SWAnalytics: Towards Insight from Social Media and Finance

Bogdan Ene
Vrije Universiteit
b.g.ene@student.vu.nl

Dusko Trajkov
Vrije Universiteit
d.trajkov@student.vu.nl

Gloria Boekhouder
Vrije Universiteit
g.boekhouder@student.vu.nl

Nicola Pieruzzini
Università di Pisa
n.pieruzzini@studenti.unipi.it

Thodoris Zois
Vrije Universiteit
t.zois@student.vu.nl

ABSTRACT

The launch of a product, a new marketing campaign, a merge or acquisition, are just some examples of brand-related events that can have a great impact on the brand itself. The analysis of those events can provide valuable insights to the company, highlighting the opportunities and the possible threats. *SWAnalytics* is a tool that performs an analysis of the events, from both sides of consumers' opinions and stock prices. It aims to help organizations or investors to understand any correlation between what people think and how the market behaves. To achieve that, *SWAnalytics* combines two of the most popular data analysis techniques, providing the distribution of the sentiments and an extensive analysis of the most relevant topics. To evaluate the impact of our work, we take as a use-case three of the most shocking events related to the stock price of *Tesla Motors* and we deeply analyze the results by the agency of *SWAnalytics*.

1 INTRODUCTION

The field of analytics is becoming more and more important for business and economics over the last few years. All the business processes are constantly monitored by means of data analysis and an increasing number of tools are developed each day for supporting the decisions of companies, investors, and every other stakeholder in many different situations.

Our work is placed into the business intelligence area, in particular in the field of *a posteriori* analysis. After a company faces an important event, like the launch of a new product, many decisions have to be taken in order to make the most of the opportunities that it brings with itself, but also to avoid the possible threats. In this context, *SWAnalytics* is born from the need to answer the following question: "*What is the impact that a certain brand-related event has on the brand itself in terms of consumers' opinions and what is the relationship between these opinions and the company's stock?*". *SWAnalytics* is a tool that determines the impact of the events from two complementary sides. From the financial point of view, stock price trends are considered and in particular, their relationship with public opinion. From the marketing side, sentiment and topic analysis are performed on the data in order to understand what are the feelings of the people around the company. This is of high interest 1) to the company itself that can have an overview of the impact of the event and 2) to other stakeholders like investors, which might decide to invest in a firm taking into account its last actions and the customers' sentiments regarding the company.

To provide useful insights, *SWAnalytics* combines information retrieved from Twitter and Yahoo Finance. To retrieve the relevant data it requires the name of a stock price and a list of events. For each event, an interested party has to provide: a title, a date range and a number of hashtags to be searched on Twitter. After the tool performs all the necessary operations it generates a web application with one HTML page per event. On each page, there are four different sections. The first section is related to aggregated statistics, the second describes the correlation between the stock price and users' sentiments. The third one is related to the global sentiment while the last one provides an overview of the most relevant topics and their composition.

One of the most shorted stocks last fiscal year on the market was Tesla. At the end of January its short interest had reached 18% of its total available shares to trade[1]. Controversial to this high volume of short positions, Tesla's stock has been rising ever since its lowest stock price since 2014, on the 3rd of June 2019 with a close of $178.97. Before this time period Tesla had followed a sharp downward trend. Some of it could be explained by Elon Musk's bad business attitude of announcing over-optimistic false information on tweeter to take Tesla private or smoking weed on the Joe Rogan show[2,3]. In addition financial reports (quarterly reports to be more exact) during different time periods and the loss of key personnel have more strongly influenced the movement of the stock price in 2019[4]. At the beginning of 2020 we see a financial chart of the company that follows an increasing parabolic trend ever since the 3rd of June 2019[5]. With controversial topics like the release and presentation of Tesla's Cybertruck, during this period Tesla has had to face many blunders and experienced great success on the stock market[6,7]. To evaluate the behavior of *SWAnalytics* to process multiple events we decide to take as a use case the three most shocking related to the stock price of *Tesla Motors* events for a period of 3 days. However, in order to evaluate the effectiveness of our tool in providing assistance to companies and investors, we elaborately discuss the results related only to the first one for clarity reasons. The events that we find interesting for the stock price of *Tesla Motors* are:

[1] https://markets.businessinsider.com/news/stocks/tesla-stock-most-shorted-companies-us-traders-betting-against-apple-2020-2-1028873641#2-apple
[2] https://www.theverge.com/tldr/2019/8/7/20758944/elon-musk-twitter-tesla-funding-secured-private-420
[3] https://edition.cnn.com/2018/10/01/tech/elon-musk-joe-rogan/index.html
[4] https://www.cnbc.com/2019/07/25/tesla-is-having-worst-day-of-2019-after-earnings-and-loss-of-cto.html
[5] https://www.cnbc.com/quotes/?symbol=TSLA
[6] https://www.businessinsider.nl/heres-everything-that-happened-when-elon-musk-unveiled-the-cybertruck-2019-11?international=true&r=US
[7] https://www.forbes.com/sites/carltonreid/2019/12/16/tesla-cybertruck-not-street-legal-in-eu/

- Tesla suffers its worst day of the year after brutal earnings report and loss of CTO.
- Tesla unveils its new Cybertruck.
- Tesla reaches its highest stock price till date of $962.86.

The shock associated with the first two events is that the company stock price did not decline much during the duration of the event but actually rose a bit in comparison to its opening. One would expect the opposite since the Cybertruck's presentation did not go as planned and the demotion of a CTO always brings some uncertainty with it. The last event is also interesting because a decline right after a global maximum is reached gives more insight into what topics/reasons made investors sell their shares. We did not choose the lowest stock price as our event, because we needed something more clear by which the event can be categorized a priori (e.g. loss of CTO), so we can possibly validate the accuracy of the analysis by confirming the presence of this categorization in the topics distribution. Moreover, by taking at least three events into account we can see what the optimal number of topics is for every event and compare this with the others. This gives users of the app more insight into how discussion intensive the event was and whether this differs between a bull (rising) and bear (declining) times. Next to this, the events' sentiment can be plotted through time to see its relationship with the financial data. Nevertheless, It is worth to mention once more that this is just a use case. *SWAnalytics* is a fully automated tool that can work with any appropriate input and provide results accordingly.

The rest of this work is organized as follows. In Section 2 we provide the motivation behind our work. Section 3 demonstrates *SWAnalytics*, it starts with an overview of the application and progressively dives deeper to the inner workings. There we are also evaluating also the effectiveness of the tool, and we list the possible limitations. In Sections 4 to 10 we discuss some of the concepts related with the Social Web and we aim to correlate our work with the lectures of the course. More specifically, we discuss the following:

- Section 4: What is the Social Web and Social Computing?
- Section 5: What does data look line on the Social Web?
- Section 6: What do people do in the Social Web?
- Section 7: How do people mine, analyze and visualize the Social Web?
- Section 8: Personalization on the Social Web
- Section 9: How can we study the Social Web?
- Section 10: What are the challenges for the Social Web?

We conclude in Section 11, and we propose some ideas for further development or improvements in Section 12. In Section 13, each one of the authors provide a section with their personal contribution to make *SWAnalytics* come true. Finally, in the Appendix we provide some more work that has been done for the topic analysis part. Specifically, in Section A.1 we provide literature related to the multiple variants of LDA, while in Section A.2, we demonstrate a comparison of the LDA we use in topic analysis with the algorithm of HDP.

2 MOTIVATION

In terms of value creation for stakeholders, sentiment and topic analysis have been at the forefront of opportunities in business[8] and naturally this sparked our interest in the project as well. There is a high utility in using both of these perspectives to tackle business questions, however, due to our time restriction we'd like to focus only on the financial side of the business application. The reason for doing so is the availability of financial reports. These are much easier to obtain than business day to day operational data. Initially, we would like to study the analyses mentioned in combination with stock data, because that gives the best general representation of a firm's financial state. Hereby it is important to note that the analyses are not restricted to only this side of a business' point of view, but they can be further extended with marketing/customer data, liquidity of a firm through time or daily operations data. Our application will enable all users to get a better understanding of a firm's performance through the combination of stock data and textual information from Twitter. This can help the firm itself to reflect on what its investors found troubling or what was actually spot on during a particular period without the need for an additional survey study. On the other hand, stakeholders of the firm are granted a more valid source of information that is resistant to confirmation bias, since it analyzes multiple conversations on Twitter and not only the ones that the stakeholders follow. Sentiment Analysis is a mixture of Natural Language Processing and Information Extraction techniques and its goal is to obtain writers' feelings expressed in a positive or negative comment, questions, and requests, by exploring or analyzing a large number of documents [11]. There are three types of sentiment analysis [3]: (1) *Document-level* classification, whereby the focus is on finding out whether emotions from an individual person are positive or negative; (2) *Sentence-level* classification, in which the sentiments in sentences are being studied in terms of subjective or objective nature, and (3) *Aspect-level* classification, where different aspects of emotions or opinions by different individuals are being studied. The sentiments can be given a score based on their degree of positivity, negativity or objectivity [11].

In the last decade, people used to do sentiment analysis via news and different type of sources. Nowadays, companies are using Twitter to determine sentiment because it has been proven to have influence on investment decisions. Multiple studies have shown that not only historical financial data of the stock market can predict the returns of a stock market, but sentiments and emotions of people can also help in predicting it [3].

Dattu & Gore [11] mentioned benefits like how Twitter contains an enormous number of text posts and that its audience is from different countries and of different kinds: regular users to celebrities, company representatives, politicians, and even country presidents. Therefore, it is possible to collect text posts of users from different social and interest groups. However, tweets that are made by important individuals can affect people and their emotions. Positive news affects people positively resulting in investing more in a firm's stocks. On the other hand, negative news leads to negative reactions in the firm's stocks [3]. It has been shown that there is a

[8] https://insidebigdata.com/2015/06/05/text-analytics-the-next-generation-of-big-data/

strong correlation between sentiments and stock prices. A study conducted by Bollen, Mao, Zeng [7] has also confirmed that Twitter moods can predict the stock market. Stock market prices are driven by new information, such as the news rather than present and past prices. They state how emotions of others have a really big effect on different people and play important significant roles in human decision-making. These researches of behavioral finance provided more proof that financial decisions are significantly driven by motion and mood. Thus, it can be assumed with this info that public mood can drive stock market values.

Another previous study concluded that the public mood collected from Twitter may be correlated with the Dow Jones Industrial Average Index (DJIA). DJIA is an index that tracks the number of large companies trading on a stock exchange. It confirmed that positive news and sentiment in tweets about a company do encourage people to invest in their stocks which in turn increases the stock price of that company. There is, for example, a strong correlation between the rise and falls in stock prices and sentiments [26]. Research by Bing [31], used a Pearson Correlation Coefficient for predicting stocks and found out that there was a high negative correlation between mood states such as hope, fear, and worry in tweets with Dow Jones Average Index. Yu et al. [3] argued that contextual distributions of words are significant in determining negative or positive aspects. They also mention that the intensity of news is also significant on the stock market returns.

The reason why changes in sentiment analysis show a correlation with stock prices is that stock markets have a random walk pattern and are easily influenced by opinions and ideas by humans, events, and news. Also, sentiment analysis has a quicker effect on short term price fluctuations [26]. Another research discovered that the best results are present when Twitter data predict the market data for 3 days [9]. A study conducted by Ahuja et al. [2] compared data with Twitter news and financial data stock returns. They concluded that there was a significant relationship between the sentiments of people because they could see that the motions of people were influenced by different news annulments.

