Text mining Elon Musk Twitter Account

Veselin Nasev
v.nasev@student.maastrichtuniversity.nl

May 6, 2022

Abstract

How do famous individuals influence people's opinions to trade on the stock market? This paper aims to analyze the Twitter account of one of the most famous people of this decade - Elon Musk. He has more than 80 million followers on Twitter and has posted more than 17 000 times. Different information retrieval and text mining techniques were used to explore his Twitter presence. Initially, the data was cleaned and new features were created. Firstly, parts of the speech tag were extracted and analyzed. The corpus was further clean after the PoS analysis. The next step was to extract NER(locations, organizations, people). Lastly, sentiment analysis was performed. Multiple libraries were compared for their explainability, performance, and drawbacks. Every step was visualized with easy-to-understand graphs to show the outcome of the performed steps. Elon Musk talks about various companies but he mainly mentions his companies Tesla and SpaceX. He uses simple and easy-to-remember words to express his opinion. The portion of positive tweets he posts grows each year. Several examples show that there is a correlation between his positive opinion about organizations(mainly cryptocurrencies) and the stock price. A lot of factors influence the movement of a stock but Elon Musk plays a role in it.

1 Introduction

Does the top 1% of people influence our decisions? Unarguably, one of the most influential people of the last decade is Elon Musk. He is the co-founder of Tesla[10], SpaceX[9], Neuralink[8], and The Boring Company[11]. Elon Musk is famous for many things, but he is especially well known for his online Twitter presence. With well over 82 million followers, Elon Musk is arguably the biggest influencer on the social media platform. Musk is very active on the platform, having more than 17 000 tweets. That is quite an amount of data that expresses his opinion about various topics. This paper aims at exploring, what entities he mentions in his tweets, what is his view about them, and how his view influences people's opinions.

The paper starts with explorations of the corpus. Continues with a discussion of what entities were found in the tweets. Afterward, it establishes a connection between the sentiment of the tweets and the recognized entities. Finally, the extracted information is compared to the stock price of the entities(organizations) found in the tweets. This comparison strives to find a correlation between what Elon says about an organization and how people react based on the tweet's sentiment.

2 Data

The dataset used in the paper was taken from Kaggle(link to dataset) and consists out of 13 files. Each file represents a year of Elon Musk tweets. The first dataset is from 2010 and the last is from 2022(until march). Before it was used for text mining, the dataset was cleaned and new features were created. Let's Explore how it was prepossess and visualize the main features to get a better understanding.

2.1 Cleaning and Prepossessing

The first step was to combine all the files from the past 13 years into one dataset. That lead to a file of 34878 records in it. Initially the dataset had 44 columns but most of them contained a lot of null values. After dropping them, 18 columns were left. The next step was to check the number of unique values per column. Based on that, only 3 columns were left - date, tweet, language.

After the initial cleaning further analysis of the tweet was performed. That lead to the creation of a script to clean the text of the tweets. The following issues were addressed: double white spaced, new rows, links, '@' symbols, '&'(ampersand) was removed. Finally, emojis were spotted in the corpus. Although, they can be used for sentiment analysis, for the purpose of this paper they were removed.

2.2 Feature Engineering

The dataset was enriched by creating new features from the tweet's text. The main purpose is to help better understand the data, clean it further if needed, and visualize it. The following features were extracted from the text: the length of the tweet(in characters), number of tokens in the tweet, and how many sentences does the tweet have.

2.3 Visualization

In this section the dataset will be explore by the help of visualisations. They were created using the Seaborn library in Python.

The first graph[1] shows a steady growth of the number of tweets Elon Musk posts per year. It is interesting to notice that the peak of the graph is when Covid-19 started.

