University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Nikolai Rol

Twitter sentiment analysis to estimate happiness level

Master's Thesis (30 ECTS)
Innovation and Technology Management Curriculum

Supervisors:

Rajesh Sharma

Tartu 2020

Abstract:

Twitter sentiment analysis to estimate happiness level

Happiness is something that people strive for. However, it has always been hard to measure and understand what happiness depends on. This paper investigates if sentiment analysis can be used to estimate how happy people are and if sentiment correlates with socioeconomic factors or with the news. For analysis, text processing techniques were applied to Twitter posts gathered over the period from November 2019 to May 2020. The study shows a weak correlation with socio-economic factors, whereas the strongest relationship was with Health Care Quality. After a closer look into the change in daily sentiment, it was found that certain topics were discussed more than others on the dates with peaks. To investigate this aspect, the correlation analysis between sentiments of Twitter posts and news was made, however, the coefficient appeared to be low. The conclusion is that the result of sentiment analysis over Twitter data does not show a high correlation with socioeconomic factors, but it might have a certain dependency on events, news, or global shocks.

CERCS: P176 Artificial intelligence

Keywords: sentiment analysis, happiness, correlation analysis, socio-economic factors

Eesti firmade palgalise ebavõrdsuse uurimine

Inimesed püüdlevad õnnelikkuse poole, kuid seda on alati olnud raske mõõta ja õnnelikkuse põhjuseid hinnata. Käesoleva töö eesmärk on uurida, kas sentimendi analüüsi saab kasutada õnnelikkuse hindamiseks ja kas sentiment korreleerub sotsiaalmajanduslike faktorite või uudistega. Analüüsiks kasutati tekstiprotsessimise tehnikaid Twitteri postituste peal perioodil november 2019 kuni mai 2020. Uuring näitab nõrka korrelatsiooni sotsiaalmajanduslike oludega, samas on tugevaim seos tervishoiu kvaliteediga. Peale põhjalikumat uurimist päevase sentimendi muutusesse, leiti et kindlaid teemasid arutati rohkem suurema postitusaktiivsusega päevadel. Korrelatsioonianalüüsi käigus leiti, et Twitteri postituste sentimendi ja uudiste vaheline seos on nõrk. Seevastu võib sentimendil olla seos igapäevaste sündmuste ja globaalsete šokkidega.

CERCS: P176 Tehisintellekt

Keywords: sentimendi analüüs, õnn, korrelatsioonianalüüs, sotsiaalmajanduslikud tegurid

Contents

1 Introduction

Happiness is something that people strive for. It is hard to measure and currently, it is mainly estimated by conducting a survey which aims to get a social opinion on how happy people are. With the rise of Online Social Networks (OSN) in recent years, it has become easier for communities to share their opinion in the form of publicly available blog posts on the internet. It has led to the rise of the field of sentiment analysis, which extracts and classifies the emotion of the text. There were several papers on this topic just in the past year: comparing and improving the performance of supervised and unsupervised methods; more papers using sentiment prediction to investigate the socio-economic aspects (more about previous studies in Chapter 4). It is still a question of whether sentiment analysis can help with measuring happiness. More research applies sentiment analysis methods specifically to learn how happy people are based on their posts in OSN.

Quercia et al (2012) [26], have conducted research that investigated the happiness of London habitants analyzing Twitter posts. They validated the findings against the Index of Multiple Deprivation which is a proxy to relative prosperity for London communities. Their research has found a significant correlation of r = 0.35, which led to the conclusion that the analysis of tweets can be an accurate and cost-effective tool for measuring happiness. A lower correlation was found in the research of Kramer, A. D. (2010) [14] conducted in the USA. The paper studied if the respondents' questionnaires correlate with the sentiment of their status update on Facebook. The researcher found that those who evaluate themselves as more satisfied with life - score higher on positivity score of Facebook status. Although the questionnaire was significant as a predictor, the correlation was 0.17 which is lower than in the previous paper.

Although some research has been carried out on the analysis of the sentiment of a single location (such as a country, city), no studies have been found which conducted cross-country analysis to identify if the sentiment is dependant on multiple socio-economic factors. It raises a question if the results of different papers addressing various countries are comparable. This paper aims to close this gap and to investigate if sentiment analysis provides results correlating with socio-economic factors or news.

To achieve this goal, blog posts from the USA, the UK, Australia, and India are collected over 6 months. The reason to pick these countries is that they are predominantly English speaking which makes generalization of findings more reliable. Posts from Twitter were collected because it is a small piece of text content saturated with the opinion or emotions of the author due to a limit to 280 characters per post. To compare the results of sentiment analysis against the real world, the data for 8 socio-economic factors such as Health, Education, Economy, Infrastructure, Opportunity, Fiscal Stability, Crime and Environment was collected, to analyse if positivity level of Twitter users' posts depends on the development of environment. Also, the news articles for 50 days were collected to understand if the sentiment of news correlates with the sentiment of Twitter posts.

