Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Mining high quality insights in social media data using machine learning methods

Early Trend Detection 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 (CHANGE)
Student (Id):	Henning Muszynski (50170)
Student (Id):	Benjamin Raethlein (CHANGE)
Academic Supervisor:	Prof. Dr. Norbert Link
Academic Supervisor:	Dr. Ingo Schwab

Declaration of Authorship

We declare that we entirely by ourselves have developed and written the enclosed report and have not used sources or means without declaration in the text. Any thoughts or quotations which were inferred from these sources are clearly marked as such.

This report was not submitted in the same or in a substantially similar version to any other authority to achieve an academic grading and was not published elsewhere. This report has been submitted exclusively to the University of Applied Sciences Karlsruhe.

Karlsruhe, January 8, 2015 ______________________

(Lukas Masuch)

Karlsruhe, January 8, 2015 ______________________

(Henning Muszynski)

Karlsruhe, January 8, 2015 ______________________

(Benjamin Raethlein)

Abstract

Contents

Abbreviations

LDA Latent Dirichlet Allocation

List of Figures

List of Tables

List of Listings

Todo list

1 Introduction

1.1 Motivation

The immense rise of social media is one of the driving forces behind the current Big Data trend. Big data creates 2.5 billion gigabytes every day and produced 90% of the worldwide data in the last two years, thus it has become a top priority for research organizations and companies [IBM12]. 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 imaginations, 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 and advanced Natural Language Processing technologies make it possible to collect and analyze those massive amounts of data and enables a fundamentally new approach for the study of society and human beings.

Rewrite this part

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.

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

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. [Check and adapt the following paragraph depending on the real content of our project] The goal of the project is the early detection and prediction of arising trends on Twitter. We assume that it is possible to predict the spreading of future trends on Twitter based on the curves of trends in the past. Therefor, we want to explore different metrics and dimensions, such as retweets, hashtag/topic occurrences, user groups and emotions. It helps to detect big headlines before they go viral and, therefore, it is very valuable in different areas such as stock market, brand awareness, political discussions and elections and the success of media (movies, music).

We suggest an architecture consisting of two systems for data collection. The first system is used to monitor the entire Twitter stream and focused on detecting on trends that are in early stage. Furthermore, it uses topic modeling to identify topics/hashtags that are correlated to the same trend. These results are then forwarded to the second system. The second system utilizes this data for observing only those topics in detail until they are not relevant anymore.

In the next step, we plan to use (unsupervised) machine learning techniques to compare the early trends with previous trend curves to predict their further course.

Additionally, we may compare the overall results with data from Google Trends to check for similarities.

1.3 Overview

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 those companies to get more information about their customers' sentiment, satisfaction or opinion by collecting and analyzing data from social media services [Ora13].

Big Data is typically distinguished into three data types: structured, unstructured or semi-structured data. **Structured data** covers information that is captured in a field within a file or database, whereas **unstructured data** covers information without a data model organization (e. g. plain text, videos). For this reason, unstructured data is hard to process and to understand by machines. **Semi-structured data** refers to data without a formal structure like a database, but it contains tags to structure semantic elements [Sin+10, pp. 2 sqq.].

Many techniques have been developed to analyze this gigantic amount of data [Ins11, p. 27]. Some of these technologies are described later in this chapter.

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]. Those 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]. That 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.

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].

2.3 Machine Learning

Machine learning describes methods that enable computer systems automatically to 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. There are a wide variety of applications for machine learning on big social data such as natural language processing, topic detection, text classification and sentiment analysis.

One of the most common approaches of machine learning is the classifier system. A classifier system can be described with a spam filter, which labels an email as "spam" or "not spam". Input for such a system might be a "Boolean vector $x = (x_1,...,x_j,...,x_d)$, where $x_j = 1$" if the word appears in the dictionary. Otherwise $x_j = 0$. "The learner now inputs a training set of examples (x_i, y_i), where $x = (x_1,...,x_{i,d})$ is an observed input and y_i is the corresponding output" (classifier). Afterwards, the learner checks whether the classifier "produces the correct output for future examples" [Dom12, p. 1].

