Erasmus
School of
Economics

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Master Thesis Data Science and Marketing Analytics

Twitter sentiment towards a brand in a crisis

Name student: Jelmer Versloot
Student ID number: 483846
Supervisor: Alfons, A.
Second assessor: Crombrugge, M. van

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second accessor, Erasmus School of Economics or Erasmus University Rotterdam

Erasmus University Rotterdam
Make it happen.

Abstract

Information, news and opinions travel at a rapid speed in the modern digital age. It is one of the reasons that makes it difficult for brands to stop the spread of negative publicity in times of a crisis. A lot of these messages are spread on Twitter. The threat of negative publicity could harm the brand reputation, which is one of the most valuable intangible assets a brand can invest in. This study investigates how a corporate crisis affects Twitter sentiment towards the corporate that is in the crisis. Two corporate crises cases are used to support the investigation. 4450 tweets that were posted in four months prior and four months after a corporate crisis has occurred were scraped for each crisis. Structural topic modelling was performed on the data to find underlying topics and the prevalence of these topics. The supervised methods for sentiment prediction comprise a naïve Bayes classifier, random forest and support-vector machine. The best performing supervised learner was used in the semi-supervised learner self-training. Local Interpretable Model-Agnostic Explanations was conducted to explain the self-trainer's prediction. This study shows that tweets related to the crises became relatively more prevalent in the period after the crises occurred. Naïve Bayes predicts the tweets' sentiment best out of the three supervised learners with an accuracy of 66.5%. Self-training obtains the highest accuracy of 75% in predicting the tweets' sentiment. A corporate crisis increases the prevalence of tweets related to the crisis and crisis related terms support a negative sentiment prediction and contradict a positive sentiment prediction.

Table of contents

1 Introduction

Corporate crises are likely to have some of the most profound and dramatic impacts (Seeger et al., 1998). Crises, triggered by faulty decisions or complex organizational structures, are often unforeseen events making it difficult for organizations to response quickly (Dean, 2004). Before the internet was used worldwide, opinions or behaviour towards a brand were based on what people heard from their own environment, like family or friends (Burnkrant & Cousineau, 1975). The consequences of word-of-mouth between people were not as harmful as they are now. Information travels at an immense speed in the modern world as we know it now, making a crisis a threat on a much more serious level than it was before the internet (Hudson et al., 2016). Twitter is such a platform where information is processed rapidly. Twitter allows customers to quickly share or read opinions on the internet, making it even more difficult for organizations to reduce negative publicity in a crisis (Cleeren et al., 2013). Negative publicity is a threat to the brands' reputation, affecting purchase decisions and overall brand evaluation. Brand reputation is a very important asset that brands can maintain and invest in. Shapiro (1982) analysed the behaviour of a monopolist when consumers are unable to observe relevant attributes of the monopolist's product. Shapiro (1982) suggested that a favourable reputation has positive effects on the monopolist's sales and market share (Nguyen & Leblanc, 2001; Shapiro, 1982). Andreassen and Lindestad (1998) examined the effect of brand reputation on customer satisfaction and customer loyalty in a service industry. Their study suggested that brand reputation was the primary driver in customer loyalty and customer retention (Andreassen & Lindestad, 1998). Smith, Smith and Wang (2010) conducted an empirical study that tested whether a positive brand reputation positively affects market value. This study proved that there are higher market-value premiums for firms with a positive brand reputation. Brand reputation has thus proven to be one of the most valuable intangible assets for a company to capitalize in (Andreassen & Lindestad, 1998; Nguyen & Leblanc, 2001; Shapiro, 1982; Smith et al., 2010). The impact of a corporate crisis on social media brand reputation is the main subject of this study. In order to understand the consequences of a corporate crisis on Twitter sentiment, the following research question is studied:

Which features are most important for the sentiment of a tweet towards a company in a crisis and how do these features affect tweet sentiment?

The importance of terms and how these terms affect sentiment implies the impact a word has on the sentiment of the tweet. Knowing which words are most vital for the tweets' sentiment is essential in understanding the consequences of the crises. This study starts by examining the topics across the tweets based on the topic modelling approach of Roberts et al. (2014). The topics of tweets are used to study the prevalence of tweets related to the crisis firstly. Secondly, it offers an opportunity to cluster tweets based on topic probability. Several machine learning algorithms are applied on the data to get an overall view on the tweets' sentiment. The machine learning algorithms will be trained both supervised and semi-supervised. The best performing model is used to predict the sentiment of a subset of unlabelled tweets. These predictions are made interpretable by visualising which words affect sentiment in what direction.

1.1 Relevance

Brand reputation can be managed well with the right resources in the right circumstances (Wartick, 2002). However, a corporate crisis can change these circumstances dramatically. Brand reputation is important for brands to invest in. Therefore, it is essential to know what the effects of a crisis are on the reputation. This paper attempts to show the negative effect of the crisis on the sentiment of tweets. By using two company crises cases, brands can understand how people react on the crisis on social media. Knowing how people respond to the crisis can help companies in constructing the right response in a crisis.

1.2 Contribution

Brand reputation and its importance is already studied extensively (i.e. Andreassen & Lindestad, 1998; Nguyen & Leblanc, 2001; Shapiro, 1982). The interest in brand reputation shifted towards online brand reputation with the growth of social media. Papers (i.e. Dijkmans et al., 2015; Hennig-Thurau et al., 2004; Kim et al., 2014) started examining the relationship between social media brand reputation and customer behaviour. Investing in brand reputation is beneficial for brand engagement of customers and non-customers (Dijkmans et al., 2015; Hennig-Thurau et al., 2004). Less favourable for brand reputation is a company crisis. Ahluwalia, Burnkrant and Unnava (2000) researched how consumers process negative information about a brand they liked. This research in social media brand reputation requires Natural Language Processing (NLP) techniques. NLP techniques are also already widely used in the existing literature to investigate social media sentiment (Ahmad et al., 2017; Kouloumpis, E., Wilson, T., & Moore, 2011). This study contributes to the literature by using NLP techniques on real Twitter data to study the effects of two company crises on online brand reputation and interpreting the results of a classifier on tweet level.

2 Theoretical Framework

2.1 Machine learning

Machine learning techniques are tools to learn and understand data, which is done by building a high-performing system (Mitchell, 1997; Quinlan, 1986; Tong & Koller, 2001). Samuel (1959) used a self-improving system to let a computer win a game of checkers against a human, thereby inventing the idea of machine learning. Machine learning, the foundation of Artificial Intelligence (AI), is the collection of computer algorithms that aims on building systems that learns and improves by using data (Mitchell, 1997). A machine learning algorithm is generally trained on a random subset of data to get experienced in the data (Mitchell, 1997; Tong & Koller, 2001). This study assesses two approaches for training a machine learning classifier on textual data: supervised learning and semi-supervised learning. Both approaches are used for a sentiment classification task. Classification is a machine learning task that learns a set of rules from training instances to generalize the learned rules and subsequently to classify new observations (Kotsiantis et al., 2007). Supervised learning trains and improves a system by using instances that have a labelled variable of interest (Kotsiantis, Zaharakis & Pintelas, 2007). In some cases, instances are not labelled. Manually assigning a label to these instances requires human efforts which can be a time consuming and difficult process (Zhu, 2005). Semi-supervised learning can create classifiers that only requires a small amount of data to be labelled together with a large amount of data that is unlabelled (Zhou, Y., & Goldman, 2004; Zhu, 2005).

2.2 NLP

This study uses text mining to extract insights from tweets posted before and after the corporate crises. Text mining implies the process of extracting patterns or knowledge from textual data. The computerized approach that handles textual data is Natural Language Processing (NLP). Natural language processing is defined as a collection of computational techniques that analyses textual data to obtain a human-like level of language processing that can be used for multiple tasks (Liddy, 2001). NLP attempts to obtain a human-like level of language processing performance which makes it appropriate to consider NLP an AI discipline (Chowdhury, 2003; Liddy, 2001). NLP is utilized for information retrieval, which allows computers to effectively search documents based on relevance (Perez-Carballo & Strzalkowski, 2000). Furthermore, NLP is also often used for machine translation and speech recognition (Chowdhury, 2003).

2.3 Topic modelling

Some tweets might be more relevant to this research than other tweets, given that opinions on internet platforms can be about many aspects of the brands (Hennig-Thurau et al., 2004). With a large collection of documents, it is very useful to cluster tweets based on topics without having to read each single tweet. A frequently used unsupervised clustering approach in text data is topic modelling (Aggarwal & Zhai, 2013). There are two sorts of topic models: single-membership models and mixed-membership models. Single-membership models restrict documents to only one topic. In mixed-membership models each document is given by a mixture of topics. So in mixed-membership models, documents are given by a vector of proportions that represents what fraction of the words belong to which topic. The most known mixed-membership is Latent Dirichlet Allocation (LDA; Blei et al., 2003). Structural topic model (STM) innovates the LDA model by allowing covariates in the construction of the topic model (Roberts et al., 2014). STM is useful for including a time covariate to study whether a topic becomes more prevalent over time.

2.4 Sentiment Analysis

Sentiment analysis is a collection of computational methods that effectively analyses opinions, attitudes and emotions towards an entity (Ahmad et al., 2017; Medhat et al., 2014). Pang, Lee and Vaithyanathan (2002) introduced a new approach by classifying documents based on overall sentiment. Their study examined the effectiveness of sentiment classification based on movie reviews (Pang et al., 2002). The literature describes three classification levels in sentiment analysis: document, sentence and aspect level (Fang & Zhan, 2015; Medhat et al., 2014). Classification on document level assumes that each document contains opinions on a single entity. Classification on a sentence level determines the sentiment of each sentence in a document. The aspect level dives a bit further into the structure of the opinions. It assumes that an opinion is divided into two components: sentiment and the target to which the opinion is aimed at (Fang & Zhan, 2015). Fang and Zhan (2015) illustrated the aspect level with a clear example. Consider the sentence "The iPhone’s call quality is good, but its battery level is short”. This sentence contains two aspects, the call quality and battery level, and the iPhone as target. The goal of classification on aspect-level is to discover these sentiments of the aspects towards the entities (Fang & Zhan, 2015).

Many studies conducted research on Twitter sentiment analysis, each with their own approach (Aggarwal & Zhai, 2013; Dey et al., 2016; Go et al., 2009; Kouloumpis, E., Wilson, T., & Moore, 2011). For example, Go, Bhayani and Huang (2009) used emoji’s to train supervised machine learning algorithms to automatically classify the sentiment of Twitter posts. Kouloumpis, Wilson and Moore (2011) tried to use Part-of-Speech tagging in a supervised setting on microblogging messages, but they concluded that part-of-speech features are not very useful for microblogging messages like tweets. Naïve bayes classifier is one of the methods that is often used in sentiment prediction (Eyheramendy et al., 2003; McCallum & Nigam, 1998; Rish, 2001). Naïve bayes assumes that the features, like words in text classification, are independent from each other given the outcome class (Rish, 2001). The independence assumption is very naïve, given that it is validated for most real-life cases. In text classification, the independence assumption is naïve because some terms often appear together. For example, battery and charger are often used together which makes these terms dependent on each other. Although the assumption is naïve, the classifier often performs well (McCallum & Nigam, 1998; Rish, 2001).

Semi-supervised classification (SSC) is an area of machine learning that deals with sparsely labelled data. SSC aims to understand the data by transductive- and inductive learning (Zhou & Li, 2005). Transductive learning focuses on predicting labels of instances that are unlabelled by taking both labelled and unlabelled instances into account in the training procedure. Inductive learning concerns the problem of using unlabelled- and labelled data to predict unseen data. Zhou and Li (2005) presented an efficient semi-supervised method, which only requires a small portion of the data to be labelled. It uses supervised algorithms that classify instances for each other and thereby enlarging the labelled dataset.

The findings of previous studies that used supervised methods for sentiment predictions (Agarwal et al., 2011; Go et al., 2009; Kouloumpis, E., Wilson, T., & Moore, 2011) and the ensemble approach of Zhou and Li (2005) are the foundation for this study to investigate the following hypothesis:

H1: Semi-supervised learning predicts the sentiment of tweets towards brands more accurately than supervised learning on pre-labelled data.

2.5 Brand reputation

The first step in investigating corporate crises on any form of brand reputation is to form a clear definition of brand reputation (Barnett et al., 2006; Wartick, 2002). A review on corporate reputation of Barnett et al. (2006) showed confusion in the literature regarding the definitions of brand identity, -image and -reputation. Chun (2005) referred to the confusion as the 'reputation paradigm', multiple approaches exist in the definition of brand reputation. Chun (2005) argued that within the reputation paradigm there exists no source that captures the concept of reputation entirely. Fombrun an van Riel (1997) showed that the definition can depend on the academic perspective. They included definitions for the following distinctive perspectives: accountancy, economics, marketing, organizational, sociological and strategic. The similar characteristics of these perspectives on reputation were used to create an integrated view. Fombrun and van Riel (1997) see image and identity as the foundation of reputation. Brand identity and -image are given by the perceptions of internal observers and external observers, respectively (Barnett et al., 2006; Fombrun & Van Riel, 1997). Examples of internal observers are employees and managers and examples of external observers are consumers and investors (Barnett et al., 2006). This study defines brand reputation similar to the definition Fombrun and van Riel (1997) proposed. Brand reputation is defined as a representation of past actions and results of a brand, which shows the ability to deliver value to external and internal stakeholders.