So, most studies have confirmed a correlation between sentiment analysis and changes in the stock market and that not only it is important to use statistical measures related to financial data, but also to use emotions of people to predict the stock market. They believe that if people are happy, they will likely invest more in the stock market than when they are not happy. In contrast, Schumaker et al. [30] showed a completely different outcome when conducting similar research about sentiment analysis and stock prices; people were selling stocks when there was good news and buying them when there was bad news. This does not support what most researches claim about sentiment analysis and the stock market. However, the difference in this study is that the source of data was not Twitter, but news on television.

Apart from sentiment analysis, applying other techniques to retrieve information from textual sources can be very valuable. Topic analysis is a Natural Language Processing (NLP) technique that allows us to automatically extract meaning from texts identifying recurrent themes or topics. Exploring the topics discussed in a corpus of tweets adds another layer of understanding what drives the sentiment of the Twitter users and what objects are associated with a certain sentiment. To illustrate, knowing that a certain analyzed event is mostly categorized with a negative sentiment, does not necessarily mean that it was directed to the negative expectation of a company's financial situation [30]. With topic analysis, one can measure what topics (e.g. product release, Loss of stakeholders confidence in the CEO, change in corporate structure) are most likely to be associated with the assigned sentiment.

3 SWANALYTICS

SWAnalytics is the tool that we develop in order to answer the following question: "*What is the impact that a certain brand-related event has on the brand itself in terms of consumers' opinions and what is the relationship between these latter and the company's stock?*".

3.1 Overview

Figure 1 demonstrates the overall design of SWAnalytics and the specific modules that are cooperating in order to provide the final outcome. In order the application to work, the user has to provide a certain input in the form of a JSON file (events.json). More specifically, the application expects the stock name, and a list of events where each one has the following attributes: title, start_date, end_date and a list of hashtags. The application starts by fetching all the necessary data, and continues with the execution of the "Data pipeline" where the results are stored permanently on the disk. The storage of the data triggers the next execution round; the "Web-app pipeline". The modules in this pipeline are responsible for generating the different graph visualizations and the various HTML pages (one for each event). Finally, it is worth mentioning that each module regardless the pipeline, produces logs that aim to inform the user about the different steps of the whole process.

After the execution of both pipelines the web application is up and running. Each page is separated into four sections. The first section "Overall Statistics", shows the input of the user for that particular event and some information related to the data cleaning process such as: the number of tweets being fetched and the number of those that are finally analyzed in the "Data pipeline" in order to produce the graphs in the next sections of the application. The second section "Correlation", contains the graphs that demonstrate the correlation of the sentiment per day in comparison with the stock price for that day of the event. The third section "Global Sentiment", exposes the overall sentiment of the tweets not only for the whole world, but also per continent. Finally, the fourth section "Most Relevant Topics", provides the necessary graphs for the user to decide on the most relevant topics related with the event and conclude for their impact.

In the following sections we dive deeper into the most interesting part of SWAnalytics, the process of data gathering and the modules related to the data pipeline. The "Web-app pipeline" is a pure engineering solution that generates the HTML content based on the processed data.

3.2 Data gathering

SWAnalytics gathers data related to (1) stock prices and (2) users' tweets, in the period of time specified by the user for the events considered. The stock price data come from *Yahoo! Finance*. The website provides the stock values of the chosen company in the desired range of dates. As the daily value, SWAnalytics considers

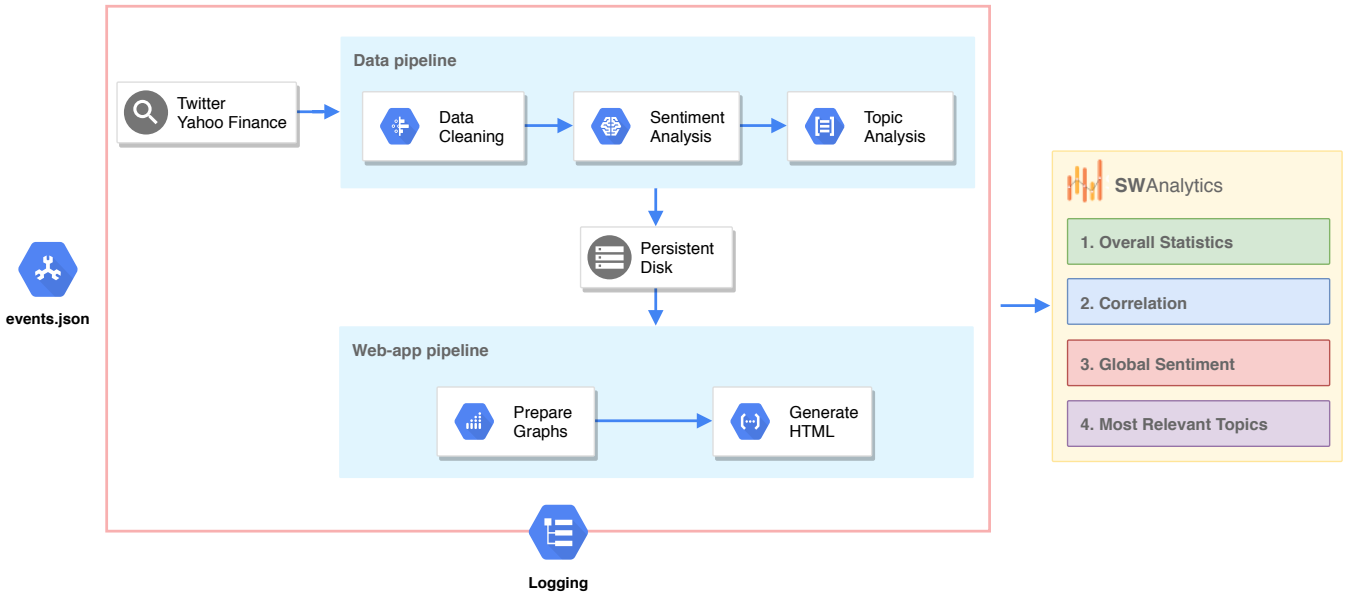


Figure 1: SWAnalytics workflow.

the closing price of the day. Tweets are gathered by the agency of Twitter API which provides SWAnalytics the necessary functionality to fetch data based on hashtags, a period of time and the language. In our work we consider only tweets written in English.

3.3 Data pre-processing

Once the data are gathered, SWAnalytics applies a set of certain filter in order to discard and clean the data. We apply those filters in the following order:

- **Column filter**: Twitter API returns many attributes along with each tweet. However, SWAnalytics is only interested in the full-text of a tweet, the date that it has been created, and the location of the user.
- **Row filter**: SWAnalytics removes all the tweets where the location of the user is (1) empty or (2) unknown. Since the location is a Twitter profile is a free-text there might be users that do not provide a certain location but instead a random string. Finally, SWAnalytics discards all the duplicate tweets made by the same user for each day of the event.

Before proceeding any further to the data pipeline and send the tweets to the topic analysis module, first we have to perform certain steps related that are vital for natural language processing. This allows for a more accurate separation between topics and better understanding by the user when the topic contents are being inspected. The elements of this process are as follows:

(1) **Tokenization**: the corpus strings are split down into a set of words. This part also removes punctuation from the text and any URL or emoji links in the text. The main algorithm used here is called "Tweet Tokenizer"[9]. Afterward, the hashtags are removed and words stuck together are split using an imported English dictionary of words.

(2) **Stopwords removal**: frequent occurring words (e.g. the, a, in) that bring little to no semantic variation to the topic analysis are removed from the data.

(3) **Spelling correction**: some words that are misspelled can be corrected using a pre-defined dictionary mapping that can be imported using known NLP packages (e.g. NLTK, textblob). After testing a number of them, "TextBlob" returned the best results, but it is not without any errors. Some words are misspelled so much that any algorithm would return the wrong match. In other cases, the spelling algorithm fails to recognize words that have meaning but are outside of its own dictionary (e.g. "amzn" is a stock tracker of the company Amazon. The algorithm would return "amen"). We cannot do anything about the former without building our own custom spelling corrector, but we could solve the latter problem by manually adding to the spelling corrector some of the most occurring words in the tokenized corpus that are excluded (seen as misspelled) in "TextBlob's" vocabulary.

(4) **Lemmatization**: using a lexical knowledge base (e.g. WordNet) the analysis obtains the base forms of all words, which ensures that words that represent the same object/thing (e.g. computer - computers) are treated as one entity. This helps with processing large data sets due to a shorter vocabulary and it can also help with precision since the frequency is being aggregated to the base. One drawback of using this method together with TextBlob is that some words can indicate another meaning after lemmatization (e.g. kidding becomes kid). However, not using this option will result in more vague and uninterpretable topics due to the many spelling mistakes present in the tokenized tweets.

(5) **Case normalization**: the tools used are sensitive to lower and upper case differences, hence the analysis must normalize these to avoid unnecessary bias.

[9] https://www.nltk.org/api/nltk.tokenize.html

3.4 Sentiment analysis

There are many algorithms and techniques that can be used for sentiment analysis on social media. One of the most popular tools is Valence Aware Dictionary and sEntiment Reasoner (VADER). VADER is a lexicon-based approach that uses a mixture of lexical characteristics usually labeled as either positive or negative according to their semantic orientation [18]. In addition, multiple studies have proven that VADER is a very effective tool for analyzing social media text [18, 13]. VADER does not only portray the positivity, neutrality or negativity score, but it also gives information about how positive, neutral or negative a sentiment is. Finally, there is no need to train a model using labeled data, because VADER relies on a dictionary to assess the sentiment of sentences. SWAnalytics feeds the full-text of each tweet to VADER which in turn, returns the corresponding results for that specific tweet. Afterwards, SWAnalytics classifies the final sentiment of a tweet based on the maximum percentage that VADER returned. For instance, if a tweet X is 80% positive, 20% neutral and 0% negative, then it is classified as positive. After all the tweets have been classified, the positivity degree for each day is computed using the following formula:

$$\begin{aligned} positivity = 1 \times (\%\ of\ positive\ tweets) \\ + 0.5 \times (\%\ of\ neutral\ tweets) \\ + 0 \times (\%\ of\ negative\ tweets) \end{aligned} \tag{1}$$

The values obtained this way are normal; if the tweets of the day are all positive the index equals 1 (maximum value); conversely, if the sentiments are entirely negative, the formula returns 0 (minimum value). All the other cases are in the range (0,1), with 0.5 being the value that represents full neutrality.

