example shows that a word cloud is a effective tool to quickly analyze the most frequent words of a trend.

Geospatial Visualization

Geospatial visualization helps to identify the location of current events and detect new events and trends that are likely to occur [KML13]. Furthermore, it is used to gain insights into the prominent locations discussing a selected trend [KML13, pp. 64-66]. All visualizations are build with CartoDB[15], a cloud-computing platform that provides mapping and visualization solutions for geolocated data.

[15] `http://cartodb.com` [Online; accessed 12-01-2015]

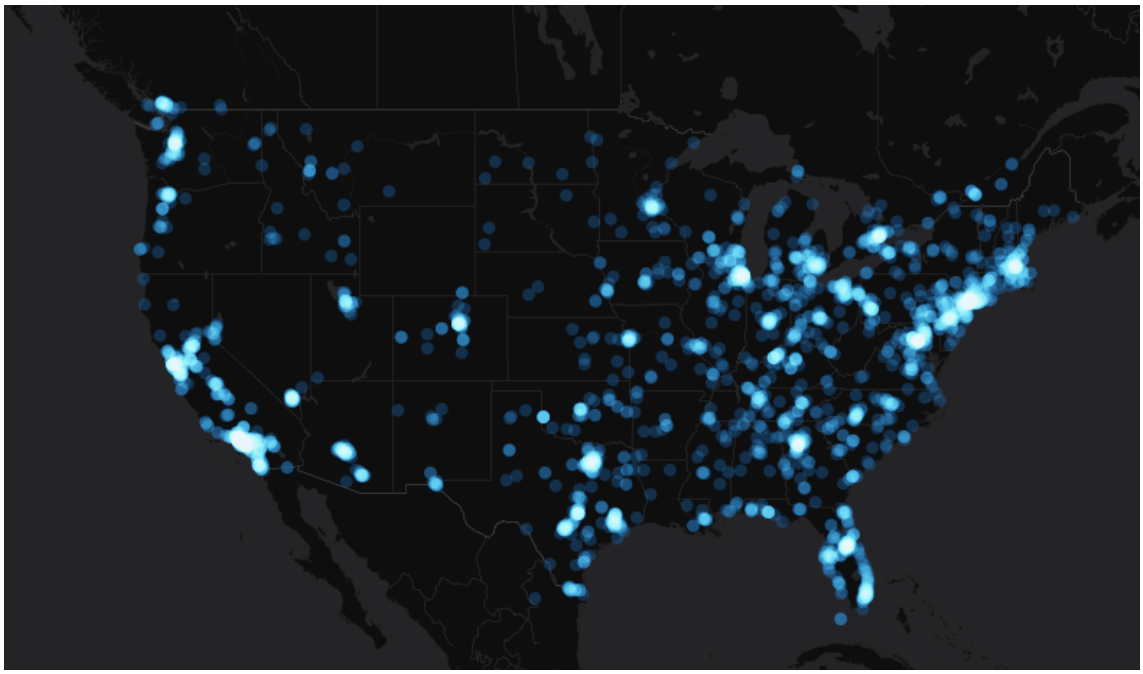

Figure 5: Geospatial analysis - Sony Hack Concerning the Movie “The Interview”

A map enhanced with geospatial data is shown in figure 5. It displays tweets related to the comedy movie “The Interview”. The tweets are distributed equally over the whole United States of America. There are visible hotspots in the biggest cities New York, Los Angeles and San Francisco. The middle west has less tweets visible on the map which may be caused by a lower population in this area.

3.4 Architecture

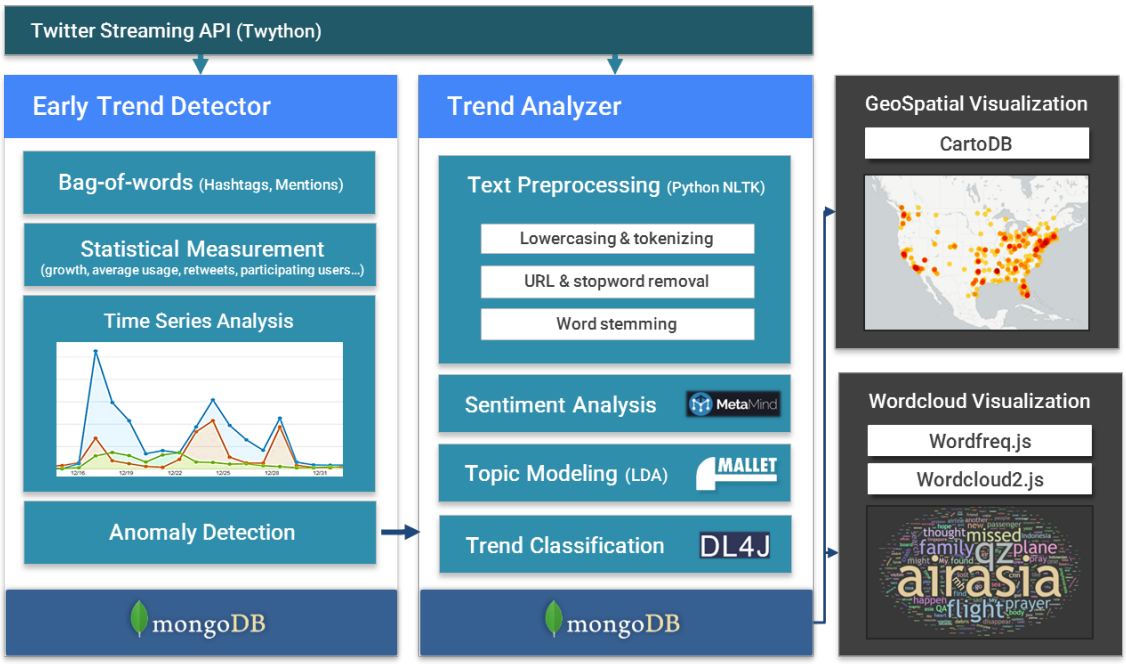

Figure 6: Trend Detection and Analysis Dataflow Architecture

The implementation of our Twitter trend analysis tool, as presented in this paper, is separated into three main components: the Early Trend Detector, the Trend Analyzer and visualization tools. Thereby, the Early Trend Detector and the Trend Analyzer run in parallel on two different servers. Figure 6 illustrates the architecture, implemented technologies and the data flow of our Twitter trend analysis tool.

The Early Trend Detector as well as the Trend Analyzer independently collect data from Twitter. Thereby, the collection of tweets from the Twitter Streaming API[16] is implemented with Python using Twython[17]. A tweet contains a 140 character text message and various metadata such as the language, location, user information, number of retweets, favorites and more. The language used in tweets is mostly informal and the correctness of grammar is often sacrificed to gain additional characters. Further, abbreviations and special characters (e.g. Emoticons) are also frequently employed [KML13, p. 67]. We decided to lowercase the tweet text and hashtags to prevent ambiguity and complexity caused by case-sensitiveness.

The main task of the **Early Trend Detector** is the early detection of upcoming Twitter trends. Therefore, this component streams data from Twitter filtered by English language and a bounding box on the USA. These tweets are stored with

[16] Push service to collect public Tweets in real time.
[17] `http://twython.readthedocs.org` [Online; accessed 12-01-2015]

an additional creation timestamp and all metadata from Twitter into MongoDB, a popular NoSQL database. In parallel, the Early Trend Detector also creates bag-of-words for hashtags and user mentions from incoming tweets and, additionally, calculates several statistics every twenty minutes such as average and total occurrences, usage growth, participating users, retweet count and more. Based on these statistics, an anomaly detection process will be started every two hours to identify a list of about ten unique hashtags and mentions with the highest trend potential. In the next step, one (or more) of those potential trends are selected and a list of correlated hashtags and mentions is generated by querying and counting other hashtags and mentions used in combination with the selected one. Finally, this list of hashtags and mentions related to a potential trend is forwarded to the Trend Analyzer component.