2.6 Word-of-mouth (WOM)

Consumers tend to use other consumers' product evaluations as a source of information for their purchase decisions (Burnkrant & Cousineau, 1975). Burnkrant and Cousineau (1975) showed that when people observe a positive product evaluation, they have a more positive perception of the product compared to a situation where the consumers' evaluation is absent. Consumers are affected by other people's product evaluations, it can even affect their attitude and behaviour towards a product or brand (Herr et al., 1991; Laczniak et al., 2001; Litvin et al., 2008). The communication between consumers about a product or brand is summarized as word-of-mouth (hereafter WOM) communications. (Herr et al., 1991; Laczniak et al., 2001; Litvin et al., 2008). More specifically, WOM is defined as consumers' interpersonal conversation that influences and forms consumers' brand and product perception (Richins, 1984). It was later established that WOM is considered to be more credible compared to brand-controlled communications such as advertisements (Ahluwalia et al., 2000; Bickart & Schindler, 2001; Dean, 2004; Gruen et al., 2006). Given the credibility and the effects of WOM communications, it is important for brands to pursue positive WOM. Rogerson (1983) suggested that positive WOM resulted in higher volumes of new customer acquisitions and a lower customer churn rate. Negative WOM communications has the ability to negatively affect product judgements and product purchases (Arndt, 1967; Brown & Reingen, 1987; Herr et al., 1991). Brands should also be very cautious with negative WOM due to the negativity effect (Mizerski, 1982). The negativity effect makes consumers weight negative information more than positive information in their evaluation (Dean, 2004; Mizerski, 1982).

2.7 Electronic word-of-mouth (eWOM)

Due to the rise of internet and social media, the original WOM shifted towards a less personal but more universal online environment, which is called the electronic word-of-mouth (eWOM; Kim et al., 2014; Litvin et al., 2008). Given that consumers' opinions are uncontrollable and eWOM forms consumer's brand perception (Bambauer-Sachse & Mangold, 2011; Dean, 2004; Kim et al., 2014; Litvin et al., 2008), brands are integrating social media into their strategies to enhance their brand image (Hudson, Huang, Roth & Madden, 2016). Marketers see social media as a valuable platform where interest and awareness is generated by spreading product experiences and -opinions (Hudson et al., 2016). This paper will therefore focus on shared experiences and opinions on Twitter. Twitter, an open information-sharing micro-blogging platform, offers a great opportunity for brand perception mining (Kim et al., 2014).

2.8 Threat of electronic word-of-mouth compared to word-of-mouth in a crisis

A corporate crisis, which is often an unforeseen event, results in negative publicity of the brands (Ahluwalia et al., 2000; Bambauer-Sachse & Mangold, 2011; Dean, 2004; Seeger et al., 1998). On social media, this negative publicity can lead to an increase in negative eWOM communications about the brand (Cleeren et al., 2013; Folkes, 1984). The increase in negative eWOM communications is more harmful for the brand reputation compared to the traditional negative WOM (Bambauer-Sachse & Mangold, 2011). The difference between

WOM and eWOM is the ability of consumers to clearly observe consensus in the polarity of posts, which follows from the attribution theory (Bambauer-Sachse & Mangold, 2011; Kelley & Michela, 1980; Laczniak et al., 2001). As mentioned before, Burnkrant and Cousineau (1975) proved that consumers use other consumer's product evaluations as a source of information for their own product purchase choice. Attribution is the process of explaining consumers' evaluations as being caused by either internal- or external factors (Kelley & Michela, 1980). Internal attribution is explaining the behaviour of a subject to be caused by a subject's internal characteristic (Kelley & Michela, 1980). External attribution is explaining behaviour to be caused by something environmental, in other words caused by something outside of the subject (Bambauer-Sachse & Mangold, 2011; Kelley & Michela, 1980). Laczniak et al. (2001) used the theory of Kelley and Michela (1980) to explain that the type of attribution depends on the consensus perception. The consensus perception is defined as the perception of the degree to which other people on a platform agree on a message or share the same experience for the same brand (Kelley & Michela, 1980; Laczniak et al., 2001).

Imagine a situation where a consumer X has a negative experience with a brand due to a company crisis. Consumer X can look on an online opinion platform for other consumers' opinions on that product (Hennig-Thurau et al., 2004; Zhang & Pennacchiotti, 2013). A low consensus, a balance in positive and negative posts, leads to consumer X explaining the behaviour to be caused by internal characteristics (Kelley & Michela, 1980). An internal reason is for example that other people were unable to evaluate the product or brand properly (Bambauer-Sachse & Mangold, 2011). In the case of high perceived negative consensus, many negative posts and few positive posts, consumer X will explain the behaviour to be caused by something external (Kelley & Michela, 1980). Laczniak et al. (2001) suggested that the external type of attribution makes consumers think that the brand is to blame rather than the people that evaluate the brand. Traditional WOM communication often occurs sharing information between one person or a few people (Arndt, 1967; Brown & Reingen, 1987; Herr et al., 1991). In contrary to the original WOM, consumers have easy access to many eWOM communications (Bambauer-Sachse & Mangold, 2011; Litvin et al., 2008). The easy access to many eWOM communications and the negativity effect of Mizerski (1982) makes it more likely that consumer X will experience a high consensus (Chiou & Cheng, 2003). The findings of these studies combined suggests that negative eWOM communication caused by a company crisis is much more likely to be harmful for the brand compared to traditional WOM communications. This study attempts to find evidence on how a corporate crisis harms the eWOM communications. In order to do so, it is first important to study whether the eWOM communications related to the increased during the crisis. This is done according to H2:

H2: During a company crisis, tweets related to the crisis are relatively more prevalent than tweets related other topics.

After studying the prevalence of tweets related to the crises, the sentiment of those tweets is analysed. The sentiment before the crisis and the sentiment after the crisis is studied to find out whether the sentiment of the eWOM was negatively affected by the crisis. This will be tested by the following hypothesis:

H3: The sentiment of tweets is negatively affected by a company crisis.

3 Data

3.1 Company crisis cases

The data scraped for this research consists of tweets posted in the period before and after two company crises occurred. The first corporate crisis analysed for this research is the Samsung Note 7 recall procedure in August 2016. What was supposed to be the best smartphone at the time became a 47 day lasting crisis (Jin-man, 2016). The batteries of some Note 7s exploded, which made the products extremely dangerous. Samsung apologized for the battery issue and halted production and sales early October 2016 (Jin-man, 2016). In this crisis the costs for Samsung were extremely high and their market value declined (Jin-man, 2016). The second crisis that will be analysed is Apple's 'Batterygate'. Early 2017, it was brought to light that recent software updates decreased the processor power of Apple's iPhones (Robertson, 2020). By slowing down the processor power, people were driven into the decision to upgrade to a newer product (Robertson, 2020). For these updates, Apple had to pay a fine of $113 million (Clayton, 2020). Both Apple and Samsung have a large assortment of products. The difference between Apple and Samsung on Twitter is that the phones of Apple are called iPhones, while the phones of Samsung have the same name as the brand. Therefore, to understand the effects of the Apple iPhone crisis, the tweets are scraped by focussing on tweets related to iPhone and thus not Apple. For the Samsung crisis, this study will use tweets related to Samsung phones.

3.2 Twitter as a platform and data source

Twitter offers a great opportunity for brand image mining (Kim et al., 2014). It is an important driver in understanding eWOM communications (Kim et al., 2014). This simplifies the extraction of direct feedback compared to other forms of customer opinions, like customer questionnaires or calls (Jansen et al., 2009). The unique communication characteristics makes Twitter an important source of eWOM and thus a source for academic research. This study uses two different datasets to find evidence for the hypotheses.

The first dataset is a Twitter dataset of Go, Bhayani and Huang (2009). Go et al. (2009) trained a model without manually labelling data. Instead they trained a model to classify the sentiment of tweets by using noisy labels (Go et al., 2009). These noisy labels were emoticons, which are features that are potentially independent of context, topic and time (Read, 2005). Go et al. (2009) used their approach to label their data and published the dataset online. This dataset contains around 1.6 million labelled tweets. From these dataset, tweets can be filtered out to be similar to the context of the tweets gathered via the Twitter API. The data for supervised learning is filtered by only using the tweets that contain the following search terms: ‘phone’, ‘iphone’ and ‘samsung’. This brings the number of tweets from down to 13998.

The second dataset is scraped specifically for this research by using the Twitter API. The Twitter API allows scraping tweets for academic purposes to analyse and understand the public conversation (Twitter, 2021). This study used full archive API access to obtain tweets that were posted in a period of four months before and after the company crises. Twitter, however, limits the monthly number of tweets and scrape requests. The Twitter scraping process resulted in a dataset of 8900 tweets, where one half is scraped for the iPhone crisis and the other half for the Samsung crisis. The scraping process does not only scrape tweets but also the usernames of the people that posted the tweet and the date a tweet was posted. The date variable is then transformed into a categorical variable ‘crisis_nrofmonths’ with 8 levels, where the levels range from 4 months prior until 4 months after a company crisis. So ‘crisis_nrofmonths’ shows which month the tweet was posted in on an interval from 4 months prior until 4 months after a company crisis.

The dataset of Go et al. (2009) is not studied but used to train supervised learning techniques. A part of the tweets scraped by the Twitter API are manually labelled, in order to evaluate the performance on tweets related to the crises of the trained supervised techniques. In total, 269 tweets scraped through the Twitter API are manually labelled. From these tweets, 69 are combined with the 13998 tweets of Go et al. (2009) to form a training set of 14067 tweets. The 200 manually labelled tweets related to the crises that are left over are used for testing the models' predictions. This leads to a training and test split for the supervised learning techniques given by Table 1. For the semi-supervised approach, the scraped tweets related to the crises are used and the data of Go et al. (2009) is ignored.

	Training	Testing
Number of tweets	14067	200

Table 1. Total number of tweets in training- and testing. Note that Training is a combination of 13098 tweets of Go et al. (2009) and 69 manually labelled tweets related to the crises.

3.3 Pre-processing

Pre-processing steps are used to reduce the noise, size and complexity of textual data (Porter, 1980). The most relevant pre-processing steps are explained in this section, all pre-processing steps are given by Appendix A. The data contained some copied tweets that were posted more than once by different users, these duplicate tweets are removed from the data. Also, some users were appearing more than once in the data. They are not likely to have different opinions in the multiple tweets they posted. To make sure that every user only appears once in the data, duplicate Twitter user names are removed. Punctuation marks are removed. They are often not meaningful or too inconsistent for particular models (Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, 2016). Emoji's in tweets are transformed into the textual description of the emoji. All characters are also transformed into lowercase characters. Tweets could be about a specific product of a company. There are multiple ways someone could mention a specific product. For example the Samsung Galaxy Note 7 could be referred to as note 7, note7 or note seven. This variance is brought back to one word by looking up these variations and transforming it into noteseven. This is done for multiple products of both brands.

The English language is a space-delimited language, which means that words are separated by white spaces (Kannan et al., 2016). In order to make textual data more accessible for computers, all documents need to be vectorized (Webster & Kit, 1992). Vectorizing a document implies that each unique word should have its own identifier, also known as a token (Gupta et al., 2009). A token is defined as a unique identifier that represents a word that occurs at least once in the corpus (Webster & Kit, 1992). An example is the sentence "I like this product, because the product is great.". After the pre-processing steps, this sentence will be "i like this product because the product is great". This document contains 8 tokens, the word product is used twice but the term product will be a single token that is counted twice. With each term as a separate token, the document-term matrix (hereafter DTM) can be created. A DTM consists of documents on the rows of the matrix and all terms over the corpus are on the columns. So, the document example above will be presented in the DTM as a vector with a length of each unique term in the corpus. The values in the vector correspond to the number of appearances of each term in the document. The labelled tweets of Go et al (2009) and the tweets of this study are combined into a DTM for the supervised learning techniques. This results in a DTM with many features, where some features only occur in one document. A minimum frequency for the features is set to decrease the dimensionality of the data, subsequently lowering the computational time. The minimum frequency a word has to appear across all documents is set to 15.

4 Methodology

Throughout the methodology section 'corpus', 'word' and 'term' are used, which asks for some clarification on the definitions. The corpus is the collection of text across all documents. A word occurs in a sentence and a term is a unique word in the corpus. This section starts by elaborating on topic modelling, which will be used to test whether tweets about the crises become more relevant than other topics. The section continues with the sentiment prediction methods that are used to study the effects of a crisis on Twitter sentiment. As described in the data section, the data gathered for this study does not include a label that represents the sentiment of the tweet. The predictions are used to get insight in which features were important in the sentiment prediction. To test the hypothesis on prediction performance, the following two approaches will be used: supervised learning on pre-labelled Twitter data and semi-supervised learning on partially manually labelled data. By partially manually labelling the scraped data and using the dataset of Go et al. (2009), the proposed techniques can be trained and evaluated. The best performing model and LIME will be applied on a subset of data to investigate the effects of the features.

4.1 Topic modelling

Topic modelling can result in useful insights regarding the general topics over all tweets that were collected for this research. First Latent Dirichlet Allocation of (Blei et al., 2003) is explained, which forms the foundation for Structural Topic Modelling of (Roberts et al., 2014) that is applied in this study. The corpus is transformed into a DTM for both Latent Dirichlet Allocation (LDA) and Structural Topic Models (STM). LDA and STM thus ignore the word order by treating documents as a bag-of-words.

4.1.1 Latent Dirichlet Allocation

LDA aims to detect the combination of topics in the documents by creating a model that captures frequency of words in a document. This leads to soft clustering the documents based on these frequencies (Blei et al., 2003). LDA assumes that each document m consists of a mixture of underlying topics. The mixture of topics in a document m is given by a topic probability for each topic k, which is notated as topic probability $\theta_{m,k}$. Each topic k then consists of a mixture of underlying words. The mixture of words in topic k is the word probability for each word n, notated as word probability $\beta_{n,k}$. Both $\theta_{m,k}$ and $\beta_{n,k}$ are taken from a Dirichlet distribution. Blei et al. (2003) start explaining their model briefly by showing how an LDA model, that has already learned the corpus, can generate number of documents M that all consist of number of words N. The graphical representation of this explanation is given in Figure 1. Topic k is picked out for document m with a probability that followed from the topic probability $\theta_{m,k}$. The word probability $\beta_{n,k}$ for this topic k is used to generate a word n in the document. A more intuitive explanation on how LDA creates a document of N words is described in Appendix B.