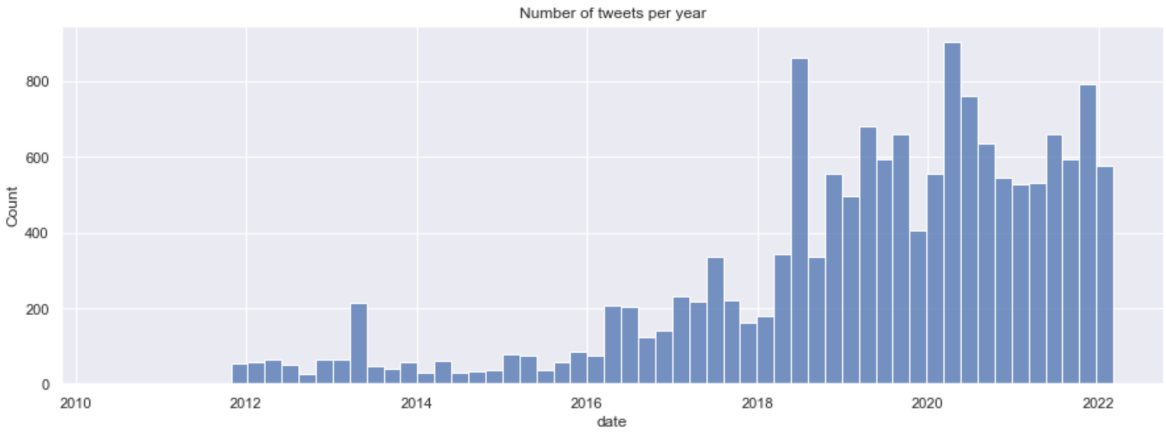

Figure 1: Number of tweets per year

The following two graphs[2] show the spread of the tweets length. The boxplot on the right shows that tweets larger than 250 characters are uncommon.

The next graphs[3] show the distribution of the newly created features(number of sentences and number of tokens). Around 75% of the tweets have less than 20 tokens and 75% of the tweets can be considered short(with not more than 2 sentences).

3 Part of Speech Tags

Famous people need to use the right words when they are talking to their audience. The verbs, noun, and adjectives will become their hallmark and people will relate certain words or phrases to a specific person. In this section, the most used words from Elon Musk will be presented. Before advancing to the exciting part, the visualizations, and the extraction process will be discussed.

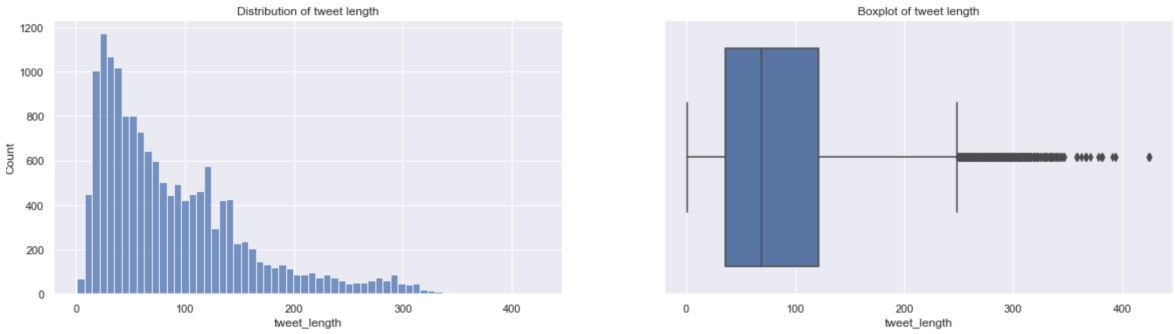

Figure 2: Spread of tweets length(in characters)

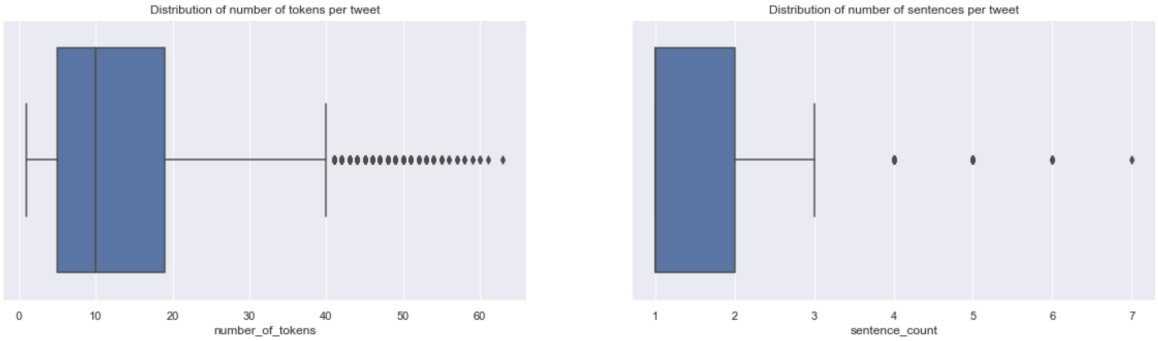

Figure 3: Distribution of number of tokens(left) and number of sentences(right) per tweet