The **Trend Analyzer** aims to further monitor, observe and analyze the potential trend for additional insights. Therefore, this component streams data from Twitter filtered by English language and the list of hashtags and mentions related to the potential trend. To simplify the analysis task, each tweet is preprocessed using common NLP text preparation techniques. In the first step, the text of a tweet is lower-cased and special characters, URLs as well as English stop words are removed. The tweet text is further simplified by using tokenizing and text stemming techniques. In the next step, the preprocessed tweet text alongside with the original tweet text, creation timestamp and all metadata is stored in a separate MongoDB database. After preprocessing, the sentiment (positive, neutral or negative) of every tweet is determined by using the sentiment classifier from MetaMind[18]. Moreover, we utilize the topic modeling algorithm LDA from Mallet[19] to discover topics (word correlations) from the collected dataset of tweets related to the selected trend.

To further analyze, understand and interpret our dataset for detected trends, we integrated a variety of visualization techniques that help to get valuable insights about certain aspects (described in section 3.3.4).

[18] `https://www.metamind.io` [Online; accessed 12-01-2015]
[19] `http://mallet.cs.umass.edu` [Online; accessed 12-01-2015]

4 Trend Analysis on Twitter: Detected Trends

4.1 Detected Trends - Overview

By using our tool, we collected and analyzed about 18 million English tweets from the USA between 15th of December 2014 and 15th of January 2015. In this period of time, we were able to detect several major events such as Christmas and New Year, the Sony Hack concerning the movie "The Interview", the Playstation Network (PSN) Hack, the Air Asia plane crash and the attack on Charlie Hebdo. Figure 7 compares three major events in this time frame based on the number of occurrences of popular hashtags.

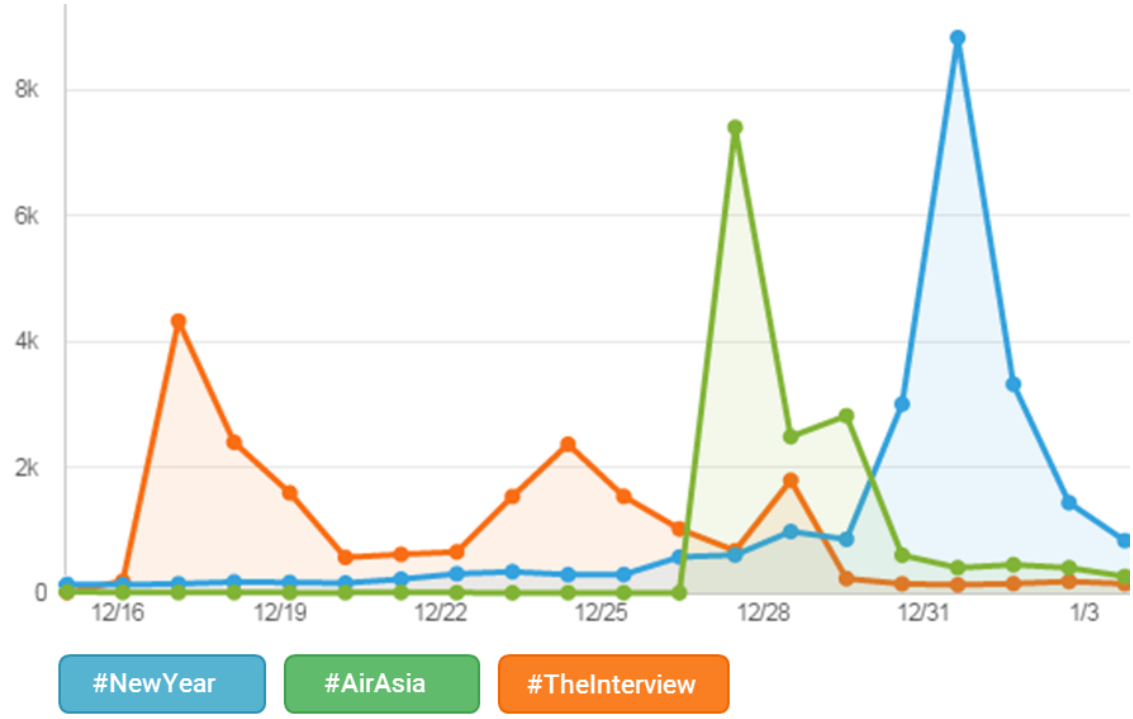

Figure 7: Detected Twitter Trends in December 2014

In the following sections, we will analyze and present some of these trends in more detail aiming to get high-quality insights.

4.2 New Year on Twitter

The most obvious events that occurred in our analysis period were Christmas and New Year. Both trends are compared in figure 8. The number of tweets mentioning Christmas already started to rise many days before Christmas Eve probably because of Advent and Christmas preparations. In contrast, the trend of New Year lasted only for about four days, but with a much higher peak on New Year Eve.

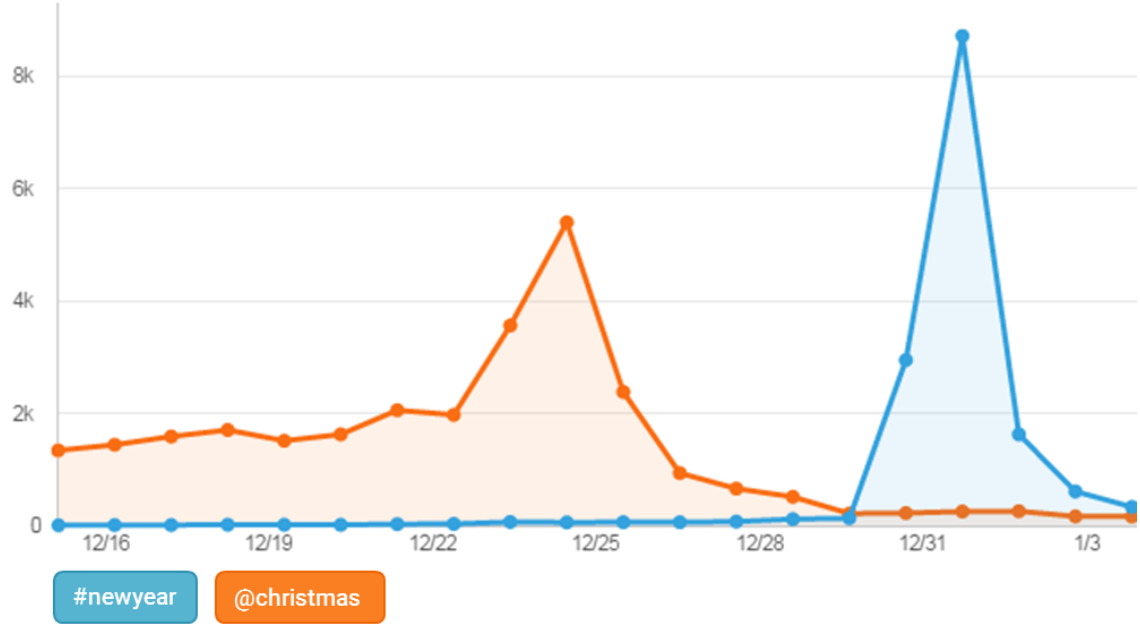

Figure 8: Time Series - Christmas & New Year

Based on the rapid increase of tweets mentioning the New Year in a short period, we could detect several hashtags related to this event. Following hashtags were discovered as being the main used ones:

#newyear	#nye2015	#hello2015
#newyearseve	#midnight	#welcome2015
#nye	#ihatenewyears	#hi2015
#hny	#newyearseveproblems	#newyears
#goodbye2014	#newyear2015	
#bye2014	#happynewyear	

Listing 4.1: New Year Hashtags and User Mentions

The word cloud depicted in figure 9 illustrates the most frequent used words from our dataset related to the New Year.