2.4 Data Mining

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]. However, compared to machine learning, it is more focused on the discovery of unknown information instead of the prediction. In general, data mining is used as a synonym for the process of discovering knowledge from data and usually includes the following iterative phases: data cleaning, data integration, data selection, data transforming, pattern detection and knowledge representation [HKP12, pp. 6 sqq.].

2.5 Natural Language Processing

Natural Language Processing (NLP) is "a set of techniques [..] to analyze human (natural) 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 (pragmatics) of a sentence [LM11].

Common application areas of NLP include stemming[1], named entity recognition[2] and sentiment analysis (explained in chapter **??**).

2.6 Trends

Trend detection methods are used to detect emerging topics or trends by using Natural Language Processing 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. Therefore, the text data is lowercased to prevent ambiguity and complexity caused by case-sensitiveness. Furthermore, a stop word removal process filters out extremely common words to speed up the processing and emphasize the important terms. In addition, stemming methods are used to merge and reduce words to a common base. After the preprocessing of 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; TR12; ZWL13, p. 1427] (e. g. #GERUSA or #WorldCup). Hashtags are well suited for trend detection by measuring the number of uses in a time interval [ZWL13, p. 1427].

[1] Process for reducing words to their stem or root form.
[2] Method to recognize well-known entities (e.g. person, location) in text.

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].

3 Use Cases

3.1 General Use Cases

3.2 Stock Market Prediction

Predicting the trends of the stock market is hugely important for today's businesses. However, a precise prediction seems to be very complex since the prices "follow a random walk pattern and cannot be predicted with more than 50 percent accuracy" [BMZ11, p. 1]. However, Twitter can predict the stock market if the right Tweets are analyzed [BMZ11]. 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].

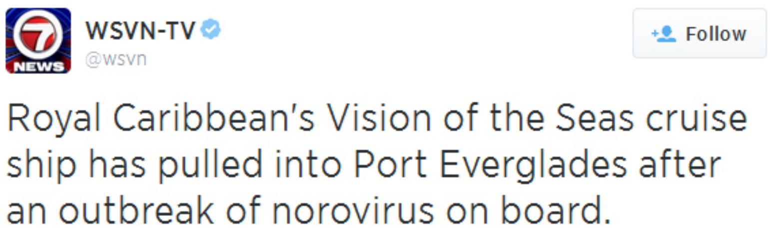

Figure 1: Tweet announcing the outbreak of norovirus on March 8, 2013 [WT13]

In 2013, a cruise ship of Royal Carribean arrived with more than 100 passengers sick with norovirus. A news agency published a Tweet announcing the outbreak of norovirus (see figure 1). Dataminrs' clients got this news two minutes later, but 48 minutes earlier than others, because their algorithm "found that words in the tweet had some resemblance to tweets in the past that had turned out to be newsworthy". According to Dataminr, the alert saved money of at least one client, due to a falling share price. Besides financial clients, also government organizations are interested in Dataminrs' Twitter analysis [Alc13].

3.3 Flu Trend Prediction

Seasonal influenza is responsible for millions of illnesses and up to 500 thousand deaths per year. Therefore, it is known as a major health issue all over the world.

An early detection of epidemics would reduce the significant effect of pandemic and seasonal influenza. The project **Google Flu Trends** aims to monitor flu cases in real time and thereby predict flu trends by analyzing social datasets [Gin+09; Tec14, p. 1].

The Google-researchers identified 45 keywords with a strong correlation to the appearance and spread of seasonal flu [Web14]. With these keywords, it should have been possible to get information about the spread of flu or even the start of a new wave of influenza [Web14; Tec14; Goo14]. Figure 2 visualizes the flu activity in the United States.

