# Twitter for failure

*The use of Twitter to detect failures in online systems*

# Twitter for failure

*The use of Twitter to detect failures in online systems.*

Reinier van den Assum

**Research paper for master Business Analytics**

**MeasureWorks**
Katernstraat 10d
1321 NE Almere
The Netherlands

**VU University Amsterdam**
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands

**Supervisor:**
Sandjai Bhulai

**June 2013**

*Sources front-page:*
*The images are derived from Google Images.*
*The words are the 50 most frequently occurring in the relevant tweets over the 21 days (see section 6).*

# 1 Acknowledgements

Part of the master Business Analytics at VU University Amsterdam, is to perform research and document this in a concise paper. After contacting my supervisor, Sandjai Bhulai, to brainstorm about the topic, I was redirected to MeasureWorks, a company that focuses on web performance. Fortunately, they had an open, unassigned project: ‘Can we combine data from social media, e.g., Twitter, with data of web traffic, e.g., number of visitors on a website, in such a way it will alert companies when their website malfunctions?’ Unfortunately, it was not possible to build the prototype requested, due to limitations in time and resources, therefore, only one of the essential parts of the overall question is developed.

My thanks in this research paper goes to Sandjai Bhulai, who helped me find and define the research question, but also supervised me throughout the process. I, also, would like to thank MeasureWorks, in particular, Jeroen Tjepkema, founder of MeasureWorks, for the opportunity to investigate and collaborate with them. Furthermore, I would like to thank André Hofman, an employee of MeasureWorks, since I was always able to call him for technical support and to gain access to particular data. Finally, I would like to thank Anne-Ruth van Leeuwen, who supported me during the process in more practical issues.

# 2 Summary

Due to the vast growth in social media, e.g., Twitter, expressions of humans become more important, and available. Currently, more companies become aware of the opportunities and start campaigns to narrow down the gap between their customers and helpdesk, e.g., creating Twitter accounts like '*INGnl_webcare*'. Next of being an extra communication channel, social media may be useful to derive the status of, for example, a website, since users may post messages (a.k.a. tweets) about related malfunctions. Deriving the right tweets and detecting whether an actual failure occurs, can save a lot of time and money for companies. Besides, the combination of failure-detection on Twitter with change-in-trend-detection on web traffic data, could prevent companies for reputation damage. ING, for example, was not yet aware of a failure, though NU.nl already published an article about it online.

The aim of this study was to develop an online alerting tool for Twitter data, applied for the case of ING, exclusively for website-performance related tweets. The research question stated was:

> '*Can we combine data from social media, e.g., Twitter, with data of web traffic, e.g., number of visitors on a website, in such a way it will alert companies when their website malfunctions?*'

To answer this research question an algorithm is developed that can be divided in three steps: (i) retrieval of Twitter data, (ii) tweet filtering, both (ambiguous) brand names and performance related topics, and (iii) trend detection, what makes a sequence of tweets a trend and when should the tool send an alert.

Based upon this automatic alerting tool developed, Data analyses revealed that the number of tweets is indeed related to the occurrence of a failure and therefore a possible measure to analyze the web-performance. Besides, manual analyses showed social media data can indeed be used in combination with web traffic data to alert companies when their website malfunctions., since the timestamps of the change in trends for both Twitter and web traffic data were relatively close to each other. Although, one should also be aware, it is possible the failure is not necessarily noticed directly by users or by ING and therefore this detection tool might not detect all failures of the website.

This paper, thus, provides insights in the possibility to combine Twitter and web traffic data to emerge malfunctions of websites. Although there are still some difficulties left, this paper provides a roadmap to develop an automatic alerting tool to detect events on Twitter and confirms the research question stated.

# 3 Contents

# 4 Introduction

Due to the vast growth in social media, e.g., Twitter[1], expressions of humans become more important, and available. Currently, more companies become aware of the opportunities and start campaigns to narrow down the gap between their customers and helpdesk, e.g., creating Twitter accounts like '*@INGnl_webcare*'. Monitoring these word-of-mouth (WOM) expressions may help companies to improve their products and services. Another major advantage is the bi-directional communication capabilities of social media, which can remind one of the functionality of Online Reputation Mechanisms (ORMs). ORMs attempt to engineer large-scale word-of-mouth networks, such as the feedback mechanisms of eBay (Dellarocas, 2003), to improve the online reputation of a brand. Social media, therefore, not only provide the ability to monitor WOM of customers, but also to preventively post messages, about for example a scheduled maintenance, or to support customers with questions. Hence, social media have become key applications in the attention economy, today (Davenport & Beck, 2002).

The research question studied in this paper is proposed by MeasureWorks[2]. MeasureWorks is a company in web performance optimization, to increase the functionalities of a website and perform analysis on web traffic data. The question proposed is to investigate the combination of available web traffic data, e.g., number of visitors on the company website, and data of social media. Combining these two datasets can result in an alerting tool that alerts when the system seems to malfunction. Using these datasets can outperform other tools, since only the company has access to the web traffic data. An example of implementation can be given via an anecdote of April 3, 2013, when customers of ING, a Dutch bank, saw wrong balances on their personal pages. This caused over ten thousand messages on social media, concerning the malfunctioning of the website and mobile applications (on an average day, a known Dutch company is mentioned in about 300 posts) (Greenberry, 2013). As consequence of the first messages posted, other customers also visited the website to verify whether their balance was incorrect as well. The result was that the servers got overloaded by requests and started to fail. In addition, news website NU.nl[3] mentioned the failure even before ING was aware of it, using a social media tool. The tool requested, should thus be able to prevent certain unfortunate situations and outperform existing social media tools.

This paper addresses one of the challenges within the development of the requested tool: live retrieval of all social media posts relevant to a specific company and the performance of the corresponding site. Therefore, the problem is restricted to web performance related posts concerning ING and posted on Twitter. Besides, it will try to answer the following question:

> '*Can we combine data from social media, e.g., Twitter, with data of web traffic, e.g., number of visitors on a website, in such a way it will alert companies when their website malfunctions*'.

Twitter is a micro-blogging service were users can send messages (a.k.a. *tweets*) into the Twittersphere[4], which are directly shown in their network of friends (a.k.a. *followers*), without any privacy conditions. Twitter can be seen as a public, real time pulse of opinions of people. One can imagine that identifying

[1] http://www.twitter.com

[2] http://www.measureworks.nl

[3] News article concerning the failure, posted by NU.nl in Dutch:
http://www.nu.nl/internet/3388298/veel-onrust-verkeerde-saldos-ing.html

[4] Twittersphere (noun): postings made on the social media website Twitter, considered collectively. By Oxford dict.

events on Twitter is thus a challenging problem due to the inherent noisy and heterogeneous data. Associated problems are caused by the short messages (140 characters) in informal language, often with specific abbreviations and incorrect grammar, which makes it hard for a computer to interpret the tweets. Thereby text-mining solutions are hard to apply, such as dealing with ambiguous company names, e.g., *Apple* and *Jaguar*, and company names containing parts used in everyday language, such as '*ING*' in 'doing'.

Despite the challenges analyzing tweets, many studies have been performed, concerning mostly different purposes, such as: finding influential tweeters, ranking tweets and users, opinion mining, categorizing and summarizing tweets, and so on (Yerva, et al., 2010). The research question in this paper, however, can be seen as a combination of previously performed studies and can be divided in the following steps:

(i) First, all tweets have to be filtered to yield only the company-related tweets, dealing with disambiguation and brand name selection. A similar task was addressed in the WePS-3 CLEF lab exercise in 2010 (Amigó, et al., 2010). The objective of this classification problem was to solve the issue of company name disambiguation in social networks. Yerva et al. (2010) sent the best algorithm constructing several profiles, most of them automatically, containing sets of positive and negative keywords. A negative keyword indicates an unrelated tweet to the company (such as '*crumble*' for *Apple*). These automatically-derived keywords were retrieved via the company homepage, metadata, a category profile and the use of Google Sets. Furthermore, some keywords were retrieved from user feedback. Afterwards they used a classification algorithm, to learn which tweets were related, based on the keywords found in the profiles. The use of (positive and negative) filtering keywords is frequently used to solve this kind of problems, e.g., other WePS-3 participants (García-Cumberas, et al., 2010; Tsagkias & Balog, 2010) and Spina (2011; 2013).

(ii) After gathering all tweets mentioning the corresponding company, it is essential to classify whether a tweet is relevant for analyzing web performance and thereby indicates a malfunctioning of the website or mobile application. This is a fairly similar assignment and can therefore also be realized using filtering keywords, though these keywords can only be derived manually by learning from the tweets. A technique that can be used is clustering tweets based on co-occurrence (Mathioudakis & Koudas, 2010) and cosine similarity (Bhulai, et al., 2012) . Based on the generated clusters a manual selection can be made to specify which clusters are web-performance related.

(iii) Finally, a measure should be introduced to indicate whether there is an event and thus an occurrence of deviation in the standard trend. This measurement is introduced by Bhulai et al., (2012) based on the speed and acceleration of consecutive tweets.

To summarize, multiple studies are performed with different purposes, though, to the best of our knowledge, no study was performed to alert the occurrence of a significant number of tweets for a particular topic of one specific company. Therefore, the relevance of this research is to provide other researchers with a grip to use social media for online reputation management. Furthermore, this study may provide companies insights in the available opportunities and their flaws of customer communication.

First, section 5 introduces the relevant work to gain insights in the current knowledge of mining company specific social media texts, reputation management and other related topics. Section 6, explains the Methods used to develop the automatic alerting tool for Twitter trends, followed by some Data analyses. The final sections are the Conclusion and discussion, and Future work (respectively section 8 and 9).

# 5 Literature background

This section introduces some jargon and provides insights in already performed studies.

## 5.1 Common terms

**Word-of-mouth (WOM)** is commonly defined as the informal communication about the characteristics of a business or product between customers (Westbrook, 1987). It is the process of conveying information from person to person, most of the time in commercial situations, involving information such as opinions and personal experiences (Jansen, et al., 2009). Ha (2002) and Jansen et al., (2009) show customers tend to trust these opinions of people outside their immediate social network and are thereby more influenced by WOM communication than for example results of online auctions or commercials. Generally the intention of recommendations is to discourage, as discovered by Ford Motor Company (Dorlin, 1985). They found that satisfied customers told 8 people about their cars, though dissatisfied customers told 22 people about their complaints.