This project provided an important opportunity to advance the understanding of sentiment analysis application, particularly in happiness prediction. It attempts to estimate the happiness levels of 4 countries; analyse the potential issues and risks associated with such study; investigate what aspects influence happiness level based on correlation analysis with socio-economic factors, topic modelling and comparison against news. It is beyond the scope of this study to propose the best performing method to estimate happiness. It can be a topic of future papers in this field.

This paper consists of the following paragraphs: Section 2, background, introduces the sentiment analysis and Online Social Networks as a source of data. Section 3, related work, discusses what methods are available in sentiment analysis and what research was done previously on the related topics. Section 4, dataset, covers the collected posts from Twitter, its structure and data preprocessing. Section 5, methodology, describes the research design and evaluation of results. Section 6, results, presents outcomes and discusses the findings.

2 Background

This section defines what is sentiment analysis, how it can be used, and what is the OSN. It starts with explaining ideas of polarity, subjectivity, level of analysis in other words scope, and different applications of sentiment analysis. Lastly, it explains why sentiment analysis became so popular in the 21st century and what is OSN.

2.1 Sentiment analysis

Sentiment analysis is a field of study which focuses on analyzing opinion or related concepts such as emotions, feelings and others from digital text. Commonly the purpose of sentiment analysis is to categorize textual documents into either positive or negative. For that sentiment analysis benefits greatly from machine learning algorithms and is driven by academic research and private companies. It got its development in early 2000 with the development of online platforms that are rich with an opinion based digital text.

As noted by Liu, B. (2012) [16] a sentiment can't be extracted from any text, only subjective, as it bears judgment, opinion or emotion of the author. Therefore sentiment analysis can also be named as subjectivity analysis or opinion mining.

B. Pang and L. Lee (2004) [23], have proposed an idea of a two-step analysis approach, where the first step is identifying objective text, not bearing any opinion, and then applying polarity analysis on remainder text, in other words, subjective text. In the same vein, B. Pang and L. Lee (2008) [24] noted that it is also possible to skip subjectivity analysis and classify opinions into three categories of positive, negative and neutral, where neutral bears no sentiment.

2.1.1 Sentiment analysis scopes

It is possible to analyze the opinion of the text on different levels, which affects the generalization of results and to what extent it is possible to track the emotion and find its association in the text. Most commonly known levels are document, sentence and aspect:

- The document level analyses an opinion on the whole body of the text. Kolkur et al (2015) [13] point out that this approach assumes that text contains opinion about only one object. Therefore can show sceptical results on forum or blog texts as they tend to contain discussion more than about one object.

- The sentence-level analyses an emotion state of each sentence in the text body. As noted by Kang Wu et al (2013) [36] this level provides enough level of detailing to analyse the blogs as they tend to be limited in characters and generally concise. Several studies thus far have used sentence-level analysis when orienting the polarity of the blog text.

- The aspect level, also named feature level, allows tracking the emotion to a specific topic. As Neha S. Joshi et al (2014) [12] state: "Instead of looking at language constructs (doc-

uments, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of sentiment (positive or negative) and a target (of opinion)." This approach allows examining the attitude toward a specific object mentioned in the text. Salas Zarate et al. (2017) [30] applied this logic in the analysis of sentiment specifically related to diabetes. Using the higher level would bias the results as tweets might have contained not the only opinion about diabetes.

2.1.2 Sentiment analysis types

There are a few types of sentiment analysis which are subject to research and application in the private sector, namely (Liu, B. (2012)) [16]:

- Fine-grained sentiment analysis is used to identify the polarity of the text, which is either positive or negative. Research such as that conducted by Chowdary, S. J. S. (2019) [11] show that this classification can be more complex as grading movie comment on a scale from 1 to 5.

- Emotion prediction aims to identify an emotional state presented in the text. A number of studies have found various applications for this type. Previous papers, relevant to this thesis, such as that of Lewis Mitchell at el. (2013) [20] analysed the tweets of the USA to identify the correlation between word use and how happy the state is.

- The intent analysis identifies what intention is present in the text. As noted by Chowdary, S. J. S. (2019) [11] this type of sentiment analysis has found its application in the customer support area to understand what the client needs help with.

- Aspect-based sentiment analysis is generally used in the product analytics field as it allows to identify the sentiment specifically about an object from the text. Al-Smadi, M. et al (2019) [5] presents how to use aspect-based analysis to identify the opinion of users towards room price and hotel location.

For the purposes of this research, emotion prediction was selected to analyze the happiness status of residents of different countries.