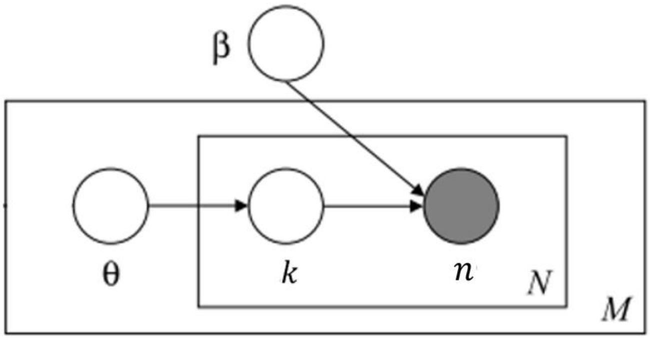


Figure 1. Graphical model representation of LDA to construct M number of documents. Topic probability θ generates a topic k. Word n follows from the word probability β of topic k, which is repeated for number of words N (Blei et al., 2003).

4.1.2 STM

LDA only uses the word frequency to determine the topics in documents (Blei et al., 2003; Yan et al., 2013). Roberts et al. (2014) innovated the LDA model by allowing covariates into the prior distributions. In other words, the prevalence of topics can be influenced by these covariates. STM is visualized in Figure 2 (Roberts et al., 2014). Starting with the left-side of Figure 2 regarding the document prevalence. Similar to LDA, STM assumes that each document consists of a mixture of topics. However, the topic probabilities $\theta_{m,k}$ in STM can be correlated. The data now consists of a matrix where the DTM is combined with additional features X, the so-called meta data. To get $\theta_{m,k}$, STM first multiplies X by penalty parameter τ. Penalty parameter τ makes sure that only highly correlated meta data will end up being influential. The topic probability vector $\theta_{m,k}$ is created by a logistic normal distribution with the outcome vector of τ multiplied by X as mean which allows for correlation between the covariates and the prevalence of topics. The topic probability vector is thus obtained by $\theta_{m,k}$ ~ $LogisticNormal(\tau X, \sum)$.

The right side of Figure 2, visualizes the covariate effect Y on word probability $\beta_{k,m}^{STM}$. The word probability $\beta_{k,m}^{STM}$ is given by $exp(\beta_{k,m} + \kappa_k + \kappa_X + \kappa_{X,k})$. It starts by the baseline LDA word probabilities $\beta_{k,m}$ across the corpus. Multiple deviation vectors are added to $\beta_{k,m}$ in order to get the covariate effect Y on word probability $\beta_{k,m}$. First, topic deviance κ_k is added to capture the non-document impact of topic k on the word frequency. Then the covariate deviation κ_X is added to capture the effects of the covariates on the word frequencies. Last addition to $\beta_{k,m}$ is deviation $\kappa_{X,k}$ that takes interaction effects, between topic k and the covariates X, on word-frequencies into account. All of these are taken together and put in a log space to obtain the word probabilities for a given topic within a document.

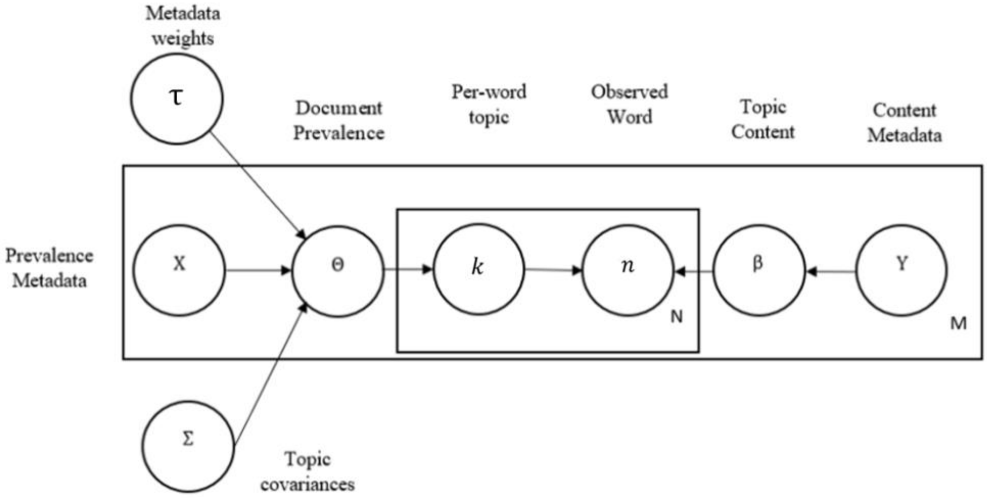


Figure 2. Graphical model representation of STM to construct M number of documents. Topic probability θ generates a topic k. n follows from the term probability β of topic k, which is repeated for the number of words N.

The number of topics K has to be determined first, which is decided based on the semantic coherence and exclusivity of all K topics (Roberts et al., 2014). A topic k is semantically cohesive when words that have a high probability for topic k often appear together in documents. The semantic coherence score is calculated according to $\sum_k \sum_{n<i} \frac{m(w_n w_i)+\alpha}{m(w_i)}$. It is the log of the probability that a document m contains a word n that is in the top 10 of the high probability words for topic k and also contains at least one lower-ranked word i. α is added to make sure that log zero issues are avoided. The semantic coherence values are negative, a large negative value indicates that high probability words do not co-occur often. A semantic coherence closer to zero indicates that high probability words do co-occur often.

Exclusivity is used for penalizing topics that are alike, which semantic coherence fails to do (Roberts et al., 2014). A topic k is exclusive when words that have a high probability for topic i have low probability under other topics. A good balance in high exclusivity and high semantic coherence is preferred. Exclusivity is calculated by dividing the probability of a high probability word in a topic by the sum of probabilities of the same word in all other topics. Exclusivity is the average of this value taken over all the high probability words.

To study the hypothesis on the prevalence of tweets about the crisis, the covariate 'crisis_nrofmonths' is included in the prior distributions. As explained in section 3.2., ‘crisis_nrofmonths’ is a 8 level categorical variable. STM is used to evaluate multiple values for the K number of topics. The number of topics with the best combination of semantic coherence and exclusivity is K_{se}, which will be the used in further evaluations. The one topic in K_{se} most related to the company crisis is selected and elaborated further, notated as k_{crisis}. Choosing k_{crisis} is based on the top words related to that topic. The model of Roberts et al. (2014) can then be used to estimate a regression of ‘crisis_nrofmonths’ on the topic prevalence of k_{crisis}. A significant positive effect, for the four months after the crisis, indicates that a tweet posted in the four months after the crisis increases the topic prevalence of k_{crisis}. In other words, k_{crisis} became relatively more prevalent in the four months after the crisis occurred. The R-package by Roberts et al. (2019) allows to test and evaluate the described STM models.

4.2 Sentiment prediction

4.2.1 Naïve Bayes

Naïve bayes classifier is the first supervision algorithm for tweet sentiment prediction. Consider a document m that consists of feature vector $X = (n_1, \dots, n_i)$ with the observed labels $Y = (y_0, y_1)$. In short, naïve bayes predicts class Y of document m based on the feature vector X. This classifying process is made simple by using the independence assumption, where features x_i are independent given class Y (Rish, 2001). The feature independence is formulated as $P(\mathrm{X}|\ Y) = \prod_{i=1}^{n} P(x_i\ |Y).$

The independence assumption makes naïve bayes able to simply learn the parameters for each feature separately. Naïve bayes classifier assigns a document with feature vector X to a class Y according to: $P(Y\ |\ X\ ) = argmax_Y \frac{P(X|Y) * P(Y)}{P(X)}$. The training data is used to calculate the prior probability $P(Y)$ of all available classes and the conditional probabilities $P(X|Y)$. The posterior probability $P(Y|X)$ is calculated for each available class Y. The predicted class for document m will be the class with the largest posterior probability. The so-called ‘evidence’ or predictor probability $P(X)$ is identical for all classes in Y and can therefore be ignored in the process.

McCallum and Nigam (1998) describe two text classification Naive bayes approaches, the Multi-variate Bernoulli Model and the Multinomial Model. In the Multi-variate Bernoulli Model, the features in the DTM are transformed into binary indicators that indicate whether a word from vocabulary V occurs at least once in the document. The independence assumption assumes independence between the probability of a word occurring in a document and the probabilities of other words occurring in a document.

The multinomial model, in contrary to the Multi-variate Bernoulli Model, does not transform the values of the features in the DTM into binary indicators. In the multinomial model, the DTM contains the number of times a word appears in each document. Both models do not take word order into account since both model use a DTM. The multinomial model, however, does capture the word frequencies within documents. Both the Multi-variate Bernoulli and the Multinomial model are trained and evaluated on the data subsets as described in the Data section.

All feature vectors in the DTM contain a lot of zero's, since most documents do contain a small fraction of all terms in the vocabulary. This issue increases the risk of zero-probabilities. Zero-probabilities occur in Naive bayes classifier when a term is not present in one of the training instances leaving the conditional probability for that term on zero. The issue can be resolved by ignoring that word in the training process or Laplace smoothing is applied in the training procedure. Laplace smoothing adds a value α to the word frequencies so that the conditional probability can no longer be zero, even though a term is not present in any of training instances (Kibriya et al., 2004). The multi-variate Bernoulli Model and the multinomial are tested and evaluated by using the naivebayes R-package of Majka (2019).

4.2.2 Random forest

Random forest is the second supervised learning algorithm used for tweet sentiment prediction. Random forest is an extension on the decision tree classifier, because it uses multiple decision trees under different circumstances to classify instances. Therefore, random forest is explained by first elaborating on the decision tree classifier. A classification decision tree classifies observations by making multiple simple decisions based on a pre-specified measure to form a tree structured flowchart (Breiman et al., 1984). In Figure 3, the circles where decisions need to be made on specific features X_m are called nodes. The data is split by each node and flows down the chart to eventually reach the terminal nodes. There are, however, a lot of features that can split the data, especially in text classification where each term is a feature. Gini Index is a splitting measures that constructs a decision tree based on feature relevance. The Gini index $I_G(c)$ observes the frequency of a feature being misclassified when randomly selected (Safavian & Landgrebe, 1991). The Gini Index is given by $1 - \sum_{i=1}^{J} c_i^2$, where J is the number of classes and c_i is the observed frequency of an instance being classified for the J^{th} class. If I_G is zero then all instances in the node belong to one class, which makes the node a pure node. An I_G of 1 indicates that the observations are randomly distributed across all classes. The Gini index thus splits the data based on node impurity. So at the start of building the decision tree, the feature with the lowest I_G, and thus the highest node impurity, will be used in the root node (Breiman et al., 1984). As the data flows down the tree, the level of impurity decreases. Consequently leading to a better classification in the terminal nodes. The decision tree determines the predicted class $\hat{Y}_m$ of a terminal node based on the class that appears the most out of the training instances that ended up in a particular terminal node. So a new instance flows down the trained decision tree to end up in a terminal node, where the class $\hat{Y}_m$ is given to the new instance.

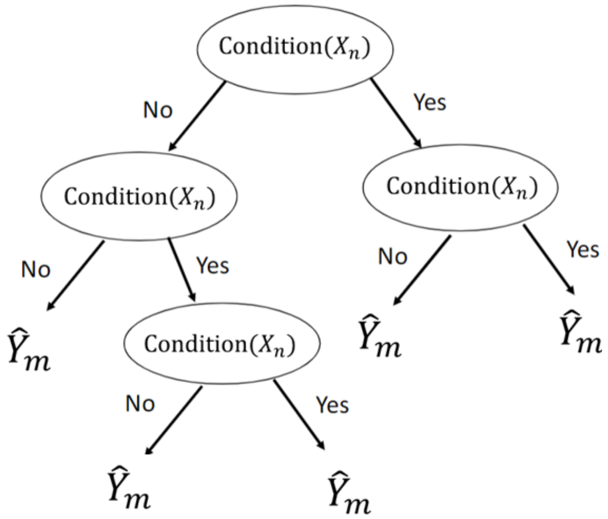


Figure 3. Example of a decision tree flow-chart structure

Branches and nodes at a lower point of a large grown decision tree are more likely to split the data based on specific noise in the training data. This leads to low prediction performance on data that was not used for training also known as overfitting (Breiman, 2001). Random forest improves the decision tree by combining the predictions of a group of decision trees into an ensemble of trees, which reduces the risk of overfitting the training data. Random forest is an extension on the bootstrap aggregating (bagging) approach of Breiman (1996). Bagging is a method that generates a number of prediction models that are combined into an aggregated model. Bagging bootstraps the original training data to create new training sets. Bootstrapping is a resampling approach that creates data sets by selecting samples of the original data with replacement (Efron & Tibshirani, 1994). Selecting samples with replacement implies that an instance is equally likely to be selected again after it was already added to the new training set. Bagging thus allows the presence of duplicate observations the new training data sets (Breiman, 1996).

Random forest is an ensemble method that ensembles decision trees. Combining multiple decision trees that are constructed according to the first part of this section, however, only leads to the same outcome as a single decision tree. All of these trees sort the data in the same way, which would not have any advantage compared to a single decision tree. The advantage of random forest is that it constructs less correlated trees by using an approach that differs from a single decision tree in two ways. It firstly uses the idea of bagging by creating a number of training data sets by bootstrapping the original training data (Breiman, 2001). Secondly, random forest uses random subsets of all available features for each node to construct each separate tree. A decision tree uses all available features in the data, equal to the number of words N in the corpus, to determine the best splits. Random forest uses a pre-defined number z of randomly chosen number of features ($z < N$) to determine the best split. Bootstrapping the training data and the random feature selection makes most trees in the random forest less correlated to each other. A different subset of the training data created by bootstrapping the original training data can result in different features being the best one to split the data, leading to new splits and thus new trees. Random feature selection results in less correlated trees, because features that have a high Gini Index might not be in the subset that the algorithm can choose from. The algorithm used for random forest is in the R-package randomForest of Liaw and Wiener (2002).