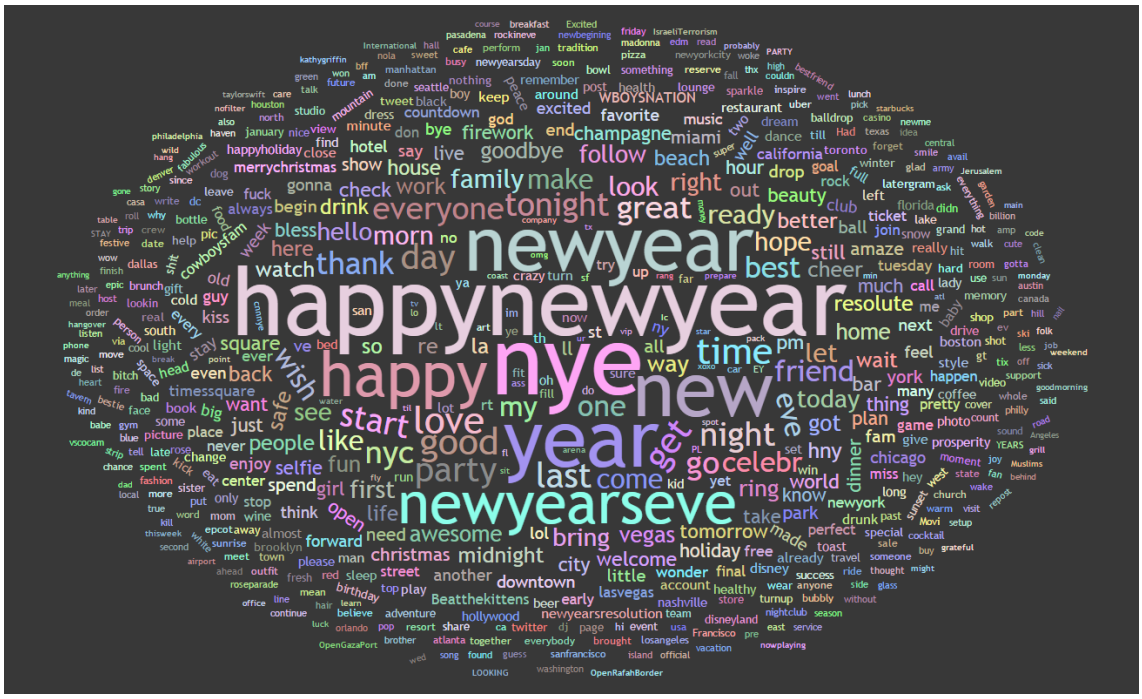

Figure 9: Word Cloud - New Year Eve

Figure 10 shows a geospatial comparison of the activity of tweets mentioning the New Year during two different points in time. The left heat map shows the west coast few minutes after midnight on 1st January 2015 (Timezone: UTC-08). In contrast, the right map visualizes the same Twitter activity three hours before on the east coast (Timezone: UTC-05). It is clearly recognizable that the New Year midnight in each time zone generates a huge amount of content related to this event in a short period of time.

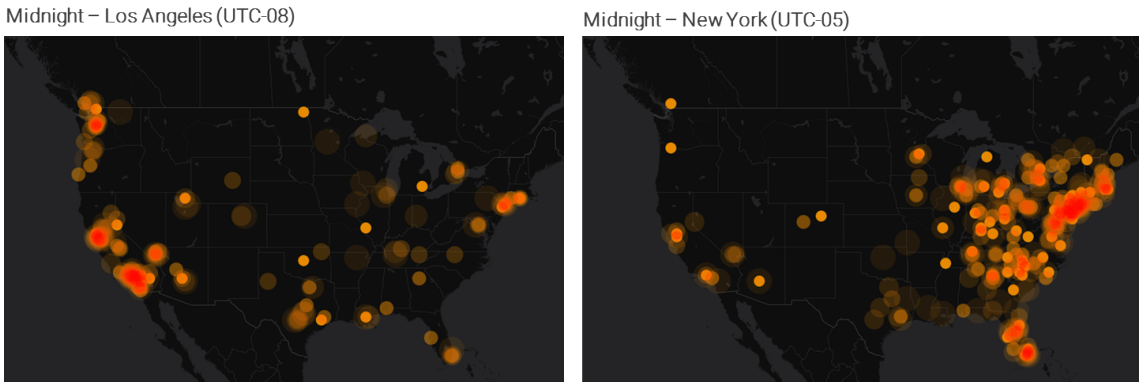

Figure 10: Geospatial Comparison - New Year Eve

Besides the preceding analysis of the geolocation and spreading of this trend, we analyzed the sentiment of the American people regarding the New Year as shown in figure 11. Thereby, the majority of tweets contains positive or at least neutral opinions and only a small number has a negative attitude.