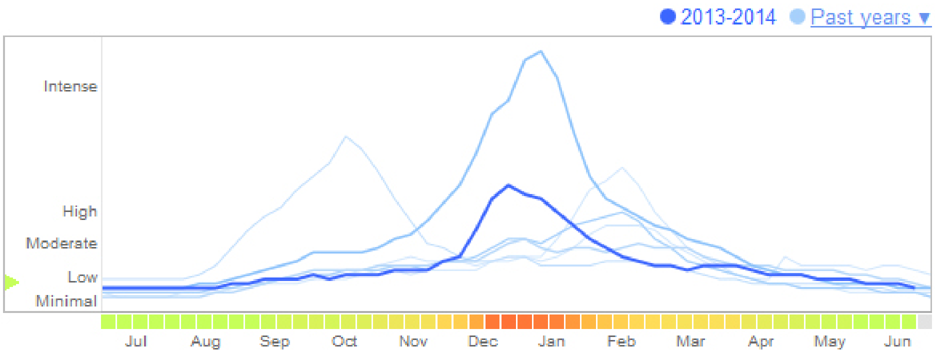

Figure 2: Flu activity in the United States [Goo14]

However, the project overestimated peak flu cases in the past two years and even failed to detect the H1N1[3] pandemic in 2009 [Tec14]. Figure 3 illustrates the estimated flu activity compared to official data. The overestimation might have happened because of not having investigated data validity or reliability [Web14; Tec14].

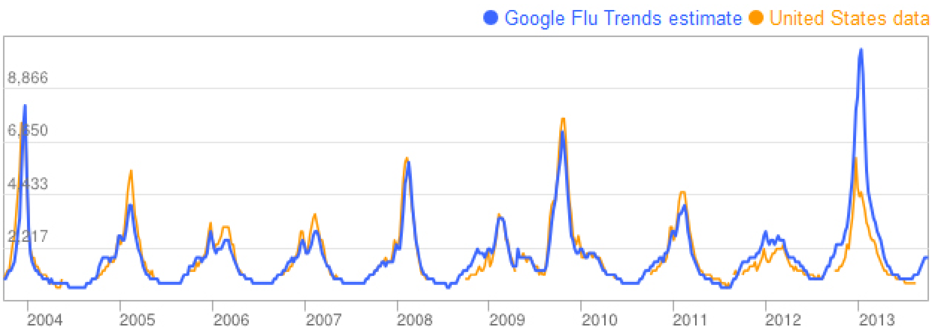

Figure 3: Google Flu Trend estimation compared to the real data[4] [**gftcomparison2014**]

[3] `http://www.cdc.gov/h1n1flu/qa.htm` [Online; accessed 07-08-2014]
[4] delivered by U.S. Centers for Disease Control `http://www.cdc.gov/` [Online; accessed 07-08-2014]

Ryan Kennedy, a professor at the University of Houston stresses, that "Google Flu Trend is an amazing piece of engineering and a very useful tool, but it also illustrates where Big Data analysis can go wrong" [Tec14]. Kennedy concludes that more accurate results could have been achieved by combining Big Data analysis with more traditional methodologies [Tec14].

4 Trend Detection on Twitter: Concept

4.1 Related Work

Not sure if this is best place for a twitter description, but the text is good

Twitter, a popular microblogging service with over 255 million active monthly users[5], 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[5]. By offering free API's to access this huge amount of unstructured data, Twitter attracted many professionals to collect and analyze Tweets to gain valuable insights on anything from stock market to natural disasters (presented in chapter 3).

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[6] 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 to track certain keywords to analyze the sentiment of corresponding Tweets in real-time and visualize 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[7], for analyzing the sentiment about the weather at a specific location, and trendsmap[8], for visualizing upcoming localized trends on a map.

