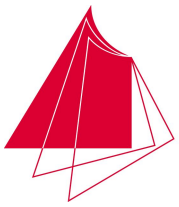

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

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

2.2 Social Media

2.3 Machine Learning

2.4 Data Mining

2.5 Trends

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].

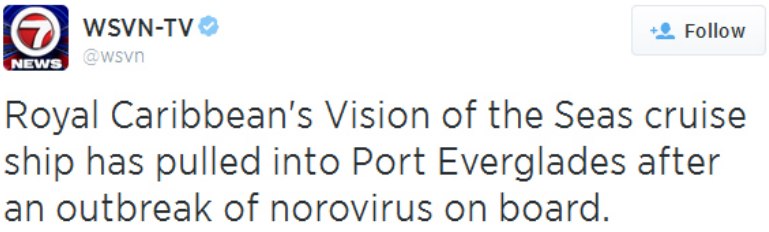

Figure 1: Tweet announcing the outbreak of norovirus on March 8, 2013 [WT13]

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.

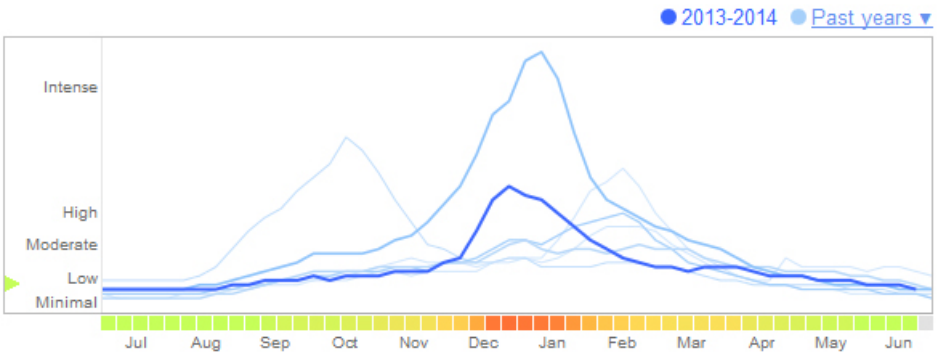

Figure 2: Flu activity in the United States [Goo14]

However, the project overestimated peak flu cases in the past two years and even failed to detect the H1N1[1] 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].

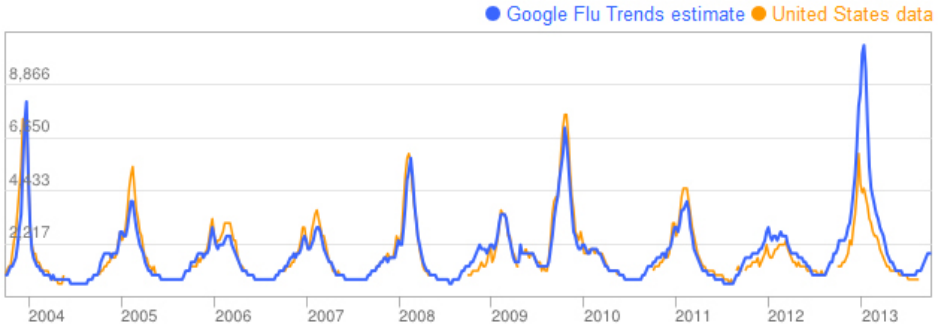

Figure 3: Google Flu Trend estimation compared to the real data[2] [**gftcomparison2014**]

[1] `http://www.cdc.gov/h1n1flu/qa.htm` [Online; accessed 07-08-2014]
[2] 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 Early Trend Detection on Twitter

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[3], 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[3]. 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.

4.2 Technologies

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[4] had been invented, such as MongoDB[5], Apache Cassandra[6] and CouchDB[7], 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[8] to distribute the data storage and processing across machine clusters.

[3] `http://about.twitter.com/company` [Online; accessed 07-08-2014]
[4] 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.
[5] `http://www.mongodb.org` [Online; accessed 07-08-2014]
[6] `http://cassandra.apache.org` [Online; accessed 07-08-2014]
[7] `http://couchdb.apache.org` [Online; accessed 07-08-2014]
[8] `http://hadoop.apache.org` [Online; accessed 07-08-2014]

4.2.1 Data Preparation

Natural Language Processing is an important part of the analysis of big social data. Toolkits such as Python NLTK[9] and Apache OpenNLP[10] 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 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]

4.2.2 Topic Modelling

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

[9] `http://nltk.org` [Online; accessed 07-08-2014]
[10] `http://opennlp.apache.org` [Online; accessed 07-08-2014]

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]. A highly simplified process description of LDA is described in appendix D.

add correct citations

reformulate to adapt to text

A highly simplified process description of LDA is described below: Requirement: collection of documents, specified number of topics, specified number of iterations

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

4.2.3 Vizualisation

Time Series Wordcloud Visualisation Geospatial Analysis

4.3 Architecture

The Twitter Stream Reader is implemented with Python using the Twython[11] library to access the Twitter Streaming API[12]. 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[13] get removed.

Check if following fits into our setup

In the next step, the preprocessed Tweet text alongside with the original Tweet text, creation timestamp and all metadata is stored into MongoDB, a popular NoSQL database that is used as the main data store for our implementation.

Fit paragraph into text

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[14].

Under Limitations sec?

[11] `http://twython.readthedocs.org` [Online; accessed 07-08-2014]
[12] Push service to collect public Tweets in realtime.
[13] Words that do not contain important significance or are extremely common (e.g. the, a, want).
[14] `http://dev.twitter.com/docs/faq` [Online; accessed 07-08-2014]

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, 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

[Alc13] Stan Alcorn. *Twitter Can Predict The Stock Market, If You're Reading The Right Tweets*. [Online; accessed 29-June-2014]. 2013. URL: http://www.fastcoexist.com/1681873/twitter-can-predict-the-stock-market-if-youre-reading-the-right-tweets.

[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).

[Coh13] Sara Estes Cohen. *Sandy Marked a Shift for Social Media Use in Disasters*. [Online; accessed 6-July-2014]. Emergency Management. 2013. URL: http://www.emergencymgmt.com/disaster/Sandy-Social-Media-Use-in-Disasters.html.

[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 6-July-2014]. 2014. URL: http://www.google.org/flutrends/intl/en_us/us/#US.

[IBM12] IBM. *IBM What is big data? - Bringing big data to the enterprise*. [Online; accessed 6-July-2014]. 2012. URL: http://www-01.ibm.com/software/data/bigdata.

[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.

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

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

[Tec14] Tech2. *Big data collection from Google, Facebook and others can be misleading, says study*. [Online; accessed 28-June-2014]. 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 28-June-2014]. 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.

[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 8-July-2014]. Twitter. 2013. URL: https://twitter.com/wsvn/status/310087727792140288.