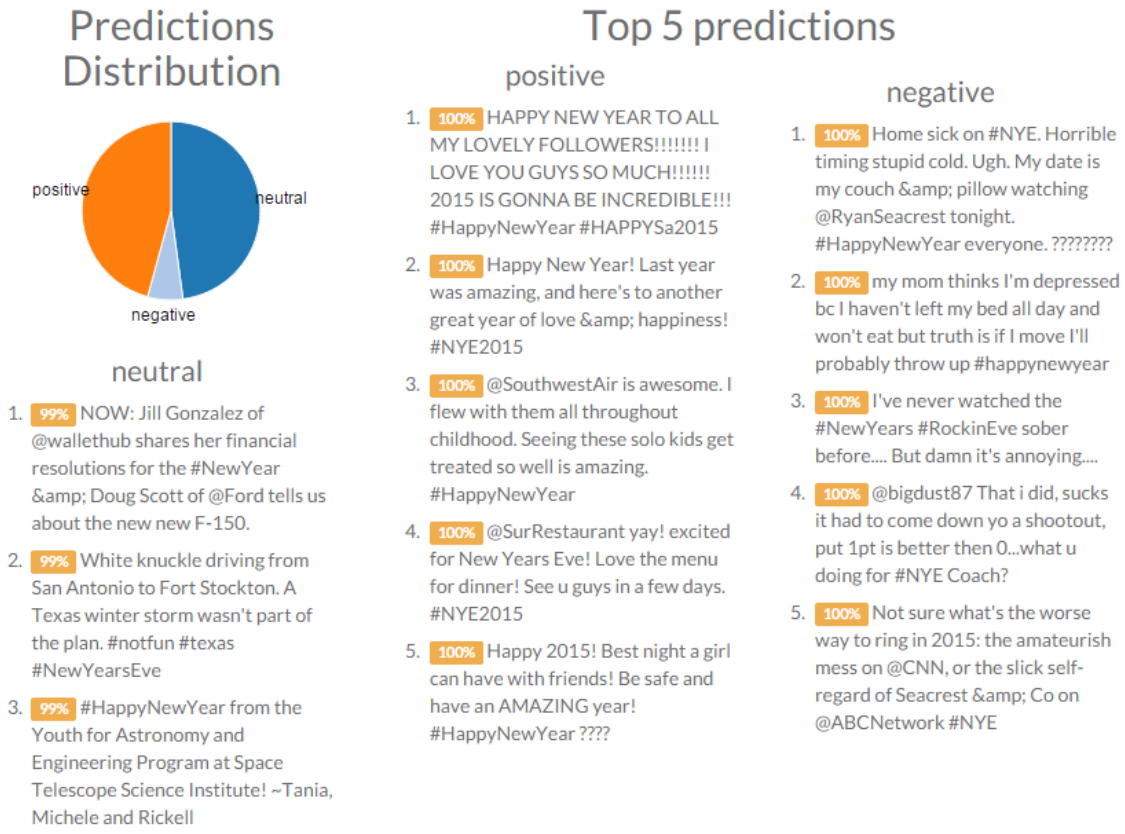

Figure 11: Sentiment - New Year Eve

4.3 Air Asia Flight Tragedy

On 28th of December, a terrible tragedy hit the news: a plane from the Air Asia carrier (QZ8501) crashed into the Java sea between Indonesia and Singapore. On board of the flight were 162 people on their way from Surabaya in Indonesia to Changi Airport in Singapore. It was around 06:12 local time when the pilot contacted air traffic control to request a change in flight altitude in order to prevent being caught by the storm clouds which are typical for that area. Air traffic control gave the permission to do so a few minutes later, but could not reach the plane anymore [Bbca]. The plane had already crashed into the sea. Many people related this event to the tragedy of flight MH370 from Malaysia Airlines, which got lost on March 8th, 2014 after the missing of flight QZ8501 was announced [Nbc].

Figure 12 illustrates the impact of this tragedy on Twitter with two distinguishable peaks on 28th of December, after it was first announced that the plane was missing, and on the 30th of December, after the first wreckage was found. As of now, the recovery effort is still ongoing as well as the discussion of this event on Twitter.

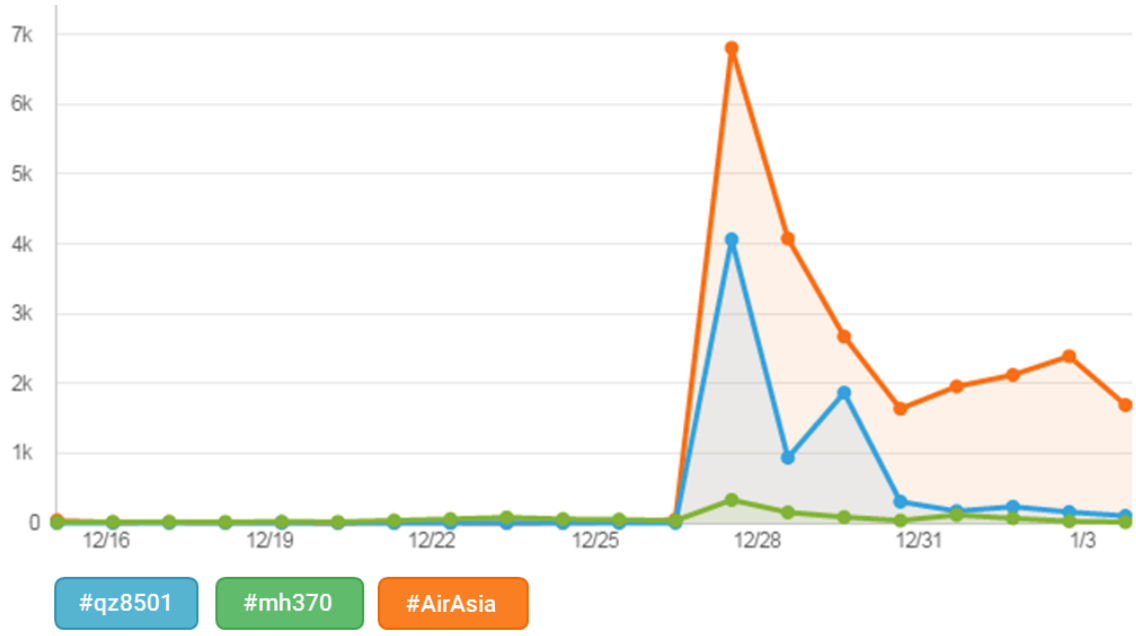

Figure 12: Time Series - AirAsia Tragedy

Our Early Trend Detector was able to identify the following list of related hashtags used by Twitter users to discuss this news or to express griefs and sympathy with the families and relatives:

#airasia	#prayforqz8501
#qz8501	#airasia8501
#prayforairasia	#mh370

Listing 4.2: Air Asia Hashtags and User Mentions

As mentioned earlier, many people related the crash of Air Asia flight QZ8501 with the disappearance of the Malaysia Air flight MH370. That explains why both flight numbers are trending topics.

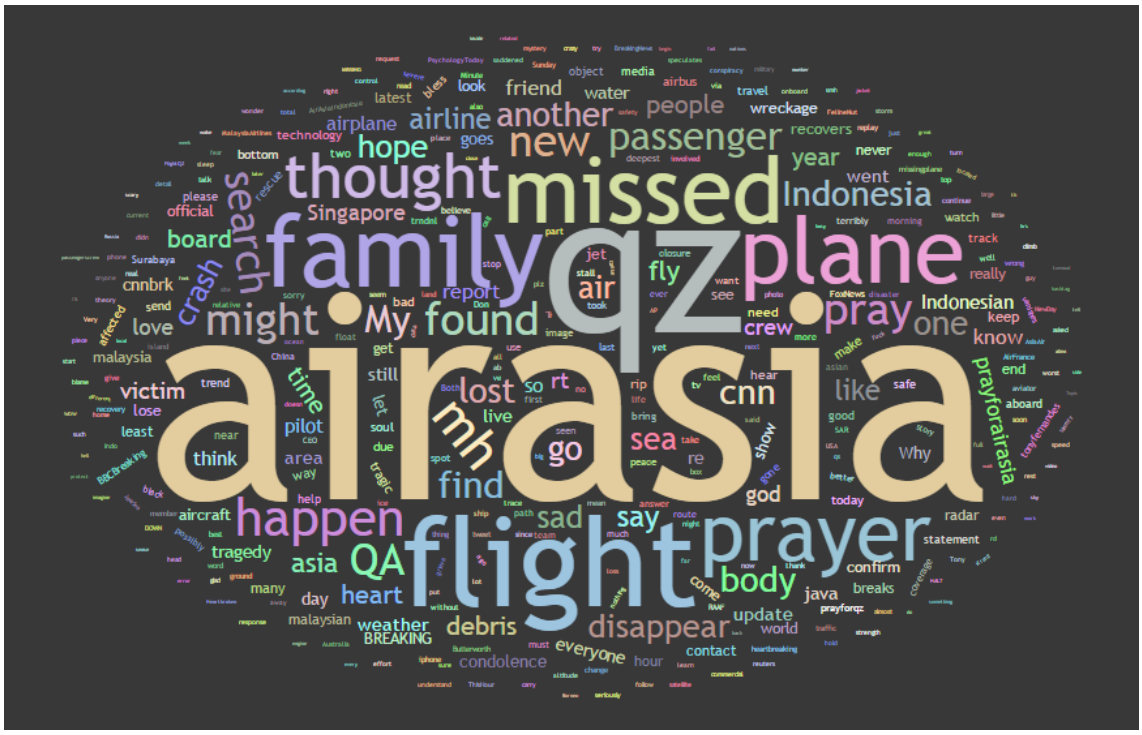

Figure 13: Word Cloud - AirAsia Flight Tragedy

The word cloud in figure 13 visualizes the most commonly used words in tweets about the plane crash. A hypothesis based on the word cloud is that the tweets have two different topics. One topic is news (discussion about the tragedy) and the other topic is emotional (showing of condolence to the families of the victims). We extracted tweets from our dataset related to this event and used LDA for topic modeling in order to further analyze our hypothesis. These topics are presented in listing 4.3.

```
1 airasia (139) missing (76) flight (55) air (39) indonesia (37)
    singapore (33) asia (31)

2 airasia (126) missing (60) planes (50) find (39) plane (36)
    world (20) technology (15)

3 prayers (86) families (81) thoughts (72) airasia (24) crash
    (14) thought (12) airfrance (8)

4 airasia (257) families (144) flight (90) passengers (69)
    prayers (58) amp (47) missing (39)

5 airasia (35) weather (23) flight (17) pilots (13) fly (12) bad
    (12) path (10)

6 raaf (8) butterworth (8) china (8) australia (5) russia (5)
    trndnl (5) trending (5)
```