2.2 Data source Online Social Networks - Twitter

The source of data for sentiment analysis is generally a text in digital form according to Liu, B. (2012) [16]. The rich-with-opinion and emotion text became more available with the development of the internet at the beginning of 2000s and has been growing since then. Over the past decades most research in sentiment analysis has used the following categories of data source:

- **News.** Many publishers have moved online presenting their overview of the day's agenda. Such information is valuable when analysing the Much of the current literature that uses

the news as a data source focused on applying a fine-grained approach. Some studies have examined this source in predicting the stock price movement (e.g. Mohan, S. et al, 2019 [21]; Li, X. et al,(2014) [15]; Ahmad, K., et al (2007) [3]).

- **Reviews.** Sharing opinion online has become a norm, many companies have found a way to adapt it to one's benefit. In recent years, there has been an increasing amount of literature on sentiment analysis using product review data (e.g Fang, X. et al, (2015) [8]; De Albornoz et al (2011) [6]). Because of it the companies can have a better understanding of what features the end-users like or dislike about the product and improve on it. However, the reviews are not limited only to product, but also movie (e.g Singh, et al, (2013) [32]; Thet, et al, (2010) [34]). In this case, the aspect-based approach is greatly practised as it gives a precise insight into what exactly the comment is about and which aspects are positive and which are negative.

- **OSN.** According to Richter, A. et al (2008) [28] Online Social Networks (OSN) virtually connect people on the grounds of their relationships on the web-based platforms, allowing users to share their experience, communicate with each other as well as consume content. It has found a broader application in the world compared to the above mentioned as it is not focused on a specific niche (movie, news, product review page). A considerable amount of literature has been published on sentiment analysis using OSN. These studies differ being in law enforcement (Subramaniyaswamy, et al (2020) [33]), health supervision (e.g Salathé, M., et al (2011) [31], Rodrigues, et al (2016) [29]), the focus of this paper - the emotional status of the communities (examples of papers in Chapter 3), and so on.

Examples of OSN can be Facebook, Twitter, Instagram, and many others, varying in their functionality. While Instagram is predominantly majors in the niche of users sharing pictures, Twitter has covered the domain of microblogging, which allows users to publish a text of 280 characters on their account page. The message that the users put into their posts can vary from sharing life-changing events to covering a global agenda. This research will mainly focus on Twitter as a source of rich-with-opinion and emotion text. Because the text is short, it has multiple advantages:

- Users' opinion is delivered straight to the point manner, due to the limited number of words;

- Short texts make it easy to analyze a large number of posts;

- Still being long enough they are more opinion and emotion descriptive compared to shorter text forms present on OSN, like a status update.

Most of OSN share common features, an important one for this paper is geolocation, it allows to locate the user or post if the user gave geo access to OSN. As many users do not use

geolocation on Twitter, it is complicated to gather multiple tweets for a specific location as the most portion of posts is not geotagged.

3 Related work

This section gives an overview of the papers done in the past on the topic of sentiment analysis. The review is given in the form of grouping papers based on the methods of sentiment analysis they applied.

3.1 Learning Algorithms

Over the span of its years, the sentiment analysis has developed a set of approaches to opinion mining. This section presents examples of papers where different approaches were used. Previous papers divided the methods into 3 distinct groups (Neha S. Joshi et al (2014) [12]; K. Ahmed et al (2015) [4]): Supervised Learning Algorithms most commonly consider sentiment analysis as a two-class classification problem where the text is either positive or negative. The fewer groups to be classified in, the easier it is to run the analysis. General supervised learning algorithms are used in this approach: Naïve Bayes, K-Nearest Neighbor, Decision Tree, and Support Vector Machines. Unsupervised Learning Algorithms refer to the fact that this approach has a specific set of rules to how the sentiment of the text is extracted. Words and phrases are assigned a certain value depicting their orientations and strength. Then based on how frequently words are used the sentiment score is computed. Hybrid Algorithms combine in itself both supervised and unsupervised learning techniques. It was shown in multiple studies that hybrid techniques tend to show better results, though it is more complex to perform.

3.1.1 Supervised learning

As sentiment analysis comes from machine learning, it applied the techniques from that field as well. After data is cleaned, the dataset is split into training and testing sets. The classifier learns from training data and builds a model to classify the text of the testing set. It is speculated that the accuracy, measure of how many texts were predicted correctly, tends to be higher than one of unsupervised learning. The classical example of supervised learning can be the Naive Bayesian text classifier. As mentioned by Vinodhini, G. et al (2012) [35]: "The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories." As the authors noted the model assumes that words are independent, which makes the classifier more efficient.

Mihalcea, R. et al (2006) [19] classify blog posts from LiveJournal into happy and sad using Naive Bayesian. They used the classifier to see how well the moods happy and sad can be separated by the content. After running an experiment it gave an accuracy of 79.13%. It showed to be a good result compared to 50% associated with the naive baseline of using one mood assignment by default. However, the same research points out that other classifiers, Support Vector Machines and Rocchio, did not show significant differences.