3.1 Techniques

Two very popular NLP libraries were compared for Part of Speech(PoS) tagging, NLTK [3] and spaCy [5]. NLTK was created in 2001 at the Department of Computer and Information Science at the University of Pennsylvania. On the other hand, spaCy was developed in 2015 by Matthew Honnibal and Ines Montani(the founders of Explosion AI).NLTK provides more algorithms for NLP tasks and it is simpler to understand the backbone of a specific algorithm. In contrast, spaCy offers more complicated algorithms that promise to have the best performance possible.

Let's first apply NLTK to find the PoS tags of the following sentence: "Please book my flight to California". The results can be found on Figure 4.

0	Please	NNP
1	book	NN
2	my	PRP$
3	flight	NN
4	to	TO
5	California	NNP

Figure 4: Results of applying PoS tagging using NLTK

It can be noticed that the word "book", was classified incorrectly as a noun, and the word "Please" was not classified as a verb. What can be concluded from this example is that the NLTK algorithm can fail to consider the context when determining the PoS tag. NLTK uses a pre-trained model called greedy averaged perceptron. The model was trained and tested on the (Wall Street Journal corpus).

Now we are going to look at the results 5 from spaCy post tagging algorithm applied to the same sentence as above.

0	Please	INTJ
1	book	VERB
2	my	PRON
3	flight	NOUN
4	to	ADP
5	California	PROPN

Figure 5: Results of applying PoS tagging using spaCy

In this case every word from the sentence was classified correctly. SpaCy uses a pre-trained pipline of statistical models to make the final prediction. The data used for training is written text (blogs, news, comments) from the web. From the example it is visible that the context is taken into account.

Although, spaCy is secretive about what are the exact models used in their PoS tagging pipeline, the different in accuracy is noticeable. More examples with hard to classify words were used to test the two libraries and the final choice was the spaCy will be used. The results presented in the visualization section will be generated using spaCy.

3.2 Visualizations

Analyzing most used nouns 6 in Elon Musk tweets we can see several words that stand out - car, rocket, year, time, etc. Most of the words are related to his two main companies - Tesla and SpaceX and year and time could be related to the announcements of new updates.

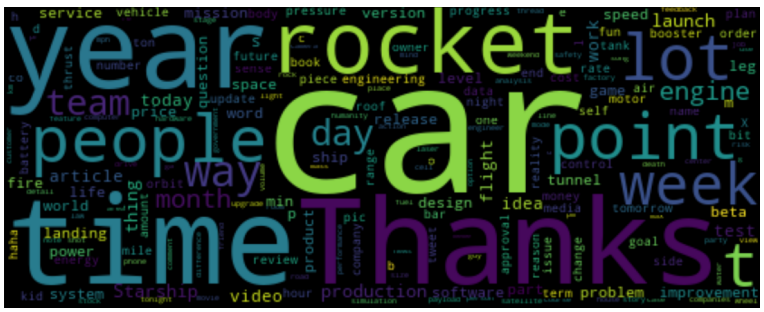

Figure 6: WordCloud of most used nouns in Elon Musk tweets

We can see that Elon focuses on using simple adjectives 7 in his tweets. That could be because using simple adjectives are easy to remember and could be more influential because of that reason.

The verbs used in Elon's tweets are also simple verbs that everybody knows and are easy to remember. Papers supporting the statement "simple/common/short words are more influential" was not found but it an interesting research topic for other faculties.

4 Named Entity Recognition

In this section we are going to look at what entities Elon Musk mentions the most in his tweets. The main focus of the paper is to discover organizations mention in the tweets and compare them to their stock prices. However, locations and people will be extracted as well. Before dive into the techniques used for Named Entity Recognition(NER), prepossessing was perform to increase the quality of NER. Personal pronouns and coreference were address and resolved.