Nowadays the number of comparison websites is growing vast, all providing the option to write and read product- or service-related reviews to enable customers to spread their experience. This type of communication can be categorized as **eWOM** (electronic Word-of-Mouth) and is a major influence for the **online reputation management (ORM)**. It can either make or break a company's reputation within short notice, like the case of Kryptonite, creator of bicycle locks, in 2004, when an online post revealed the possibility to open an expensive Kryptonite lock with only a ballpoint in a few seconds[5]. The result was a replacement project of 400,000 locks for free and a project to redesign all products made in the past 9 years (Hoffman, 2008). Furthermore, one can imagine people will always remain careful in buying a lock of Kryptonite. Therefore, ORM is important for a company to 'control the message', instead of repairing afterwards. ORM gives the opportunity to be pro-active and to reach a large community at very low cost (Dellarocas, 2003). Since people trust online reviews (Jansen, et al., 2009), companies can monitor and analyze eWOM to detect problems and control these by interacting with the user, via bi-directional communication capabilities (Amigó, et al., 2010).

To monitor and analyze eWOM for a specific brand, it is necessary to be able to gather the corresponding data from all real-time posts on the web. This is quite similar to the task of the **WePS-3** (Third web people search evaluation campaign) competition, which was to automatically filter out tweets that do not refer to a certain company, especially focused on ambiguous brand names. This classification problem was given for 28 specifically chosen brand names and was, restricted to only Twitter posts (a.k.a. *tweets*), ignoring blog posts and other social media, such as Facebook and Instagram. For each brand name both the homepage URL and the first 100 tweets were provided. In order to create a classification problem, all tweets were manually classified by 902 humans, using Mechanical Turk[6]. Each annotator received a number of tweets to classify it as 'related', 'non related' or 'undecidable', for tweets like: "I love my MacBook of Apple", "Just made a nice apple crumble!" and "Apple is a good choice", respectively. Based on these classifications, it was possible to create a ´test set´ and build classification models. The best approaches will be discussed later on.

[5] Article: '*Twist a pen, open a lock'*, September 14th 2004 http://www.wired.com/culture/lifestyle/news/2004/09/64987
[6] Market place for intelligent jobs. https://www.mturk.com/mturk/

## 5.2 What we can say about Twitter

**Twitter** is a micro-blogging service, launched on March 21, 2006[7], which allows users to express themselves through text-based messages of 140-characters, named *tweets*. Tweets are public and can thus be read by anyone. To structure tweets, users can '*follow*' friends, co-workers or idols, without the need of mutual permission (Jansen, et al., 2009), whereby only the tweets of those *following users* are shown on the users' dashboard. Another way one can specifically select tweets of interest is using Twitter Search[8]. Searching tweets is made easy by the well-defined markup vocabulary, where 'RT' stands for retweet, '#' indicates a topic of the corresponding tweet and '@' can be used to address a user (Kwak, et al., 2010). An example of a (re)tweet can be:

> *"RT @INGnl_Nieuws: #ING introduces innovative bank card (…)"*

Twitter has been growing rapidly: after 3 years, 2 months and 1 day the billionth tweets was posted and each day the 288 million regular users send on average 400 million new tweets into the Twittersphere[4] (Holt, 2013). Due to this large number of opinions shared every day, it is hard to monitor what users are tweeting about. Therefore Twitter Trends provides insight in the real-time emerging topics (Twitter, 2010; Bhulai, et al., 2012). A side-effect of this helpful tool is the large number of 'spam tweets'. Danny Sullivan (2009) notice spammers use the Twitter Trends feature to increase their target group. Adding hashtags (#) of trending topics will show the spam tweets among relevant tweets and thereby ensure spammers of a large audience. When analyzing tweets, one should thus be very aware of spam and use a coarse filter that, e.g., ignores tweets naming three or more trending topics (Sullivan, 2009).

As consequence of the large number of activities and the public access many studies have already been performed to, e.g., analyze the habits of users, which users and tweets are most influential, opinion mining, how to categorize and summarize tweets and so on (Yerva, et al., 2010). Furthermore, the Twittersphere also drew the attention of companies and journalists. According to an investigation performed by ING and Social Embassy (2013), it is apparent that PR and journalism, today, are inseparable of social media and the role of social media is only increasing. Though PR professionals and journalists still use their traditional (offline) resources, these resources become less relevant for your daily job. Besides, the speed of emerging news increased, which can be detrimental to the verification and thus the reliability. Furthermore, social media placed the journalists and PR professionals in a triangular relationship with the consumers in key role. Hence, a more transparent information retrieval can be build, as consumers can participate in the newsgathering and can be contacted for further information.

## 5.3 What we can derive from social media?

Next to the possibilities of using Twitter to share experiences and interests and look for others, Twitter can, thus, as mentioned above, also be used to emerge news topics and be a tool within Online Reputation Management (ORM).

[7] Dorsey, Jack (March 21, 2006). First tweet: "just setting up my twttr". Twitter. Retrieved June 22, 2013.

[8] Twitter Search https://twitter.com/search-home

In 2010 TwitterMonitor (Mathioudakis & Koudas, 2010) was developed to detect trends over the Twitter stream. Trends are generally driven by emerging events and general topics that attract the attention of a large fraction of Twitter users. Thereby, the trends can be of large value for news reporters and analysts to find fast-evolving news stories (Mathioudakis & Koudas, 2010). A majority of Twitter topics (over 85%) are headline news or persistent news in nature (Kwak, et al., 2010). Also when generalizing to social media (including blog, etc.), many discussions are inspired by news (Tsagkias, et al., 2011), as are the search requests (Mishne & de Rijke, 2006). A better insight in the relation between bloggers and news events was provided by Thelwall (2006) by analyzing blog-posts at the time of the London attacks. Blogging provides fewer insights in real-time events than tweets as there is no limitation to the length of the message and therefore blogs are generally longer than three paragraphs. On the other hand, it provides better insight into contemporary discussions. Also in contrast with Twitter, blogs are scattered across the Web and are less public since one should be familiar with the specific hyperlink to find and read the blog. Thelwall's major finding is the difference in posting frequency. During the attacks in London, the posting bloggers were found to be 'atypical, linking and posting much more frequently than general bloggers' (Thelwall, 2006). The blogs, thereby, became (slightly) 'more real-time'. Another finding is the use of links within blog-posts. During the attacks, fewer links to 'top linked' sites were posted, which can be equivalent to the previous finding. In general, bloggers write about other websites they ran into, including their opinion. During the attack, they did not have the possibility to do 'research' before sharing their opinion.

Trend detection can also be important for online marketing professionals and opinion tracking companies, as trends point to topics that capture the public's attention (Mathioudakis & Koudas, 2010). Tweets can, for example, serve as a significant repository for companies to monitor their online reputation (Qureshi, et al., 2012). Next to obtaining reputation-related information, and thus also study product-related opinions (Yerva, et al., 2010), micro-blogging services can also be used to immediately intervene with unsatisfied customers and retrieve near real-time feedback, by, for example, setting up an corporate user-account (Jansen, et al., 2009), such as the Twitter account 'INGnl_webcare'. Currently, already 60% of large companies say they did acquire customers through Twitter (Lockwood, 2013).

However, social media is not only a medium to acquire customers, as is mentioned before. It is evident negative opinions in online media can seriously harm the reputation of a company (Amigó, et al., 2010), as is mentioned above by the example of Kryptonite (Section 5.1 - Common terms, ORM). To emerge such negatively toned posts, one should identify messages related to the corresponding company, which is quite a challenge since company- and product-names often are homonyms (ambiguous) (Yerva, et al., 2010). Another phenomenon, that creates a difficulty, is called '*power law*'. This means humans only talk about things others talk about. Therefore, online topics are not completely random, but strongly determined by actively spoken topics (Greenberry, 2013). Therefore, companies sometimes call social media 'earned media', as it is hard to initiate a conversation about your brand (Barabási & Bonabeau, 2003). Additionally, it is hard for companies to prevent discussions with negative sentiment, once started. Though, it should be mentioned sentiment analysis of a brand is relatively fluctuating, related to events happening. If a server malfunctions this may cause an explosion of negative social media messages. Therefore the online reputation decreases fast. However, the online communication agency Greenberry (2013) shows that one event of malfunctioning of service did not have a structural effect on the online sentiment of ING, as within two weeks the sentimental score was recovered to the score before malfunctioning (Greenberry, 2013). Though, their conclusion is not verified over a longer period, and is only concerning one failure. Perhaps multiple failures in a short period will harm the long-term reputation. Furthermore, it is possible different users are captured. There is no guarantee the current sentiment score can be applied to all users, as it is only derived from a subset of users who are expressing themselves. Therefore, it is possible the sentimental score is normal, though there is an increase in the number of unsatisfied users that are not tweeting about the product or brand.

## 5.4 140 Characters challenge

Thus, micro-blogging can be seen as a 'promising competitive intelligent source' (Jansen, et al., 2009), to help companies and news reporters obtain reputation- and product-related information from social media. However, identifying posts related to a company is quite challenging, as is mentioned earlier. This section elaborates on these challenges.

Since tweets have a maximum of 140 characters, there is little information and thus little context for possible text-mining applications. As a consequence a specific 'chat-speak' or 'SMS style' language is used, full of phonetic spelling and acronyms for common phrases, such as 'lol' ('laughing out loud'), 'omg' ('oh my god') and 'gonna' ('going to'). Thereby, tweets often are purported to be, e.g., English, though they are not (Gouws, et al., 2011). Hence, Twitter data is often described as inherently noisy and heterogeneous and thus also misleading or ambiguous, because of the not uniform quality (Becker, et al., 2010). Besides the specific use of language, ´spam´ tweets often occur, which promote certain products, or refer to malware websites (Sullivan, 2009). Therefore, Twitter data should be filtered coarsely before analyzing. Another obstacle gathering the correct tweets is no learning algorithm can be set up perfectly, as tweets are unknown in advance and only historical data can be used (Yerva, et al., 2010). Unfortunately, one can imagine the language used on Twitter is constantly changing, as are the topics users talk about. Hence, filters for a specific topic on Twitter should be maintained often and will never cover all tweets.