The other method, Support Vector Machines (SVM) even though it gave lower results in

the study above, still is considered to be a robust approach in sentiment analysis. As given by Vinodhini, G. et al (2012) [35]: ”Based on the structural risk minimization principle from computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set.”

In their research, Abdelwahab, O. et al (2015) [1] compared the performance of SVM against Naive Bayes in terms of accuracy of sentiment analysis performance on different training set sizes. The results showed that SVM has achieved an average of 75% accuracy across all 10 training sets, whereas Naive Bayes similar results only on 9 training sets. Interestingly, Naive Bayes showed the lowest result on the size of the training set equaling 10%. The results imply that SVM can be used on the data sets not requiring a lot of manual labelling.

3.1.2 Unsupervised learning

Unsupervised learning, also called lexicon analysis, requires to have a lexical dictionary containing a set of words with an orientation of polarity and its magnitude. The words in the text are evaluated based on the dictionary and then the text is given a sentiment score based on an average of the word scores. The advantage of it is it does not require a huge set of labelled training data on contrary to supervised learning.

Commonly the results of the unsupervised approach depend on which dictionary is used in the research. For this reason, papers are comparing the available dictionaries. As an example Cataldo Musto et al. (2014) [22] have conducted a comparative analysis on 4 lexical resources: SentiWordNet, WordNet-Affect, MPQA, and SenticNet. This research did not show the dominance of any of the resources.

One of the lexical resources listed above, WordNet, was used in the study of Godbole N. et al. (2007) [9] where they have used unsupervised sentiment analysis to classify Greek Tweets to following categories: “Anger”, “Disgust”, “Fear”, “Happiness”, “Sadness”, and “Surprise”. Since they studied if sentiment analysis can identify the intensity of emotion, their results showed acceptable results mainly to “Happiness” and “Fear”. They have validated the results using the judgment of human raters.

There is another way to evaluate the performance of the unsupervised approach, such as comparing the results to publicly available statistics. One study by Lewis Mitchell et al (2013) [20] examined the results of their experiment comparing against public measures of well-being. The research analysed the happiness of states in the USA using tweets. Using Pearson correlation the authors have reached a significant correlation between wellbeing and happiness scores over most states.

3.1.3 Hybrid

It is believed that Supervised learning most often provides better results. It can be evident in studies such as that of Abdulla, N. A. et al. (2013) [2] where SVM and Naive Bayes models outperform the self-built dictionary. The accuracy of the models was on average above 80%, whilst the lexical approach gave a result of about 50% accuracy. However, it is worth noting that it was an Arabic dictionary built specifically for that study. English dictionaries tend to perform better.

As there are advantages to each method, papers are analyzing the option to combine both. In their paper, Zhang, L. et al (2011) [37], propose the following method: "The method first adopts a lexicon-based approach to perform entity-level sentiment analysis. This method can give high precision, but low recall. To improve recall, additional tweets that are likely to be opinionated are identified automatically by exploiting the information in the result of the lexicon-based method. A classifier is then trained to assign polarities to the entities in the newly identified tweets. Instead of being labelled manually, the training examples are given by the lexicon-based approach." The accuracy of the model appeared to be 85.4% compared to other methods where only one has reached 80%.

Based on the above studies it is evident that even though supervised algorithms seem to outperform the unsupervised approach - models must be chosen based on the goal and condition of the research. Unsupervised learning requires less manual work to label the training data. It is especially important on large data sets. Supervised learning is believed to provide more accurate results. Combination of both approaches is still not a fixed area. There is still room for development and no method is undoubtedly preferred over another.

3.2 Gap

A considerable amount of literature has been published on the focus of this paper, namely sentiment analysis to estimate happiness. One study by Kramer, A. D. (2010) [14] focused on estimating happiness of the USA by analysing Facebook user updates. Collected required dataset was analysed using unsupervised learning. The results were compared against the results of the survey, filled out by the same Facebook users. The Pearson correlation showed a significant correlation of 0.17, which is generally perceived as low.

Lewis, M. et al (2013) [20] examined nearly the same topic. They analysed happiness in the USA, however, made a step more and made a study on the level of each particular state. The results of lexicon-based sentiment analysis were compared against indexes consisting of multiple socio-economic factors, which showed Spearman's correlation being significant in 4 out of 5 factors shown in Figure 1.

Building upon the study of Kramer, A. D. (2010) [14], Quercia et al (2012) [26] have made a similar research, however, the location was changed to the city of London. They have used a similar approach, unsupervised learning, however additionally Maximum Entropy was used for

comparing. The results were compared against socio-economic factors similar to how Lewis, M. et al (2013) [20] validated the outcomes. The correlations were significant at 0.35 and 0.365 for unsupervised and supervised approaches respectively.