3.5 Topic analysis

Multiple methods have been proposed to represent, mine and finally retrieve information from text data: *boolean* models representing documents as sets of words and sentences, *algebraic* models that transform documents to vectors, e.g. Latent Semantic Analysis (LSI) [12] and *probabilistic* models that represent documents as probabilities of words, e.g. Latent Dirichlet allocation (LDA) [6]. LDA, in particular, has been used extensively in topic analysis to uncover latent topics that best represent what is discussed in a pool of text [33, 20]. SWAnalytics incorporates only vanilla LDA, although a hierarchical version of it called the Hierarchical Dirichlet Process [34] is also tested against the results of LDA. In Appendix A.1 a short summary is conducted on a number of variants for LDA and the most novel approach in the field of topic analysis. LDA takes on a corpus of a certain size and extracts a representation of these documents in terms of latent topics. The latent topics themselves which are represented by a multinomial distribution[10] of words are statistically inferred from the documents based on a predefined number of K latent topics and a vocabulary of words.

$$P(\theta_{1:M}, Z_{1:M}, \phi_{1:M} \setminus D; \alpha_{1:M}, \beta_{1:K}) \tag{2}$$

$$\theta^{t+1}, \phi^{t+1} = \arg\max_{\theta,\phi} E_{Z|W,\theta^t,\phi^t}[\log L(\theta, \phi; W, Z)] \tag{3}$$

Figures 2 and 3 demonstrate the LDA process in SWAnalytics, where M is a collection of documents, N represents the frequency

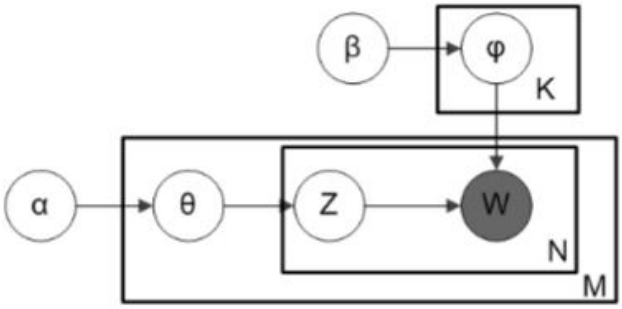


Figure 2: Illustration of the LDA process

W	Words in documents (observed)
Z	Topics for words (not observed)
θ	Distributions of topics in documents (parameters)
ϕ	Distributions of words in topics (parameters)
$L(\theta, \phi; \mathbf{W}, \mathbf{Z}) =$	Log-likelihood of observed data **W** and unobserved
$p(\mathbf{X}, \mathbf{Z} \mid \theta, \phi)$	random variables **Z**, given parameters θ, ϕ

Figure 3: Terms in the LDA process

of word and W and K holds the frequency of topics a document belongs to. Equation 2 shows this process in mathematical terms, where it tries to find the joint posterior probability of parameters θ, ϕ and Z given some distribution-related parameters that govern what the distribution of topics (α) and words (β) in a document and topic respectively looks like. After the initialization of random variable Z, the aim of this algorithm is to find the optimal θ_d for every document d and ϕ_z for every topic z. LDA can do this iteratively using the expectation-maximization method[11] and the log-likelihood of observed data as described in equation 3. This equation is accompanied by two generative processes, one for the parameters α and β that follows the Dirichlet distribution[12] (distribution of distributions) and one for θ and ϕ that follows the multinomial distribution. From here it follows that LDA is linear in the total number of words in the corpus and the number of topics to be generated, with official time complexity of $O(N * K)$ and space complexity of $(N + K(D + W))$ [6, 25].

SWAnalytics then, evaluates the proposed topic models derived from tweets according to:

(1) **Quality of the topics**: The evaluation metric used for the analysis is derived from fairly new research [28] which outperforms earlier metrics and correlates very strongly with human judgments. This score ranges between 0 and 1, where a higher score means a better separation between the topics. The inner quality of the topic is restricted to the most occurring words within the topic itself. This is visualized and calculated using a package called *pyLDAvis* as described by [32]. The most important parameter in ranking the terms within the topics is λ. This weight parameter $\lambda \in [0, 1]$ adds more precision to the topic analysis by adjusting between

[10] http://onlinestatbook.com/2/probability/multinomial.html

[11] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

[12] https://en.wikipedia.org/wiki/Dirichlet_distribution

Impact of Data Cleaning	
	# of Tweets
Fetched	5144
Empty location	1248
Unknown location	762
Duplicate	21
Tweets analyzed: 3113	

Table 1: The impact of the data cleaning process. A total of 5144 tweets are fetched but finally only 3113 are used to generate the results related with the loss of Tesla Motors CTO.

the importance of a term's probability within a topic to its marginal probability across all of the processed tweets in the event (λ set to 0) and the initial topic-specific probability (λ set to 1). The authors of [32] propose an optimal value of 0.6 for λ.

(2) **Speed at which parameter K converges**: For the iterative method (LDA) that does not automatically infer K from the data, simulations are run with a steady increase (step of 3) of topics until the highest CV is reached before flattening out. Afterwards, the time at which this model reaches the highest point is subtracted from the starting time of the simulations in order to get the execution time. In this work, we are more interested in the interpretability of the topics than the maximum separation of the topics. In general, the algorithms will return higher scores of CV for more sparse topics, thus we limit the applications incremental runs to a maximum of 32 topics.

3.6 Evaluation

To evaluate SWAnalytics, we take a closer look to the event: "Tesla suffers its worst day of the year after brutal earnings report and loss of CTO", for which we get data for the period of 24-07-2019 until 26-07-2019 included; 3 days in total. Apart from evaluating SWAnalytics, this section aims to provide some insight for the charts that are generated in the final application. We include everything, except from the world map where the sentiments are distributed per continent, in the "Global Sentiment" section of SWAnalytics.

Table 1 is the most significant part of section "Overall Statistics" and demonstrates the impact of the data cleaning process. As we can observe from the results, a total of 5144 tweets are fetched but finally only 3113 are used in the next graphs. There are 1248 with an empty location, from those 762 have a location that is not known and from the remaining 21 are considered as duplicate.

Figure 4 is part of the "Correlation" section of the application. The purpose is to demonstrate the relationship between the trends of the stock price and public sentiment around that particular event.

From the comparison we can observe that the trend-lines are slightly different, the sentiment is almost a straight line while the stock price follows a downward trend. The correlation coefficient that SWAnalytics reports is -0.97, which indicates a negative relationship between the positivity of the event and the company's stock price. We can clearly notice that the outcomes of the graph and of the correlation index do not match. The trendlines of the

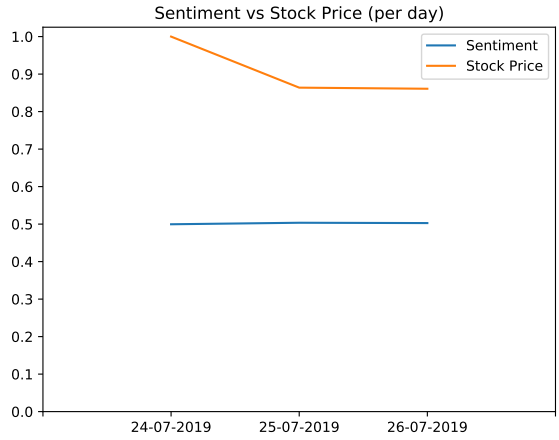


Figure 4: Correlation between sentiment and stock price for the loss of Tesla Motors CTO.

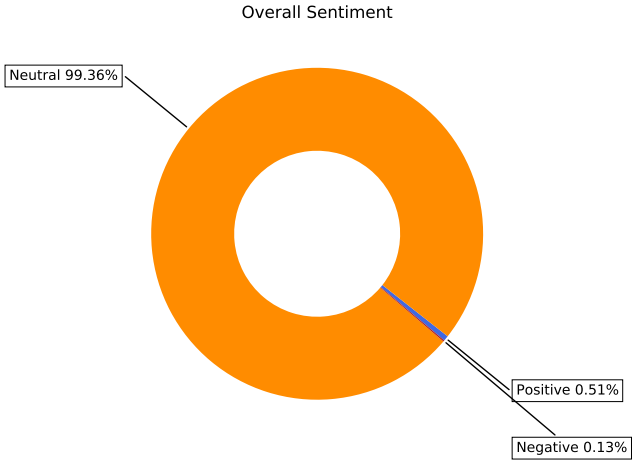


Figure 5: Overall sentiment across the globe for the loss of Tesla Motors CTO.

graph let us understand the absence of correlation, while the index exhibits a high negative correlation. The latter is the result of a data problem. Indeed it would not be possible to calculate the Pearson correlation coefficient when one of the two-time series is a constant, like in this case. However, the very small percentages of positives and negatives tweets, influence the values in a way that the sentiment time series is not actually a numerical constant; even though it is made of very close values. This leads to an erroneous calculation of the correlation, which should not be misunderstood.

Figure 5 is part of the "Global Sentiment" section of the application. Its goal is to show the distribution of the sentiment across the globe. The figure highlights the predominance of neutral tweets with 99.36% against positives with 0.51% and negatives with 0.13%.

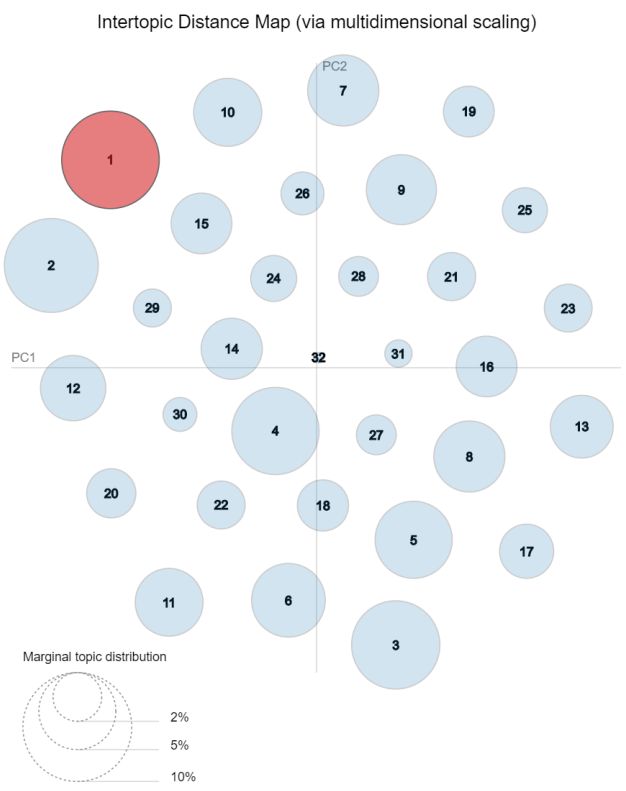


Figure 6: Topics distribution for the loss of Tesla Motors CTO.

Figure 6 and 7 are part of "Most Relevant Topics" section of the application. Getting a closer look into the topics from the visualizations we can clearly see that not all topics are as prominent in the tweets, but the algorithm did manage to separate a number of useful topics from the noisy data. The first topic describes the demand problem of Tesla but it is also talking about an amazing story that has been published during that period. The second topic is close to the first one and this one is clearly encompassing the negative sentiment of Tesla's stakeholders. Perhaps these two are connected in a way. This connection could be topic 20 which clearly describes the news of Mr. Straubel's, Tesla's CTO at the time, being demoted to an advisory role. On the other side of the 2D illustration in Figure 6 we see topic 16. This one takes on a more positive tone with stakeholders talking about a profitable future for the company.