As mentioned, filtering tweets for a specific company can, thus, be quite challenging. In particular because many tweets compare companies (Yerva, et al., 2010), such as:

> *"A failure at #ING turns the complete population topsy-turvy. Currently #Rabobank has a failure, though nobody is talking about that! Weird!!!"*

On the other hand, social media data does include information, traditional documents, such as books, do not contain, e.g., tags, user information, time of creation, etc., which can also be used when mining texts (Becker, et al., 2010). For example, the creation time of a tweet can give insights in the influence of that particular tweet. Yang and Counts (2010) discovered a disempowerment for the assumption that the earliest tweet of a topic is most important. Therefore, one should continue watching a topic, as tweets arriving later in the observation often are more influential.

Furthermore, the importance of a tweet can be derived by the feature whether it has been retweeted. Any retweeted tweet is likely to reach an average of 1,000 users, independent of the number of followers of the original tweet. This is caused by the phenomenon that a retweeted tweet is almost immediately retweeted by others (Kwak, et al., 2010). Hence, a retweeted tweet containing negative sentiment can cause more damage to the reputation of a brand than a single not-retweeted tweet. Finally, it is possible to rank users to indicate the influence of the tweet. Kwak et al. (2010), for example discovered a similarity in rankings by number of followers and PageRank (calculated by Twitter), but a difference in ranking by retweets. The number of followers is, thus, not directly the best indicator of user influence, as the number of average retweets performs a more accurate measure.

When specifically filtering tweets for a certain company or topic, one should also be aware of the possibility of ambiguous brand names (e.g., *Blackberry* and *Ford*). Besides, a tweet can contain words related to a topic or brand, though not be actually related, such as the tweet comparing Rabobank and ING. The tweets contain the tags 'ING' and 'failure', though, do not indicate an actual failure at the ING. However, these problems already have been investigated quite thoroughly (Hearst, 1991; Sarmento, et al., 2009; Mihalcea & Moldovan, 1991), though these algorithms cannot be applied in this study, because of the limited context within the tweets and the language used. The algorithm of Hearst (1991) for example, 'CATCHworD: Corpus-based automatically trained coarse homograph disambiguator', does only perform correctly on 'large text corpora' (Hearst, 1991).

## 5.5 Filtering tweets

The first intuitive response to filter ambiguous brand names is to refine the search query. For example when looking for tweets about cars of Ford and to exclude Ford Models, one can search for "Ford car" instead of only "Ford". Though, Spina et al. (2013), mention the user then has to put in extra effort and furthermore it will harm recall over the mentions of the brand in the Web, which can be misleading in an online reputation management scenario. Hence, most researchers use filtering keywords, categorized negative or positive (Spina, et al., 2013; Yerva, et al., 2010). The occurrence of positive keywords confirms a related tweet, e.g., "iPod" for the ambiguous brand name "Apple". A negative keyword is logically the opposite, and thus discarding a reference to the company, e.g., "crumble".

Filtering keywords to yield only relevant tweets for analyzing the topic or brand, is equivalent to Named Entity Disambiguation (NED), which involves the association of mentioning an entity and the concrete object to which it actually refers (Spina, et al., 2013). A confusing sentence, for example, can be:

> *"Went to Big Apple's new Apple retail store with my bestie today!"*

Filtering keywords should, thus, not be too general as this would cause multiple false positives (wrongly classified as positive/related). On the other hand, keywords should not narrow the filter down too much, as this could filter out relevant tweets (false negatives) (Yerva, et al., 2010) or may cause overfitting.

Yerva et al. (2010), who won the WePS-3 challenge (Section 5.1 - Common terms, WePS-3), first create entity profiles for each entity mentioned in the WePS-3 challenge and categorize these. Each profile is a set of keywords related to the company in a certain way. These profiles are set up as follows: (i) automatically gather frequently occurring keywords from the homepage, (ii) metadata from HTML pages[9], (iii) category related keywords, such as 'software, program, bug and virus' for a software company,

[9] http://www.w3schools.com/html/html_meta.asp

(iv) GoogleSet[10] [11], feature of Google Labs which return a set of words related to the input keyword, (v) user feedback, both positive and negative keywords are requested, furthermore they use wiki disambiguation pages[12], when available. Next, they pre-process all tweets and remove stop-words ('the', 'and', etc.), emoticons and Twitter syntax('RT', '@' and '#'). Afterwards they perform a SVM classifier (Cristianini & Shawe-Taylor, 2000) for generic learning, which can be seen as in-between supervised (predicting/classifying based on training data) and unsupervised (to determine how data is organized) learning. This result in, at best, an accuracy of 83%, where accuracy is calculated using:

$$accuracy = \frac{TP(True\ Positive)+TN(True\ Negative)}{TP+TN+FP(False\ Positive)+FN(False\ Negative)},$$

whereby TP stands for True Positive and indicates the correctly classified, relevant (positively related) tweets. In contrast, FN means False Negative and indicates all tweets classified as irrelevant, though they should have been classified as relevant. The formula of accuracy, stated above, thus, calculates the percentage of correctly classified tweets.

Other WePS-3 participants used slightly similar approaches, however, the SINAI system (García-Cumberas, et al., 2010) additionally include heuristic rules based on occurrence of entities in both tweets and external resources like Wikipedia, DBPedia and the company website. In contrast, the UvA system (Tsagkias & Balog, 2010) only uses features involving the collection of tweets, e.g., URLs, hashtags, etc.

Spina et al. (2013) focus on deriving and analyzing the effect of correct keywords. Using manually selected keywords from relevant web sources results in only 15% accuracy. While the best five keywords of their final solution already realized an accuracy of 28%, based on their test tweet collection. This significant difference in accuracy and the failure of manually selected keywords can be attributed to the "vocabulary gap", which was mentioned first by Tsagkias et al. (2011) and means the different vocabulary and topics used on the Web and in micro-blogging posts. A solution for this vocabulary gab can be to use co-occurrence between web-based features, manually selected keywords and historical Twitter occurrence, to select only the most occuring terms (Spina, et al., 2013).

Thus, automatic detection of keywords is plausible and can be a powerful tool for filtering tweets, though the keywords are not easy to find, as tweets use a different vocabulary and language.

## 5.6 Trendy topics

Mentioned in section 4 and 5.3, Twitter can be used in our attention economy (Davenport & Beck, 2002), to monitor the opinions and experiences of humans. Hence, it would be easy to be able to monitor topics of which a significant number of Twitter users is talking about with an exploding growth. In other words, gathering the topics of which more than X users are talking about, within the last Y minutes.

TwitterMonitor (Mathioudakis & Koudas, 2010) seem to be the first public, academic oriented algorithm to detect trends on Twitter. First it checks for 'bursty' keywords, with an unusually high arrival rate. These

[10] Deprecated since July 20, 2011: http://googleblog.blogspot.nl/2011/07/more-wood-behind-fewer-arrows.html
[11] though still semi-accessible: http://googlesystem.blogspot.nl/2012/11/google-sets-still-available.html
[12] Wikipedia page enumerating different 'Apple' - entities: http://en.wikipedia.org/wiki/Apple_(disambiguation)

keywords are clustered in groups, based on co-occurrence, to verify whether keywords belong to the same topic. Finally, additional information is gathered from all tweets for extra aspect discovery.

Quite similar, though with more mathematical background, Bhulai et al. (2012) show a speed- and smoothed acceleration- function to present the evaluation for each topic. Besides tokenizing, filtering and clustering, they base their conclusions whether a topic is hot on a mathematical measure. Furthermore they describe the possibility to visualize these clustered topics using dynamic squarified treemaps.

## 5.7 Summary & relevance

One may note many studies have been performed in the three discussed topics (ORM, filtering and trend detection). However, no research investigated the combination of all three, using Twitter. Besides, all filtering studies were performed based on a classified test dataset and could therefore apply classifying learning algorithms. This research does not have a test dataset as input and is therefore not able to apply learning algorithms, since it is unknown whether a tweet is indeed relevant or not. Therefore, this research performs a more adjusted approach, based on ideas derived from previous research. The relevance of this research is to provide a tool/roadmap for further research, by combining the three topics, and furthermore help companies being able to communicate more adequately with their customers and therefore become more customer friendly.

# 6 Methods

The goal of this research is to develop a tool that sends an alert when a trend occurs for a specific brand about a specific topic. The algorithm developed can be divided in three steps: (i) retrieval of Twitter data, (ii) tweet filtering, both (ambiguous) brand names and performance related topics, and (iii) trend detection, what makes a sequence of tweets a trend and when should the tool, thus, send an alert. This section will elaborate on each step.

## 6.1 Retrieval of Twitter data

Twitter offers different methods to retrieve all tweets. One can use the API ("Application Programming Interface"), which is more for single, simple requests. For example, retrieval of historical tweets, users' information, or to post a tweet. Next, it is possible to use the Streaming API. The latter, constructs an infinite, real-time stream to Twitter, to retrieve all tweets containing some tags/tokens, immediately when posted. Before, one could use the Basic Rest API v1[13] to stream, without the need of authentication, though this was shut down, June 11, 2013[14]. Currently, one should use OAuth[15] to connect to Twitter[16]. Fortunately, several third-party software is available as wrapper for the API's.

Using AnyEvent[17] for Perl[18], one of the third-party software, one can easily create a stream to retrieve all tweets containing particular words or word-combinations. Once a tweet is posted online, Twitter will immediately push the tweet into your queue. For this paper, concerning web performance related tweets for ING, the token used is '**ing**' for only **Dutch users**, users who set their account language to Dutch, or **Dutch tweets**, tweets that are classified as Dutch. Unfortunately, not all users define their language, or do it incorrectly. On the other hand, the use of tweet language is currently not really accurate, as is mentioned in section 8 (Conclusion).