Listing 4.3: Topic Model for Air Asia Flight Tragedy

We used LDA to model 6 different topics showing the 7 most relevant words of each topic. There is an observable difference between reporting tweets (like topic 1, 2 and 5) and emotional tweets (like topic 3 and 4). Topic 8 stand out from the other, it is about RAAF Butterworth airport in Malaysia, this airport is used by Australia and others to coordinate the search for the missing wreckage of the airplane. This shows that our initial hypothesis is true. There are two different subjects tweeting about the airplane crash of flight QZ8501.

The sentimental analysis of Air Asia related tweets in figure 14 shows a significant difference compared to the sentiment of New Year Eve in figure 11. Only a small portion of tweets is classified as positive, mostly containing words of hope, compassion and prayer. A larger portion of tweets are labeled with a negative sentiment by people referencing this terrible tragedy and condemning the end of this year. The biggest portion of tweets contains neutral information such as objective news updates and discussion.

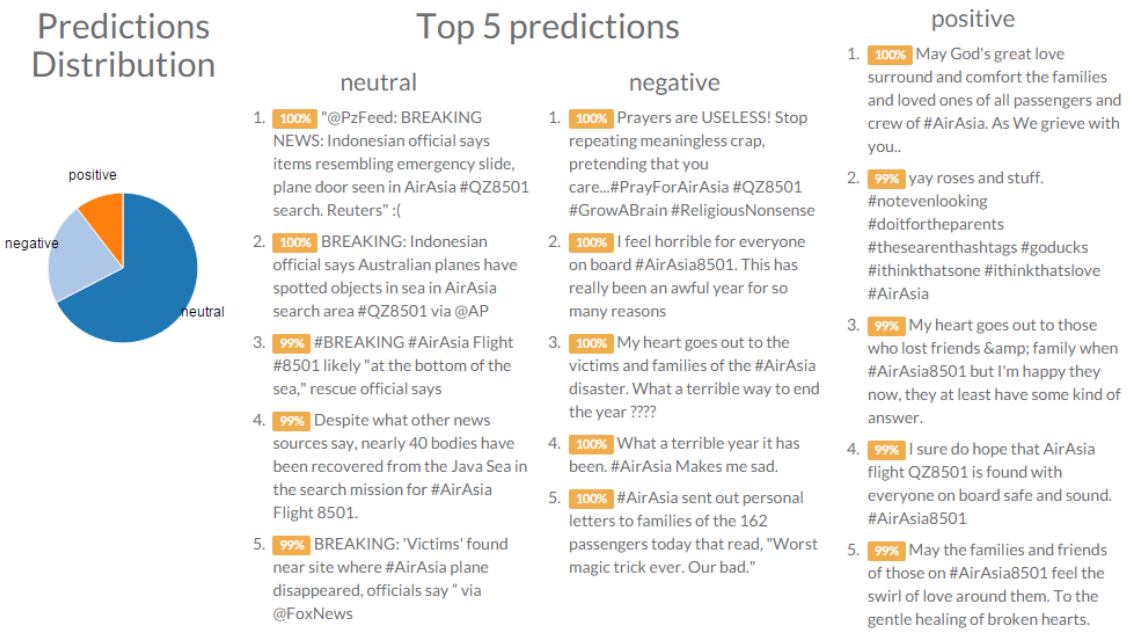

Figure 14: Sentiment - AirAsia Tragedy

Comparison with Google Trends

In this section, we compare the trend data related to the Air Asia tragedy with statistics from Google Trends to check for similarities and the validity of detecting trends on Twitter based on a small statistical sample. Further, we want to get additional insights into the character of Twitter trends and how close those trends resemble real world activities. Google Trends makes it possible to analyze the search-volume for specified search terms entered in the Google Search. Kwak et al. argue

that search keywords from Google "represent topics users are interested in and popular keywords represent hot trends" [Kwa+10, p. 6]. Therefore, those keyword trends "have become a good indicator to understand activities in the real world" [Kwa+10, p. 6].

Figure 15 illustrates the comparison between the time series visualization of the Air Asia flight tragedy on Twitter and a Google Trends statistic for similar search terms.