Random forest attempts to find a balance in overfitting and underfitting the training data. As explained before, decision trees are likely to focus on the variance in the training data and are thus likely to overfit the data (Quinlan, 1986). The use of different training samples and features puts less focus of the model on the smaller patterns in the data, making the model more generalizable on unseen data (Breiman, 1996). Reducing the variance of the random forest model thus limits the overfitting issues (Breiman, 1996). This comes at a cost of bias, because small but important patterns can be overlooked.

Each individual decision tree predicts the class of an instance, these predictions are transformed into one prediction based on the class that was predicted the most, also known as the majority-vote. Instances that are not selected for the training procedure are out-of-bag instances. The prediction error on these instances is the out-of-bag error (OOB-error), which can be used for tuning the hyperparameters. The random forest parameters that are tuned in this study are the total number of trees that need to be constructed and the number of features that are randomly selected for each decision tree. The number of trees should be high enough to ensure that every instance will be predicted a few times (Breiman, 2001). Therefore, the number of trees will be based on the point where the OOB-error stabilizes in the number of trees. The number of randomly selected features r_{try} are found by a step-wise OOB-error validation. r_{try} is defaulted to the square root of the original number of parameters (Liaw & Wiener, 2002). The step-wise search starts with the OOB-error of the default value, the new options for r_{try} are found by in- and decreasing the r_{try} based on a step factor (Liaw & Wiener, 2002). The function first looks left of the default by dividing the current r_{try} by a pre-defined step-factor. Decreasing r_{try} will be done until the OOB-error does not meet a pre-defined minimum improvement threshold. The same is done by increasing the default value by the step-factor. The value for r_{try} with the lowest OOB-error is used for the final random forest model.

4.2.3 Support Vector Machine

Support Vector Machine (SVM) seeks to maximize the margin between the patterns in the training data and a class boundary (Boser, Guyon & Vapnik, 1992). In linear separable cases there could be an infinite number of possible class boundaries (Boser, Guyon & Vapnik, 1992). The aim of SVM is to find the Maximal Margin Hyperplane (MMH) based on the perpendicular distance from the training instances to the hyperplane. A DTM consists of vectors for each document m with a length of the number of unique words, therefore the number of dimensions in textual data is equal to the number of words N. In a dimension space N, the hyperplane is a boundary of dimension $N-1$. The $N-1$ dimensional flat surface hyperplane is defined as $\Phi_0 + \Phi_1 n_1 + \Phi_2 n_2 + \ldots + \Phi_N n_{N\text{-}1} = 0$. With $\Phi_1, \Phi_2, \ldots, \Phi_p$ as parameters and an instance in the N dimensional space. Any document m that satisfies this equation lies on the hyperplane. Suppose that in Figure 4, the class label Y_m can be either 1 (circles) or -1 (squares). The mathematical expression for documents with label $Y_m = 1$ is then $\Phi_0 + \Phi_1 n_1 + \Phi_2 n_2 + \ldots + \Phi_N n_{N\text{-}1} > 0$. Class $Y_m = 1$ is above the hyperplane and thus larger than 0. The mathematical expression for documents with label $Y_m = -1$ is $\Phi_0 + \Phi_1 n_1 + \Phi_2 n_2 + \ldots + \Phi_N n_{N\text{-}1} < 0$. Class Y_m = -1 is smaller than 0 since it lies below the hyperplane in Figure 4.

The separating hyperplane in Figure 4 is the maximum margin hyperplane. It maximizes the margins to instances in the N-dimensional space. Generalizing the previous two equations while looking for the maximum margin leads to: $Y_m * (\Phi_0 + \Phi_1 n_1 + \Phi_2 n_2 + \ldots + \Phi_N n_{N\text{-}1}) \geq V$. With V as the margin width to the instance and Y_m as the class. The instances with equal distances to the hyperplane defines the margin, these instances are support vectors. Therefore, the classification strongly depends on supporting vectors (Boser, Guyon & Vapnik, 1992).

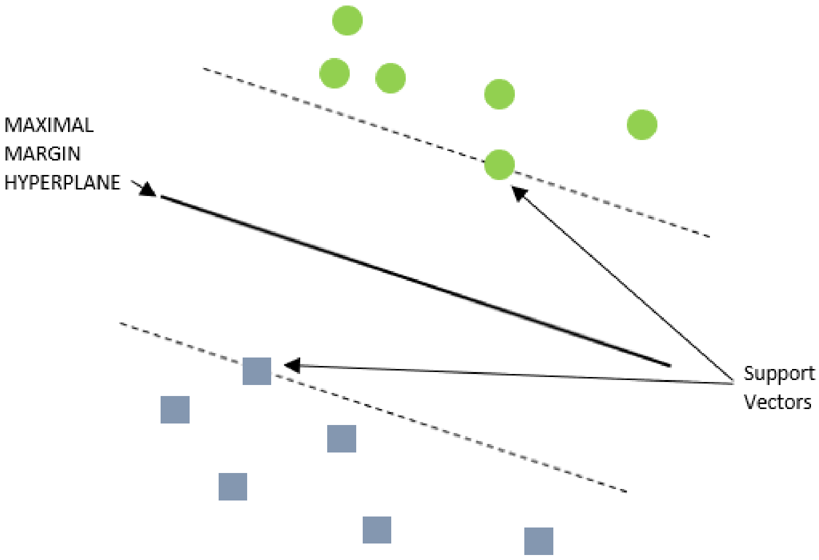


Figure 4. Maximum margin decision hyperplane example.

However, data with a lot of features makes a linear separation unlikely (Boser et al., 1992; Schölkopf, 2001). For cases where data is not linear separable, SVM can also use a soft-margin. This soft-margin allows for some misclassification. While looking for the maximum margin hyperplane, a slack variable ϵ_m is added that allows for constraints to be violated. It can penetrate the margin and the boundary but it is favourable to keep the occurrence of penetration as small as possible. ϵ_m is larger or equal to 0 $(0 \leq \epsilon_m)$ and can be larger or smaller than a value u, where $u > 0$. This leads to the following equation for the hyperplane $Y_m * (\Phi_0 + \Phi_1 n_1 + \Phi_2 n_2 + \ldots + \Phi_N n_{N\text{-}1}) \geq V(1 - \epsilon_m)$. Take for example

u = 1 in Figure 5. If ϵ_m is larger than 1, the instance is misclassified and on the wrong side of the boundary. An ϵ_m smaller than 1 but larger than 0 lies between the boundary and the margin. A smaller value for u leads to a narrow margin that is not often violated, which leads to a lower bias but a higher variance. The opposite is the case for a larger value for u.

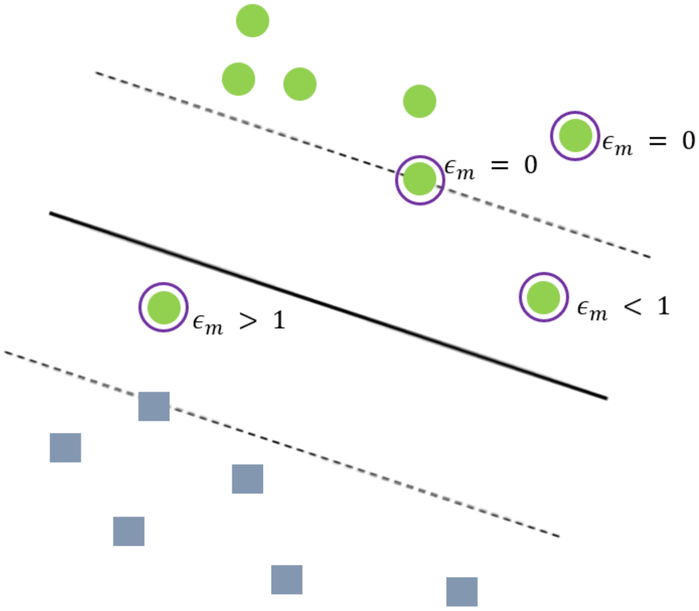


Figure 5. The slack variable ϵ_m with u = 1. $\epsilon_m = 0$ for a correct classification, $\epsilon_m > 1$ for misclassifications, and $0 < \epsilon_m < 1$ for misclassifications close to the margin line.

The class boundary in text classification is a high-dimensional hyperplane, where the number of dimensions is influenced by the number of features in the data. The Kernel trick can be the solution to improve prediction performance for non-linear class boundaries. The Kernel trick applies non-linear transformations on the existing features to create new features (Schölkopf, 2001). These features will be the solution for SVM to find a better boundary fit that is nonlinear (Schölkopf, 2001). One of the most frequently used radial basis function kernel is the Gaussian Radial Basis Function $R(x, x') = exp(-\gamma \sum_{n=1}^{N}(x_{m,n} - x'_{m,n})^2)$. With hyperparameter γ that controls the effect of new features $x'_{m,n}$ on the boundary. It accounts for the boundary smoothness and controls model variance, as can be seen in Figure 6. A higher γ implies higher focus on the variance of new features, which can lead to overfitting. The boundary has lower variance for a lower γ.

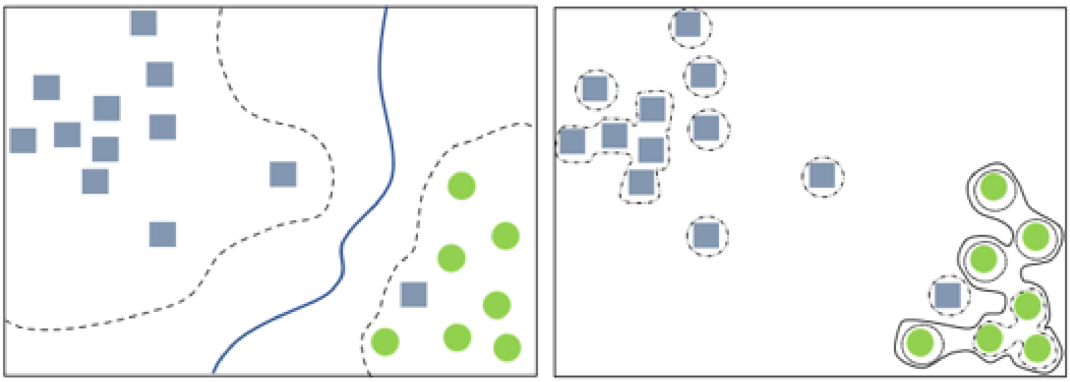


Figure 6. SVM non-linear boundaries for a smaller value of γ (left) and a higher value of γ (right)

The SVM models are tested and evaluated by the e1071 R-package of Meyer et al. (2020). The optimal value for γ is found by K-Fold cross-validating multiple possible values for γ. K-Fold cross-validation is done by the caret package of Kuhn (2020). Cross validation refers to the process of validating a model's performance on data that was not used for the training procedure of the model. In K-Fold cross validation, the training data is randomly divided into K number of equal training data subsets. For each fold K, each subset will serve as a test set once to validate the model while the other subsets are used for training the model. With the optimal value γ that followed from K-fold cross-validation, the SVM model is evaluated on the 200 labelled instances related to the crisis (Boser et al., 1992; Breiman, 2001).

4.2.4 Semi-supervised learning

The best performing supervised learner out of the ones described so far will be used for the semi-supervised learning approach self-training (González et al., 2009; Yarowsky, 1995). González et al. (2009, 2019) created the ssc R-package that can be used for self-training. The data contains the labelled instances L and the unlabelled instances U. The self-training classifier is trained on the small subset of labelled instances L to classify instances that are unlabelled. The model is retrained with the most confident predictions, which enlarges the labelled training dataset. The most confident prediction is defined as the predicted class probability threshold that decides whether the instance will be labelled and added to the training set. Self-training assumes that the most confident predictions are very likely to be the correct predictions. This process of continuously adding labelled instances into the training set of the classifiers repeats itself until a threshold is met. The threshold is a percentage of new labelled instances, which stops the process of adding newly labelled instances in the training set if met.

4.2.5 Local Interpretable Model-agnostic Explanations

To extract insights from how a crisis affects tweet sentiment during the crisis, the predictions of the applied models will be made more explainable. An explanation technique for the classifiers is Local Interpretable Model-agnostic Explanations (LIME). LIME learns a local interpretable model around the predictions of a model (Ribeiro et al., 2016). LIME produces an explanation for a model following function: $explanation\ (x) \ = \ \underset{g\,\in\,G}{\operatorname{argmin}}\ \mathcal{L}(f, g, \pi_m) \ + \ \Omega(g)$. With $g \ \in \ G$ defined as an explanation and where G is the collection of potential interpretable models.