[5] `http://about.twitter.com/company` [Online; accessed 07-08-2014]
[6] `http://mapd.csail.mit.edu/tweetmap-desktop` [Online; accessed 07-08-2014]
[7] `http://www.sproutloop.com/prediction_demo` [Online; accessed 07-08-2014]
[8] `http://trendsmap.com` [Online; accessed 07-08-2014]

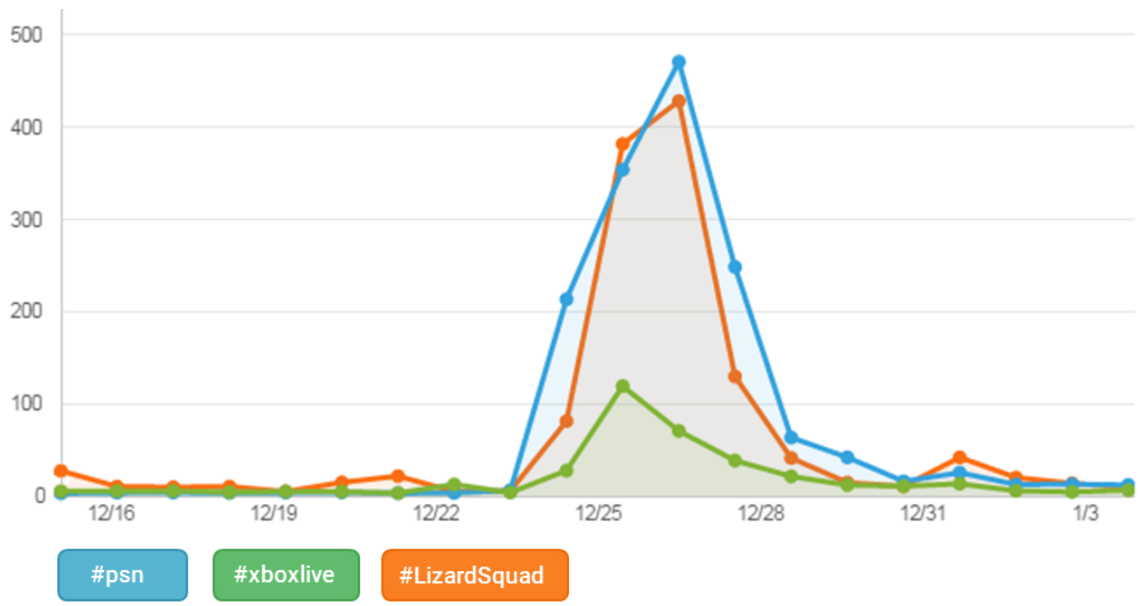

Figure 4: Christmas Network Outage Word Cloud

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 previously collected dataset of 48 million tweets while we try to achieve results on a much smaller dataset and with live data instead of historic data. Zubiaga et al. focus on the classification problem 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]. We however try to identify trends without knowing that Twitter declared them as trending.

4.2 Setup / Limitations

decide on section title

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 collecting 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 1% of the total number of Tweets at any given moment[9].

The use of commercial sentiment analysis API's would be too expensive for this huge number of Tweets. Therefore, we utilized free machine learning technologies

[9] `http://dev.twitter.com/docs/faq` [Online; accessed 07-08-2014]

for this task. However, only the first five million Tweets have been classified with a sentiment by machine learning techniques due to the expensive computing power that is required for such a huge data set. We restricted our analysis not only on tweets in English but also to tweets from anywhere in the United States of America. This makes it easier to use geospatial visualisations techniques.

4.3 Analysis Methods

Mining, storing, analyzing and visualizing terabytes of unstructured data requires optimized and new cutting edge technologies.

Since traditional relational **databases** cannot meet these requirements [KML13], new NoSQL databases[10] had been invented, such as MongoDB[11], Apache Cassandra[12] and CouchDB[13], that makes it possible to store, manage and analyze the huge amount of unstructured data in real time. Further optimization can be achieved by using Apache Hadoop[14] to distribute the data storage and processing across machine clusters.