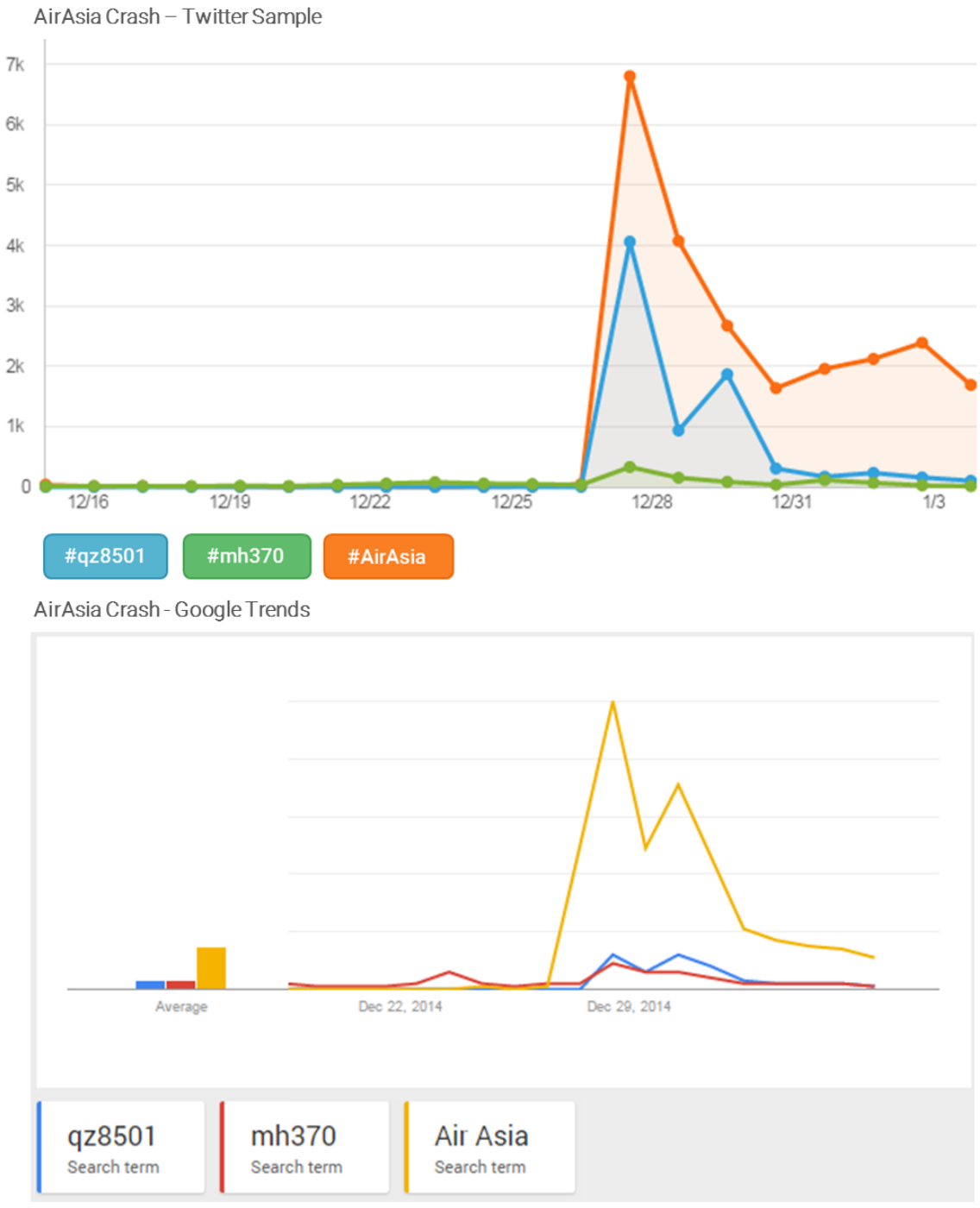

Figure 15: AirAsia Time Series - Comparison with Google Trends

There is a clear resemblance between the trend curve based on our dataset of tweets and the curve on Google Trends for the used search terms. Moreover, figure 16 compares the regional interest about the Air Asia tragedy in the US on Twitter with the relative search volume for each state on Google Search.

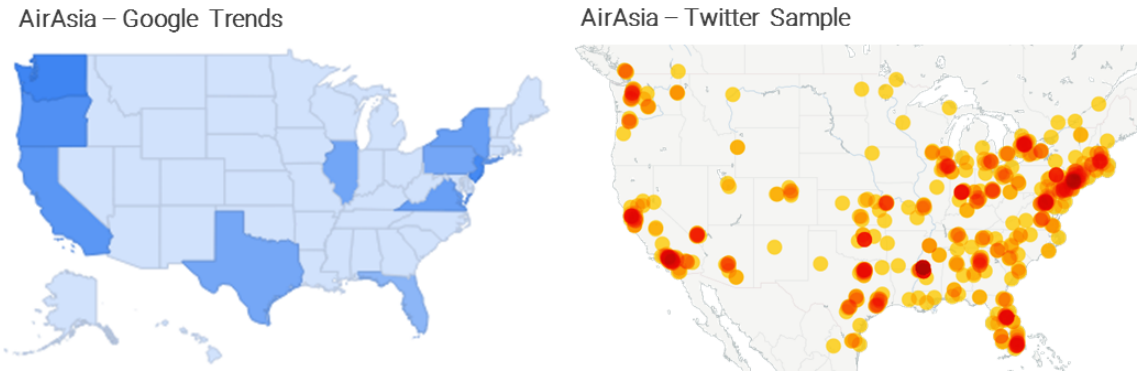

Figure 16: AirAsia Geospatial - Comparison with Google Trends

Both comparisons demonstrate a high similarity between Twitter trends and Google Search trends. We conclude that detecting and analyzing Twitter for trends provides high-quality insights into the world's interest in global events. Twitter trends even resemble regional interests in certain topics with a high accuracy. Further, Twitter is one of the fastest ways to detect and predict global trends, due to its open access to real-time data. However, Kwak et al. mentions that interactions such as retweet and reply "might be a factor to keep trending topics persist" for a longer time on Twitter.

4.4 Gaming Network Outage

On the 24th of December in 2014, hackers started to attack the Playstation Network (PSN) and the Microsoft Xbox Live Network. The hacker attacks brought the networks down for several days. The gamer community was outraged not to be able to play games during this period of time [Woo+]. After a few days, a hacker group called *Lizard Squad* claimed credit for the attack. In the end, the popular german internet entrepreneur Kim Dotcom paid Lizard Squad with vouchers for his web platform *MEGA* [Dot14]. In return, Lizard Squad decided to stop further attacks on the gaming networks. After the network recovered, Sony announced to give discounts to PSN users as a compensation. Figure 17 shows how the Twitter activity regarding this topic increased rapidly during the period of attacks.

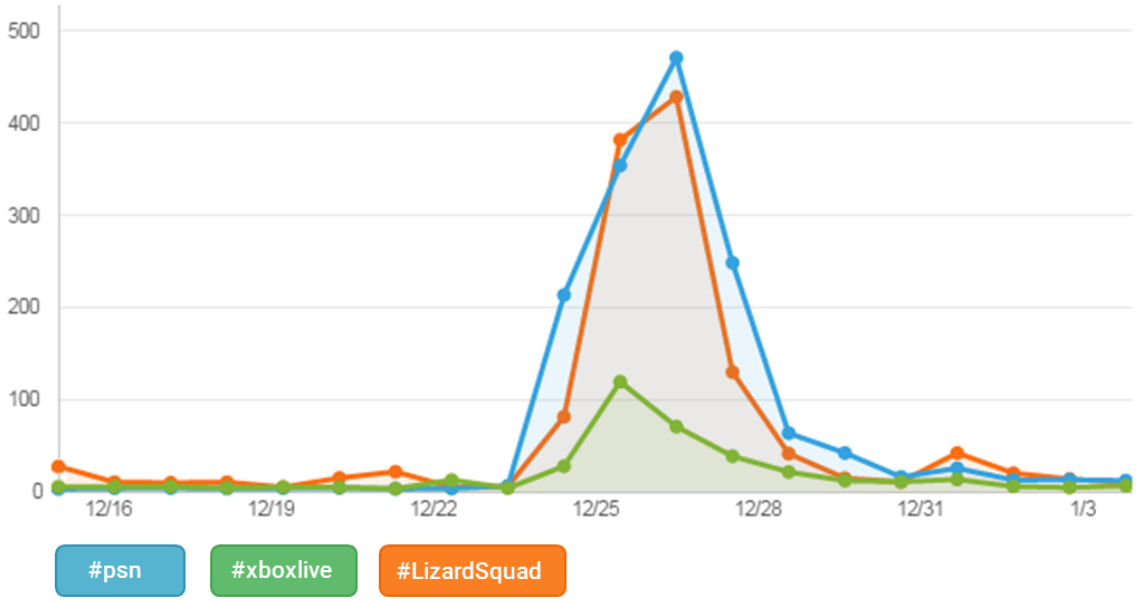

Figure 17: Time Series - Gaming Network Outage

The involved persons and topics are reflected in the following list of hashtags and user mentions that we were able to identify for this trend:

#finestsquad	#psn	@AskPlayStation
#lizardpatrol	#psndown	@KimDotcom
#lizardsquad	#PSNDownTime	@LizardMafia
#payingfornothing	#psnup	@MEGAprivacy
#playstationnetwork	#xboxlivedown	@PlayStation
#playstationsucks	#xboxsupport	

Listing 4.4: Gaming Network Outage Hashtags and User Mentions

We fetched tweets containing at least one of the listed hashtags or user mentions and created a word cloud, depicted in figure 18.

Figure 18: Word Cloud - Gaming Network Outage

In the next step, we wanted to find out which of the words in the word cloud are used together most of the time. Therefore, we performed topic modeling on the trend related tweets by using LDA. The number of inspected documents was 7447, the topics and iterations parameters were set to 10 and 2000, respectively. The most distinctive topics are displayed in listing 4.5.

```
1 xbox (101) playstation (50) watch (44) movie (32) fuckcrucifix
    (31) north (29) korea (27) interview (27)

2 xbox (310) christmas (178) play (81) xboxlivedown (72) live
    (71) xboxlive (68) xboxsupport (66) day (63)

3 playstation (55) dollar (27) psn (20) company (19) lizardsquad
    (18) sony (17) billion (16) multi (12)

4 fuckcrucifix (204) lizardmafia (172) lizardsquad (125) fuck
    (116) lizard (108) squad (102) finestsquad (95) stop (94)

5 psn (468) back (461) playstation (324) online (246)
    askplaystation (205) network (173) psndown (89) working (88)


6 xbox (250) psn (230) sign (215) connect (143) live (110) error
    (103) account (93) issues (82)
```

Listing 4.5: Topic Model for Gaming Network Outage

The words that were clearly visible in the word cloud are also dominating the detected topics. Furthermore, the detected topics reflect the real events in a reasonably good way.