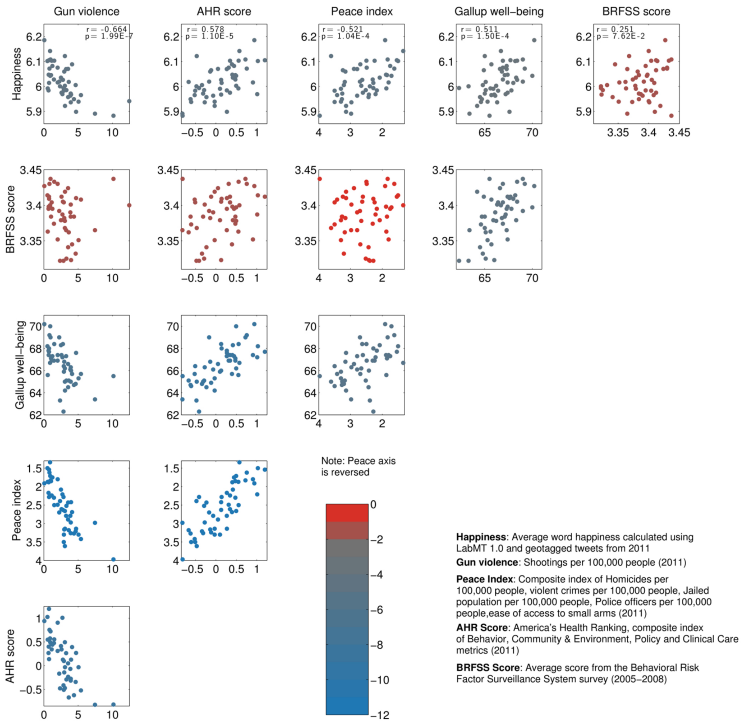

Figure 1: Scatter plot matrix of correlations between different socio-economic factors. Retrieved from: Mitchell, Lewis Frank, Morgan Harris, Kameron Dodds, Peter Danforth, Christopher. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. PloS one. 8. e64417. 10.1371/journal.pone.0064417.

Based on the all above mentioned, it is evident that the majority of the previous research on the estimation of happiness by sentiment analysis, and in the sentiment analysis field in general, focus on one location. There is a lack of cross-country analysis. As previous work approached the experiments differently - it is hard to compare the results of different countries made from different studies. This paper aims to do a comparative study of sentiment analysis applied to Twitter posts in different countries to measure happiness.

4 Dataset

This chapter describes the data that will be analyzed. It gives an understanding of what were the criteria for data selection, what is the format of data and what steps were taken to preprocess text.

4.1 Data Collection

The guiding principles for data collection were based on previous papers, with some additions. The aim of this paper and the main difference from previous papers are data was collected over a period of approximately 6 months.

The research focuses mainly on the USA, as Twitter is widely used in this country, but also such countries as, the UK and Australia were also included. The reason for such a sample is that the English language is widely spread across these countries and there is a high degree of cultural proximity. Also, they are representative of different regions of the globe, such as being North America, Europe, Asia and Oceania.

A number of previous papers have used the keywords to narrow the selection process of tweets. For this reason, a list of keywords was compiled to narrow the list of tweets. The main keywords were happy, sad, peaceful and anxiety. The other words were picked as synonyms to the previously mentioned and which have the most frequent use in the past decade according to Google Books Ngram Viewer. Such words are as following: cheerful, glad, delighted, blessed, bitter, melancholy, pessimistic, sorry, calm, neutral, quiet, smooth, steady, concern, doubt, panic, suffering.

The source of data is Twitter posts, tweets, as they are full of opinion rich text. The data from Twitter was extracted using Twitter Streaming API. The API has certain limitations: not all existing tweets that contain given keywords are allowed to be collected; a number of allowed filters are 400 keywords, 5000 user ids and 25 location boxes; one filter rule is applied on one allowed connection, therefore some part of messages was not recorded.

The data was collected over the period from 10.11.2019 to 26.05.2020, with total dataset size 6 127 583 tweets. Which makes it one of the largest datasets analysed on this topic to the author's knowledge. The number of tweets is divided among 4 countries as shown in Table 1.

Country	**Absolute Number of Tweets**	**Relative Number of Tweets**
USA	4 770 091	77.8%
UK	1 193 556	19.4%
Australia	121 744	1.9%
India	42 192	0.6%

Table 1: Distribution of tweets among 4 countries

As Table 1 shows, 77.8% of tweets were collected from the USA, whereas the UK, Australia

and India have smaller shares. The reason for it is that Twitter is predominantly used in the USA and has lower penetration in other countries. However, it allows investigating and comparing if sentiment analysis was able to spot large emotional shift over time in countries with a smaller number of tweets for the same period.