After some pre-processing, all tweets are stored in a (SQL) database. The pre-processing allows the database to save UTF-8 encoded data, whereby the dates are in corresponding (SQL) format and all fields are escaped, to prevent SQL-injection. The structure of the database table can be found in section 11.1 (Appendix A :: Database tables).

## 6.2 Filtering tweets

Section 4 (Introduction) and 5 (Literature background) clearly showed, use of filtering keywords is currently the best approach to extract emerging topics from inherently noisy tweets. Therefore, the algorithm used is based on a combination of multiple filters. The filters are performed in series and will be discussed below. When building the filters, one should be aware of the evolutionary aspect of Twitter language and therefore try to not narrow down the filters too much, as this can cause over fitting.

[13] Documentation Twitter API v1 https://dev.twitter.com/docs/api/1

[14] Expiration date for API v1 https://dev.twitter.com/blog/api-v1-retirement-date-extended-to-june-11

[15] Documentation for use OAuth https://dev.twitter.com/docs/auth/oauth

[16] Blog about consequences change https://dev.twitter.com/blog/changes-coming-to-twitter-api

[17] Documentation AnyEvent http://search.cpan.org/~punytan/AnyEvent-Twitter-0.64/lib/AnyEvent/Twitter.pm

[18] Perl programming language http://www.perl.org

So far, the dataset is filled by real-time retrieval of tweets containing the trigram (three-letter combination) '*ing*'. Unfortunately, this trigram occurs often in everyday languages, e.g., doing and going in English or ing as prefix for an engineer in Dutch. However, 'ing' also occurs in product names, e.g., Boeing .

A PHP (Hypertext Preprocessor)[19] script is developed to filter/classify tweets either relevant or irrelevant. The implementation for this script is elaborated on in section 11.2 (Appendix B :: Implementation of alerting tool). The implemented filters are case-insensitive and elaborated on below. Each filter is explained by a simple example. Green markup indicates keywords or tweets passed through to the next filter and the red examples indicate tweets marked as irrelevant. This approach, thus, narrows down the number of 'potentially relevant' tweets, whereby a tweet is seen as a 'bag-of-words', since the structure of the sentence is unimportant. The examples are translated to English, though are developed in Dutch.

1. **Coarse filter**
   Like the name may imply, this filters most of the irrelevant tweets. Only tweets without letters, or non-allowed special characters in front or behind the trigram ('ing') remain.

   > *Ing, #ING, @ING, "opened a bank account at ING!!"*
   >
   > *Doing, rt'ing, Inge*

2. **Other brands filter**
   Often tweets are posted to compare companies (Yerva, et al., 2010). Tweets with multiple brand names of concurring banks are therefore, most likely, irrelevant for the web performance of ING. However, an extra filter can verify whether there are words that indicate simultaneously performance failure. Furthermore, a negative extra filter is introduced with words as 'like' and 'similar'.

   > *"Technical problems at* ***both*** *Rabobank and ING"*
   >
   > *"A failure at #ING turns the complete population topsy-turvy. Currently #Rabobank has a failure, though nobody is talking about that! Weird!!!"*

3. **Brand post filter**
   (Retweets of) tweets posted by the brand are not preferred to monitor, since these tweets often mention scheduled maintenances, provide support or confirm a failure. Therefore, these posts are no indicators of an unknown malfunctioning. Therefore, all tweets whereby the username contain grams equal to the company usernames are classified irrelevant. For ING Twitter-accounts all usernames start with 'INGnl_...' or 'ING_...'.

   > Usernames: *Inge, Bing*
   >
   > "**RT @INGnl_webcare** : tomorrow a maintenance is scheduled from 9:00 – 11:00 AM."

[19] PHP programming language http://www.php.net

4. **Negative performance keyword filter**

   Similar to ambiguous brand names, it is possible to classify a tweet ‘performance related’ based on positive and negative keywords. Since this algorithm eliminates irrelevant tweets and afterwards look whether the remaining tweet is relevant, it is decided to start with the negative performance keywords filter. The keywords are derived from manual selection by tweet analyses, attentively looking for different interpretations. A tweet that contains a negative keyword, is marked as irrelevant.

   > Keywords: vacancy, donation, sponsor, app, commercial, sells
   > “I really love the guy playing in the new ING commercial!!!! <3”

5. **Positive performance keyword filter**

   In contrast to the 4th filter, **Negative performance keyword filter**, tweets that contain positive keywords are directly classified as relevant.

   > Keywords: internet, payment, site, failure, iDeal, 403, slow
   > “Can’t pay using iDeal, still retrieve 403 failure, is the ING server down?”

6. **Positive performance double keywords filter**

   Word order is another challenge when applying filters based upon keywords. Especially when filtering on tweets, since all users express themselves differently. Therefore ‘double keywords’ are introduced. This mean both keywords should be in the tweet before it will be classified as relevant. This filter is more feasible in the Dutch language and hard to explain in English.

   > Keywords: poorly => { accessible }, not => { found, available}
   > “ING is poorly accessible all the time!” & “ING has been accessible poorly since April 3th. ”

7. **Remaining filter**

   Not all tweets will be filtered relevant or irrelevant, since they may contain ING in the right syntax, though not concern the website or any performance related keywords. Therefore, all tweets that remain after the filters discussed above, are classified irrelevant.

   > Just received my awesome new bank card from ING with my own picture on it!

After the filters are performed, the ignored and relevant tweets are moved into separate databases, which were optimized for this particular data. The structure of these databases is shown in section 11.1 (Appendix A :: Database tables).

[*Technical*] Because the tweets are moved, no redundant data is saved. Besides, two tables make it easier to analyze the relevant and ignored tweets, in correlation with the filter and tag that classified the tweet, since no *join* is needed and no irrelevant columns are stored. The extra ‘performance costs’ to move (insert and delete) the tweet, is negligible, since it is only performed for the tweets of one minute.

### 6.2.1 Not applied filters

Next to the filters applied, some filters were created during the development process, however, were not implemented in the final version. One of these filters classified all tweets as relevant that are addressed to the online support team of ING (tweets containing '@INGnl_webcare').

> "@INGnl_webcare is it possible that I can't check my balance?"
>
> "@INGnl_webcare is it possible to deposit coins at every ING office?"

Caused by this filter, all tweets, concerning other topics than web performance, were also analyzed in the trend detection calculation. Therefore, this filter was skipped and all tweets addressed to the online support team are only classified as relevant when they mention at least one of the (double) positive keywords. Furthermore, a double negative keywords filter was tried, though this resulted in several False Negatives.

## 6.3 Trend detection

Now all relevant tweets are classified, it is possible to detect events. An event occurs when multiple tweets are posted in short notice. Off course, this is dependent on the average number of tweets occurring. A topic with generally low occurrence can have an event, while the frequency is lower than the average tweets of a highly occurring topic. Therefore, the algorithms derived by Bhulai, et al. (2012) can be used to calculate speed (frequency) and acceleration, as acceleration can be used as measure for a change in the trend.

### 6.3.1 Speed calculation

To calculate speed, it is necessary to order tweets on arrival/creation time (provided by Twitter in seconds). The interarrival time of two consecutive tweets ($a_i = t_i - t_{i-1}$), is zero, when $t_i$, the creation time of tweet $i$, is equal to the creation time of the previous tweet ($t_{i-1}$). Therefore, $a_i'$ is introduced, to uniformly spread tweets over one second. To define $a_i'$, first $z_i = |\{n\, |t_i = t_k\}|$ is defined, indicating the number of tweets arriving at the same second. Then, $a_i$ is transformed to $a_i'$ by:

$$a_i' = \begin{cases} a_i - \left(1 - \dfrac{1}{z_i + 1} - \dfrac{1}{z_{i-1} + 1}\right), & a_i > 0 \\ \dfrac{1}{z_i + 1}, & a_i = 0 \end{cases}$$

Next, the uniformly spread interarrival time is exponentially smoothed to decrease the effect of outliers:

$$b_i = \alpha b_{i-1} + (1 - \alpha) a_i', \qquad 0 \leq \alpha \leq 1$$

This new time series $b_i$, whereby $b_1 = a_1'$, can still be too volatile. Therefore it is extra smoothed by the average over the past $k$ values of $b_i$. The speed (number of tweets per second), $v_i$, then, is defined by:

$$v_i = \frac{k}{\sum_{j=i-k+1}^{i} b_j}$$

Due lack of test datasets and time, it was impossible to thoroughly verify all parameters chosen by Bhulai et al. (2012). However, one can intuitively conclude, since the number of tweets occurring in this study is significantly smaller than the number in the research of Bhulai et al., $k$ should be set lower, since Bhulai et al. analyzed all tweets available and this study only analyzes tweets for ING, related to web-performance. Therefore, $\alpha$ is chosen equally to the study of Bhulai et al. ($\alpha$ = 0.8) and $k$ is chosen smaller ($k$ = 2, instead of 10). This is also based on section 7.5 (Data analyses :: Influence of k (history parameter)).

### 6.3.2 Acceleration

To calculate the acceleration for each minute and thereby determine the change in speed of tweet-occurrence, two variables are introduced: (i) $\overline{z_t} = max\{i \mid t_i < t + 1\}$, the index of the last tweet before the end of minute $t$, and (ii) $\underline{z_t} = \max\{\min\{i \mid t \leq t_i < t_{\overline{z_t}}\}, \overline{z_t} - 1\}$, the first tweet in minute $t$. When no tweet is posted in minute $t$, $\overline{z}_{t-1} = \underline{z_t} = \overline{z_t}$. Thus, both $z_t$ values are equal to the index of the last tweet in the previous minute. The acceleration, $w_t$, then can be defined by:

$$w_t = \frac{v_{\overline{z_t}} - v_{\underline{z_t}}}{t_{\overline{z_t}} - t_{\underline{z_t}}}$$

Acceleration, thus, implies the difference in speed between the last and first tweet of a minute, divided by the timespan between these tweets, what is relatively similar to the derivative of speed. When in one minute multiple tweets arrive at the exact same timestamp, this would result in a mathematical error, since the denominator would than become zero. Bhulai et al. (2012) also mention this problem and state the solution to use $a'_i$ instead of $a_i$. This seems odd, since it is possible more than two tweets arrive at the exact same time and furthermore, $a'$ is not used in the general formula. Hence, is not logical to derive the difference in time by summing over all $a'_i$. Therefore, the $a'_{\overline{z_t}}$ is used instead of $t_{\overline{z_t}} - t_{\underline{z_t}}$.