4.3.1 Data Preparation

Natural Language Processing is an important part of the analysis of big social data. Toolkits such as Python NLTK[15] and Apache OpenNLP[16] offer a rich set of algorithm for tokenization, stemming, named entity recognition, stop word removal and more.

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

[10] NoSQL ('Not Only SQL') represents a new type of data management technologies created to meet the new requirements to process, store and analyze Big Data.
[11] `http://www.mongodb.org` [Online; accessed 07-08-2014]
[12] `http://cassandra.apache.org` [Online; accessed 07-08-2014]
[13] `http://couchdb.apache.org` [Online; accessed 07-08-2014]
[14] `http://hadoop.apache.org` [Online; accessed 07-08-2014]
[15] `http://nltk.org` [Online; accessed 07-08-2014]
[16] `http://opennlp.apache.org` [Online; accessed 07-08-2014]

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 analysed and words in different forms are used like *am*, *are* or *is* a stemming algorithm would then find the common base form as *be*. There exist different approaches for stemming like just cutting off the ends of words and hoping for a good result. More advanced approaches try to find 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 where documents are analysed 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. A bag of words model is much simpler than applying topic modelling algorithms described in the next section and therefore more convenient in some cases. [MRS08, p. 117]

4.3.2 Sentiment Analysis

Sentiment Analysis is a widely used NLP technique to analyze social media data. Therefore, many companies, such as AlchemyAPI[17], ViralHeat[18] and TextAlytics[19], 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[20] and Mallet[21], offer similar algorithm 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. Apache UIMA[22] and Gate[23] provide extensible frameworks to combine and manage these technologies for the analysis of unstructured information.

Insert figure and reference it

[17] `http://alchemyapi.com` [Online; accessed 07-08-2014]
[18] `http://viralheat.com` [Online; accessed 07-08-2014]
[19] `http://textalytics.com` [Online; accessed 07-08-2014]
[20] `http://cs.waikato.ac.nz/ml/weka` [Online; accessed 07-08-2014]
[21] `http://mallet.cs.umass.edu` [Online; accessed 07-08-2014]
[22] `http://uima.apache.org` [Online; accessed 07-08-2014]
[23] `http://gate.ac.uk` [Online; accessed 07-08-2014]

4.3.3 Topic Modelling

add correct citations

Topic modeling is a statistical machine learning model for 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). Latent Dirichlet Allocation (LDA) is a common method of topic modeling introduced by Blei et al. [BNJ03]. The LDA method assumes that each document contains a mixture of topics where each word attributes to one of these topics [BNJ03]. The requirements for executing LDA for topic modelling are a collection of documents, a specified number of topics and specified number of iterations. A highly simplified process description of LDA is described below:

1. Go through every document and assign a random topic (from the specified number of topics) to each word occurrence
2. Count up the number of assignments of each word occurrence for every topic (how many times for each topic does a word appear)
3. Resemble the topic assignment for one selected word occurrence in a document:
    - Delete the assignment of the selected word occurrence
    - Compute conditional distribution over all possible topic assignment of the selected word:
        - A = number of assigned word occurrence categorized by their topics for the document
        - B = number of times each topic appeared in the document
        - C = number of assignments of each word for every topic
        - (A+B)*C = topic with highest value is new topic assignment for the selected word occurrence
4. Repeat with step two for the specified number of iterations

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

Insert figure and reference it

4.3.4 Visualization

To understand and interpret the results of this big social data 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 on the vertical axis. The time series evaluation are particularly good when it comes to detecting new trends. Most trending topics will not show up in previous data at all but as soon as they begin to trend they show as clear peaks in the times series graph.

Insert figure and reference it

Figure XX shows a time series visualization of the hashtags *XXX* and *XXX*. There is an observable peak of both hashtags on XXXXXXXXXX which is very hard to spot when solely looking at the data without any visualization. The big advantage of the time series visualization over the other visualization techniques is that it considers the time. That allows us to see when a topic begins to trend.