The previous research did not have a big pool of data in terms of a number of tweets and such a long period. This will allow us to have a wider overview of the change in sentiment score over time.

4.2 Structure of Tweet

The tweets are collected using Twitter Streaming API, where they are provided in JSON format consisting of multiple parameters, the text is one of them. Other parameters are mainly metadata of tweet, which includes the location of tweets origin, user profile ID, language, and many more. There are nested JSON such as 'entities', 'metadata', 'place' and 'user'. The more detailed structure of the tweet is below.

The parameters that are most relevant for this research are: 'created_at', 'text', 'place_name'. The text itself is needed for the sentiment analysis; 'created_at' parameter contains a time and date when a tweet was posted and it is needed to see the change of the sentiment result over time; the 'place_name' parameter is self-explanatory and contains the geolocation of the tweet and is useful for allocating tweet sentiment result to its own country. A more detailed structure of tweet can be seen in Figure 2


```
{
   "contributors":null,
   "coordinates":null,
   "created_at":"Tue Feb 18 23:59:56 +0000 2020",
   "entities":{},
   "favorite_count":0,
   "favorited":false,
   "geo":null,
   "id":1229918502489874434,
   "id_str":"1229918502489874434",
   "in_reply_to_screen_name":"HTF_RTF",
   "in_reply_to_status_id":1229882029921259522,
   "in_reply_to_status_id_str":"1229882029921259522",
   "in_reply_to_user_id":1091040397311266818,
   "in_reply_to_user_id_str":"1091040397311266818",
   "is_quote_status":false,
   "lang":"en",
   "metadata":{},
   "place":{},
   "retweet_count":0,
   "retweeted":false,
   "source":"<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
   "text":"@HTF_RTF @Mandalorian_Ren So very sorry for their families.  RIP brave fellows.",
   "truncated":false,
   "user":{}
}
```


Figure 2: Structure of tweet from Twitter API in JSON format

4.3 Data Preprocessing

Data preprocessing is an important step in text analysis. During the phase of data collection, it is possible to collect the data which is most relevant for the research. However, the data is still raw, unstructured and noisy which is not suitable for text analysis. Therefore to produce quality data which can be useful for the analysis it requires certain methods or cleaning techniques to happen at first. For purposes of this thesis, a number of certain cleaning techniques were performed which are common for tweets processing.

The data preprocessing can be divided into the following steps:

- **Data Flattening.** As data was collected from Twitter API in JSON format with nesting, it was converted to CSV format for quicker work with data frame;
- **English language.** Only tweets in the English language were selected for further sentiment analysis. The reason for it was that there are much more packages for English text processing, which makes the results more robust and reliable;
- **Remove links.** The collected tweets contained URLs which can bring the bias to text analysis, therefore, all 'https' links were removed with the regular expressions;
- **Remove mentions.** The tweet also collected '@' mentions of other users. They were removed with regular expressions;
- **Remove hashtags.** As part of standard text cleaning hashtags '' were removed as well with regular expressions;

The data preprocessing did not include such steps as are removing capitalisation or removing repeating punctuation. For instance, if a sentence was "I LOVE YOU!!!" it was not converted to "I love you!". The reason for it is that they reflect the emotion of the author of post better.

In the Table 2 are shown a couple of examples of preprocessing result.

Before preprocessing	**After preprocessing**
@SDS now has a crap TON of material for tomorrow's "Sad Fans Are Sad"! #HappinessIsaBamaBeatdown https://t.co/u2nk947Wu8'	now has a crap TON of material for tomorrow's "Sad Fans Are Sad"!
@spacegoat04 @business I wouldn't even waste the gas on the travel lol. Don't worry.	I wouldn't even waste the gas on the travel lol. Don't worry.

Table 2: Preprocessing of raw tweets into clean text

5 Methodology

This section explains what method is used for sentiment analysis as well as how to evaluate if the results of the analysis closely reflect how happy people are.

5.1 Sentiment analysis

To date, various methods have been developed and introduced for sentiment analysis. Some of them were discussed in the Related work of this paper. For the purposes of this paper, Vader is used as a tool for sentiment analysis. Vader is a part of the NLTK package and is a lexicon and rule-based sentiment analysis tool that works great with the data from social media (C. J. Hutto et al (2014)) [10].

Since Vader uses a lexical based sentiment analysis it does not require prior training. Also, Vader makes it easy to interpret the results. It's analyses how negative positive and neutral the text is and presents the results as a compound of all the three categories as well as separately. The outcome is normalised between -1 and 1 which is great when it comes to comparing the results.