These interpretable models can visualized in a plot. These plots can, however, still vary in interpretation complexity. Therefore, LIME uses a fidelity-interpretability trade-off by including a complexity measure $\Omega(g)$ into the function. For classification tasks, $f(m)$ is the probability of m belonging to a certain class. To define the locality around an instance, π_m defines the size of the neighbourhood around the instance m that is explained. Some of the parameters explained above are taken together into $\mathcal{L}(f, g, \pi_m)$, which represents how close the explanation model g is to the estimation of model f in the locality of π_m. LIME ensures the fidelity-interpretability trade-off by minimizing $\mathcal{L}(f, g, \pi_m)$ while keeping $\Omega(g)$ low enough to keep the explanation interpretable. An instance of interest m is selected to explain the prediction of m by a model f. LIME then perturbs predicted observations to get new predictions from a classifier, subsequently creating a new dataset with its own labels (Ribeiro, Singh & Guestrin, 2016). LIME trains an interpretable model g on this new dataset and approximates the local weights for the features (Ribeiro et al., 2016). In the context of text classification, these local weights show which words support LIME’s prediction and which words contradict LIME’s prediction and how strongly they affect the prediction. Knowing which words support or contradict the prediction offers an interpretable opportunity to identify the most important words that determine the tweets’ sentiment. This approach is limited to a local explanation though, which is not necessarily a good global explanation (Ribeiro et al., 2016). LIME is applied on the instances that have the highest topic probability for the crises topics of STM to extract important features from the tweets. It is done so with the R-package called lime of Pedersen and Benesty (2019).

5 Results

This section explains the results and performances of the methods described in the methodology. The supervised- and semi-supervised methods are all evaluated on a matrix, displayed in Table 2 (Boser et al., 1992; Breiman, 2001). Table 2 shows the actual labels compared to predicted labels. From this table, the accuracy of a model can be obtained by: $\frac{TP+TN}{TP + FP + TN + FN}$. The accuracy metric will be important in studying which approach results in the highest prediction performance of tweet sentiment towards brands.

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	True Positives (TP)	False Positives (FP)
Predicted negative (0)	False Negatives (FN)	True Negatives (TN)

Table 2. Confusion matrix of actual labels compared to predicted labels.

5.1 Structural Topic Modelling

The two company crises are different from each other, which may result in different topics after both crises occurred. Therefore, STM is applied on the two crises separately. A covariate 'crisis_nrofmonths' is added to the prior distributions of word probabilities θ and β.

5.1.1 Choosing number of topics K

The exclusivity and semantic coherence of each topic is calculated separately different values of K = (5, 10, 15,20, 25). This relationship between exclusivity and semantic coherence is visualized in Figure 7 and averaged for each K in Table 3.

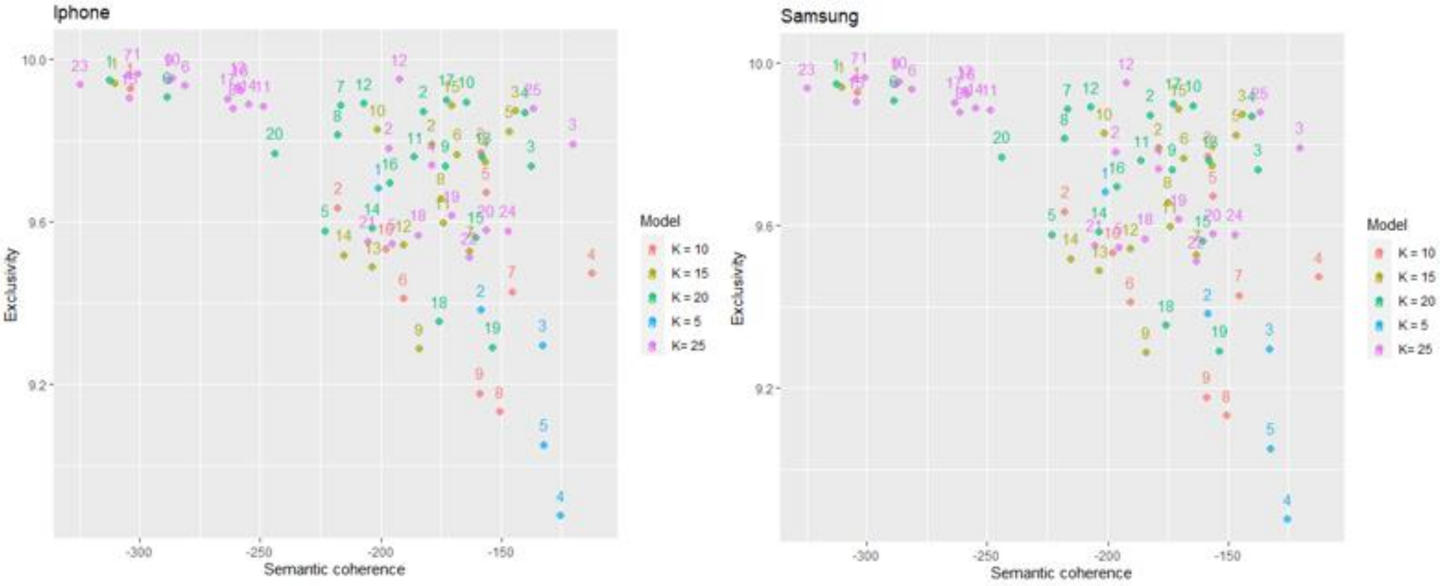


Figure 7. Exclusivity plotted against the semantic coherence of each topic for different values of K

iPhone			**Samsung**		
Number of topics	**Exclusivity**	**Semantic Coherence**	**Number of topics**	**Exclusivity**	**Semantic Coherence**
K = 5	9.2586	-153.1061	K = 5	9.3149	-153.6233
K = 10	9.5920	-187.8621	K = 10	9.5690	-182.4879
K = 15	9.6617	-183.4737	K = 15	9.6464	-174.9026
K = 20	9.7395	-203.7798	K = 20	9.7142	-198.7879
K = 25	9.7828	-217.8964	K = 25	9.7585	-203.6782

Table 3. Average STM exclusivity and semantic coherence for different values of K for iPhone (left) and Samsung (right).

Both iPhone and Samsung have similar values of exclusivity and semantic coherence in terms of increasing and decreasing in the number of topics K. The decision on the number of topics K for both cases thus follow the same argumentation. For K = 5, the topics have high semantic coherence, indicating that the words that have a high probability for these topics often appear together in a document. The topics are, however, similar to each other given that the topics for K = 5 perform worse on exclusivity. Table 3 shows that 25 topics results in the highest average exclusivity, but the average semantic coherence is the lowest for all options of K. A group of topics underperform in semantic coherence as can be seen in the upper-left side of Figure 7. Setting K to 15 makes the topics more exclusive and cohesive compared to K = 10. The decision between the options K = 15 and K = 20 depends on the relative importance of exclusivity and semantic coherence and the interpretation of the topics. Topic modelling is used to prove that tweets about the crisis become more prevalent after the crisis. To make sure that a topic is really about the crises and not the brand in general, exclusivity is made slightly more important than semantic coherence in this case. Therefore, K_{se} = 20 will be used in the STM construction for both cases.

5.1.2 Topic interpretation

The topic probabilities for both cases are presented in Figure 8 together with the top three words based on word probability. For the Iphone STM model, the topic with the highest probability is Topic 4. Though, the topic that addresses the crisis is Topic 15, where words with high probability are ‘battery’ and ‘life’. Topic 15 is quite a relevant topic with a topic probability larger than 12 other topics. The Samsung STM model shows high topic probability for Topic 3. The topic relevant for the prevalence hypothesis is Topic 17 since ‘noteseven’ is the word with the second highest probability for this topic. Topic 17 has the second highest topic probability making it one of the most relevant topics.

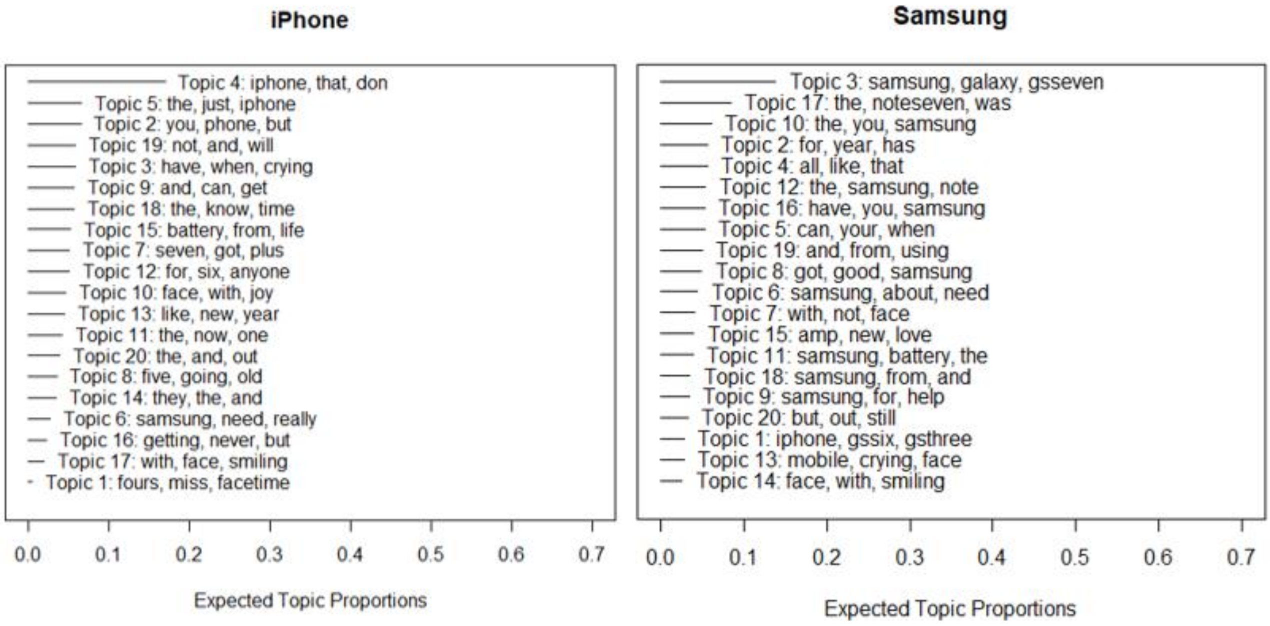


Figure 8. Expected topic proportions for STM model with K=20 and the top 3 words with the highest probability for each topic

Figure 9 zooms in on these topics by showing all relevant words in a wordcloud. In the wordcloud for iPhone Topic 15, there are many words that correspond to an iPhone battery draining faster than expected. Especially words or combinations like ‘replace’, ‘life died’ and ‘randomly drains faster’ indicate that this topic covers opinions regarding the battery life of an iPhone. The wordcloud of Samsung Topic 17 in Figure 9 also specifically covers the issues people tweet about regarding the Samsung Note 7. Issues regarding the crisis captured by the wordcloud are the banishment of the device from ‘airports’, the ‘recall’ procedure of ‘samsung’ and foremost the relation of ‘noteseven’ and ‘samsunggalaxynoteseven’ to ‘exploding’. K_{crisis} for Samsung is thus Topic 17 and K_{crisis} for iPhone is Topic 15.

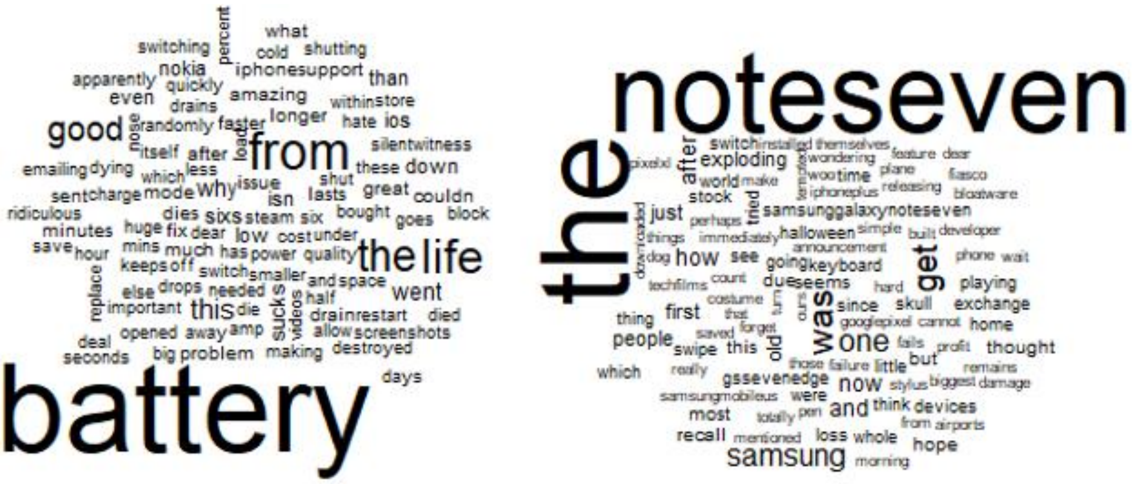


Figure 9. Wordclouds that contain words highly associated with Topic 15 of iPhone (left) and Topic 17 of Samsung (right)

Topic 15 for the iPhone case and Topic 17 for the Samsung case are thus identified as the topics that capture the company crises. The covariate ‘crisis_nrofmonths’ was added to the prior distributions before constructing the STM model, which allows to estimate a regression of this covariate on the topic prevalence. Table 4 shows such a regression of all levels in ‘crisis_nrofmonths’ on the topic prevalence of Topic 15 of the iPhone STM. All months after the iPhone crisis occurred show positive significant effects of around 0.05 for all months after the crisis on the topic prevalence of Topic 15 on a significance level of at least 0.01. These effects suggest that Topic 15, which captured aspects of the crisis, becomes more prevalent in the months after the crisis.

iPhone Topic 15

Coefficients	Estimate	Std. Error	Pr (> |t|)	significance
Intercept	0.0233	0.0127	0.0679	.
Crisis_nrofmonths3 months prior	-0.0036	0.0140	0.7966	
Crisis_nrofmonths2 months prior	0.0048	0.0138	0.7245	
Crisis_nrofmonths1 month prior	0.0056	0.0143	0.6936	
Crisis_nrofmonths1 month after	0.0623	0.0143	0.0000	***
Crisis_nrofmonths2 months after	0.0553	0.0142	0.0001	***
Crisis_nrofmonths3 months after	0.0509	0.0144	0.0004	***
Crisis_nrofmonths4 months after	0.0435	0.0143	0.0143	**

*Table 4. Regression summary of crisis_nrofmonths on the topic prevalence of iPhone Topic 15. Significant codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1*

Table 5 shows the regression of all levels in 'crisis_nrofmonths' on the topic prevalence of Topic 17 of the Samsung STM. The months after the crisis of Samsung show positive significant effects of the months after the crisis on the topic probability of Topic 17. In this case, two- and three months after the crisis show higher effects than 1- and 4 months after the crisis. This indicates that the tweets about Topic 17 appeared mostly in two- and three months after the crisis. Overall, all months after the crisis significantly increases the topic prevalence of Topic 17.