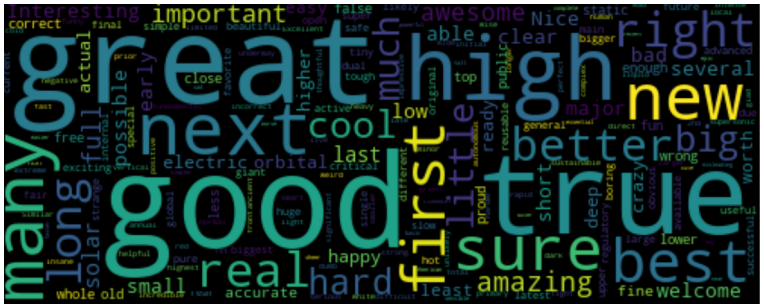

Figure 7: WordCloud of most used adjectives in Elon Musk tweets

Figure 8: WordCloud of most used verbs in Elon Musk tweets

4.1 Resolve Personal Pronouns

Performing PoS tagging not only helped us explore what words Elon Musk uses but it was a useful step to find possible issues with the corpus. The personal pronouns were explored and it was noticeable that Elon uses them quite often 9.

Figure 9: WordCloud of most used pronouns in Elon Musk tweets

A decision was taken to replace "I" and "me" with "Elon Musk".

4.2 Resolve Coreference

The second issue and the one that is harder to resolve is coreference. In the Data Section we explored that substantial portion of the tweets have more than two sentences. With further exploration it was observed that Elon Musk often uses the pronoun, when referring to an entity in the second sentence of his tweet. Based on that finding a technique to resolve coreference was explored.

The library used to resolve the issue is AllenNLP. AllenNLP is PyTorch based library developed by the Allen Institute of AI. The model they use to resolve coreference is called "End-to-end Neural Coreference Resolution" [7]. The model computes span embeddings that combine context-dependent boundary representations with a head-finding attention mechanism. It is trained to maximize the marginal likelihood of gold antecedent spans from coreference clusters and is factored to enable aggressive pruning of potential mentions.

Let's take an example and see how AllenNLP coreference resolver performs. The following sequence of sentence "Veselin is a master's student at DKE. He is currently participating in the IRTM course. His passion is NER." has two coreferece issues - he and his. The following sequence is the output of the model: "Veselin is a master's student at DKE. Veselin is currently participating in the IRTM course. Veselin's passion is NER.". The model was able to flawlessly solve all the issues.

4.3 Techniques

After the two prepossessing steps the data was ready for NER. Two NLP libraries were compared for their NER models - Flair [2] and Hugging Face. Hugging Face is open-source and platform provider of machine learning technologies. The company is based in New York and it was found in 2016. Flair is simple framework for state-of-the-art NLP. It was developed by by Humboldt University of Berlin.

Let's take one of Elon Musk most popular tweets in 2020 as an example "You can now buy a Tesla with Bitcoin". The most essential part for this research is to extract the organizations, which in this case are "Tesla" and "Bitcoin". Let's compare the libraries and inspect their output.

Flair uses a pre-trained model on the Conll-03 corpus. It uses it own embeddings called Flair embeddings [1]. The model can predict 4 classes - organization, person, location, misc. The model has an F1 score of 93% on the test data. The result from the model on the example sentence can be found on Figure 10.

```
Span[5:6]: "Tesla" → ORG (0.9929)
Span[7:8]: "Bitcoin" → ORG (0.6285)
```

Figure 10: Flair NER prediction for "You can now buy a Tesla with Bitcoin"

Hugging Face uses Bert Model with a token classification head on top (a linear layer on top of the hidden-states output). The model is trained and tested on the same corpus - Conll-03 and predicts the same 4 classes. The result form the model on the example sentence can be found on Figure 11.

	entity	score	index	word	start	end
0	I-MISC	0.992528	6	Te	18	20
1	I-MISC	0.959386	7	##sla	20	23
2	I-MISC	0.842477	9	Bit	29	32
3	I-ORG	0.487922	10	##co	32	34

Figure 11: Hugging Face NER prediction for "You can now buy a Tesla with Bitcoin"

The two models were compared with more examples and they tend to yell similar results. The main difference is what embeddings they use. BERT embeddings are better in general because they divide longer words in smaller sub parts. That can help the model to find words that are not in the initial vocabulary. However, for the purpose of NER we don't expect non existing organizations, people, and locations. Both libraries were used but Flair yelled better results.