Word Clouds

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 read from the database and then the frequency of each word in the text is counted using wordfreq.js[24]. Finally, every unique word and the associated frequency is forwarded to wordcloud2.js[25], a JavaScript visualization library, to render the corresponding word cloud.

Insert figure and reference it

[24] `http://timdream.org/wordfreq` [Online; accessed 07-08-2014]
[25] `http://timdream.org/wordcloud2.js` [Online; accessed 07-08-2014]

An example word cloud is depicted in figure XXX. ABCDE and FGHIJK are the most common terms for the detected 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 event [KML13, pp. 64-66]. As mentioned in section 4.2, the geospatial visualization is limited only to English and geolocated Tweets. All visualizations are build with CartoDB[26], a cloud-computing platform that provides mapping and visualization solutions for geolocated data.

Insert figure and reference it

Lorem ipsum figure XXX describes flow of topic around the usa with major impact in new york ....

4.4 Architecture

The Twitter Stream Reader is implemented with Python using the Twython[27] library to access the Twitter Streaming API[28]. The streaming data from Twitter is filtered based on . A Tweet contains a 140 character text message and various metadata such as the language, location, user information, number of retweets and favorites and more.

finish sentence!

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]. Therefore, each Tweet is preprocessed in the Data Analysis Module using common NLP text preparation techniques to remove these elements. In the first step, the text of a Tweet is lowercased and special characters, URLs as well as English stop words[29] get removed.

Check if following fits into our setup

In the next step, the preprocessed Tweet text alongside with the original Tweet text,

Fit paragraph into text

[26] `http://cartodb.com` [Online; accessed 07-08-2014]
[27] `http://twython.readthedocs.org` [Online; accessed 07-08-2014]
[28] Push service to collect public Tweets in realtime.
[29] Words that do not contain important significance or are extremely common (e.g. the, a, want).

creation timestamp and all metadata is stored into MongoDB, a popular NoSQL database that is used as the main data store for our implementation.

5 Trend Stories

5.1 Air Asia Flight Tragedy

addTOPSY BILD

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 Singapore. It was around 06:12 local time when the pilot contacted air traffic control to request a change in flight altitude. The pilot wanted to climb from 9.500 metres up to 11.500 metres 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.[Bbcb]

Most of the families and relatives of the passengers are still in a deep grief since only 40 victims have been found by now. Experts assume that most passengers are still strapped to their seats in the missing main body of the airplane. As today no survivor has been found and the search is still being continued.[Bbca]

When first hearing from the awful tragedy many people thought of the flight 370 from Malaysia Airlines (MH370) which got lost on march 8th. On board of the flight were 239 people including passengers and crew. The search for the plane or its black box have been unsuccessful until today.[Nbc]

As shocking this news it we were able to identify an uprise of related tweets on twitter. People were using the following hashtags to discuss this news or to express griefs and sympathy with the families and relatives:

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

As mentioned earlier many people connected the crash of Air Asia flight 8501 with the disappearance of the Malaysia Air flight 370 that is why both flight numbers are trending topics.

link to image word cloud air asia!