Samsung Topic 17

Coefficients	Estimate	Std. Error	Pr (> |t|)	significance
Intercept	0.0452	0.0102	0.0000	***
Crisis_nrofmonths3 months prior	-0.0109	0.0113	0.3342	
Crisis_nrofmonths2 months prior	-0.0038	0.0110	0.7264	
Crisis_nrofmonths1 month prior	-0.0045	0.0118	0.6986	
Crisis_nrofmonths1 month after	0.0266	0.0112	0.0183	*
Crisis_nrofmonths2 months after	0.0455	0.0109	0.0000	***
Crisis_nrofmonths3 months after	0.0573	0..0116	0.0000	***
Crisis_nrofmonths4 months after	0.0243	0.0119	0.0415	*

*Table 5. Regression summary of crisis_nrofmonths on the topic prevalence of Samsung Topic 17. Significant codes: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1*

The subset of unlabeled data that will be used in LIME is created by selecting Samsung tweets that have the highest probability for Topic 17 and iPhone tweets with the highest probability for Topic 15. This leads to a subset of 396 unlabeled tweets related to the crisis used for further analysis.

5.2 Sentiment prediction

5.2.1 Naïve Bayes model

The confusion matrix in Table 6 shows the performance of naïve bayes on the labelled tweets. The model correctly classifies 133 tweets out of the 200 tweets resulting in an accuracy of 66,5%. The model's prediction results in more false negatives compared to false positives.

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	65	29
Predicted negative (0)	38	68

Table 6. Confusion matrix of naïve bayes classifier

5.2.2 Random Forest

In order to find the optimal random forest model, the number of trees and the number of random features have to be found. Starting with the optimal number of trees, Figure 10 shows the OOB-error plotted for each number of trees until a number of trees equal to 400. The OOB-error starts at 35% and decreases until it stabilizes at 26% for any number of trees larger than 75. As Breiman (2001) suggested, the number of trees does not result in overfitting the model. Therefore, finding the optimal number of random features will be based on a random forest with 75 trees.

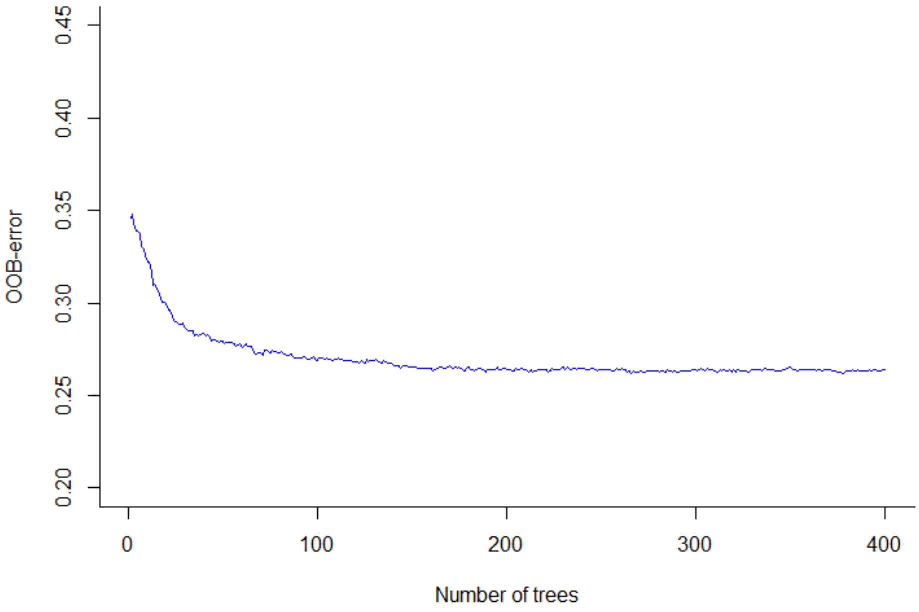


Figure 10. Plot of OOB-error for number of trees ranging from 1 until 400

Figure 11 shows the step-wise search for the r_{try}. In this search, the step-factor is 1.5 and the relative improvement has to be larger than 0.01. The search started at 34, the rounded square root of the number of features, which was 1162 features. The combination of step-factor and relative improvement results in a minimum OOB-error for r_{try}= 16. The final random forest is thus constructed with 75 trees and r_{try} = 16.

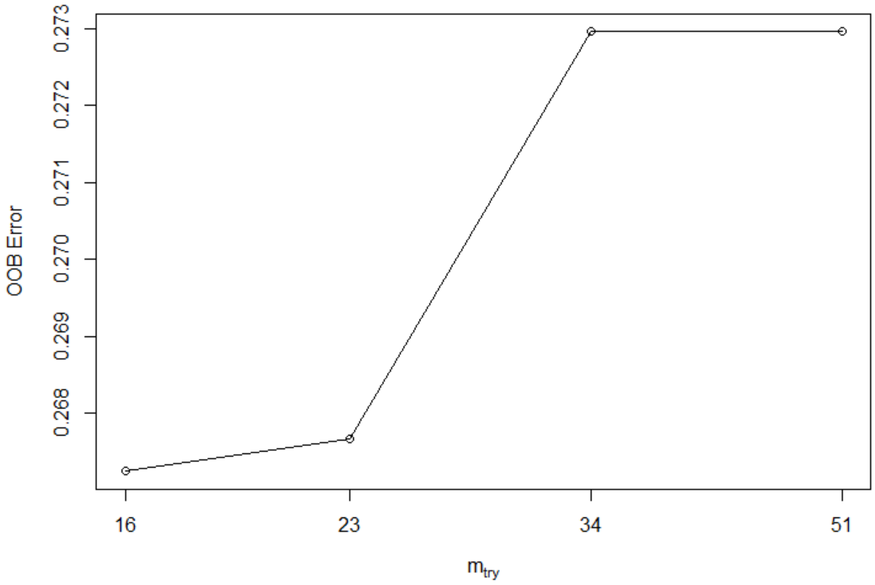


Figure 11. Plot of OOB-error for different values of m_{try}

The final random forest model performs similarly to the naïve Bayes classifier. Random forest correctly classifies 130 tweets and misclassifies the remaining 70 tweets, resulting in an accuracy of 65% in Table 7.

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	64	34
Predicted negative (0)	36	66

Table 7. Confusion matrix of random forest classifier

5.2.3 Support Vector Machine

The optimal SVM model with radial basis kernel function is found by 3-Fold cross-validating multiple values of γ. The left plot in Figure 12 shows the accuracy for different values of γ between 0.1 and 1 with steps of 0.1. The accuracy on the cross validated testing sets is highest for 0.1. Since no values smaller than 0.1 were tested in the 3-Fold cross-validation, it could be possible that there is a better value for γ that is not tested by the left plot in Figure 12. Therefore, a new 3-Fold cross-validation is executed for values of γ on an interval 0.01-0.1 with steps of 0.01. The cross-validated accuracy is visualized in the right plot of Figure 12. It becomes clear that a value of 0.07 leads to a higher accuracy than a value of 0.1 for γ. A value of 0.07 for γ implies lower focus on the variance of new features, lowering overfitting the data.

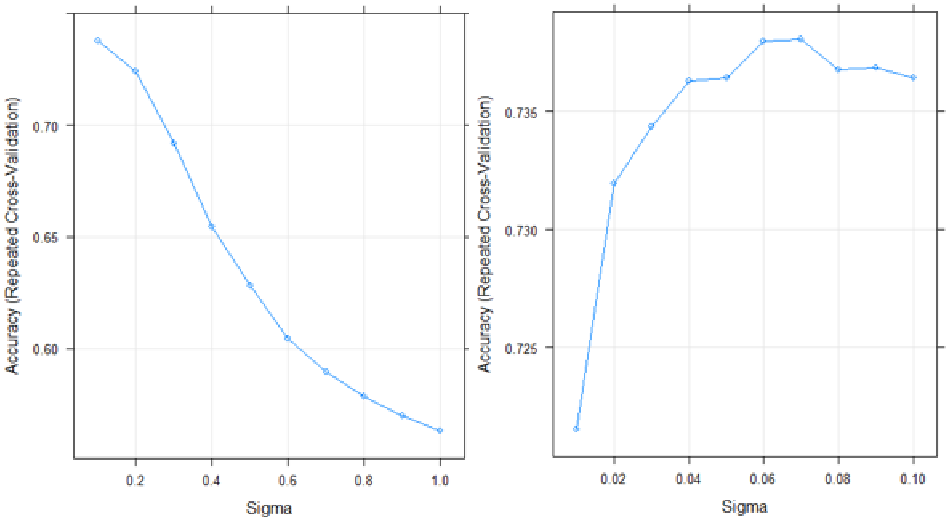


Figure 12. 3K-fold cross validated accuracy for SVM model with sigma on interval 0.01-0.1 (left) and 0.1-1 (right)

The SVM model with γ = *0.07* is used to predict the labelled part of the Twitter data. SVM performs best on true positives compared to naïve bayes and random forest. This increased performance in true positives comes with a decrease in performance on true negatives. SVM also results in more false positives. The SVM model classifies 65% of the tweets correctly as can be seen in Table 8, similar to naïve bayes and random forest.

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	70	37
Predicted negative (0)	33	60

Table 8. Confusion matrix of support vector machine classifier

5.2.4 Self-training

As mentioned in section 4.2.4., the best performing supervised algorithm is used for the self-training approach. As can be seen from the performance overview in Table 9, the naïve Bayes classifier correctly predicted 66.5% of the 200 labelled instances. Random forest and SVM both correctly predicted 65% of the instances. Therefore, naïve Bayes classifier is used for the self-training model.

	Naïve Bayes	Random forest	Support-Vector Machine
Accuracy	66.5%	65%	65%

Table 9. Performance overview of naïve Bayes, random forest and SVM measured in accuracy on the 200 labelled instances in the test dataset

The Self-training model requires two values for the confidence threshold and the percentage of unlabelled instances that become labelled and added to the training sets. Table 10 shows the optimal values out of a grid search of all combinations for confidence threshold interval 0.1-1 and percentage labelled on an interval 0.1-1.

	Confidence threshold	Percentage labelled
Value	0.7	0.7

Table 10. Optimal values for confidence threshold and percentage labelled for the self-training model

The, on the grid search based, optimal self-training model is accurate for 75% of all labelled instances used for testing the predictor as can be seen in Table 11. Compared to previous techniques, the number of false negatives is strongly reduced and the amount of false positives does not differ much. The self-training is applied on the unlabelled instances related to the crisis, these results are given by Table 12. According to the self-training predictions, the number of negative tweets exceeds the positive tweets even more than predicted by naïve bayes, random forest and SVM. 75,8% of all unlabelled instances related to the crises has negative sentiment. LIME is applied in the next section to be sure that the predictions on the unlabelled instances related to the crisis are meaningful.

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	82	35
Predicted negative (0)	15	68

Table 11. Confusion matrix of self-training

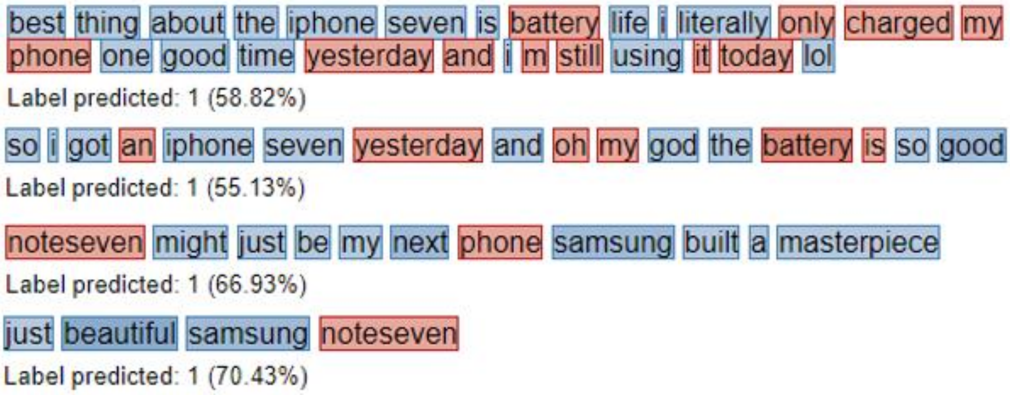
Self-Training	Predicted positive (1)	Predicted negative (0)
Number of tweets	300	86

Table 12. Number of positive and negative tweets in the filtered unlabelled data for support vector machine

5.3 LIME

LIME shows feature importance together with the predicted label and the probability of the predicted label. The features can be either blue or red, where blue suggests that the feature supports the predicted label and red contradicts the predicted label. The brightness of the colours vary too, darker colours have higher feature weight compared to lighter colours. This section is divided into tweets that are observed as true positives, true negatives, false positives and negatives.