To reduce the number of calculations and, thereby, increase the performance, only the accelerations of minutes where more than one tweet arrived are calculated and saved in the database (see section 11.1 (Appendix A :: Database tables)). Note that the acceleration of a minute where only one tweet arrived is equal to zero.

# 7 Data analyses

For 20 days (from June 13 till 3th of July 2013), all tweets about ING are gathered for analysis, according to the section 6 (Methods). Therefore, all calculations are based on the code available at the moment of retrieval. However, during this period, the filters were frequently adjusted, based upon tweet analyses and other improvements. This causes incorrect data, since some tweets should have been classified differently after updating the filters. Therefore, an offline tool was developed, that is able to re-classify and -calculate all tweets gathered. Furthermore, it enables the feature to gather tweets from an earlier date, where for example, a failure is recorded, using the Twitter API instead of the stream. The following data analyses are based on all tweets – containing the trigram 'ing' – for the dates: {June 9, June 13 – July 3 2013}. June 9th 2013 is chosen because of a known failure[20].

## 7.1 Tweets per day

Primarily, one can be curious whether known failures can be observed in the Twitter data. Figure 1 shows the number of tweets and one can easily detect the days of failure, by looking at peaks for the relevant tweets (June 9 and 24). Thereby, one can state tweets can indeed be used to detect failures, although this is only using offline calculation. The extreme value for the number of ignored tweets at June 20 can be explained after some word occurrence analysis. This analysis shows the words 'confidence' and 'housing market' occur most frequently. This can be related to the study of ING, published the 20th of June 2013: "Confidence in housing market rises sharply"[21].

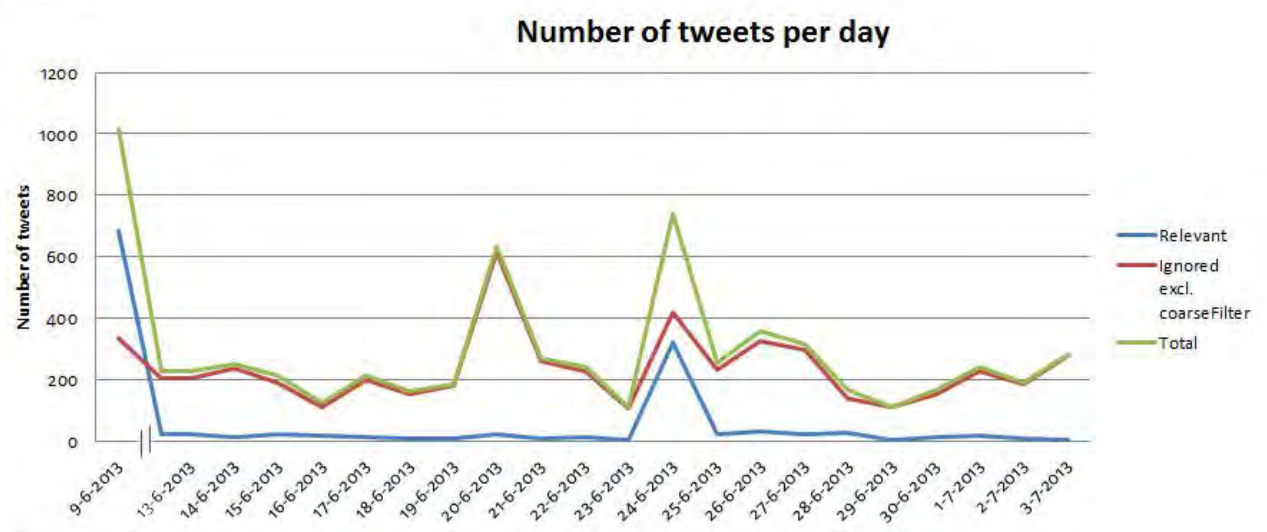

Figure 1 - Number of tweets per day. A clear difference is visible for days with a malfunction (June 9th and June 24th)

On average, each day, without failure, 6% of the tweets are classified as relevant, against 57% on a day with failure. This significant difference, does not imply a change in topic, but is caused by the extra tweets posted. In general 236 tweets are posted each day concerning the bank (tweets passing the coarse filter), in comparison with 879 tweets posted when a failure occurs.

[20] Overview of all known malfunctions of ING: http://allestoringen.nl/storing/ing/overzicht

[21] Announcement of results of study by ING at 20th June 8:00 AM. Article is in Dutch. http://www.ing.nl/nieuws/nieuws_en_persberichten/2013/06/vertrouwen_in_woningmarkt_neemt_fors_toe.aspx

## 7.2 Tweets per hour and the acceleration

To verify whether, in the short period, one can detect some daily trends and which threshold for acceleration should be used, one can plot the number of tweets occurring each 5-minute interval. Figure 2 shows the number of tweets for four dates, to show the difference between two days with, and two days without failure. The failure of June 9 seems to evolve after 8:30, the failure of June 24 after 7:50. Figure 3 shows the acceleration and number of tweets per minute for the 24th of June. Note, there is a large difference in interpreting the 1- and 5-minute(s) interval. Although, there may be a growth of tweets in a 5-minutes interval, this is not necessarily visible in the acceleration of one minute, since, e.g., 5 tweets can arrive in either one distinct minute, whereby the acceleration of each minute remains zero. Therefore, the number of tweets in the 5-minutes interval cannot directly be compared to the one-minute intervals.

Note, Figure 3 shows a little spike at 7:50, however, the event really seems to evolve at 8:51 AM. This continues at 9:09 and from 9:40 till 11:28 AM. A similar plot, as Figure 3, for the 9th of June resulted in an evolving-time of 8:53 (where the acceleration exceeds an acceleration of 0.001).

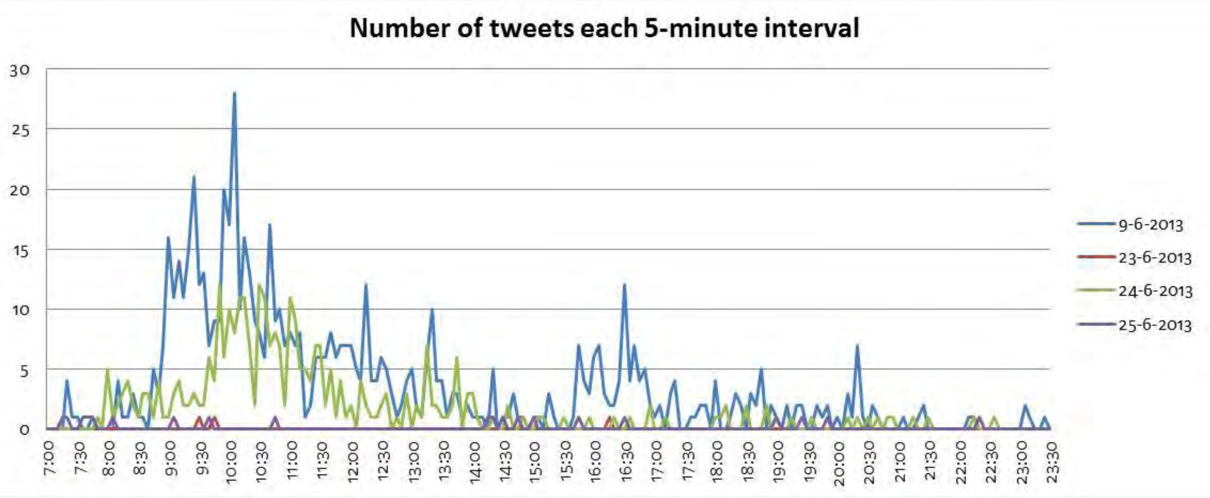

Figure 2 - Number of tweets arriving grouped by 5 minutes

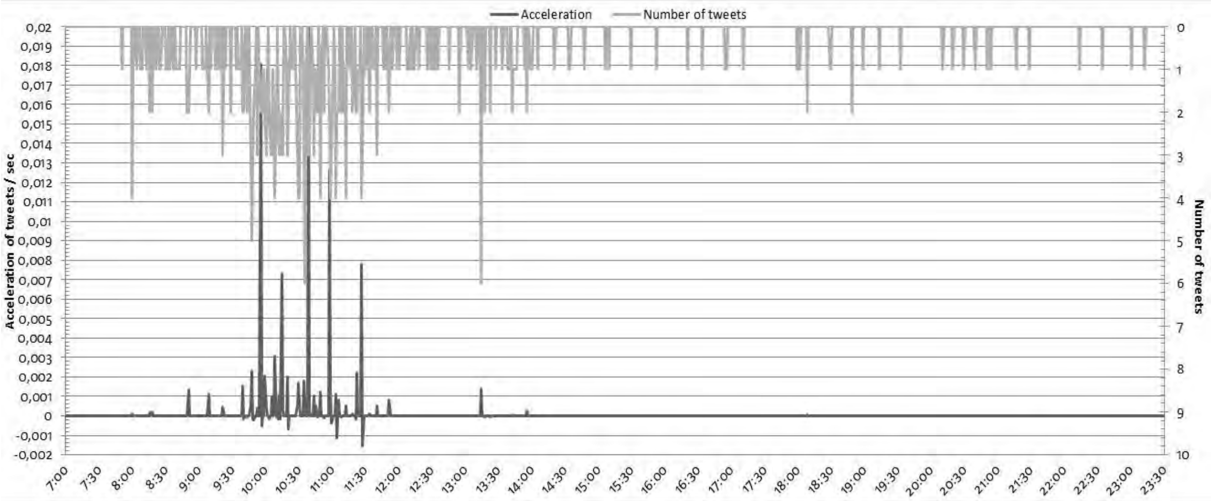

Figure 3 - Calculated acceleration for each minute of June 24, 2013, and the corresponding number of tweets

The difficulty is to estimate a threshold applicable for all events. Each trend evolves in its own way, as one can see in Table 1. Unfortunately, only for June 24 the real start time of failure is known and therefore, it is likely the threshold of 0.001 will overfit the data of other (unanalyzed) failure dates.