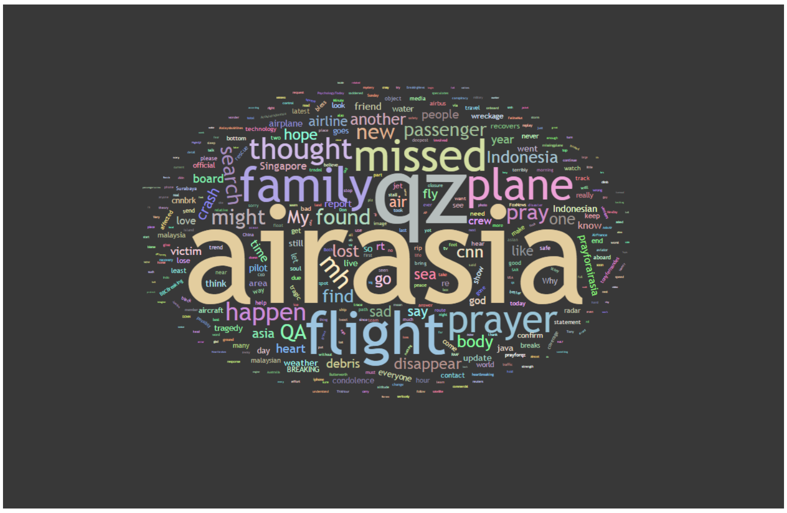

Figure 5: Air Asia Flight Tragedy

The word cloud above depicts the most commonly used words in tweets about the plane crash. A hypothesis based on the wordcloud is that the tweets have two different subjects. One subject is news and tweets are there to inform others about the tragedy and the other subject is emotional and shows condolence to families of the victims. We extracted all hashtags from our dataset and used LDA for topic modelling in order to further analyze our hypothesis.

link to listing!

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

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

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

cnn (13) put (7) speculation (6) ground (6) airasia (6) speed
   (5) stop (5)

airasia (140) found (65) plane (53) sea (51) bodies (49) search
    (49) debris (40)

airasia (146) flight (122) amp (99) happened (87) disappearance
    (14) malaysia (7) trends (6)

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

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

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

Listing 5.1: Topic Model for Air Asia Flight Tragedy

We used LDA to model nine different topics showing the 7 most relevant words of each topic. There is an observable difference between reporting tweets (like topic 0, 1, 4 and 7) and emotional tweets (like topic 2 and 6). Topics 3 and 8 stand out from the other, topic 3 is about the famous news network CNN which was one of the first to bring coverage about the crashed plane. Topic 8 on the other hand 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.

5.2 Christmas Network Outage

On the 24th of December in 2014, hackers started to attack the Playstation Network and the Microsoft Xbox Live Network. The DDoS attacks brought the networks down for several days. The gamer community was infuriated 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 of his web platform Mega [Dot14]. In return, Lizard Squad stopped the attacks letting the gamers play again. Twitter was used by the companies Microsoft and Sony, the gamers, the attackers and Dotcom for discussion, asking for support and negotiating. After the network recovered, Sony announced to give discounts to PSN users. The involved persons and instances and events are reflected in the following list of hashtags and user mentions that were used to tag the related tweets:

#finestsquad
#lizardpatrol
#lizardsquad
#payingfornothing
#playstationnetwork
#playstationsucks

#psn
#psndown
#PSNDownTime
#psnup
#xboxlivedown
#xboxsupport

@AskPlayStation
@KimDotcom
@LizardMafia
@MEGAprivacy
@PlayStation

We fetched all tweets containing at least one of the listed hashtags or user mentions and created a word cloud. The resulting word cloud consists of the words used in the fetched tweets.

link to word cloud image!

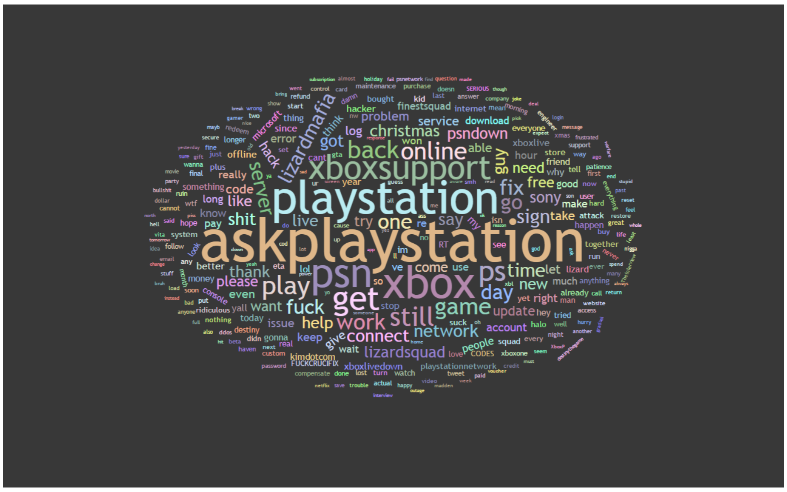

Figure 6: Christmas Network Outage Word Cloud

The words ‘askplaystation’, ‘playstation’, ‘xbox’, ‘xboxsupport’, ‘psn’ form the core of the cloud.