3.7 Limitations

We consider the app as a fully operational and stable prototype and we have tested it numerous times in order to reach this conclusion. However, all of our tests were performed with a small amount of data which means that we cannot accurately predict the efficiency of our algorithms over a large dataset. In the era of big-data, processing 300 MBs of data can only serve as a proof-of-concept. The issue arises from the fact that we are interested in historical data; thus, it is vital for our app to use the premium version of the Twitter API, which comes with a cost. The *free* version of the premium

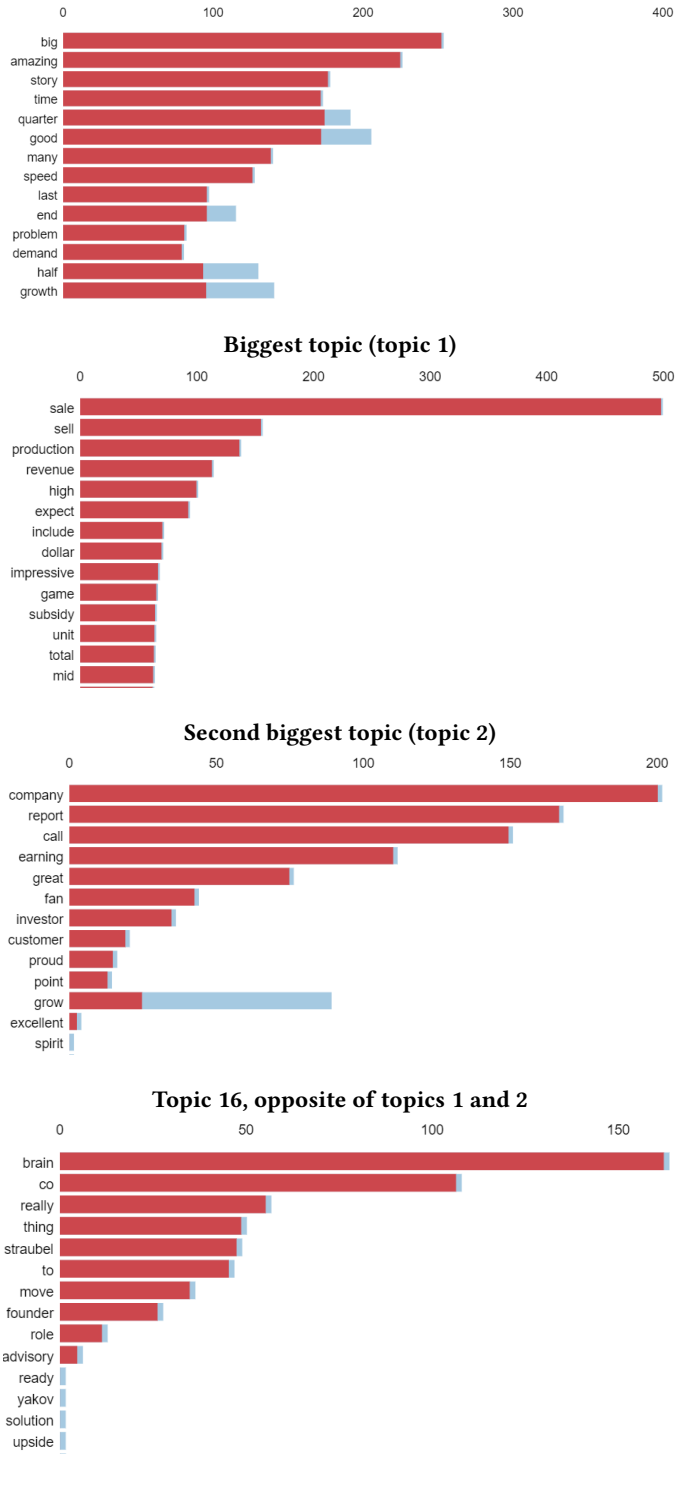


Figure 7: Topic composition for the loss of Tesla Motors CTO. Blue denotes the overall term frequency while red represents the estimated term frequency within the selected topic, $\lambda = 0.6$

API allows 50 requests per month and a maximum of 5.000 tweets. Although we faced this limitation we tried to minimize it as much as we could by applying for 5 different Twitter developer accounts, so we can fetch at least 25.000 tweets. We split those tweets equally to each day of all the events but still, our dataset was not large enough. For that reason, we developed the code keeping in mind that we parse huge datasets. We analyzed the complexity of the code over and over again and we tried to find efficient ways to manipulate the data. Even though we are delighted with the final result, we would really like to have the opportunity to test SWAnalytics in a real-world environment.

From the evaluation, we can observe that the results for sentiment analysis are not so interesting, and most of the tweets are characterized as neutral. In particular, we observed that the percentage of neutral tweets reaches 99.36%. Indeed, this accentuates the weight that the neutral tweets have on the final outcome. This makes practically irrelevant both the positives and the negatives posts, and it is reflected in our graphs. However, we do believe that processing a larger dataset will provide much better results. If that is not the case, then we most probably have to rethink the data-preprocessing and possibly apply some of the techniques that we followed for topic analysis (i.e stopwords removal, spelling correction, and lemmatization).

Finally, even if the results of sentiment analysis process were fair enough, there are other more general limitations when it comes to sentiment analysis for stock prices via Twitter. Even though it provides a lot of benefits, the correlation is valid to a certain extent. Using computer programs to recognize emoticons, sarcasm, exaggerations or jokes comes with a cost. Sometimes the sentiments may arguably be really irrational and it is up to the analyst to decide if they should be included in the sentiment analysis.

3.8 Public version

Our code is publicly available: https://github.com/thzois/sw_2020. Apart from making our code open-source, we thought that instead of inserting figures in the report that demonstrate the web application, it is a better idea to publish the results of our Tesla use case online. However, since the amount of data is small, the live version of SWAnalytics, aims to serve as a proof-of-concept and give the opportunity to any interested party to experiment with the interactive graphs. You can find the live version here: https://thzois.com/swanalytics/.

4 WHAT IS THE SOCIAL WEB AND SOCIAL COMPUTING?

In social computing, the Internet enables users to interact through several media such as blogs, social bookmarking, instant messaging, social communities, and online business networks. People are involved in social computing and are interacting in a broad spectrum of social and commercial activities [29]. Social Web platforms are in essence online communities. A long time has passed since the birth of *SixDegrees*, the first social networking platform, in 1997. In this more than 20 years, social networks have grown exponentially in number, as well as the quantity of the users and the time they spend on them. A social network is defined as "*A dedicated website or other application which enables users to communicate with each other by posting information, comments, messages, images etc.*". Social networks play a dominant role in today's everyday life; there are 3.8 billion of active social media users in 2020[13], meaning that these platforms are used by one-in-three people in the world. With 2.5 billion monthly active users[14], Facebook is the largest one, but many others are widely used from the now established Instagram and Twitter to the younger TikTok.

Along with the wide use of these new forms of interaction comes an abundance of data, with all the opportunities and threats that it brings. More than 100 million pictures are uploaded on Instagram everyday[15]. Every 20 minutes, 1 million links are shared, 20 million friend requests are made, and 3 million messages are sent on Facebook. That information has great value for companies, governments, and research in general. In this context, the field of data analysis is constantly improving and the developed technologies allow experts to extract knowledge from this huge amount of data, which in turn can be used in many different ways and sectors. SWAnalytics is one of these technologies related to the field of commercial analysis. The data it uses are gathered from Twitter, which counts about 336 million monthly active users and 500 million tweets posted everyday[16]. There are several reasons under the choice of this social network as a base for SWAnalytics Social Web analysis. The main one is that it provides, unlike Facebook or Instagram, an API for fetching the data. This search functionality returns a collection of tweets matching a specified query. Facebook and Instagram do not have any API for this purpose, thus scraping data from them is considered illegal. However, it is worth specifying that this does not mean that users' privacy and security are not preserved on Twitter. Conversely, Twitter policies are in line with GDPR and the collection process does not lack limitations, as we further discuss in Section 6.4.

Other motivations regard the user-generated content on Twitter. The mainly textual form of tweets is perfect for the sentiment analysis that our tool performs and the widespread use of hashtags facilitates the gathering by helping our tool to find posts related to a specific topic or content. Moreover, the normally discussed topics on Twitter are more appropriate for our investigation compared for example with the Facebook ones. Indeed, Facebook's priorities are relationships between users and storytelling, while Twitter's posting is more focused on current news and opinions' sharing. To confirm this, we report the self-introduction of the social network in its main page: *"From breaking news and entertainment to sports and politics, get the full story with all the live commentary"*. This "live commentary" on whatever it takes place around the world makes Twitter the perfect source of information for understanding the public opinion on the desired topics, and it is just what SWAnalytics needs for its purpose. Apart from that, Twitter is used on a regular basis by politicians, celebrities or other people who have a big influence to the public.

[13] https://www.statista.com/statistics/617136/digital-population-worldwide/
[14] https://www.omnicoreagency.com/facebook-statistics/
[15] https://www.omnicoreagency.com/instagram-statistics/
[16] https://www.websitehostingrating.com/twitter-statistics/

5 WHAT DOES DATA LOOK LIKE ON THE SOCIAL WEB?

In this section, we discuss the content that is directly generated by Internet users.This is done by SWAnalytics, which analyses what Twitter users post about the client company in order to understand their feelings. The structuring of these Social Web data facilitates the work of our tool.

5.1 User contribution

More than 50% of the world population has access to the Internet, with this percentage growing to more than 85% if only developed countries are considered[17]. These people together produce a huge amount of data everyday. In particular, attention should be paid to what is called "user-generated content", i.e. any kind of content that is posted by users on online platforms. Since the birth of blogs the users' tendency to publish personal opinions on the web has grown exponentially, and this has a great value for companies and other stakeholders. SWAnalytics exploits user-generated content on Twitter (tweets) for helping companies to understand the public opinion about them. Gathering data from this social network is a form of *crowdsourcing*, the practice of obtaining information from services used by a large number of people. However, the research would not be so easy without the help of *folksonomy*. Folksonomy is the user-generated system of using tags to classify contents. In the context of social networks, hashtags are widely used by users to relate their posts to certain topics and this is something that SWAnalytics takes advantage of.

5.2 Data structures

The rapid growth of Social Web data brought attention to the necessity to structure these data on the web. Nowadays, they are not only text or images but they are designed as very specific entities. Indeed, various and different types of information can be found on the Social Web; from users' features to their activities, passing through relationships between entities and content descriptions. Depending on the type, data can be represented differently. This "evolution" of the web towards a more structured frame is known as *Semantic Web*.

In the Semantic Web data are well-structured for being easily read by both machines and users. An important concept in this type of web is *ontologies*. In philosophy, ontology deals with the basic description of things in the world [15]. It is defined by the Oxford dictionary as *"A set of concepts and categories in a subject area or domain that shows their properties and the relations between them"*. In Computer Science, ontology is a formal representation of the knowledge by a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain and may be used to describe the domain itself [22]. Important examples of Semantic Web ontologies are Semantically Interlinked Online Communities (SIOC), which describes user-generated content of online communities, and Friend of a Friend (FOAF), which deals with users and their relationships. SWAnalytics does not use these ontologies, the main reason is that the concept of a 'user account' does not exist and the system does not use a relational database for permanent storage; thus, there not any entities. For the same reason, Open Graph protocol and RDF are not either appealing approaches to be used in SWAnalytics. On the other hand, microformats could be applied, even though they do not really fit with our context. Most of the HTML classes of our user interface are part of Bootstrap 4 and serve to style our webpages.

However, data on Twitter are also structured. Its API provides the tweets with all the related information to SWAnalytics in the form of JSON objects. JavaScript Object Notation (JSON) is a language-independent and lightweight data-interchange format. It is easy for humans to read and write and for machines to parse and generate. Given its characteristics it has become very popular and appreciated in data science over the last few years and nowadays it can be considered the main format for the exchange of information in web applications. Each tweet is represented as a distinct object and it comes with various information such as the author, the message of the tweet, a timestamp and many others. The structure of a JSON data object is represented in Figure 8.

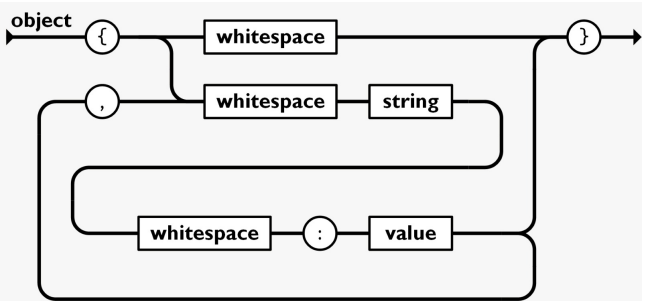


Figure 8: Demonstration of a JSON object

6 WHAT DO PEOPLE DO IN THE SOCIAL WEB?

This section describes how particular topics of Social Web are expressed in SWAnalytics. The topics we discuss in this section are social media accounts, W3C standards as well as privacy. We support some of these topics with literature in order to provide more insight into how SWAnalytics can relate to them.