Starting with Figure 13, which shows four tweets out of Table 12 for the LIME interpretation of true positives. The first two tweets related to iPhone in Figure 13 have a lower label prediction probability compared to the last two Samsung related tweets, indicating less prediction confidence that the label is correct. The first two tweets are positive towards the iPhone seven but the model sees ‘battery’ and ‘charged’ as negative features. The negative feature weight is higher for ‘battery’ in the second tweet making the prediction probability even lower than the first tweet. The feature ‘noteseven’ contradicts the probability of the third tweet and fourth tweet being positive. However, ‘samsung’ and ‘beautiful’ compensate this in the third and fourth tweet, respectively.



best thing about the iphone seven is battery life i literally only charged my phone one good time yesterday and i m still using it today lol

Label predicted: 1 (58.82%)

so i got an iphone seven yesterday and oh my god the battery is so good

Label predicted: 1 (55.13%)

noteseven might just be my next phone samsung built a masterpiece

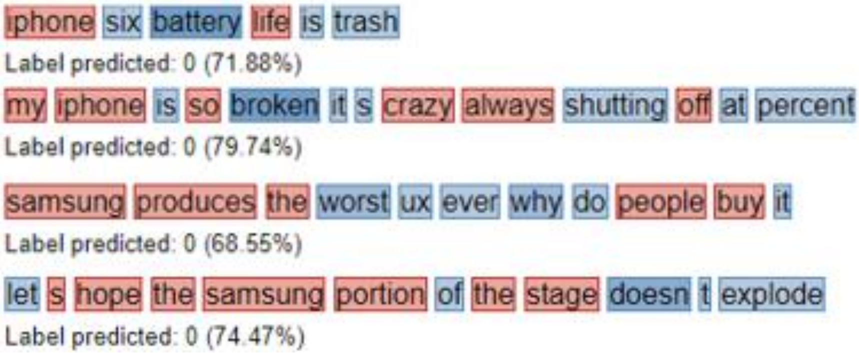
Label predicted: 1 (66.93%)

just beautiful samsung noteseven

Label predicted: 1 (70.43%)

Figure 13. LIME interpretations of true positives in the filtered unlabelled data

A similar pattern is shown in the true negatives in Figure 14, but this time a higher label prediction probability for all four tweets. The feature 'battery' strongly supports a negative prediction. Also, 'iphone' and 'samsung' contradict the negative label prediction in the tweets. However, while iPhone tends to be more related to positive tweets, the product iPhone 'six' is associated with negative sentiment. The model does show that 'explode' supports negative sentiment in a tweet.



iphone six battery life is trash

Label predicted: 0 (71.88%)

my iphone is so broken it s crazy always shutting off at percent

Label predicted: 0 (79.74%)

samsung produces the worst ux ever why do people buy it

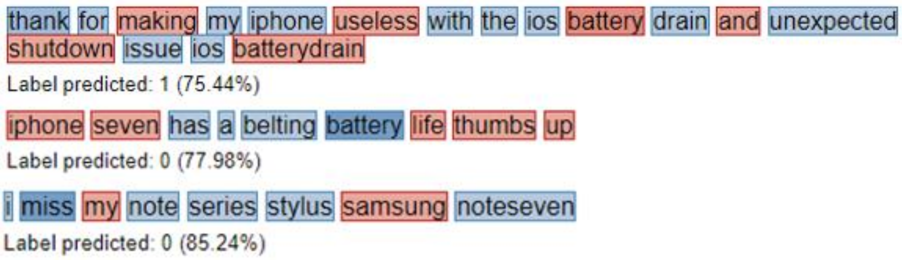
Label predicted: 0 (68.55%)

let s hope the samsung portion of the stage doesn t explode

Label predicted: 0 (74.47%)

Figure 14. LIME interpretations of true negatives in the filtered unlabelled data

Figure 15 shows a set of misclassified tweets. The first tweet is sarcastic, which is difficult for the model to recognize. The feature 'thank', the misspelled version of thanks, is strongly associated with positive sentiment. So even though 'battery', 'shutdown' and 'batterydrain' contradict the prediction, it is still labelled as a positive tweet. Again, iPhone has positive feature weight for positive sentiment. The second tweet in Figure 15 really shows the impact of the word 'battery'. The negative feature weight is so impactful that a positive tweet is labelled as negative with a label probability of almost 78%. The last tweet is negative according to the model, mostly due to the impact of ‘miss’. In this tweet ‘samsung’ does not support the negative label while ‘noteseven’ does.



thank for making my iphone useless with the ios battery drain and unexpected shutdown issue ios batterydrain

Label predicted: 1 (75.44%)

iphone seven has a belting battery life thumbs up

Label predicted: 0 (77.98%)

i miss my note series stylus samsung noteseven

Label predicted: 0 (85.24%)

Figure 15. LIME interpretations of misclassifications in the filtered unlabelled data

6 Conclusion

6.1 Summary of findings

This paper aimed to prove that a crisis has a negative effect on Twitter messages. The Samsung Note Seven exploding battery crisis and the iPhone update crisis were used as cases to measure the effects. The results of these crises are studied on a collection of tweets and on tweet level. It studied the effects of the crises on sentiment by the following research question:

Which features are most important for the sentiment of a tweet towards a company in a crisis and how do these features affect tweet sentiment?

In order to find an answer to the research question, it was first important to know whether the crises were mentioned in the tweets. Structural Topic Modelling was used to find evidence for hypothesis 2:

H2: During a company crisis, tweets related to the crisis are relatively more prevalent than tweets related other topics.

The Structural Topic Modelling model generated 20 topics for both cases, where for each case one topic was associated to the crisis. For iPhone, the associated words or word combinations to this topic were ‘battery’, ‘life died’ and ‘randomly drains faster’. ‘Noteseven’, ‘exploding’ and ‘samsunggalaxynoteseven’ were associated to the crisis topic for the Samsung case. The covariate 'crisis_nrofmonths' was added to the prior distributions to estimate the effect of the number of months prior and after the crisis on the topic prevalence. For both crisis cases, the months after the crisis occurred showed significant positive coefficients on the topic prevalence. Evidence was thus found for hypothesis 2.

Machine learning techniques were used to predict the sentiment of the tweets. However, only a small part of the data was labelled. Therefore, supervised learning techniques were used and additional to these techniques a semi-supervised learning technique was tested. The accuracy metric was used to evaluate the methods and to find evidence for hypothesis 1:

H1: Semi-supervised learning predicts the sentiment of tweets towards brands more accurately than supervised learning on pre-labelled data.

The classifiers of naïve Bayes, random forest and support-vector machine were trained on pre-labelled data of Go et al. (2009) and tested on the labelled subset of the scraped data. Self-training used the small labelled subset of 69 instances and the unlabelled instances to retrain the model a number of times while enlarging the labelled subset with the most confident predictions. The supervised techniques showed similar metrics in the confusion matrices with an accuracy of 66.5% for naïve Bayes and 65% for SVM and random forest. The self-training model correctly predicted 75% of the labelled instances of the scraped data, which was the highest across all used methods. In this case, the semi-supervised technique outperformed the supervised techniques.

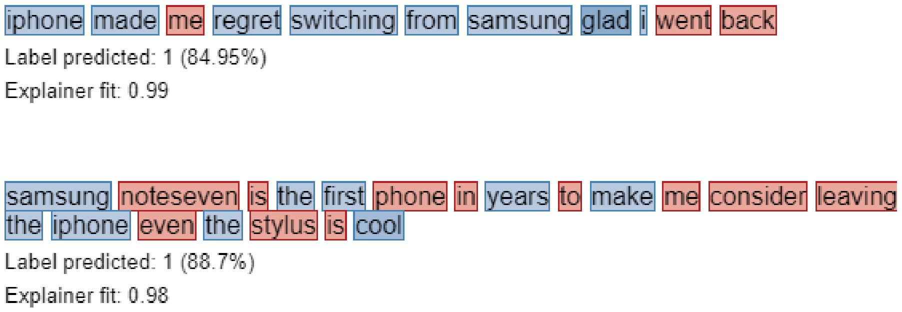
With the tweets about the crisis becoming more prevalent and the best performing classification method, the effects of the features were studied on tweet level. LIME was used together with the self-training method to examine hypothesis 3:

H3: The sentiment of tweets is negatively affected by a company crisis.

Across all tweet examples, the brands Samsung and iPhone had feature weights that supported the positive label prediction probability. However, the products iPhone 6 and Samsung Note 7, given by 'iphone six' and 'noteseven' , contradicted the positive label prediction probability. The iPhone crisis topic was related to the words 'battery', 'life died' and 'randomly drains faster', all of these supported the negative label prediction. The same goes for the Samsung crisis topic, the features 'Noteseven', 'exploding' and ‘samsunggalaxynoteseven’ supported the negative label classification. So the brand names themselves were not necessarily negatively affected by the company crisis but the products of those brands were. To summarize conclusions for hypothesis 3, the company crisis increased the prevalence of tweets about the crisis and the features related to the crisis did affect sentiment negatively. To answer the research question, the most important features that mostly determined tweet sentiment were related to the crisis. These features did affect the tweet sentiment negatively.

6.2 Limitation and further studies

The results of this study are of limited scope due to the following factors. Firstly, negations can easily affect the polarity of a sentence (Farooq, 2017; Liu, 2012). Negation handling was not taken into account when constructing the document-term-matrix. The second complex task that this study lacks is to solve coreference resolution in comparative sentences. Comparative opinions are sentences in which the opinion is given by a comparison between two or more entities, like the ones in Figure 16 (Jindal & Liu, 2006). The challenge in NLP is to find out what expressions belong to what entity (Aggarwal & Zhai, 2013). Lastly, some words can have different meaning when used in different contexts. This issue, known as word sense disambiguation, makes it difficult for machines to determine the underlying meaning of words (Navigli, 2009). Again, the context of words was not taken into account when constructing the document-term-matrix. Also, tweets are not a good reflection of the overall population sentiment (Sehl, 2020). Twitter users are often younger and more affluent people compared to the population average and the number of tweets gathered for this study is also only a small fraction of the total number of tweets that are posted towards the brands. (Sehl, 2020).



iphone made me regret switching from samsung glad i went back

Label predicted: 1 (84.95%)

Explainer fit: 0.99

samsung noteseven is the first phone in years to make me consider leaving the iphone even the stylus is cool

Label predicted: 1 (88.7%)

Explainer fit: 0.98

Figure 16. LIME interpretations of comparative sentences in the filtered unlabelled data

In the sentiment prediction section of this study, the data of Go et al. (2009) were combined with 69 manually labelled corporate crises tweets and used for training the supervised learning algorithms. The labels of Go et al. (2009) are, however, not observed labels since they were predicted by their distant supervision algorithm. This algorithm does not predict sentiment perfectly, which results in measurement errors in the labels in the training set. Subsequently, the prediction performance on the labelled instances is not entirely reliable (Osborne & Waters, 2003). The STM regressions and the conclusions that followed from these are also limited in a few ways. Firstly, the response variable topic prevalence is an result from a prior analysis and is thus not observed, which also violates the assumption that the response variable is measured without measurement error (Osborne & Waters, 2003; Poole & O'Farrell, 1971). Secondly, there are regressions for each unique topic that is found by STM. The hypotheses of these regressions are tested simultaneously and that leads to an increased risk of the multiple testing problem (Schochet, 2009). The multiple testing problem is the increased probability of finding a significant result in one of the regressions while this conclusion can be unfounded (Schochet, 2009). It decreases the reliability of the significance of the coefficients in the regressions. Besides the multiple testing problem, this many regressions makes it more likely that the assumption of independent conditional distributions of the response variable is violated (Poole & O'Farrell, 1971).

This study thus has room for improvement that can be addressed by future studies. Firstly, the challenges of NLP like negation handling, coreference resolution and word sense disambiguation can be addressed. This could be done by part-of-speech taggers, N-grams or more complex machine learning techniques like neural networks with word embeddings. Secondly, more tweets could be scraped and manually labelled to make sure that the labels are observed and not a result from a prior analysis. Thirdly, future research can focus on making sure that the regressions that follow from the STM model conducted in this do not violate regression assumptions. This would make the conclusions that followed from the regression more reliable. In the marketing context, it would be advisable for future studies to use NLP techniques on tweets posted after a brand has responded to the crisis. The crisis itself is often an unforeseen event, it is therefore important to know which response shows the most effect in restoring Twitter sentiment towards the brand.

7 References

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. In *Mining Text Data* (Vol. 9781461432). https://doi.org/10.1007/978-1-4614-3223-4

Ahluwalia, R., Burnkrant, R. E., & Unnava, H. R. (2000). Consumer response to negative publicity: The moderating role of commitment. *Journal of Marketing Research*, *37*(2), 203–214. https://doi.org/10.1509/jmkr.37.2.203.18734

Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment Analysis of Tweets using SVM. *International Journal of Computer Applications*, *177*(5), 25–29. https://doi.org/10.5120/ijca2017915758

Andreassen, T. W., & Lindestad, B. (1998). Customer loyalty and complex services. The impact of corporate image on quality, customer satisfaction and loyalty for customers with varying degrees of service expertise. In *International Journal of Service Industry Management* (Vol. 9, Issue 1). https://doi.org/10.1108/09564239810199923

Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. In *Journal of marketing Research* (Vol. 4, Issue 3, pp. 291–295). https://doi.org/10.4324/9781003125518-4

Bambauer-Sachse, S., & Mangold, S. (2011). Brand equity dilution through negative online word-of-mouth communication. *Journal of Retailing and Consumer Services*, *18*(1), 38–45. https://doi.org/10.1016/j.jretconser.2010.09.003

Barnett, M. L., Jermier, J. M., & Lafferty, B. A. (2006). Corporate Reputation: The Definitional Landscape. *Corporate Reputation Review*, *9*(1), 26–38. https://doi.org/10.1057/palgrave.crr.1550012

Bickart, B., & Schindler, R. M. (2001). Internet Forums As Influential Sources of Consumer Information. *Journal of Interactive Marketing*, *15*(3), 31–40.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. *3*, 993–1022.