Hugging Face classified big part of the entities under the category MISC. On the other hand, Flair was able to recognize more organizations and it was even able to recognize Tesla model names(Model S, Model 3, Cybertruck) and SpaceX rocket names(Dragon, Falcon).

4.4 Visualizations

The visualizations in this section will be the result of applying Flair model for NER on the corpus. The first graph 12 shows the most frequent entities classified as MISC. Most of the words represent Tesla models, SpaceX rockets and other products that Elon Companies produce.

The next graph 13 show the most frequent locations that Elon mentions in his tweets. It is interesting to see that he mentions Mars as often as Earth. That is expected because Elon's biggest dream is to land on Mars before he passes away.

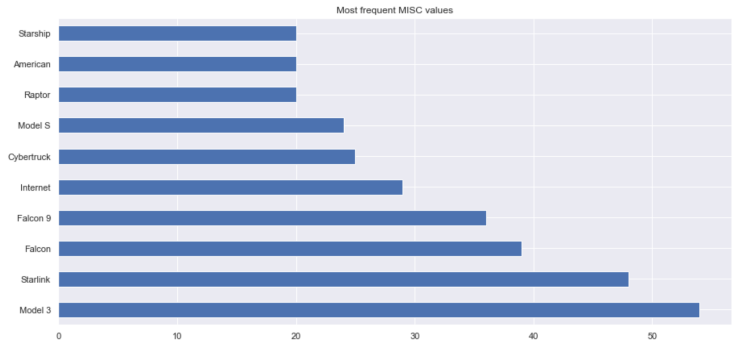

Figure 12: Most frequent entities classified as MISC

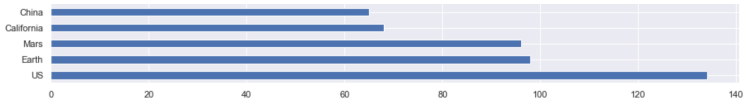

Figure 13: Most frequent entities classified as LOC

The last graph 14, and the most important one for this paper shows the most frequent organizations. Elon mainly talks about his two prime companies Tesla and SpaceX. However SpaceX is a private company and it is not traded on the stock market. Because of that, in the next sections, only Tesla stock will be compared to the extracted sentiment.

Figure 14: Most frequent entities classified as ORG

5 Sentiment Analysis

The last text mining step in this paper is to extract the sentiment of ELon's tweets. Two method will be used for that purpose - Vader [6] and Flair. The sentiment of the tweets will be classified as negative, positive or neutral. The above mentioned methods will be compared and Flair's predictions will be manually validated.

5.1 Techniques

Vader is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. One benefit of Vader is that it works a lot faster compared to other pre-trained machine learning models. However, the model has a difficulty dealing with negation. This can be seen when analyzing the results 15. Only 14% of the tweets were classified as negative. Later on we will compare this ratio with Flair's results and see the difference.

The second technique used for sentiment is Flair a pre-trained classifier. Flair's sentiment classifier is based on a character-level LSTM neural network which takes sequences of letters and words into account when predicting. The model was trained on the IMDB movie review data, that consist of 50,000 reviews. The dataset contains an even number of positive and negative reviews. Now, we can compare the results of Flair to Vader. As we previously explored using Vader the ratio of negative tweets was 14%. On Figure 16 we can see what Flair predicted. The ratio is completely different - the negative tweet ratio is around 40%. For further comparison of how much the two techniques agree with each other we will use Kappa [4] statistic. The score is merely 0.16, that shows that the two models barely agree with each other.

```
pos     7250
neu     5225
neg     2080
```

Figure 15: Predictions summary of Vader sentiment analysis

```
pos     8849
neg     5706
```