It is possible to retrieve data for SWAnalytics because of users' data on social media accounts such as Twitter. Users on social media are important factors for analyzing trends based on personal information, but their openness concerning their personal data may affect the results of the analysis. However, linking accounts to SWAnalytics does not fit within our context. The reason behind this is that SWAnalytics serves as an analytics tool rather than an actual application; thus, it is our conviction that it is not deemed necessary to link social accounts. The purpose of SWAnalytics is to provide straight access to business/data analyst teams of an organization and to investors who are interested in a quick overview of correlations between stock prices and events of interest. A way to attract this target group is by avoiding any link to a social account, to prevent any form of data spread amongst third parties' websites (e.g. Facebook, Twitter, Google account). Our idea is that

[17] https://en.wikipedia.org/wiki/Global_Internet_usage

if authentication is required to access SWAnalytics, the targeted audience should access the tool through our own portal to preserve and control privacy.

6.1 Social media account

What has been discussed, among others, how users may have different social networks and how this can be utilized in various ways. Users may utilize social media platforms for different purposes such as work-related activities, networking or dating. Those social networks are made of systems that contain personal or work-related information that the user provided. There are even situations where users can log on diverse social platforms with the same digital access, which results in an easier approach to access a social media platform. Users are participating in these networks because it comes with a lot of advantages, such as staying connected with multiple individuals while being abroad, managing relationships and receiving news about organizations and more.

6.2 Social Web W3C standards

W3C consists of international member organizations that produce or recommend web standards that are considered as protocols, guidelines, and technologies that are created to help the Web reach its maximum potential. In order to achieve this, it is crucial that web technologies are compatible with each other. If not, this can result in interoperability issues where, for example, elements of a website will not work properly. If a website does not meet a certain standard, different software elements cannot interact with one another to produce robust applications or websites [8].

Two important standards that are related to SWAnalytics are JSON-based syntax and the responsive user interface. Our analytics tool heavily uses JSON objects not only to manipulate the data internally but also to expose them to the interface so that the graphs can be generated. HTML5 is used for the main interface, while various JavaScript libraries are used to visualize the results of SWAnalytics in a presentable or structured way. The interface for SWAnalytics is not created by us, but it is a template found online that was picked carefully. Both the template and the libraries ensure that SWAnalytics is compatible with computers and mobile devices so that it is possible for users to view the results on any device. Apart from some standards related to the Social Web, our tool complies with other quality attributes of the W3C standard as well.

6.3 OAuth

Nowadays, data of individual users are increasingly spread on the Social Web due to the growth in Social Web activities. There may even be situations whereby making use of third-party websites would lead to exposing passwords and granting them full access to do everything they want with the users' data. Here is where OAuth comes into place. OAuth tackles this issue by providing access to users' data without sharing identity or passwords.

OAuth is an open-standard delegated authorization framework (e.g. application, websites) that developers can use to retrieve data. It enables applications to access data of a user on third-party websites, like Facebook and Twitter. The application uses a token that allows access to an API server. Afterward, it makes an API request to retrieve data from a data source.

In our project OAuth plays also an important role, specifically we make use of OAuth 2.0. In order to request the data from Twitter API SWAnalytics needs to be authorized. The required authentication allows the developers of SWAnalytics to access publicly available information on Twitter.

6.4 Privacy

It is very crucial to take privacy into consideration when processing personal data. General Data Protection Regulation (GDPR), a set of rules for collecting and processing information from individual or data-subject who live in the European Union, does protect users against certain privacy violations. Users do always have the right to let their personal data be erased or forgotten [27]. Having a great understanding of the web semantics of Twitter is important to analyze and process the data correctly. However, processing data, whether it is tweets or opinions, can come with some privacy concerns as these come from individuals. Some may arguably agree that Twitter is an open platform but contemplate if it is a reasonable ethical justification to collect sentiments via Application Programming Interface (API) for research purposes [36]. Because data is public it does not necessarily mean it makes it justifiable. For example, individuals may not be under the idea that their tweets are being gathered, evaluated and reported or have limited understanding of how their public content might be used [14]. A study conducted by Williams et al. [38] about whether users find it important to be informed about their data being used, concluded that 20% of the respondents did want information beforehand. Demartini et al. [1] explained that with social media it is not always easy to determine what online spaces people perceive as a private matter or a public matter. However, it will also be an unrealistic task to ask every Twitter user for consent to use their tweets for the analyses.

Twitter has a privacy policy where it is mentioned that tweets may be viewed and shared around the whole world and that by agreeing to their services, they do agree that their data can be collected and used by third parties[18]. Thus, Twitter does cover the consent part of the GDPR. Also, research does propose that after users have agreed to the terms, that the data can be considered in the public domain [35]. On top of this, as Beninger et al. [5] suggested, researchers should not assume that Twitter users always read and understand the terms of privacy and consent of the government takes into account. Though, the question is to what extent these rules from GDPR cover the spectrum of privacy. If researchers are only using data without personal identification can the data still be considered personal?

People think that even though data may not be identifiable, there is always a chance to make sensitive personal information identifiable beyond the context it was intended for, and that under some conditions the publication of these data may expose users to harm [38]. Scraping tweets could arguably be in contradiction to Twitter terms because they do demotivate people from doing it. Nonetheless, it is not said anywhere that it is not allowed to scrape data and use them for research purposes. A study conducted by

[18] https://twitter.com/en/privacy

Townsend & Wall [35] suggests to take anonymity into consideration in such research practices and to find ways to secure users' privacy and to prevent breaches. Moreover, the researcher should read through all relevant terms, conditions, and guidelines before using the data for research purposes.

Thus, based on the aforementioned information and literature, privacy for SWAnalytics is preserved by discarding almost all the personal data of the Twitter user accounts. We consider only the text and the date of the tweet so that can we can perform our analysis and match it with the Tesla stock price events, as well as the location of the user. From the location, we keep only the country and then we match each country with the corresponding continent. All the results on the graphs related to the location, that SWAnalytics produces are aggregated so that there cannot be any direct linking with the users. Aggregating results is a technique to preserve privacy. The reason behind this method is that data without personal identification makes it harder to expose a particular user of Twitter. In any case, no sentiment can be traced back to a user, which is strongly stimulated by GDPR.

7 HOW DO PEOPLE MINE, ANALYZE AND VISUALIZE THE SOCIAL WEB?

Big Data deriving from the Social Web need to be processed in order to return valuable knowledge. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The great variety of data on the web brings the possibility of gaining insights in many different areas. SWAnalytics relates with the area of businesses and brands. The goal of this section is to scan the methodology applied by our web tool for *mining*, *analysing* and *visualising* the gathered data.

7.1 Data mining

Mining the data is a complex task. The first step is to analyze the research question in order to understand which are the necessary data and where to find them. SWAnalytics finds its Social Web data on Twitter. Once the data are gathered, they should be examined for understanding the characteristics of the dataset and its attributes. This is crucial for deciding which kind of pre-processing to apply. SWAnalytics filters the data both by discarding some records and by removing some features which are irrelevant for the analysis. Apart from that, reporting some basic statistics is also important for understanding the data; thus, SWAnalytics provides relevant information to the tweets before and after the pre-processing phase.

7.2 Data analysis

Research can be classified according to the type of analysis it provides. *Quantitative* analysis has as its main purpose the quantification of data. This allows generalizations of results from a sample to an entire population of interest and the measurement of the incidence of various views and opinions in a given sample. On the other hand, *qualitative* analysis aims to gain an in-depth understanding of underlying reasons and motivations. It provides insights into the setting of a problem. Yet, the two types are frequently combined in order to have more complete results. This is the case also for SWAnalytics. It performs a sentiment analysis that provides qualitative information about the tweets (i.e. if they are positives, negatives, or neutrals). However, aggregating its results for calculating the distribution of the sentiments, as well as the computation of the correlation index is clearly a quantitative task. Moreover, topic analysis can be considered quantitative, too, as it deals with calculating frequencies of the recurrent terms and clustering of them.

The next step is the transition to *deep data*, which is completed at the end of the analysis. We are referring to deep data when the information has already been processed for being relevant, high quality, and ready to provide answers to the problem. In SWAnalytics, the deep data are those that are ready to visualized in the output web pages.

Finally, once the analysis is concluded and the visualizations are ready, the results have to be evaluated and interpreted. This is a task for the user of SWAnalytics. However, in further developments of our tool, we could insert some metrics for evaluating the outcome of the whole process of analysis.

7.3 Data visualisation

Data visualization is the task of transforming the ultimate raw data into user-friendly charts. It is a very important task in Data Science because it allows the users to understand the outcomes of the analysis. The output page of SWAnalytics is completely made up of interactive graphs. Different types of visualizations have been chosen to show the obtained results. Specifically, we use a multi-line graph to demonstrate the correlation of the sentiment with the stock price, as well as a doughnut chart and a world map, to demonstrate the overall and the sentiment per continent respectively. Finally, we also use two other types of graphs to describe the results of topic analysis; a cluster diagram where each cluster represents a different topic and its size and a bar chart that demonstrates the decomposition of each topic with the relative frequencies of each term.

8 PERSONALIZATION ON THE SOCIAL WEB

This section describes how organizations have the possibility to personalize application based preferences of users on Social Web. The application infers information about the user. Based on this information, an application can be tailored as users wished. Implementing content-based-filtering algorithms in the web application can even result in providing recommendations based on the interest of a user. However, personalization is not encompassed in SWAnalytics and we discuss the motivation behind this decision.

8.1 Personalization

SWAnalytics does not give users the possibility to personalize the application based on their preferences. In the early stages of developing this tool, it did not cross our mind to provide the possibility of personalization. SWAnalytics is developed mostly as a tool that processes data and produces graphs and visualizations related to tweets and stock prices and provide users instant insight into results. So, we prioritized an instant overview over personalization. Nonetheless, we do agree that personalization may have made SWAnalytics more versatile.

A possibility to personalize SWAnalytics is by providing registration with the system. In that sense, it would be possible for every

user to analyze its own events, store the results of old analyzed events and be able to restore results from a past analysis. Another approach to personalize this web application is by using session cookies. In such a way, we can provide different access levels for the users. For instance, an "Admin" can view more content than someone with the role "User", or the administrator can grant permission to the user to access data that the administrator has fetched. Another reason that session cookies are useful is for improving the overall user experience. Since the system will provide a registration page we can use the cookies to track open sessions in a sense that if someone logs in the system and closes the browser, then the user does not have to re-login. However, regardless of the reason that we use cookies, privacy should always be our first priority. Thus, we should allow users to know the data that we are using, for which reasons and provide them the option to toggle the settings related to the usage of cookies. In short, there could have been many ways to personalize SWAnalytics as an extra service for the users.

9 HOW CAN WE STUDY THE SOCIAL WEB?

In this section, we discuss how we can study the Social Web. The topics related to SWAnalytics are the 3 Vs of Big Data, Web Sciences and web requirements of the target audience.

9.1 Science Data analysis: 3 Vs of Big data

Extracting a huge amount of data to analyze data is in most cases necessary to study social problems. A way of studying the Social Web is, like in our case, analyzing tweets and sentiments of the tweets to determine if there is any correlation between the sentiment and the events of Tesla stock prices. In addition, SWAnalytics applied topic analysis to find frequent topics discussed. Semantic analysis is used to categorize the tweets into positive, neutral or negative topics and relate it to Tesla stock market events.