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 queried tweets by using LDA. The identified topics are displayed in the following:

The words that formed the core of the cloud are also dominating the detected topics. Furthermore, the detected topics reflect the real events in a very good way.

Topic 1 covers words indicating a discussion about a connection between the DDoS attack during Christmas and a previous hack against Sony. The earlier attack happened in the end of November, 2014 concerning the movie ‘The Interview’.[Bbcc]

Maybe explain a few more details

The second topic is about the hack affecting the Xbox Live Network. Obviously, a lot of people tweeted to Microsoft’s support. Topic 8 on is similar to topic 2, however, this time the words are concerning Playstation. Topic 4 covers words about fixing the problem for Playstation as well as Xbox.

Topics 4, 7, 10 cover general terms concerning the hack and the inability to connect to the networks or to play a game.

Topic 5 covers words of tweets that are to or about Lizard Squad. The words indicate that the gamer community was not very amused about the hack.

Topic 3, 6 contain words about the financial impact of such a hack and the claim for redemption for the lost hours of being able to use the networks.

Interestingly, topic 9 contains the word ‘Halo’, which is a game series developed by Microsoft. In another attack in December 2014, parts of the source code of the

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

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

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

playstation (467) askplaystation (362) shit (279) psn (273)
    xbox (270) play (246) fix (245) guys (197)

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

psn (223) play (217) free (184) games (166) game (153) online
    (145) xbox (134) codes (93)

xbox (95) game (58) warfare (29) controller (24) advanced (24)
    wait (22) copy (22) party (20)

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


halo (61) xbox (45) beta (42) guardians (20) multiplayer (19)
    xboxsupport (15) live (13) xboxp (12)

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

Listing 5.2: Topic Model for Christmas Network Outage

newest game of the series were stolen. Either the twitter community discussed a possible relation between the two hacks or they were upset not being able to play the current version of the game.[Gri]

6 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, natural language processing and data mining technologies are necessary to gain valuable insights into large amounts of social content.

Appendix

A Additional Tables and Graphics

Literature

[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.

[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`.

[Bbca] *AirAsia QZ8501: Tail of crashed plane found*. BBC, 7 January 2015. [Online; accessed 07-January-2015]. URL: `http://www.bbc.com/news/world-asia-30706298`.

[Bbcb] *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`.

[Bbcc] *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`.

[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).

[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`.

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

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

[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.

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

[Gri] Andrew Griffin. *Unreleased Xbox games could be leaked after hack.* Independent, 31 December 2014. [Online; accessed 07-January-2015]. URL: `http://www.independent.co.uk/life-style/gadgets-and-tech/gaming/unreleased-xbox-games-could-be-leaked-after-hack-9951880.html`.

[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`.

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

[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`.

[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.

[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`.

[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`.

[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).

[Sin+10] Rolf Sint, Stephanie Stroka, Sebastian Schaffert, and Roland Ferstl. "Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis." In: CEUR Workshop Proceedings 464 (Jan. 26, 2010). Ed. by Christoph Lange 0002, Sebastian Schaffert, Hala Skaf-Molli, and Max Völkel.

[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`.

[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.

[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`.

[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`.

[WT13] WSVN-TV. *Royal Caribbean's Vision of the Seas cruise ship has pulled into Port Everglades after an outbreak of norovirus on board.* [Online; accessed 08-January-2015]. Twitter. 2013. URL: `https://twitter.com/wsvn/status/310087727792140288`.

[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.