Figure 16: Predictions summary of Flair sentiment analysis

5.2 Validation

In this section Flair sentiment classifier results will be validated for accuracy. For this purpose a sample of 50 sentences were manually labeled and the precision, recall, specificity, and F1 scores were calculated as evaluation metrics. From the table 5.2 below we can see that precision is higher than recall which shows that the high percentage of the predictions were correct. On the other hand, specificity is significantly lower which show that the model struggles with predicting the negative class. Further manual exploration showed that when a tweet consist out of several sentences and it starts with a negative sentence but eventually ends on a positive note, is classified wrong. For example, *'As mentioned earlier this year, cost of the Tesla FSD option will increase every few months. Those who buy the Tesla FSD option earlier will see the benefit.'*, the following sentence starts negative(the cost of option will increase) but ends with a positive statement that people who buy them early will benefit. The sentence was classified as negative but it can be argued that it is actually positive.

Metric	Score
Precision	0.94
Recall	0.85
Specificity	0.8
F1	0.89

6 Results

In this section we will explore how Elon Musk tweets influence peoples opinion. Several examples will be shown where Elon tweeted something positive about an organization and the price increased significantly for a short time. Lastly, the sentiment over time will be compared to the Tesla stock price.

Elon Musk himself is a bug fan of cryptocurrencies, Tesla acquired 1.5 billion of Bitcoin in 2021. He often tweets about cryptocurrencies and let's look at two example how it influenced them. After acquiring such a big amount of Bitcoin he tweeted that Tesla will accept Bitcoin as payment method. On figure 17 we can see that the price increase shortly and then dropped again. There is a correlation but we can not conclude that it is the only cause of the increase.

The second example is from a meme cryptocurrencies, Elon himself is a big fan of meme cryptocurrencies, especially Dogecoin. In April 2021 he tweeted that Dogecoin is the people's crypto. That lead to a huge increase of the price from 0.35 to 0.5 cents. That is roughly 43% increase in just one day. Figure 18 shows the price movement before and after the tweet.

Next let's have a look at the price movement of Tesla price and the sentiment ratio of positive/negative. Both graphs will be from 2016 until 2022. The purpose that it is not from 2012 when Elon Musk started using Twitter is that, he was not really active before 2016. He tweeted around 100 times before 2016 and suddenly he reached 617 tweets in 2016. Since then his Twitter presence was growing and in 2021 he over 3000 tweets. The first graph 19 shows a steady increase from 2020 and the second graph 20 shows that the ratio of positive tweets started increasing around 2018. Could it

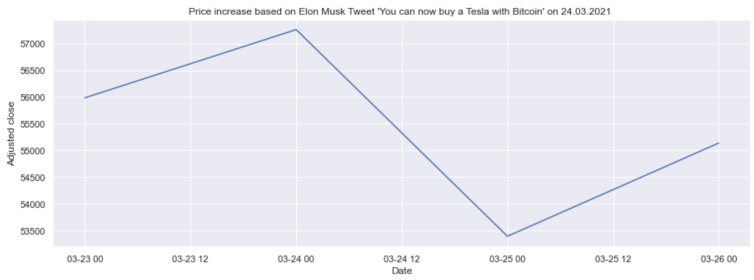

Figure 17: Price increase based on Elon Musk Tweet 'You can now buy a Tesla with Bitcoin' on 24.03.2021

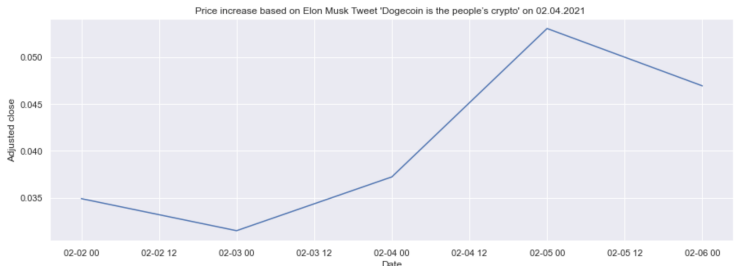

Figure 18: Price increase based on Elon Musk Tweet 'Dogecoin is the people's crypto' on 02.04.2021

be that in 2018 Elon knew that his company will be successful and started to post more positive tweets or is it a random correlation?