Topic 1 covers words indicating a discussion about a connection between the DDoS attack during Christmas and a previous hack against Sony concerning the movie "The Interview" [Bbcb]. The second topic is about the hack affecting the Xbox Live Network. Obviously, a lot of people tweeted to Microsoft's support. Topic 5 is similar to topic 2, however, these terms are related to the Playstation Network. Topic 6 covers general terms concerning the hack and the instability to connect to the networks or to play a game. Topic 4 is focused on terms about *Lizard Squad*. The words indicate that the gamer community was not very amused about the hack. Topic 3 contains terms about the financial impact of such a hack and the claim for redemption for the lost hours of not being able to use the networks.

Figure 19 displays the sentiment of the user community regarding the gaming consoles and their hacks. Obviously, some people were quite upset. Interestingly, the anger was directed against the companies and not against the hacker group. Some people, however, were thankful for the past year and backed up Sony and Microsoft.

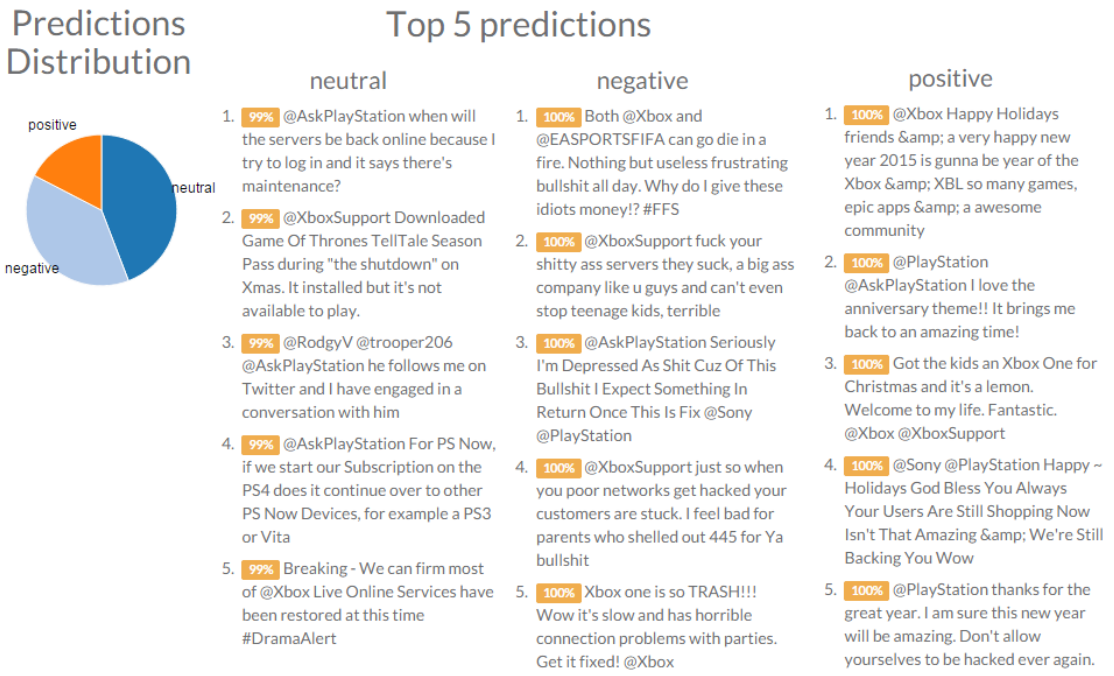

Figure 19: Sentiment - Gaming Network Outage

5 Conclusion and Future Work

Big social data analysis has grown into a serious business over the past several years with important use cases not just for research projects, but also in commercial products. Social Data analysis techniques are applied to predict terrorist attacks, stock performance, election results or the spread of diseases. Further, it is utilized by companies to analyze their customer's satisfaction and the public opinion about their products. Cutting edge machine-learning, NLP and data mining technologies are necessary to gain valuable insights into large amounts of social content.

In this paper, we presented our concept and implementation of several components to collect, analyze and visualize Twitter data. Thereby, we collected about 18 million English tweets from the USA. Based on this dataset, we built a component to detect the spreading of trends at an early stage based on occurrences of hashtags and user mentions. With this system, we were able to detect several major events at an early stage such as the Air Asia tragedy and the PSN Hack, only based on the usage frequency of hashtags and mentions from a statistical sample of the Twitter dataset. The early detection can be further improved by analyzing all words of a tweet, since hashtags mostly get created and used within a short time delay to the actual event and only a small number[20] of tweets even contain hashtags.

We used several machine-learning and data mining techniques to analyze the data, such as bag-of-words, sentiment analysis, topic modeling and data classification. Finally, we visualized the trends with various visualization techniques. The development of a trend and the characteristic of events were analyzed by using time series visualizations. Further, we used geospatial visualizations to show the regional activity and emphasize different aspects about the local spreading of trends. Moreover, word clouds were used to highlight commonly used terms in the detected trend.

We conclude that detecting and analyzing Twitter for trends provides high-quality insights into the worlds interest in global events. Twitter trends even resemble regional interests in certain topics with a high accuracy. Further, Twitter is one of the fastest ways to detect and predict global trends, due to its open access to real-time data. However, to get more overall accurate results, it would be necessary to eliminate the current limitations by collecting a larger portion of the Twitter stream and offering multi-language support.

[20] Only about 13% of tweets in our dataset included hashtags

Appendix

A Project History

In the beginning of the project, we wanted to analyze Stack Overflow. Stack Overflow is one of the biggest Q&A pages of the today's web and the flagship of the Stack Exchange Network. Our goal was to get high-quality insights into trending topics of developers around the globe. After identifying current hot topics people write about, we wanted to search Twitter messages for the same topics. As a result, we wanted to find out if it is possible to discover trends we identified on Stack Overflow also on Twitter. In the next step, we wanted to categorize and analyze detected intersection on both media platforms. The project was supposed to answer among other possible questions the following ones: Is Twitter used to ask questions? Is there a chronological difference between the uprising of a trend on Stack Overflow and Twitter? Are there opinion leaders in one of the sources? [People who ask a lot of questions / tweet a lot about a topic]

After a renewed validation of the project's purpose we shifted the direction. We had the assumption that we would find only a few intersections between topics discussed on Stack Overflow and Twitter, if any. Additionally, Stack Overflow already offers quite sophisticated statistics about its data, including topics. These statistics make an own analysis redundant.

As a consequence, we changed the project's objective, which is depicted in the following.

Bibliography

[Agi+08] Eugene Agichtein et al. “Finding High-quality Content in Social Media”. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. WSDM ’08. Palo Alto, California, USA: ACM, 2008, pp. 183–194.

[Bbca] *Flight QZ8501: What we know about the AirAsia plane crash*. BBC, 7 January 2015. [Online; accessed 07-January-2015]. URL: `http://www.bbc.com/news/world-asia-30632735`.

[Bbcb] *The Interview: A guide to the cyber attack on Hollywood*. BBC, 29 December 2014. [Last updated at 07 January 2015]. [Online; accessed 07-January-2015]. URL: `http://www.bbc.com/news/entertainment-arts-30512032`.