In regards to the volume of data, 3113- 4750 from a total of 21.819 tweets are eventually analyzed for each Tesla stock Prices event. The total amount of data is similar to 300 MBs, which is definitely not huge. This has more to do with Twitter API and limitations attached to the amount of data that can be extracted. Assuming that large companies have a paid version of the API and may use a large amount of data when they want to adjust SWAnalytics based on their own data, the algorithm used to develop SWAnalytics is continuously optimized during implementation. This way, SWAnalytics can perform well when a huge amount of data is being extracted.

Regarding the variety of data, the attributes of a tweet that we consider in our tool, are not many. As mentioned in a previous section SWAnalytics takes into consideration the full text of a tweet, the created date, and the location of the user. Some of the attributes that are not considered are, for example, the time zone of the user, the username, the creation date of the Twitter account, images and URLs. These types of data do not provide any value for the users of SWAnalytics and hence we discard them.

Finally, the velocity of data refers to the speed of which data is generated. Through Twitter API, it is possible to communicate with the Twitter server so that the data can be retrieved. There is a rate limit with regard to the amount of data requested for SWAnalytics. So, every time a certain amount of data is retrieved from Twitter, there is a timeout of 15 minutes, until we receive finally 5.000 tweets per month.

9.2 Web sciences

Web Sciences consist of researches from different disciplines, such as Computer Science, Mathematics, Economics and so on, in order to study and explain the web. SWAnalytics surely relates to this. In order to generate the results of SWAnalytics, different fields have to come together to produce the final result. The following types of disciplines are fundamental:

- **Artificial intelligence (AI)**: topic analysis and sentiment analysis. Both of these analyses stem from AI tools because they involve Machine Learning algorithms to find patterns in the data.
- **Web Engineering**: Makes it possible to develop SWAnalytics and to display the outcome of the analyzed data in graphs.
- **Computer Science**: Our tool is designed to handle a huge volume of retrieved data. Therefore, it is important to find efficient ways for data-processing.
- **Economics**: There are already data results about stock prices published on the web via Yahoo Finance. These data are used to find correlations within sentiments.
- **Mathematics**: For SWAnalytics, it is necessary to come up with a formula that calculates 'the positivity' of the tweets for a given event in order to answer the research question. By applying this formula and aggregating the different types of sentiment, analysts can evaluate the correlation of the tweets with the stock price.

So the intersections of these fields makes it possible to study the results on SWAnalytics.

9.3 Society diversity and Web requirements

Different parts of the web society have their own objectives with regard to the usage of the web. Objectives as transparency, privacy and openness are considered in SWAnalytics. The target audiences of SWAnalytics are basically companies and/or investors. With that in mind, the application is developed so that the companies interested in SWAnalytics, have access to the open-source code ensuring transparency and openness. Thus, they can either modify the code themselves to add the desired features and graphs.

10 WHAT ARE THE CHALLENGES FOR THE SOCIAL WEB?

This section discusses what types of challenges there are concerning Social Web and how they affect SWAnalytics. With regard to those challenges, fake news is increasingly spread around on the Social Web and how this influences individuals to gain for instance political advantages. This may, of course, affect the accuracy of web data analyzed, because inaccurate data does not present valid results. We are aware that fake tweets may have been a part of the data analyzed for SWAnalytics and we tackled the issue to a certain extent.

10.1 Fake tweets and bots

One of the Social Web issues tackled is the account of bots. Bots are fake accounts and form a big problem on Twitter. In SWAnalytics, bots are detected by removing duplicate tweets from the same user, the same day, in the preprocessing stage. Some of the extracted tweets are posted every 10 seconds, where it is illogical to think that a real user will post the same continuously. This approach does not fully cover the bots issue. It is still possible that a bot has two accounts or more, where one account tweets and the other account retweets. In this case, we have not been capable of detecting them.

There have been some application tools created to detect bots by different people, but most of it has not been qualified as great sources to use [17]. Thus, it is still risky to use such existing detection platforms. Another approach a study suggests is to use a fingerprint of user behaviors and a set of statistical measures that describe different aspects of those behaviors. Therefore, it can be observed whether the behavior of a bot is imitated [19].

Fake tweets are also considered as challenges on the Social Web. A study explored how to detect fake tweets by training machine learning algorithms and more specifically by using decision trees [21]. Another approach this study proposes is an application where the end-user firstly provides an URL or ID of a tweet of interest. The application then evaluates the retrieved tweet, user-based features and verification results of the tweets being real or fake based on the color. The validity of this application is, however, questioned.

To conclude, there is ongoing research when it comes to detecting fake news on Twitter or bots, but the existing approaches are not trustworthy to use. There is always a possibility that bots or fake news are part of the analyzed results. Though, we have made the necessary efforts to minimize these issues in the results of the data.

11 CONCLUSION

In this work we have successfully managed to design an app that takes textual data from Twitter about a certain company (e.g. Tesla) and analyzes it together with financial data (i.e. stock price) to try and give more insight into the "*Why?*" of a company's performance. In our use-case we find some results that were expected and some that are inconclusive and controversial. For instance, the sentiment analysis returned a strong negative correlation between the sentiment of users in the event analysed and the performance of the company. This correlation is mainly due to the big percentage of neutral tweets. From a research perspective this finding is inconclusive because of the small sample size per event, thus also lack in statistical inference. From a practical point of view, the application gives insight into the sentiment behind the raw data and it allows for user validation through topic analysis. The topic analysis returned coherence scores of approximately 50% which is neither good nor bad. Upon further look, we clearly saw a good separation between positive and negative topics. However, many of the topics were hard to distinguish in terms of sentiment even from a human perspective, hence the abundant number of neutral tweets in the analysed corpus. Nonetheless, for the designed proof-of-concept in this work the information obtained meets the requirements and expectations that were envisioned at the beginning of this project. Aside from the technical perspective, we also managed to link SWAnalytics with the social aspects of the Social Web quite clearly. As we explain in the next section, our tool is by no means tuned to perfection and it can be improved further.

12 FUTURE WORK

In terms of improvements, we would like to improve the way that the user interacts with the system. Instead of modifying the "events.json" and setting an environment variable in order to restart the whole process, we would like to provide a more user-friendly way. For instance, we can provide a web-interface where the user can provide all the various configurations needed, as well as a button that initiates the whole procedure. The implementation of such an idea does not require any significant effort, but the time was limited and we had to deal with other difficulties. Apart from that, we would like to introduce the concept of personalization. In that sense, we can provide registration with the system where each account is associated with different events. A multi-user environment allows users to work on different events of their company and investigate the potential impact of Twitter for each case. Apart from that, we could also introduce various account levels. In that sense, a user with the role "Admin" is capable of sharing the results of the analysis with other users so that multiple users collaboratively reason on the impact of sentiments.

Regarding our data cleaning procedure, we would like to investigate more ways to detect bots through network analysis. One possible way that has the potential to detect a bot is by observing the 'followers' of an account as well as the number of users that the account 'follows'. For example, if a user has only two or three 'followers' and two or three users that 'follows' it is likely to be a bot or low impact on the overall result. This is not a 100% accurate way to detect bots but it has the potential to change the results significantly since the application will take into account users with bigger influence.

As the process of the topic analysis is concerned, we would like to test and implement faster algorithms in the app, such as the SMH or versions of LDA that can parallelize the number of topics selection step. The hyperparameters of the model we are using are not optimized, because this is out of the scope of this study, however, we do stress that when further updates are made to the app (e.g. parallelizing the methods) a fast search method (e.g. Random Search) should be implemented that models the optimal hyperparameters as well. In addition, we would like to improve sentiment analysis and combine it with topic analysis. Future updates could incorporate both by segmenting all the tweets by sentiment first and then run the topic analysis pipeline to see whether a better topic separation and interpretability can be obtained.

13 INDIVIDUAL CONTRIBUTION

13.1 Bogdan Ene

First of all, during this project I was involved in the decisions we took regarding the topic of our project as well as the research question and the goal that we plan to achieve. We debated on different topics and laid down all the positive and negative points. In the end, we concluded that it is best to create an application that is capable of automatically analyzing twitter data regarding some

given events and correlate people's opinion with the stock price fluctuation during the events times. We took as proof of concept events that concern Tesla Motors. We all debated on the strategy we want to follow in the means of data that we should use, data cleaning techniques, as well as the graphs that we want to display as results of our analysis.

Secondly, I focused on the development of the software application together with Thodoris. We decided to use Twitter APIs to retrieve users' tweets regarding Tesla's events. Moreover, we decided to create python scripts for our data retrieval and processing, as well as for the generation of the HTML web pages. In this case, I researched few python libraries that help in retrieving the tweets, as well as techniques to retrieve as much data as possible, since only a limited number of tweets can be retrieved in a period of time. In this way, I chose to make use of the Tweepy Python library which also provides a mechanism that retrieves tweets periodically. Having this setup, I created a first version of the script to retrieve the twitter data, which later on Thodoris updated to fit the rest of the application. During our meetings, we decided that we want to show people's sentiment on Tesla's events categorized in the user's countries. In this sense, while Thodoris worked on cleaning the data, I researched and implemented the algorithm that matches the users based on their country. During this process, I made use of the pycountry library which helps in finding out what is the users' country based on its name or abbreviation. After we cleaned the data and categorized the tweets based on the provenience country, I worked on the script that generates the HTML files. In this sense, I created a first version of the script which was updated and integrated into our application by Thodoris.

As we had our data and Thodoris created the HTML template that is used to generate the custom HTML web pages which display our results, we moved on to the creation of the graphs. In the beginning, we wanted to create a bar chart where we display the number of tweets that we analyzed based on their provenience country. In this sense, I created a piece of javascript code and made use of the Charts.js library in order to generate the bar chart. Later on, we decided to change it to a world map where we display the sentiments of the users based on the provenience continents. At first, we based it on countries, but since there were a lot of countries and a lot of data to display, the performance of the application was low. I worked on the world map creation and I made use of the Amcharts javascript library. Moreover, after Dusko finished the topic analysis, I integrated his python script into the application.

Finally, I was involved in the presentation regarding the paper review. I collaborated with my colleagues and especially with Thodoris to create the slides for the last three chapters of the paper regarding the fake news during Trump's elections (last three chapters: Exposure to fake news, Who sees the fake news and Conclusion).

13.2 Dusko Trajkov

During the early stages of the project, I coordinated how the project would look like. I did this by engaging my fellow teammates in a research proposal about Tesla, where we combine natural language processing techniques and data mining with finance data to extract insights about the company. Being a big fan of the company I've been following Tesla for a long time and I noticed that there has been quite some turmoil and exciting days for the company during the past year. This makes it a very viable candidate for a use-case in our app to test how well our analyses perform. After writing the proposal and handing it in, we got feedback to slightly change our way of thinking and design an application that is able to take any event (range of dates) about a company and extract information from the raw textual data. At the beginning of this plan, we encountered the issue of how to best download the data. Here I helped research the most appropriate option for us and set up the tweeter developer accounts to download the data. When the data was obtained, I focused solely on writing the code, doing the literature review and evaluating the results of the topic analysis part. Here I spent quite some time finding the best topic analysis algorithms and trying to integrate them into the data. There are many ways of doing topic analysis and just the LDA algorithm itself has many number of implementations. First I tried to understand how to implement the Mallet version, since quite a lot of blogs and papers suggested that it performs best when measured up against the coherence score. Since Mallet is written in Java and I was working in google colab these were really hard to link. After many trials and errors I decided to go with the more intuitive implementation of "Gensim" and use the vanilla LDA. Much of the research around topic selection and LDA proposed HDP as the most intuitive algorithm to try and automatically select the number of topics for the user. For the purposes of the app we all agreed that it'd be best to just use one algorithm, however it is also important to deliver evidence for our reasoning behind the usage of LDA. This reasoning is shortly included in the main text and more in the appendix.