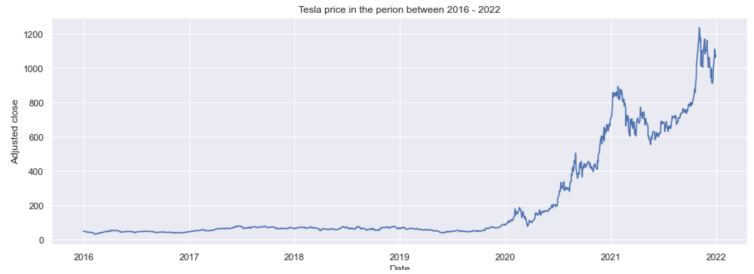

Figure 19: Tesla price in the period between 2016 - 2022

7 Conclusion

The paper showed different information extraction and text mining techniques to analyze tweets. The paper focused on Elon Musk's tweets from 2012 to 2022 but the techniques can be used for any other corpus. Initially, simpler information was extracted from the tweets. The paper gradually moved to more complicated techniques and compared different tools for each step. The techniques that yelled best result were used at the end and multiple visualizations were used to show the results of every step.

The result shows that there is some correlation between Elon's positive opinion about trending organizations, especially cryptocurrencies, and the rise of the stock price of the organizations. However, it was noticeable that the effect of his tweets are for a short time because the stock price quickly went to its previous price. Based on the analysis it would be hardly possible to predict the stock price of an organization solely on what Elon Musk tweets. He influences people's opinions but the stock price is influenced by various factors and he is one of those factors.

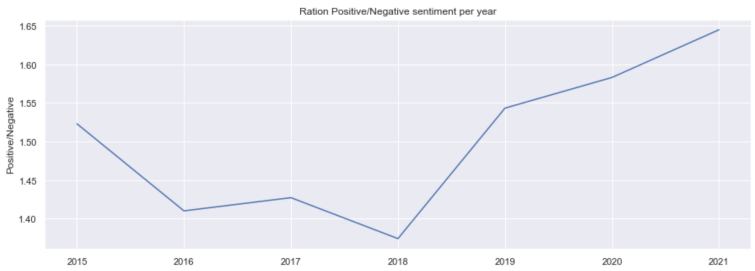

Figure 20: Ration Positive/Negative sentiment per year between 2016 - 2022

8 Future Work

In future work, other sources of information can be analyzed and compared to stock prices. When several main sources of information that correlate with stock prices are identified then a machine learning model could be built. With enough sources and proper text mining techniques to extract important features from them, a dataset could be built. This dataset can be used to train a regression model that can predict the stock price. It will be a difficult challenge but it will be a worthwhile experience to attempt to build such a model that can predict the stock price of a certain organization based only on online text sources. Is it possible to build such a model, it is interesting to test if it can be used to help people to trade more efficiently. The model can be used as a helper tool because it will make informed decisions based on the latest information on the internet.

References

[1] A. Akbik, D. Blythe, and R. Vollgraf. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649, 2018.

[2] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, 2019.

[3] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[4] B. D. Eugenio and M. Glass. The kappa statistic: a second look. *Computational Linguistics*, 30 (1):95–101, 2004. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/089120104773633402. URL http://portal.acm.org/citation.cfm?id=1005385&dl=#.

[5] M. Honnibal and I. Montani. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. "", 2017.

[6] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar, P. Resnick, M. D. Choudhury, B. Hogan, and A. H. Oh, editors, *ICWSM*. The AAAI Press, 2014. ISBN 978-1-57735-659-2. URL http://dblp.uni-trier.de/db/conf/icwsm/icwsm2014.html#HuttoG14.

[7] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *EMNLP*, 2017.

[8] Wikipedia. Neuralink — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Neuralink&oldid=1085160595, 2022.

[9] Wikipedia. SpaceX — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=SpaceX&oldid=1085322081, 2022.

[10] Wikipedia. Tesla, Inc. — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Tesla%2C%20Inc.&oldid=1085358178, 2022.

[11] Wikipedia. The Boring Company — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=The%20Boring%20Company&oldid=1085314333, 2022.