Boser, B. E., Vapnik, V. N., & Guyon, I. M. (1992). Training Algorithm Margin for Optimal Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.3390/risks8030083

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1201/9780429469275-8

Brown, J., & Reingen, P. (1987). Social Ties and Word-of -Mouth Referral Behavior. *Journal of Consumer Research*, *14*(3), 350–362. https://www.jstor.org/stable/pdf/2489496.pdf?casa_token=6rsyxHprOBsAAAAA:bbH3jKk8oRixr6GmiClfpQVDuLukzBXzf0KDk-4ZT26DAm7PbPBXrqVX6ub0I1wy6Mw9pe3DVn__AxxTJX2NkHrArUTdulqjNCI-8MCIRceV3KPFhwQ

Burnkrant, R. E., & Cousineau, A. (1975). Informational and Normative Social Influence in Buyer Behavior. *Journal of Consumer Research*, *2*(3), 206. https://doi.org/10.1086/208633

Chiou, J. S., & Cheng, C. (2003). Should a company have message boards on its web sites? *Journal of Interactive Marketing*, *17*(3), 50–61. https://doi.org/10.1002/dir.10059

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and*

Technology, *37*(1), 51–89. https://doi.org/10.1016/0004-3702(82)90032-7

Chun, R. (2005). Corporate reputation: Meaning and measurement. *International Journal of Management Reviews*, *7*(2), 91–109. https://doi.org/10.1111/j.1468-2370.2005.00109.x

Cleeren, K., Van Heerde, H. J., & Dekimpe, M. G. (2013). Rising from the ashes: How brands and categories can overcome product-harm crises. *Journal of Marketing*, *77*(2), 58–77. https://doi.org/10.1509/jm.10.0414

Dean, D. H. (2004). Consumer reaction to negative publicity: Effects of corporate reputation, response, and responsibility for a crisis event. *Journal of Business Communication*, *41*(2), 192–211. https://doi.org/10.1177/0021943603261748

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets Using Naïve Bayes‘ and K-NN Classifier. *International Journal of Information Engineering and Electronic Business*, *8*(4), 54–62. https://doi.org/10.5815/ijieeb.2016.04.07

Dijkmans, C., Kerkhof, P., & Beukeboom, C. J. (2015). A stage to engage: Social media use and corporate reputation. *Tourism Management*, *47*, 58–67. https://doi.org/10.1016/j.tourman.2014.09.005

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. In *CRC press*.

Eyheramendy, S., Lewis, D. D., & Madigan, D. (2003). On the naive bayes model for text categorization. *International Workshop on Artificial Intelligence and Statistics*, 93–100.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0015-2

Folkes, V. S. (1984). Consumer Reactions to Product Failure: An Attributional Approach. *Journal of Consumer Research*, *91*(2010), 291–310. https://academic.oup.com/jcr/article-abstract/10/4/398/1822424?redirectedFrom=fulltext

Fombrun, C., & Van Riel, C. (1997). The Reputational Landscape. *Corporate Reputation Review*, *1*, 5–14.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. *CS224N Project Report, Stanford*, 1–6.

González, M., Rosado, O., Rodríguez, J. D., Bergmeir, C., Triguero, I., & Benítez, J. M. (2009). *ssc : An R Package for Semi-Supervised Classification*. 1–13.

Gruen, T. W., Osmonbekov, T., & Czaplewski, A. J. (2006). eWOM: The impact of customer-to-customer online know-how exchange on customer value and loyalty. *Journal of Business Research*, *59*(4), 449–456. https://doi.org/10.1016/j.jbusres.2005.10.004

Gupta, V., Science, L. C., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52. https://doi.org/10.1002/dir.10073

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of Word-of-Mouth and Product-Attribute Information on Persuasion: An Accessibility-Diagnosticity Perspective. *Journal of Consumer Research*, *17*(4), 454. https://doi.org/10.1086/208570

Hudson, S., Huang, L., Roth, M. S., & Madden, T. J. (2016). The influence of social media interactions on consumer-brand relationships: A three-country study of brand perceptions and marketing behaviors. *International Journal of Research in Marketing*, *33*(1), 27–41. https://doi.org/10.1016/j.ijresmar.2015.06.004

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter Power:Tweets as ElectronicWord of Mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169–2188. https://doi.org/10.1002/asi

Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J. & Nithya, M. (2016). *Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining*. *5*(October 2014), 7–16.

Kelley, H. H., & Michela, J. L. (1980). Attribution Theory and Research. *Annual Review of Psychology*, *31*(1), 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, *3339*, 488–499. https://doi.org/10.1007/978-3-540-30549-1_43

Kim, E., Sung, Y., & Kang, H. (2014). Brand followers' retweeting behavior on Twitter: How brand relationships influence brand electronic word-of-mouth. *Computers in Human Behavior*, *37*, 18–25. https://doi.org/10.1016/j.chb.2014.04.020

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24. https://doi.org/10.1007/s10751-016-1232-6

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1). https://doi.org/10.1016/0378-1097(92)90668-E

Laczniak, R. N., DeCarlo, T. E., & Ramaswami, S. N. (2001). Consumers' Responses to Negative Word-of-Mouth Communication: An Attribution Theory Perspective. *Journal of Consumer Psychology*, *11*(1), 57–73. https://doi.org/10.1207/15327660152054049

Liddy, E. D. (2001). *Natural language processing.* https://doi.org/10.1016/0004-3702(82)90032-7

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, *29*(3), 458–468. https://doi.org/10.1016/j.tourman.2007.05.011

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on Learning for Text Categorization*, *752*(1), 41–48. https://doi.org/10.1002/em.22125

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mitchell, T. M. (1997). *Machine Learning*.

Mizerski, R. W. (1982). An Attribution Explanation of the Disproportionate Influence of Unfavorable Information. *Journal of Consumer Research*, *9*(3), 301. https://doi.org/10.1086/208925

Nguyen, N., & Leblanc, G. (2001). Corporate image and corporate reputation in customers' retention decisions in services. *Journal of Retailing and Consumer Services*, *8*(4), 227–236. https://doi.org/10.1080/03031853.1990.9524161

Osborne, J. W., & Waters, E. (2003). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, *8*(2), 2002–2003.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using Machine Learning Techniques*. https://doi.org/10.3115/1118693.1118704

Perez-Carballo, J., & Strzalkowski, T. (2000). Natural language information retrieval: Progress report. *Information Processing and Management*, *36*(1), 155–178. https://doi.org/10.1016/S0306-4573(99)00049-7

Poole, M. A., & O'Farrell, P. N. (1971). The Assumptions of the Linear Regression Model. *Transactions of the Institute of British Geographers*, *52*(52), 145. https://doi.org/10.2307/621706

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/bf00116251

Read, J. (2005). Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. *Proceedings of the ACL Student Research Workshop*, 43–48. https://doi.org/10.1002/9781118347379.ch11

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46. https://doi.org/10.1039/b104835j

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rogerson, W. P. (1983). Reputation and Product Quality Author. *The Bell Journal of Economics*, *14*(2), 508–516.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, *21*(3), 660–674. https://doi.org/10.1007/978-3-540-71441-5_104

Schochet, P. Z. (2009). An approach for addressing the multiple testing problem in social policy impact evaluations. *Evaluation Review*, *33*(6), 539–567. https://doi.org/10.1177/0193841X09350590

Schölkopf, B. (2001). The kernel trick for distances. *Advances in Neural Information Processing Systems*, *May*.

Seeger, M. W., Sellnow, T. L., & Ulmer, R. R. (1998). Communication, Organization, and Crisis. *Annals of the International Communication Association*, *21*(1), 231–276. https://doi.org/10.1080/23808985.1998.11678952

Shapiro, C. (1982). Consumer Information , Product Quality , and Seller Reputation Author. *The Bell Journal of Economics*, *13*(1), 20–35.

Smith, K. T., Smith, M., & Wang, K. (2010). Does brand management of corporate reputation translate into higher market value? *Journal of Strategic Marketing*, *18*(3), 201–221. https://doi.org/10.1080/09652540903537030

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, *2*(Nov), 45–66.

https://doi.org/10.1353/aq.0.0077

Wartick, S. L. (2002). Measuring Corporate Reputation: Definition and Data. *Business & Society*, *41*(4), 371–392. https://doi.org/10.1177/0007650302238774

Webster, J., & Kit, C. (1992). Tokenization as the initial phase in NLP. *The 15th International Conference on Computational Linguistics*, *4*, 6–10.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A Biterm Topic Model for Short Texts. *Proceedings of the 22nd International Conference on World Wide Web.*, 1445–1456.

Yarowsky, D. (1995). *Unsupervised word sense disambiguation rivaling supervised methods*. 189–196. https://doi.org/10.3115/981658.981684

Zhang, Y., & Pennacchiotti, M. (2013). Predicting Purchase Behaviors From Social Media. *Proceedings of the 22nd International Conference on World Wide Web.*, 1521–1531.

Zhou, Y., & Goldman, S. (2004). Democratic Co-Learning. *16th IEEE International Conference on Tools with Artificial Intelligence*, 594–602.

Zhou, Z. H., & Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, *17*(11), 1529–1541. https://doi.org/10.1109/TKDE.2005.186

Zhu, X. (2005). *Semi-supervised learning literature survey.*

8 Appendix

8.1 Appendix A

The following pre-processing steps are applied on the tweets besides the ones explained in section 3.3. First the cleaning steps of Table 13 are used. Then each year that is written in numbers is replaced by ‘year’, any number is hereafter removed.

Originally	New (pre-processed)
Iphone 3	iphone three
Iphone 4	iphone three
Iphone 5	iphone three
Iphone 7	iphone three
Note 8 or Note8	noteeight
S3 or galaxy S3	gsseven
S4 or galaxy S4	gsfour
S5 or galaxy S5	gsfive
S6 or galaxy S6	gssix
S7 or galaxy S7	gsseven
S8 or galaxy S8	gseight
:) or :-)	happysmile
:(or :-(	sadsmile

Table 13. Intuitive graphical representation of LDA.

8.2 Appendix B

Blei et al. (2003) start explaining their model by showing how an LDA model, that has already learned the corpus, can generate documents. Figure 17 visualizes the document very intuitively. In Figure 17, the triangle contains the topic probabilities (θ) and the octagons contain the word probabilities (β), these probabilities are given by the number of balls. The chosen topic from the triangle leads to one of the octagons that contains the probabilities over all words for that specific topic (β). The topic for the first word could be from topic ‘Animals’. This leads to the middle octagon, where topic probabilities $\beta_{Animals}$ are shown. The first word in the document, which is Cats in Figure 17, could be any of the words with a probability larger than zero. This process can be iterated for N number of words to create a document from a LDA model that has learned the corpus.

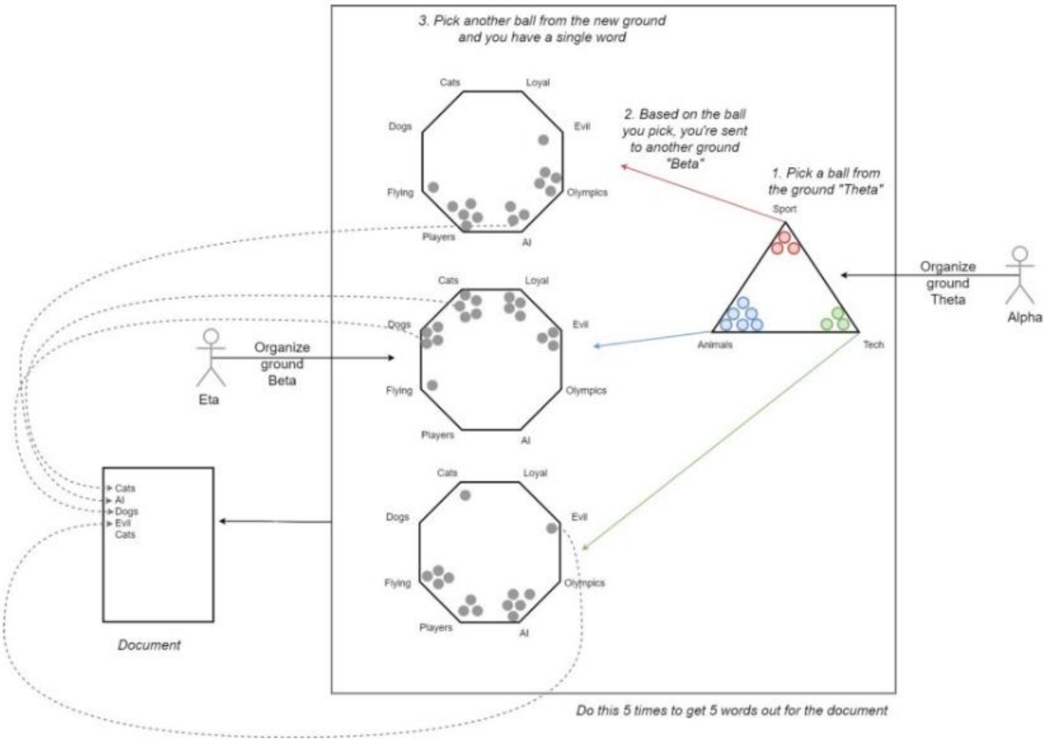

3. Pick another ball from the new ground and you have a single word
Cats
Loyal
Dogs
Evil
Flying
Olympics
Players
AI
2. Based on the ball you pick, you're sent to another ground "Beta"
1. Pick a ball from the ground "Theta"
Sport
Animals
Tech
Organize ground Theta
Alpha
Cats
Loyal
Dogs
Evil
Flying
Olympics
Players
AI
Organize ground Beta
Eta
Cats
Loyal
Dogs
Evil
Flying
Olympics
Players
AI
Cats
AI
Dogs
Evil
Cats
Document
Do this 5 times to get 5 words out for the document

Figure 17. Intuitive graphical representation of LDA.