By this day, Vader was used in a number of papers on sentiment analysis and proved to be a reliable tool (C. J. Hutto et al (2014))[10]. It also offers certain advantages to the user such as listed below:

- typical negations, such as "not good"
- use of contractions as negations, such as "wasn't very good"
- conventional use of punctuation to signal increased sentiment intensity, such as "Good!!!"
- conventional use of word-shape to signal emphasis, such as using capital letters for words/phrases
- using degree modifiers to alter sentiment intensity, such as words "very" and "kind of"
- understanding many sentiment-laden slang words, such as 'sux'
- understanding many sentiment-laden slang words as modifiers, such as 'uber' or 'friggin' or 'kinda'
- understanding many sentiment-laden emoticons such as :) and :D
- translating utf-8 encoded emojis
- understanding sentiment-laden initialisms and acronyms, such as 'lol'

5.2 Evaluation of results

Two analyses were conducted to establish whether the sentiment analysis results provide a close estimate of how happy people are. The first analysis aimed at understanding if the socio-economic factors, in other words, wellbeing, can explain or correlate with happiness or positiveness. The second analysis focused on interpreting the sentiment score based on world events and shocks.

5.2.1 Evaluation based on socio-economic factors

To understand whether socio-economic factors affect happiness derived from tweets, a correlation analysis was performed. For this analysis, a number of indexes were collected from the U.S. News and World Report for the year 2019 [27]. The platform contains the data on Health, Education, Economy, Infrastructure, Opportunity, Fiscal Stability, Crime and Environment for 50 states of the USA.

The analysis will use Pearson correlation to between sentiment score grouped by the state of the USA and each of indexes separately. As the period from November 2019 to May 2020 was full of events and it could affect the sentiment of people. Therefore aside of running correlation against total sentiment score, it was run for monthly sentiment against indexes.

The intention for such analysis is to understand if the sentiment or emotions can be explained or at least show significantly high enough correlation with the socio-economic factors.

5.2.2 Evaluation based on world events or shocks

It was considered that analysis of change in sentiment over the period of study would usefully supplement and extend the understanding of sentiment result. For this part, the analysis of how news, events and world shocks impact the sentiment was conducted. The world news was collected and the text was processed to extract the sentiment, using the same tool. Afterwards, the correlation coefficient between the sentiment of news and sentiment of tweeds was calculated.

Also, it was interesting to see if the sentiment analysis of tweet was able to spot large events and shocks its biggest news of the day. For this part, the analysis of large fluctuations was conducted. The scores of tweets were grouped by dates and the largest deviations from the norm were analysed. Namely topic modelling was run at this phase.

This analysis intends to see if the sentiment is affected by news more compared to socio-economic factors.

6 Results

This section reviews the results of the sentiment analysis and how it correlates with different indexes. This chapter also reviews how well the analysis identified the big events or shocks which happened during the period of study.

6.1 Correlation with socio-economic factors

The correlation between sentiment score of the U.S. states and socio-economic indexes was tested. The indexes were taken from the U.S. News and World Report for the year 2019 [27] and present in itself 8 categories such as Health, Education, Economy, Infrastructure, Opportunity, Fiscal Stability, Crime and Environment. Each of the factors arguably affects how happy the person can be.

It can be seen in the Table 3 that the sentiment analysis does not have a significantly high correlation with any of the factors. The highest correlation is with Health Care Quality, which is -0.468. This implies that the higher the health care quality in the state, the less happy people are.

Socio-Economic factors	Sub-category	Full period	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020	May 2020
Health	HEALTH CARE ACCESS	0.003	-0.004	0.151	-0.164	-0.075	0.020	-0.023	-0.047
	HEALTH CARE QUALITY	-0.468	-0.307	-0.286	-0.441	-0.515	-0.468	-0.546	-0.376
	PUBLIC HEALTH	-0.238	-0.147	-0.053	-0.314	-0.306	-0.286	-0.274	-0.249
Education	HIGHER EDUCATION	-0.236	-0.262	-0.258	-0.344	-0.128	-0.254	-0.216	-0.160
	PRE-K-12	-0.069	-0.029	0.033	-0.229	-0.155	0.013	-0.105	-0.136
Economy	BUSINESS ENV.	-0.293	-0.190	-0.210	-0.261	-0.350	-0.398	-0.376	-0.283
	EMPLOYMENT	-0.075	-0.060	-0.100	-0.163	-0.069	-0.124	-0.082	-0.043
	GROWTH	-0.198	-0.202	-0.208	-0.205	-0.114	-0.220	-0.217	-0.203
Infrastructure	ENERGY	-0.248	-0.356	-0.402	-0.244	-0.166	-0.192	-0.169	-0.264
	INTERNET ACCESS	-0.008	-0.016	0.015	-0.046	0.021	-0.098	-0.016	0.040
	TRANSPORTATION	-0.291	-0.299	-0.325	-0.290	-0.228	-0.239	-0.345	-0.219
Opportunity	AFFORDABILITY	0.378	0.138	0.130	0.349	0.476	0.413	0.432	0.341
	ECONOMIC OPPORTUNITY	-0.284	-0.099	-0.139	-0.402	-0.334	-0.307	-0.269	-0.256
	EQUALITY	-0.003	0.033	0.050	-0.030	-0.145	0.062	0.053	-0.084
Fiscal Stability	LONG-TERM	0.083	-0.045	0.027	-0.039	0.123	0.091	0.059	0.124
	SHORT-TERM	0.009	0.023	-0.026	0.014	0.068	0.056	0.004	0.042
Crime	CORRECTIONS	-0.073	0.070	0.063	-0.118	-0.178	-0.074	-0.042	-0.049
	PUBLIC SAFETY	-0.143	-0.044	-0.008	-0.202	-0.173	-0.122	-0.117	-0.176
Environment	AIR AND WATER QUALITY	0.231	0.191	0.302	0.168	0.154	0.215	0.243	0.088
	POLLUTION	-0.241	-0.153	-0.123	-0.275	-0.312	-0.248	-0.093	-0.304
Overall	OVERALL	-0.307	-0.230	-0.202	-0.184	-0.188	-0.137	-0.072	-0.173