Table 1 – Summary of the acceleration for the two days with the most- and one day with the fewest relevant tweets (# Tw). For the average notation: x (y), x is the average for the complete day; y only for the minutes with at least 1 tweet arriving

| | Minimum | | Average | | Maximum | | Total |
|---|---|---|---|---|---|---|---|
| | Accel. | # Tw | Accel. | # Tw | Accel. | # Tw | # Tw |
| 9-6-2013 | -0,0099 | 1 | 0 (0,0032) | 0,0055 (1,8665) | 0,4654 | 10 | 685 |
| 24-6-2013 | -0,0015 | 1 | 0 (0,0005) | 0 (1,5784) | 0,02 | 6 | 322 |
| 29-6-2013 | 0 | 1 | 0 (0) | 0 (1) | 0 | 1 | 3 |

## 7.3 Speed of tweet

One may note, the acceleration is most reliable when many tweets arrive, as acceleration will only be non-zero when in at least two sequent minutes more than one tweet arrives. Bhulai et al. (2012) performed a research with fewer filters and therefore the acceleration was more accurate, since the number of tweets was significantly bigger. Hence, one should be aware of the misinterpreting of acceleration with fewer tweets. Therefore, also the speed of tweets is plotted against time, shown in Figure 4.

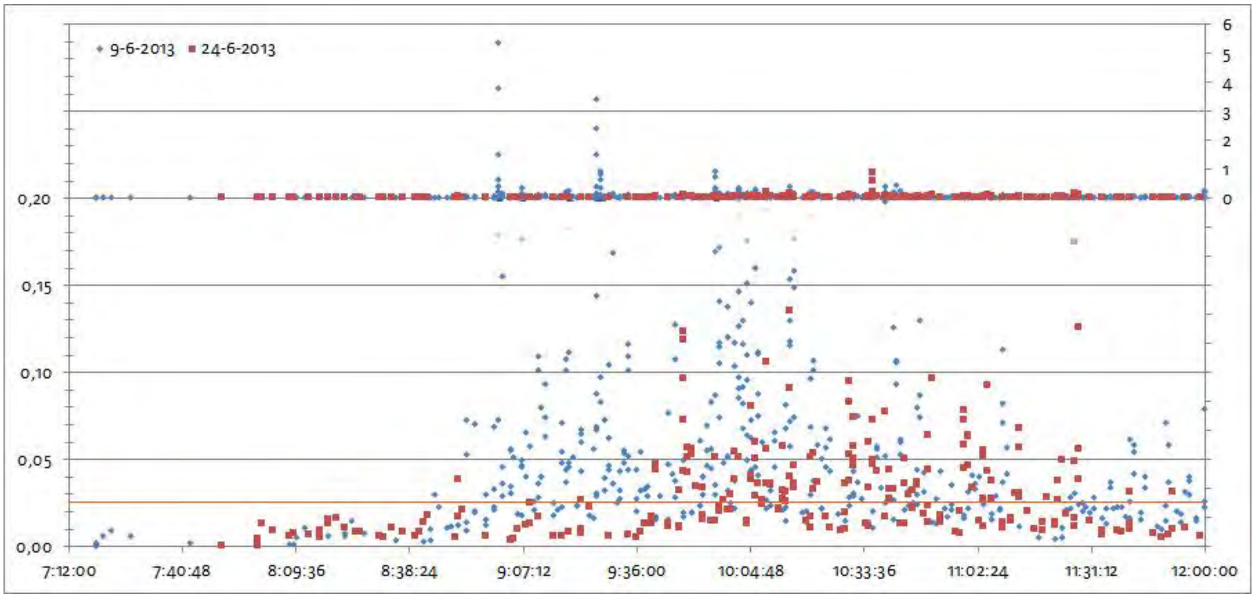

Figure 4 - Speed of tweets plotted against time for June 9 and June 24, 2013. The top pane shows the complete results, the bottom pane is scaled for more detailed analyses and is cropped at 0.2. The orange line indicates a threshold of 0.025

Speed is a measure indicating the number of tweets arriving in one second. The speed of 5.36 at 9:01, thus, implies 320 tweets per minute. The average speed calculated, based on the dataset with 2/21 days of failure, is 0.051 (implying 180 tweets per hour).

One may note, speed is a quite fluctuating measure, therefore, Bhulai et al. (2012) chose the acceleration. However, to use the speed, one should define a threshold, at which the tool should alert when the speed exceeds it (the threshold). This task is quite difficult, since only for one day the real time of failure is known. The time of failure for the 24th of June is around 9:10 AM (see section 7.4). The right threshold, then, would become 0.025 (at 9:09), although this is also exceeded at 8:51 AM. Assume, the threshold is set at 0.025, then the tool would have alerted at 8:45 AM for the data of June 9, and again at 8:53 AM. Besides, a threshold of 0.025 seems odd when the maximum speed reached is over 5.

## 7.4 How accurate is the analysis?

To verify the results of the different measures, the same timestamps are analyzed in the webpage data for June 24, 2013. Figure 5 shows the increase in errors, loading time and decrease of good page views at 9:10 AM. Note, analyses of the Twitter data showed a significant increase in the acceleration at 9:06 AM and a exceeding of the speed-threshold at 9:09 AM. Hence, we can say the Twitter and web traffic data do at least not contradict each other. However, one should be aware these thresholds and measures are currently based upon two failures, of which only one date is available in the webpage data.

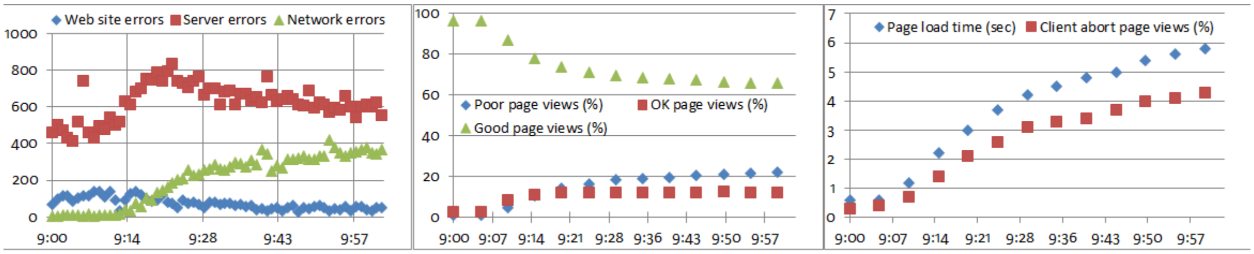

Figure 5 - Webpage data for June 24, 2013, (i) failure categories – type of failure, (ii) page satisfaction rates – whether the loading time is in the threshold interval, (iii) page loading time and client aborts. All show a change in trend at 9:10 AM

Furthermore, the website 'Allestoringen.nl'[22], that provides users with information whether an online brand is in failure, based on Twitter and Facebook data, and information provided by the company, mentioned failures at similarly timestamps. For June 24, they emerged the failure at 10:10 AM, one hour later. The malfunction of June 9 was mentioned at 8:26 AM, the speed-threshold exceeded at 8:45 AM.

Table 2 - Summary of timestamps of change in trends in tweet arrivals. Using the following parameters, respectively: more than 4 tweets (in one minute), manually selected, acceleration $\geq 0.001$, speed $\geq 0.025$

| | # tweets in 1 min | # tweets in 5 min | Acceleration | Speed | Allestoringen.nl | Web traffic |
|---|---|---|---|---|---|---|
| 9-6-2013 | 8:53 | 8:30 | 8:53 | 8:45 | 8:26 | - |
| 24-6-2013 | 8:00 | 7:50 | 9:06 | 9:09 | 10:10 | 9:10 |

Table 2 summarizes the different timestamps. One may notice every measure results in a slightly different time of event occurrence. Therefore, it is hard to draw a conclusion which measure should be preferred, on the current available datasets. One should also be aware that it is possible the Twitter events occur at a different time as the web traffic data or the announcement of ING (allestoringen.nl), since a failure is not necessarily noticed directly by users or by ING.

[22] www.allestoringen.nl, also available in English via: www.downdetector.com

## 7.5 Influence of k (history parameter)

All analyses above, are executed with $k$ = 2. $k$, is the number of historical tweets taken into account when calculating the speed. For a large number of tweets, a larger $k$ is preferred, than for a smaller number of tweets. Bhulai et al. (2012) set $k$ = 10, after some analyses. However, for this study, too few tweets are classified relevant to use such large $k$. Figure 6 shows that the larger $k$, the less skewed the speed data, as the average, the standard deviation and number of extremes decreases. The use of $k$ = 1, is relatively similar to ignore all past tweets and only take the interarrival time into account, when calculating the speed and acceleration. $k$ is therefore preferred to be set between 2 and 4, since $k$ = 5, already deals with too centralized data, whereby it becomes harder to find the optimal threshold. Due to lack of valid tests and training data, $k$ is only plotted, but not verified. Besides, more dates of failure-tweets should be analyzed, in comparison to webpage data, to define a general threshold, to prevent overfitting.

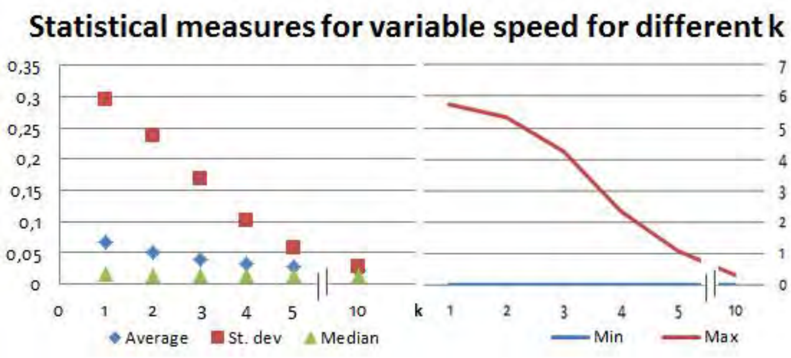

Figure 6 - Statistical measures for different k (number of historical tweets taken into account when calculating speed)

## 7.6 Influence of filters and keywords

To verify what the coverage of a filter is, one can look at the percentage of irrelevant classifications. After classification, 34% of all ignored tweets (total number of ignored tweets: 7,869) is classified by the coarseFilter, as is shown by Figure 7. This indicates 34% of tweets forwarded by Twitter is not about the bank ING, but contains the trigram ('ing') within an everyday word (e.g., 'flying'). For the keywords, 'failure' is occurring most frequently in tweets (33% of all, 1,208, relevant tweets). Note, one tweet can be classified as relevant based on multiple tags, those tweets are counted for each tag separately.