Finding the right pre-processing tools for the raw data was evenly matched in effort as doing analysis on it. The tweets were very messy. I tried a number of technics to make the vocabulary for LDA more valid: spelling correction algorithms (e.g. TextBlob, which worked the best), splitting words stuck together, removing redundancies etc. During the first number of trials of topic analysis I noticed that some words were misspelled by TextBlob and some words needed not to be corrected by the algorithm. There was no way to account for this automatically, since NLP is still a very young field of research and current technologies are not robust enough to account for all cases. Hence, I needed to make a manual list of exceptions for all NLP tasks before feeding the data to the app. Once the topic analysis part was finished, I worked together with Bogdan to integrate it into the app and spent the remainder of my time in the project reviewing my teammate's work, correcting it and writing the reflecting chapters (e.g. Conclusion) in the report. Since I have a background in Data Science and a background in Business Economics it was quite easy to contribute to the technical and the non-technical side of this project.

13.3 Gloria Boekhouder

During this whole project, I was responsible for different aspects. Since my main background in education is Business Management, I wanted to focus more on the literature, lectures and the presentation rather than the technicalities behind the web application. I also believe that this was crucial for me to be able to comprehend the course. Before this project, I did not even know what sentiment

analysis was. I heard of it but did not thoroughly understand it. It was, therefore, necessary to do literature review on sentiment analysis and the correlation between stock prices in order for me to comprehend its meaning. Firstly I had to collect different papers that only had to do with sentiment analysis. Then I had to retrieve papers that were only related to stock prices. Ultimately I gathered literature papers that combined these topics. After collecting the literature papers, I had to read through a lot of these papers to assess whether they were valuable for the project or not. After doing so, I made a list of the papers I wanted to use for our project.

Subsequently, I conducted a mini-literature review on these topics whereby I also considered the pro and the cons that come with sentiment analysis and the VADER algorithm we used. This was not only necessary for creating a good background for the project, but also for the implementation of our web application. Also, not only to make us aware of what type of implications it could bring but also to manage our expectations of the results for our project.

I was also responsible for some sections of the report such as linking the results of the project to the lectures. As for some of these topics discussed in the lecture, I already had a good insight into them (e.g. privacy, personalization) which gave me the advantage to discuss some topics of the lectures and elaborate on it. Not all topics in lectures were encompassed in the web application, so I had to think of creative ways to fill in this gap.

Thus, I conducted a literature review again and discussed my proposals with the rest of my team members. We discussed how we could encompass some of the lectures in this project, based on these proposals. In order to link all the lectures to our web application, it was important for me to be in the loop of some of the technical decisions that were made for the web application.

Additionally, it was also important to communicate with Thodoris, who was most responsible for the implementation of the web application. I also discussed further improvements about the results of the web application with my team members.

Even though attending the lectures provided more clarity on topics such as W3 web standards, the veracity of data or Web Sciences, for someone that did not have a solid background in web applications, I did have to do extra research and communicate with my teammates continuously to comprehend these topics thoroughly. I would have not been able to make a link between the app and the lecture if I did not understand the lectures. Nonetheless, it was a great learning process to be responsible for this part of the project.

I also participated in the paper presentation about fake news. so I had to read the paper and present my part. With all that was previously mentioned, I am satisfied with the dynamics of the team. Because of our different backgrounds, the contributions of every team member were essential to succeed in this project.

13.4 Nicola Pieruzzini

Having an economic background I preferred to work on the settings of the project instead of the coding and more technical part. At the very beginning of the work, I took part in the process of defining our research. It was my idea to focus the analysis on brand-related events and on the impact that they have on the companies. I formalized the research question and I contributed to design the layout of SWAnalytics. In particular, I designed the correlation section of our web application. I came out with the idea of the graph comparing the two trendlines and I developed the formula for the normalization of the sentiments' values. I also suggested the use of the correlation index. Moreover, I contributed to define the data cleaning rules.

With Thodoris, I managed the presentations. We worked together on the slides and I gave the pitches for the 1-minute presentation in the middle of the course as well as the final one. I also worked on the fakes news' paper review, where I designed the slides, presented one part and answered the question at the end of the presentation.

I wrote the contents of the introduction (Section 1) and of Section 4 (What is the Social Web and social computing?), 5 (What does data look like on the Social Web?) and 7 (How do people mine, analyze and visualize the Social Web?). For doing that I needed the help of the colleagues that worked on the code: they gave me the technical information that I needed for linking SWAnalytics to the course lectures. I made some research from online sources and literature in order to elaborate on some topics of the lectures that better fit the contents of our application. My groupmates also provided me with some tips on the motivations for which we did not link certain arguments and if we could have done it.

Regarding the app report (Section SWAnalytics) I wrote some contents that my colleagues could elaborate with more technical details. I contributed to defining the structure of the Section (as well as the global structure of the report) first, and then I described the steps of the SWAnalytics workflow. I also analyzed the results of our case study regarding the sections of correlation and global sentiments, taking part in the discussion over the motivations for which the outcome is not as interesting as we would like it to be.

Concerning our group, the presence of different backgrounds has been very important for this work. From a personal side, I learned a lot of technicalities which will be very useful for my career as a data scientist. Despite knowing what they are, I dealt with sentiment and topic analysis in a practical way for the first time. From the team point of view, our differences brought to a continuous exchange of opinions and feedbacks, sometimes also conflicting. Yet, this has been important for analyzing our choices from different sides and for taking more responsible decisions.

13.5 Thodoris Zois

Along with the other members of the team, I participated actively in the discussion in order to decide not only the research topic that we would investigate but also the way to reach the final outcome. After we came to the conclusion that we will develop a unified analytics application that can be used by any company or investor to gain valuable insight, we brainstormed in order to find an interesting use case and determine the periods of the various events.

Due to my technical background, the part that I was mostly involved in was the core development of the application. Initially, I conducted a small research in order to determine all the different technologies that we needed to use in order to build an application that has a distinct behavior depending on the user input. After designing the whole structure of the system (what is the internal functionality? what kind of scripts do we need to succeed? where the application is going to store the data? how the back-end and

the front-end are going to communicate over the data without the use of an API?) I came up with the implementation of a simple way that allows any party to interact with the system and determine the various events (events.json file). After that, I worked on the backend of the application. Firstly, I wrote a piece of code that fetches the stock prices for each day of each event from Yahoo finance. After Bogdan finished with his script that fetches Twitter data, I combined both pieces of code in a single script, since both parts were related to data-fetching. However, after fetching some tweets we saw that we cannot perform a historical search. The reason was that Twitter offers old data and the ability to search between a dates-range only by using the premium API. Hence, I started investigating what are the possibilities in the library that Bogdan was using (Tweepy), or the alternatives. Fortunately, Tweepy is an open-source library and after some search on the public GitHub repository, I found out that we can use a version of the library that is not released yet but exists on a branch other than master. I also contacted the developers of the library to make sure that the version is stable enough to be used in a production environment.

After fetching all the data we had to perform some cleaning. After a team meeting, we discussed the interactive charts that we are interested to make and thus we concluded on which data we should keep and which should be cleaned out. I worked on the cleaning process along with Bogdan, were each one of us had a different part. My personal contribution was the detection of the full text of a tweet, which proved to be tricky since I had to investigate what kind of different tweet-structures were returned. Apart from that, I also wrote a small piece of code that keeps some statistics related to data cleaning. My idea was that it would be nice for the end-user to see the amount of data that was fetched, those that were finally analyzed to produce the graphs and also demonstrate the reason why some data had to be removed. Since our data was ready, I wrote a Python script that uses Vader and performs sentiment analysis on the data that we would use to create the final result. When Bogdan finished with the integration of topic-analysis in our codebase, it was time to work on the front-end of the application.

Initially, I searched for an HTML5 template that uses Bootstrap, so our app can be visible in mobile devices also. Then I performed some modifications to the code so it can fit the structure that we desired, as discussed in a team meeting, and created our “brand” logo. However, the real challenge for the front-end was the way to generate from Python, one HTML page for each event. I came up with the idea to have one .html file as a template with placeholders. For instance, the navbar menu depends on the content of events.json since each navlink is related to one event in the events file. Thus, in the HTML template file, I added an entry “ SWAnalytics NAV ”, so that with Python we can parse the template file and generate the necessary HTML code in certain parts. The same procedure was followed for everything else that had to be shown in the front-end part. Bogdan created the initial Python script that copies the event template, and eventually, we were both populating it with more code as each one was building the different parts for the front-end. Regarding the final result of the user interface, I implemented the first two sections. For the first section, I added code to the script that generates the HTML files in order to make visible to the front-end the user input related to a particular event as well as the various statistics. For the second section, I had to build the sentiment vs stock price graph as well as the correlation coefficient. Before visualizing the data, I had to perform some processing in order to compute the positivity for each day of the event, normalize the stock prices and compute the correlation coefficient. All of these three were ideas that belong to Nicola who was of a great help, especially for because he made the mathematical formula that computes the positivity. After implementing the necessary code, I wrote a javascript function that reads all the necessary data and creates the graph by using Chart.js. Then, I populated the generate HTML Python script, to generate the calls for the necessary functions in the front-end.

After Bogdan finished his part of the graphs, the core code for the application was finally ready. However, there was not any way to inform the user about the data processing steps. So, I had the idea to add some logging so that the user knows what the application is doing while preparing the data. I used the “logging” library from Python and shared the logger between all our scripts in order to produce a single log file (swanalytics.log). Even though our first prototype was working, we had many dependencies that someone had to install before running our application, and apart from that, there was not an easy way to publish the application online. Thus, I made two Docker containers. The first one was related to the actual application. Instead of the user installing all the dependencies, the container takes care of all of them. Then, I created a second one that uses Nginx to serve the web content. In that sense, it is easy for anyone to deploy the application in a Cloud service and make it accessible within the organization. Since both containers were vital for the application to work, I wrapped both containers in a docker-compose file. All the burden to run the application was gone straight away and the only thing that someone who is interested to run it is 1) Install Docker 2) run “docker-compose up -d –build” and 3) visit localhost:8080. However, there was still something missing. There was not any way to configure the app in such a way that it does not repeat the procedure of data fetching, cleaning, and analysis. It can be the case that the data have already been processed one wants to restart the container but without restarting the whole procedure. To achieve that, I introduced an environment variable in the docker-compose file that is used to determine if the back-end procedure should be skipped. After that modification, I performed all the necessary tests in order to determine the behavior of the app in any case. For the tests I allowed the code to fetch and process only 6 tweets per event (2 tweets per event day), in order to make the testing procedure faster. Finally, I maintained the README file in our GitHub repository and uploaded our work in my private domain.

Apart from the technical context, I tried to help the team as much as possible in order to achieve the best possible result, collaboration is always beneficial. From the non-technical tasks, I helped with the styling of the paper presentation. Due to some last-minute issues that Bogdan was facing, I also presented his slides, however, the presenter notes were made completely by him. Regarding the final project presentation, I also helped with the styling of the slides as well as the presenter notes related to the technical details of the app, so that Nicola could easily present the slides. For the final report, I contributed to the whole document. Firstly, I organized the whole structure of the report and eventually I made the necessary modifications to preserve the coherency for each and between

sections. I also combined all the parts that each one of my teammates wrote, placing them in the relevant sections. I went through the text of each one and modified parts that were inaccurate, removed parts that were unnecessary, elaborated in some cases, and moved certain paragraphs to other sections because they did not fit the context of their section. Then, I did the first review of the report and took care of the document styling and formatting (spaces, capital letters, italics, bold words, quotes, broken references, broken URL links). Regarding my personal contribution in terms of text, I wrote some parts of the section related to SWAnalytics, and in some cases I borrowed some text that Nicola wrote before me. However, anything related there with topic analysis is work been done by Dusko. Finally, I also made the diagram that demonstrates the inner-workings of SWAnalytics, the line graph, the doughnut chart and the table that demonstrates the impact of the data cleaning process.