[Ble+03] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. “Latent dirichlet allocation”. In: *Journal of Machine Learning Research* 3 (2003), p. 2003.

[BMZ11] J. Bollen, H. Mao, and X. Zeng. “Twitter mood predicts the stock market”. In: *Journal of Computational Science* (2011).

[BS11] Adam Bermingham and Alan F Smeaton. “On using Twitter to monitor political sentiment and predict election results”. In: (2011).

[Cam+13a] Erik Cambria, Dheeraj Rajagopal, Daniel Olsher, and Dipankar Das. “Big social data analysis”. In: *Big Data Computing* (2013), pp. 401–414.

[Cam+13b] Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. “New Avenues in Opinion Mining and Sentiment Analysis.” In: *IEEE Intelligent Systems* 28.2 (2013), pp. 15–21.

[Car+14] Cornelia Caragea et al. “Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy”. In: *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Managements* (2014).

[Dom12] Pedro Domingos. “A Few Useful Things to Know About Machine Learning”. In: *Commun. ACM* 55.10 (Oct. 2012), pp. 78–87.

[Fla+12] Ilias Flaounas et al. “Big Data Analysis of News and Social Media Content”. In: (2012).

[Gin+09] Jeremy Ginsberg et al. “Detecting influenza epidemics using search engine query data”. In: *Nature* 457 (2009). doi:10.1038/nature07634, pp. 1012–1014.

[HKP12] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Waltham, Mass.: Morgan Kaufmann Publishers, 2012.

[Ins11] McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity*. Tech. rep. 2011.

[Kim+13] Daehoon Kim, Daeyong Kim, Seungmin Rho, and Eenjun Hwang. “Detecting Trend and Bursty Keywords Using Characteristics of Twitter Stream Data”. In: *International Journal of Smart Home* 7.1 (2013), pp. 209–220.

[KML13] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter Data Analytics*. New York, NY, USA: Springer, 2013.

[Kum+11] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. “TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief.” In: *ICWSM*. 2011.

[Kwa+10] 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. Raleigh, North Carolina, USA: ACM, 2010, pp. 591–600. ISBN: 978-1-60558-799-8.

[LM11] Serge Linckels and Christoph Meinel. *E-Librarian Service: User-Friendly Semantic Search in Digital Libraries*. 1st. Springer Publishing Company, Incorporated, 2011.

[Mor+13] Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. “Understanding twitter data with tweetxplorer”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 1482–1485.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008.

[Nbc] *By the Numbers: Malaysia Airlines Flight 370*. BBC, 27 December 2014. [Online; accessed 07-January-2015]. URL: http://www.nbcnews.com/storyline/missing-jet/numbers-malaysia-airlines-flight-370-n275136.

[NBG11] Mor Naaman, Hila Becker, and Luis Gravano. "Hip and Trendy: Characterizing Emerging Trends on Twitter". In: *J. Am. Soc. Inf. Sci. Technol.* 62.5 (May 2011), pp. 902–918. ISSN: 1532-2882. DOI: 10.1002/asi.21489. URL: http://dx.doi.org/10.1002/asi.21489.

[RDL10] Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. "Characterizing Microblogs with Topic Models." In: *ICWSM*. Ed. by William W. Cohen and Samuel Gosling. The AAAI Press, 2010.

[Sec13] Homeland Security. "Lessons Learned: Social Media and Hurricane Sandy". In: (2013).

[TR12] Oren Tsur and Ari Rappoport. "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities." In: *WSDM*. Ed. by Eytan Adar, Jaime Teevan, Eugene Agichtein, and Yoelle Maarek. ACM, 2012, pp. 643–652.

[WGB12] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. "Automatic crime prediction using events extracted from twitter posts". In: *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2012, pp. 231–238.

[Woo+] Victoria Woolaston, Julian Robinson, Darren Boyle, and Rachel Burnett. *Sony extends the PlayStation Plus memberships of gamers affected by the Lizard Squad hack*. Dailymail, 2 January 2015. [Online; accessed 07-January-2015]. URL: http://www.dailymail.co.uk/sciencetech/article-2894191/Sony-extends-PlayStation-Plus-memberships-gamers-affected-Lizard-Squad-hack.html.

[Zub+11] Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, and Raquel Martínez. "Classifying Trending Topics: A Typology of Conversation Triggers on Twitter". In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. CIKM '11. Glasgow, Scotland, UK: ACM, 2011, pp. 2461–2464. ISBN: 978-1-4503-0717-8. DOI: 10.1145/2063576.2063992. URL: http://doi.acm.org/10.1145/2063576.2063992.

[ZWL13] Peng Zhang, Xufei Wang, and Baoxin Li. "On predicting Twitter trend: factors and models." In: *ASONAM*. Ed. by Jon G. Rokne and Christos Faloutsos. ACM, 2013, pp. 1427–1429.

Online Resources

[Alc13] Stan Alcorn. *Twitter Can Predict The Stock Market, If You're Reading The Right Tweets*. [Online; accessed 08-January-2015]. 2013. URL: http://www.fastcoexist.com/1681873/twitter-can-predict-the-stock-market-if-youre-reading-the-right-tweets.

[Cen14] Pew Research Center. *Social Networking Fact Sheet*. [Online; accessed 08-January-2015]. 2014. URL: http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/.

[Coh13] Sara Estes Cohen. *Sandy Marked a Shift for Social Media Use in Disasters*. [Online; accessed 08-January-2015]. Emergency Management. 2013. URL: http://www.emergencymgmt.com/disaster/Sandy-Social-Media-Use-in-Disasters.html.

[Com11] Semantic Web Company. *Big Data and Linked Data*. [Online; accessed 08-January-2015]. 2011. URL: http://www.semantic-web.at/big-data-linked-data.

[Dot14] Kim Dotcom. *A Christmas Miracle*. [Online; accessed 08-January-2015]. Twitter. 2014. URL: https://twitter.com/kimdotcom/status/548305704776241152.

[Goo14] Google. *Explore flu trends - United States*. [Online; accessed 08-January-2015]. 2014. URL: http://www.google.org/flutrends/intl/en_us/us/#US.

[GS13] Konstantinos Giannakouris and Maria Smihily. *Social media - statistics on the use by enterprises*. [Online; accessed 08-January-2015]. European Commission. 2013. URL: http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Social_media_-_statistics_on_the_use_by_enterprises.

[HR13] Healy and Ramaswarmy. *Visualizing Twitter Sentiment*. [Online; accessed 08-January-2015]. NC State University. 2013. URL: http://www.csc.ncsu.edu/faculty/healey/tweet_viz/.

[IBM12] IBM. *IBM What is big data? - Bringing big data to the enterprise*. [Online; accessed 08-January-2015]. 2012. URL: http://www-01.ibm.com/software/data/bigdata.

[Ora13] Oracle. *Oracle White Paper - Big Data for the Enterprise*. [Online; accessed 08-January-2015]. 2013. URL: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

[Rob14] David Robb. *Sony Hack: A Timeline*. [Online; accessed 08-January-2015]. Dec. 2014. URL: http://deadline.com/2014/12/sony-hack-timeline-any-pascal-the-interview-north-korea-1201325501/.

[Tec14] Tech2. *Big data collection from Google, Facebook and others can be misleading, says study*. [Online; accessed 08-January-2015]. Mar. 2014. URL: http://tech.firstpost.com/news-analysis/big-data-collection-google-facebook-others-can-misleading-says-study-219931.html.

[Web14] Christian Weber. *Google versagt bei Grippe-Vorhersagen*. [Online; accessed 08-January-2015]. Mar. 2014. URL: http://www.sueddeutsche.de/wissen/big-data-google-versagt-bei-grippe-vorhersagen-1.1912226.