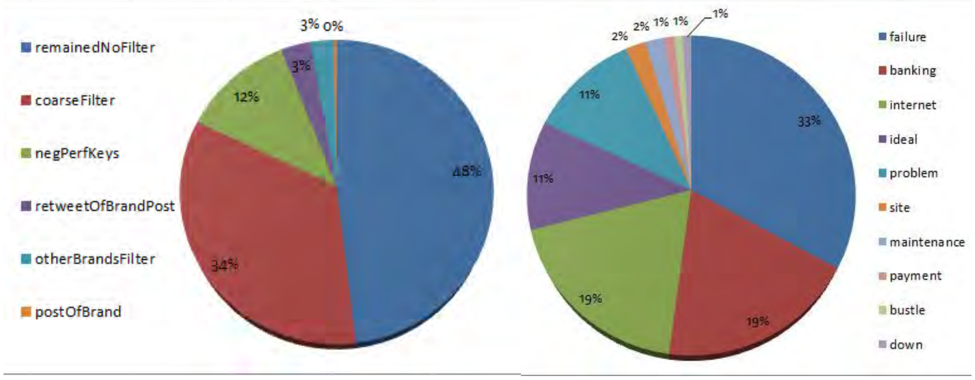

Figure 7 - Coverage of filters (irrelevant tweets) and coverage of keywords (relevant tweets)

# 8 Conclusion

To answer the research question: '*Can we combine data from social media, e.g., Twitter, with data of web traffic, e.g., number of visitors on a website, in such a way it will alert companies when their website malfunctions?*', this study tried to develop an online alerting tool for Twitter data, applied for the case of ING, a Dutch bank, exclusively for website-performance related tweets.

Data analyses revealed the number of tweets is indeed related to the occurrence of a failure and therefore a possible measure to analyze the web-performance. When a failure occurs, more tweets are posted, mentioning ING. Interesting is that there is no movement in tweet topics, but an increase in the number of tweets. This means users who normally do not post about ING, do post when a failure occurs, as a kind of '*disaster tweeters*'.

Furthermore, manual analyses showed social media data can indeed be used in combination with web traffic data to alert companies when their website malfunctions, since the timestamps of a change in trends for both Twitter and web traffic data were relatively close to each other. Currently, no fully automatic alerting tool is developed, though this paper provides the roadmap to develop a certain tool.

Before one is able to automatically monitor tweets for a particular brand or product, specifically for a certain topic, several difficulties should be challenged. This paper, describes the following three-step-process that is performed. Firstly, all tweets are gathered real-time from Twitter, that contain the trigram ('ing') and are posted by a Dutch user, or classified by Twitter as a Dutch tweet. Currently, the language classification of Twitter is not accurate enough to use as a hard restriction (ignore all tweets that are non-Dutch), since Twitter can often not (correctly) determine the language of the tweet. However, this feature is in development and could be used more strictly in the future.

Secondly, multiple filters are applied in series, to select the relevant tweets. Most important is the Coarse filter (classified 34% of all irrelevant tweets) that ignores tweets wherein the trigram ('ing') is just a subset of an everyday word, e.g., 'ing' in 'doing' or 'Boeing'. Filters to handle ambiguous brand names, e.g., *Apple* or *Jaguar*, were not applied, since ING is not an ambiguous brand name, though those are mentioned in both section 5 (Literature background) and section 9 (Future work). Furthermore, one should be aware of the fact that the keywords should be maintained consequently, as the language used on Twitter is evolving every day. Therefore, one could use a kind of word cloud (like on the front-page) to monitor most occurring words in both relevant and irrelevant classified tweets.

Thirdly, multiple measures are calculated for the relevant tweets to check whether a change in trend occur, such as the speed and acceleration. However, the acceleration is more applicable when there is a constant stream of tweets, instead of one or two tweets per hour. The speed, on the other hand, is quite fluctuating and makes it therefore hard to define a general threshold. Because of the different pros and cons of measures, it is hard to draw a conclusion which measure should be preferred and how the parameters should be defined, also, largely, caused by the lack of test sets. Additionally, it is possible Twitter events occur at a different time as the web traffic data or the awareness of ING, since a failure is not necessarily noticed, directly, by users or ING.

Thus, this paper provides insights in the possibility to combine Twitter and web traffic data to emerge malfunctions of websites. Unfortunately, no test or training set was available to verify the proceedings, though this could be done in further research, as this paper also provides a roadmap to develop an automatic alerting tool to detect events on Twitter. We can thus confirm the research question stated.

# 9 Future work

This study is performed within 1,5 months, therefore several opportunities were not applied or verified. As mentioned in section 4 (Introduction), the main scientific relevance is to introduce the combination of multiple techniques and the possibilities for companies to use Twitter/social media data to analyze user opinions. Thereby, the goal is to provide a grip for further research. This section will discuss some opportunities for future work and what topics should be elaborated on more.

The original research question studied in this paper was: '*Can we combine data from social media, e.g., Twitter, with data of web traffic, e.g., number of visitors on a website, in such a way it will alert companies when their website malfunctions*'. This paper shows the possibility and provides a roadmap to improve the alerting tool, already developed. For the future, it would be useful to investigate in the other part of the original research question: '*Can we check whether there is a malfunctioning based on a change in trend for the number of visitors, or the loading time of a certain webpage*'. When both questions are confirmed and implemented, this could result in an alerting tool that could support companies to detect malfunctions immediately and thereby, improve the customer satisfaction.

One of the main difficulties in this paper was the lack of valid test and training data, thus at what time the failures occurred and whether a tweet is relevant or not. When certain data is provided, it is possible to validate the filters and derive the accuracy of the algorithm. Besides, every improvement, then, can be compared to the previous version and one would be able to determine the effect of the improvement, such as a new chosen keyword or filter. Furthermore, valid test data would give possibilities to use machine learning algorithms to derive extra keywords, or detect other (unknown) trends in the tweets.

When more dates with failures are gathered, all measures can be verified more and improved, for example, the $k$. Currently, $k$ was not investigated thoroughly and can probably be defined better as it is currently based on two days of failure and therefore may overfit the current data. Another possibility is to introduce a dynamic $k$ that is dependent on the number of tweets in the past hour. In time of failure, the number of tweets rises significantly and therefore the speed can be calculated more precisely by adjusting the $k$.

Furthermore, it is possible to introduce new and better filters:

- *Remove retweets*. A retweet is currently captured and counted as a tweet. One could investigate whether this is valid. The effect of a retweet is equal to, possibly bigger than, a single post (Kwak, et al., 2010). Therefore, it could be useful to verify the consequences of omitting retweets. Especially since the 'RT' tag is found in 36% and 45% of all tweets for relevant and ignored tweets, respectively. Imagine the influence of a wrongly classified tweet.
- *Sentiment analysis*. By performing sentiment analysis, the algorithm can be aware whether the tweet is positive or negative. Most of the time, the sentiment of a tweet about a failure is negative. Therefore, this could help to improve the relevance filters.

- *User importance.* By estimating the influence of a user, one can calculate the effect of one tweet. A post of an (almost) unknown user can be taken less into account, than a user which is widely known. To estimate the importance, one can use the `user_followers_count`, or Kred[23] statistics.
- *Past tense detection*. Newspapers often tweet about failures that have been solved. Those tweets can be filtered based upon their past tense, though one should be aware of special situations:

> *"When will the malfunction finally be resolved?"*
>
> *"The problem for customers of ING not being able to login, is resolved."*

- *Future tense detection.* Similar to *past tense*, some tweets are written in advance to, e.g., announce a maintanance. Those are not necessarily initiated by the company and therefore retweeted, but sometimes also initiated by a company that uses, for example, iDeal and therefore warns the customers of ING.

> *"Just heard the weather forecast and a new failure of ING and Rabobank is coming."*
>
> *"#MAINTENANCE #ING commits to maintain their #iDEAL system on 7-7 from 01:00 till 03:00. Transactions via ING are then not possible."*

Furthermore, it would be able to improve the keywords, and how they are derived:

- *Automatic derive keywords* from, e.g., webpages, Wikipedia, and so on, as is suggested by several researchers (Yerva, et al., 2010; García-Cumberas, et al., 2010). This may create a more general algorithm, which can be applied more easily for other brands.
- *Clustering words* based on co-occurrence and/or cosine-similarity (Mathioudakis & Koudas, 2010; Spina, 2011; Bhulai, et al., 2012) to gain insights in the clusters of words used and possible new feasible keywords.

Finally, it would be useful to develop a dashboard that shows whether there is a change in trend in the Twitter or webpage data and on which criteria this is based. This could, for example, be realized by showing a word cloud of all tweets occurring in time of the failure (to also prevent claims of copyright).

[23] Kred estimates the influence of users by ranking them and comparing their activity and statistics of the complete WWW, including Twitter, Facebook, and so on. http://www.kred.com/

# 10 Bibliography

Amigó, E. et al., 2010. *WePS-3 evaluation campaign: Overview of the online reputation management task.* Padua, Italy, WePS-3, CLEF.

Barabási, A.-L. & Bonabeau, E., 2003. Scale-free networks. *Scientific American,* Volume 288, pp. 50-59.

Becker, H., Naaman, M. & Gravano, L., 2010. *Learning similarity metrics for event identification in social media.* New York, ACM.

Bhulai, S. et al., 2012. Trend visualization: What's hot and what's not?. *Proceedings of Data Analytics*, pp. 43-48.

Cristianini, N. & Shawe-Taylor, J., 2000. *An introduction to support vector machines and other kernal-based learning machines..* Cambridge: Cambridge University Press.