REFERENCES

[1] Ahmed, W., Bath, P., and Demartini, G. Using twitter as a data source: An overview of ethical, legal, and methodological challenges. *Advances in Research Ethics and Integrity 2* (2017), 79–107.

[2] Ahuja, R., Rastogi, H., Choudhuri, A., and Garg, B. Stock market forecast using sentiment analysis. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (2015), IEEE, pp. 1008–1010.

[3] Alkubaisi, G., Kamaruddin, S., and Husni, H. A systematic review on the relationship between stock market prediction model using sentiment analysis on twitter based on machine learning method and features selection. *Journal of Theoretical and Applied Information Technology 95* (12 2017), 6924–6933.

[4] Asuncion, A. U., Welling, M., Smyth, P., and Teh, Y. W. On smoothing and inference for topic models. *CoRR abs/1205.2662* (2012).

[5] Beninger, K., Fry, A., Jago, N., Lepps, H., Nass, L., and Silvester, H. Research using social media; users' views. *NatCen Social Research* (2014), 1–40.

[6] Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, null (Mar. 2003), 993–1022.

[7] Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (2011), 1 – 8.

[8] Caldwell, B., Cooper, M., Reid, L. G., and Vanderheiden, G. Web content accessibility guidelines (wcag) 2.0. *WWW Consortium (W3C)* (2008).

[9] Chen, R., and Lazer, M. Sentiment analysis of twitter feeds for the prediction of stock market movement. *stanford edu Retrieved January 25* (2013), 2013.

[10] Cvitanic, T., Lee, B., Song, H. I., Fu, K., and Rosen, D. Lda v. lsa: A comparison of two computational text analysis tools for the functional categorization of patents. *International Conference on Case-Based Reasoning* (Jan 2016).

[11] Dattu, B. S., and Gore, D. V. A survey on sentiment analysis on twitter data using different techniques. *International Journal of Computer Science and Information Technologies 6*, 6 (2015), 5358–5362.

[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. 391–407.

[13] Elbagir, S., and Yang, J. Twitter sentiment analysis using natural language toolkit and vader sentiment. In *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019, 13-15 March, 2019, Hong Kong, pp12*, vol. 16.

[14] Fiesler, C., and Proferes, N. "participant" perceptions of twitter research ethics. *Social Media+ Society 4*, 1 (2018), 2056305118763366.

[15] Fonseca, F. The double role of ontologies in information science research. *Journal of the American Society for Information Science and Technology 58*, 6 (2007), 786–793.

[16] Fuentes-Pineda, G., and Meza-Ruiz, I. V. Topic discovery in massive text corpora based on min-hashing. *Expert Systems with Applications 136* (Dec. 2019), 62–72.

[17] Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., and Larivière, V. Tweets as impact indicators: Examining the implications of automated "bot" accounts on t witter. *Journal of the Association for Information Science and Technology 67*, 1 (2016), 232–238.

[18] Hutto, C. J., and Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media* (2014).

[19] Kosmajac, D., and Keselj, V. Twitter bot detection using diversity measures. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing* (Trento, Italy, Sept. 2019), Association for Computational Linguistics, pp. 1–8.

[20] Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus 5*, 1 (Sept. 2016).

[21] Looijenga, M. S. The detection of fake messages using machine learning. B.S. thesis, University of Twente, 2018.

[22] Man, D. Ontologies in computer science. *Didactica mathematica 31*, 1 (2013), 43.

[23] McCallum, A. K. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[24] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., and et al. Mllib: Machine learning in apache spark. *J. Mach. Learn. Res. 17*, 1 (Jan. 2016), 1235–1241.

[25] Newman, D., Asuncion, A., Smyth, P., and Welling, M. Distributed algorithms for topic models. *J. Mach. Learn. Res. 10* (Dec. 2009), 1801–1828.

[26] Pagolu, V. S., Reddy, K. N., Panda, G., and Majhi, B. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (2016), IEEE, pp. 1345–1350.

[27] Regulation, G. D. P. Consent url =https://gdpr-info.eu/issues/consent/ month =march, lastaccessed =March, 2020„ 2018.

[28] Röder, M., Both, A., and Hinneburg, A. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (2015), ACM Press.

[29] Sadiku, M., Shadare, A., and Musa, S. Social computing. *International Journal of Innovative Science, Engineering Technology 4* (01 2017), 117–119.

[30] Schumaker, R. P., Zhang, Y., Huang, C.-N., and Chen, H. Evaluating sentiment in financial news articles. *Decision Support Systems 53*, 3 (2012), 458–464.

[31] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. Exploiting topic based twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2013), pp. 24–29.

[32] Sievert, C., and Shirley, K. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (Baltimore, Maryland, USA, June 2014), Association for Computational Linguistics, pp. 63–70.

[33] Sukhija, N., Tatineni, M., Brown, N., Moer, M. V., Rodriguez, P., and Callicott, S. Topic modeling and visualization for big data in social sciences. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (July 2016), IEEE.

[34] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, 2005, pp. 1385–1392.

[35] Townsend, L., and Wallace, C. The ethics of using social media data in research: A new framework. *The Ethics of Online Research (Advances in Research Ethics and Integrity, Volume 2). Bingley: Emerald Publishing Limited* (2017), 189–207.

[36] Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., and Burnap, P. The ethical challenges of publishing twitter data for research dissemination. In *Proceedings of the 2017 ACM on Web Science Conference* (2017), pp. 339–348.

[37] Wei, X., and Croft, W. B. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06* (2006), ACM Press.

[38] Williams, M. L., Burnap, P., and Sloan, L. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology 51*, 6 (2017), 1149–1168.

[39] Yut, L., Zhang, C., Shao, Y., and Cui, B. Lda*: a robust and large-scale topic modeling system. *Proceedings of the VLDB Endowment 10*, 11 (Aug. 2017), 1406–1417.

A APPENDIX

A.1 Variants of LDA

Latent Dirichlet Allocation is associated with a significant increase in computational complexity when applied to large data sets. To deal with this issue well-tested variants of the original LDA have been proposed by Apache Spark [24] and Mallet [23], while other less replicated research has been continuously trying to build and reinvent the algorithm for distributed topic modeling using numerous machines in a real-world production environment [39, 16]. In particular, the most recent method invented by [16] called Sampled Min-Hashing (SMH) for topic analysis takes on a different approach to the vanilla LDA and its successors. It samples directly from the co-occurrence space of words to identify those belonging to the same topic, while LDA and its alternatives [39, 20] acquire the topics as distributions over the words of a predefined vocabulary and documents are likewise distributions over K topics. Another key difference besides solving the computational problem associated with the inference process to extract topics using Min-Hashing, is that SMH does not require the user to specify a priori the number of K topics. The authors of [33] emphasize the fact that naive methods such as LDA can output different results for the different numbers of topics and configurations. Especially in large data sets defining K can be a very daunting task since the vocabulary and the number of variables to be processed become large as well [10]. The LDA adjustment for estimating the K topics statistically is called the Hierarchical Dirichlet Process (HDP) [34].

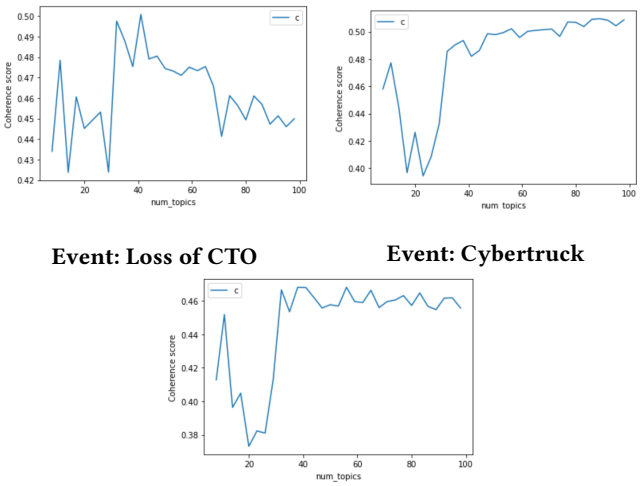
The optimal parameters for LDA can be reached in numerous ways. The ones of interest are the variational inference with expectation-maximization as described previously and Gibbs sampling [37]. Gibbs sampling is a bit more mathematically challenging to describe, so in simple terms explained what it does is successively sampling conditional distributions of the parameters such that it converges to the true distribution over time. The method achieves this by continuously searching for the conditional probability distribution of every single word's topic allocation conditioned on the remaining topic allocations by slightly changing parameters θ and ϕ. For this reason, there is a trade-off between quality and speed when it comes to these two methods. Hence, the general expectation excluding the tuning of hyperparameters is that *Mallet (Gibbs use)* would produce more accurate estimates than *Apache Sparks' or Gensim's LDA (expectation-maximization (EM))* since, contrary to EM, it does not easily get stuck in local optima [4]. General speed improvements using distributed computing techniques have been proposed before Apache Sparks' introduction, mainly with the invention of approximation LDA and HDP variants. As first proposed by [25] these types of variants assume a weak dependency between parallel updates of two topic assignments across a large N in a given corpus. Their results show that this assumption holds, but it does come with a slight decrease in quality.

A.2 LDA vs HDP

In order to decide on if we should use LDA or HDP we compared both approaches. Table 2 showcases the results of the topic analysis modeling. Here we see that in general HDP outperforms LDA in terms of coherence score, but it returns a lot more topics of which many are very sparse and a few are so large that they are hard to interpret. HDP tends to increase the number of topics in favor of a higher coherence score as the ceiling of the maximum number of topics is lifted even higher. After running HDP on a ceiling of 150 topics the coherence score increased just slightly (approx. 0.03 points) with a number of topics to equal the ceiling for every event.

Tesla event (title)	Model	Number of topics (limit of 35)	Coherence score
Loss of CTO	LDA	32	0.50
Loss of CTO	HDP	100	0.74
Cybertruck	LDA	32	0.49
Cybertruck	HDP	100	0.70
Highest stock price	LDA	32	0.47
Highest stock price	HDP	100	0.73

Table 2: LDA vs HDP: Optimal results of topic analysis for all events



Event: Loss of CTO **Event: Cybertruck**

Event: Highest stock price

Figure 9: Coherence score of LDA, runs for max 100 topics for all three events related to Tesla Motors

Nonetheless, the HDP shows that adding many folds (e.g. 3 times) of LDA's number of topics does not result in a big gain of useful topic separation. In Table 2 we can also see that within the constraints of the algorithm (max 32 topics), it also selected the ceiling as its optimal number of topics. Figure 9 shows that LDA does reach the optimum number of topics for every event in the range between 8 and 32. In two of the events the curve seems to increase indefinitely, but the incremental gain of coherence score is extremely small and not worth the extra computing power. The exact average running time of the LDA algorithm is approximately 6 minutes, which takes 5 minutes longer than HDP. However, due to the loss of interpretability associated with HDP, LDA remains the better option for our application. Using this empirical analysis the team decided to put the ceiling for LDA in the designed application to 32 topics.