Davenport, T. H. & Beck, J. C., 2002. *The attention economy: Understanding the new currency of business.* Cambridge: Harvard Business Press.

Dellarocas, C., 2003. The digitization of word-of-mouth: promise and challenges of online feedback mechanisms. *Management Science,* October, 49(10), pp. 1407-1424.

Dorlin, D., 1985. Although U.S. cars are improved, imports still win quality survey.. *The wall street journal*, 16 December, p. 27.

García-Cumberas, M. A., García-Vega, M., Martínez-Santiago, F. & Peréa-Ortega, J. M., 2010. *SINAI at WePS-3: Online reputation management.* Pedua, Italy, WePS-3, CLEF.

Gouws, S., Metzler, D., Cai, C. & Hovy, E., 2011. Contextual bearing on linguistic variation in social media.. *Proceedings of the workshops of languages in social media,* pp. 20-29.

Greenberry, 2013. *Social Insights Top-100 Merken onderzoek,* Utrecht: Tijdschrift vor Marketing.

Ha, H.-Y., 2002. The effects of consumer risk perception on pre-purchase information in online auctions: brand, word-of-mouth and customized information. *J. Computer-Mediated communication,* 8(1).

Hearst, M. A., 1991. *Noun homograph disambiguation using local context in large text corpora.* Oxford, Corpora.

[Online] Hoffman, T., 2008. *Online reputation management is hot -- but is it ethical.*
Available at: [Accessed June 20, 2013]
http://www.computerworld.com/s/article/9060960/Online_reputation_management_is_hot_but_is_it_ethical_.

Holt, R., 2013. Twitter in numbers. *The Telegraph*, 21 March.

ING & Social Embassy, 2013. *PR & Social media in 2013 onlosmakelijk verbonden (English: ... inseparable) ,* Amsterdam: ING Nederland.

Jansen, B. J., Zhang, M., Sobel, K. & Chowdury, A., 2009. *Micro-blogging as online word of mouth branding.* New York, ACM.

Kwak, H., Lee, C., Park, H. & Moon, S., 2010. *What is Twitter, a social network or a news media?.* New York, WWW'10.

Lockwood, M., 2013. *2013 state of inbound marketing,* sl: Hubspot.

Mathioudakis, M. & Koudas, N., 2010. *TwitterMonitor: Trend detection over the twitter stream.* New York, ACM.

Mihalcea, R. & Moldovan, D. I., 1991. *A method for word sense disambiguation of unrestricted text.* Stroudsburg, Association for computational linguistics.

Mishne, G. & de Rijke, M., 2006. *A study of blog search.* Heidelberg, Springer-Verlag Berlin.

Qureshi, M. A., O'Riordan, C. & Pasi, G., 2012. *Concept term expansion for monitoring reputation of companies of Twitter.* Online, CLEF.

Sarmento, L., Kehlenbeck, A., Oliveira, E. & Ungar, L., 2009. *An approach to web-scale named-entity disambiguation.* Berlin, ACM.

Spina, D., 2011. *Filter keywords and majority class strategies for company name disambiguation in Twitter,* Madrid: Universidad Complutense de Madrid.

Spina, D., Gonzalo, J. & Amigó, E., 2013. Discovering filter keywords for company name disambiguation in Twitter. *Expert Systems with Applications,* Volume 40, pp. 4986-5003.

[Online] Sullivan, D., 2009. *Twitter's real time spam problem.*
Available at: [Accessed June 22, 2013]
http://searchengineland.com/twitters-real-time-spam-problem-20614.

Thelwall, M., 2006. Bloggers during the London attacks: Top information sources and topics. *Proceedings of the 15th International WWW Conference: 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics.*

Tsagkias, M. & Balog, K., 2010. *The University of Amsterdam at WePS-3..* Pedua, Italy, WePS-3, CLEF.

Tsagkias, M., de Rijke, M. & Weerkamp, W., 2011. *Linking online news and social media.* Honk Kong, ACM.

[Online] Twitter, 2010. *Twitter blog: To trend or not to trend.*
Available at: [Accessed June 22, 2013]
https://blog.twitter.com/2010/trend-or-not-trend.

Westbrook, R. A., 1987. Product/consumption-based affective responses and postpurchase processes. *Journal of marketing research,* Volume 24, pp. 258-270.

Yang, J. & Counts, S., 2010. *Predicting the speed, scale and range of information diffusion in Twitter.* Washington, ICWSM'10.

Yerva, S. R., Miklós, Z. & Aberer, K., 2010. *It was easy, when apples and blackberries were only fruits.* Padua, Italy, WePS-3, CLEF.

# 11 Appendices

## 11.1 Appendix A :: Database tables

This appendix provides the structure of database tables used. For each table the attribute -name, -collation and -description is given. Furthermore, the default value and a boolean, whether the column is allowed to be left empty are provided. A bold attribute name indicate a primary key, underlined indicate indices. An index can be used to search, filter and group large datasets with better performance. The primary keys are defined in a not chronologically order, to decrease the cardinality of the database, whereby the most frequently requested column is set primarily. Furthermore, the collations are defined for this particular data, saving every byte possible, since the content of these datasets will grow vast.

Table 3 - Structure of database table 'tweets' containing all (new) unparsed tweets
Primary key: (tweet_time, user_name, tweet_text, id)

| ATTRIBUTE | COLLATION | EMPTY | DEFAULT | EXTRA / DESCRIPTION |
|---|---|---|---|---|
| **id** | int(11) | No | - | AUTO_INCREMENT,<br>provided by algorithm |
| tweet_id | bigint(18) | No | - | UNSIGNED<br>ID provided by Twitter |
| **tweet_text** | varchar(180) | No | NULL | Tweet text |
| **tweet_time** | datetime | No | - | Tweet created at |
| tweet_lang | char(2) | Yes | NULL | Tweet language |
| **user_name** | varchar(15) | No | - | User name |
| user_id | int(10) | No | - | UNSIGNED<br>User id in Twitter |
| user_lang | char(2) | No | - | User language |
| user_followers_count | mediumint | No | - | UNSIGNED<br>Number of followers,<br>users following user/author |
| inserted | timestamp | No | - | Time of insertion in database |

Table 4 - Structure of database table ‘tweets_ignored’ containing all irrelevant tweets
Primary key: (created_at, tweet_text, user_name, id); index: (filterThrown, tag)

| ATTRIBUTE | COLLATION | EMPTY | DEFAULT | EXTRA / DESCRIPTION |
|---|---|---|---|---|
| **id** | int(11) | No | - | AUTO_INCREMENT,<br>provided by algorithm |
| **created_at** | datetime | No | - | Tweet created at |
| **tweet_text** | varchar(180) | No | - | Tweet text |
| filterThrown | varchar(32) | No | - | Filter which classified the tweet as irrelevant |
| tag | varchar(150) | Yes | NULL | The tag on which the filter classified the filter |
| **user_name** | varchar(15) | No | - | User name |
| user_lang | char(2) | No | - | User language |
| user_followers_count | mediumint | No | - | UNSIGNED<br>Number of followers,<br>users following user/author |

Table 5 - Structure of database table ‘tweets_relevent’ containing all relevant tweets
Primary key: (created_at, tweet_text, user_name, id)

| ATTRIBUTE | COLLATION | EMPTY | DEFAULT | EXTRA / DESCRIPTION |
|---|---|---|---|---|
| **id** | int(11) | No | - | AUTO_INCREMENT,<br>provided by algorithm |
| **created_at** | datetime | No | - | Tweet created at |
| **tweet_text** | varchar(180) | No | - | Tweet text |
| tag | varchar(150) | NO | - | The tag on which the filter classified the filter |
| **user_name** | varchar(15) | No | - | User name |
| user_lang | var_char(2) | No | - | User language |
| user_followers_count | int(11) | No | - | Number of followers,<br>users following user/author |
| b | Float | No | - | Exponentially smoothed inter-arrival time between tweets |
| speed | Double | No | - | Speed of tweet |

Table 6 – Structure of database table ‘acceleration_history’ containing the history of tweet -accelerations

| ATTRIBUTE | COLLATION | EMPTY | DEFAULT | EXTRA / DESCRIPTION |
|---|---|---|---|---|
| **timeStamp** | datetime | No | - | Timestamp of minute the acceleration is calculated for |
| w | Float | No | - | Acceleration |
| nTweets | mediumint | No | - | UNSIGNED - Number of tweets in corresponding minute |

## 11.2 Appendix B :: Implementation of alerting tool

The PHP script developed is requested each minute, using CronJobs[24]. Each whole minute, for example at 03:14:00 (H:m:s), the script will check the database for new tweets. When new (unclassified) tweets have arrived, the script is executed, otherwise, the script will sleep again untill the next whole minute.

One of the difficulties is to handle tweets that arrive at the exact beginning of a minute, thus, e.g., at 03:14:00. Since, the tweet is in the database when the script is requested, the tweet is intended to be classified. Unfortunately, this would force the algorithm to save an entry for 3:14 in the `acceleration_history` table, what would cause a duplicate entry when another tweet arrives at 03:14:15, and thus is in the next minute.

To prevent this, the algorithm will ignore a tweet that is the last in the retrieved 'new tweets', thus the latest arrived tweet, with an arrival time whereby the number of seconds is zero. When the latest tweet is ignored, the tweet before is also checked whether it has the same timestamp and then, also should be ignored. Although this will put the tweet in hold, the assumption is made, this do not harm the general functionality of the algorithm.

Assume, a tweet is posted at 03:14:00, let's call this tweet *P.* Tweet P is then put on hold in the script request of 3:14. When tweet P is followed by a tweet posted at 03:14:15, the script request of 3:15 will classify both tweets, since the latest tweet is not posted at a whole minute. Another situation can be that a similar tweet P is posted at 03:14:00, however the next tweet arrives at 09:26:53, let's call this one tweet *i*. Then, tweet P has to wait 373 minutes (6 hours and 13 minutes) to be classified. The assumption then made, is that this will not harm the functionality of the algorithm since a difference between tweet P and i, of more than 6 hours, would imply there is no (unknown) event occurring.

[24] CronJobs explained by Wikipedia as a time-based shell-command: http://en.wikipedia.org/wiki/Cron