Table 3: Correlation of Socio-Economics

In the bottom of the Table 3 the 'Overall' factor consists of the weighted average of other factors. And it on its own does not show signs of significantly high correlation. The value of -0.307 of Pearson correlation means that the better the socio-economic situation in the state the less positive people are.

The Figure 18 and Figure 19 depict the difference in rankings between the 'Overall' factor and Sentiment result. Interesting that according to sentiment analysis southeastern part of the USA is happier than west. Whereas according to overall socio-economic factor south and mid-east have a lower ranking than others.

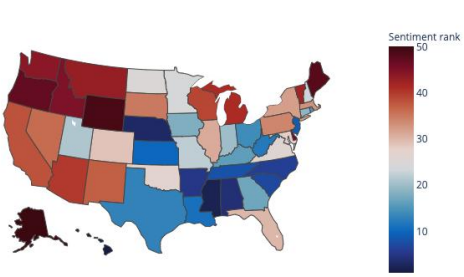

Figure 3: Ranking USA states by Happiness

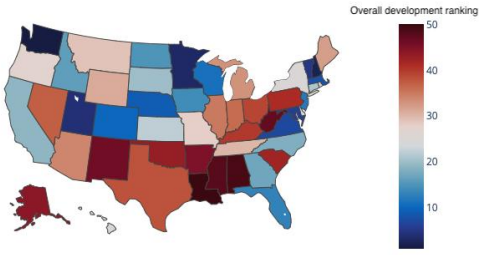

Figure 4: Ranking USA states by Overall socio economic factor

To see if sentiment score can be explained by socio-economic factors but on a more broad level, country to country comparison, the sentiment analysis results were compared to the Human Development Index (HDI). HDI is formed from multiple demographic factors and is reported in the Human Development Report 2018 by United Nations Development Programme [25]. As it is seen the Table 4 the results seem to be reverse. India is the happiest of all according to sentiment analysis, is 129th in HDI Ranking. However, Australia, which is the least happy of all, appears to be 6th in HDI Ranking, the top out of four countries of this research.

Country	Sentiment	HDI Index (2018)	Sentiment Rank	HDI Rank (2018)
India	0.457	0.647	1	129
United Kingdom	0.320	0.92	2	15
USA	0.311	0.92	3	15
Australia	0.267	0.938	4	6

Table 4: Countries' Sentiment results against HDI

6.2 Analysis of sentiment change

As the correlation analysis between the sentiment of tweets and socio-economic factors did not show a high correlation, it is decided to further analyse if sentiment could depend on other factors. For such it was decided to analyse if dependency on the sentiment of tweets on stock market volatility. For this purpose, the correlation between the VIX index and the sentiment of tweets was calculated.

The CBOE Volatility Index (VIX) is a market index derived from the price of SP500 index options, which provides an estimate of current market sentiment and risk. Thus it was picked as market volatility to see the if sentiment analysis results of this paper correlated with a financial market drop in the first half of 2020.

As it is seen from the Table 5, the correlation of -0.452 is low but high enough to be considered significant. It means that the higher the VIX, which signals high volatility and fear on the market, the more negative the sentiment of tweets was. It can be seen from the table that over months December, January, March and April showed the highest correlation of all as they bear the most significant events that happened in the period.

Correlation	Full period	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020	May 2020
VIX/Sentiment	-0.452	-0.113	-0.416	-0.482	0.053	-0.357	-0.351	0.248

Table 5: Correlation between VIX and Sentiment score

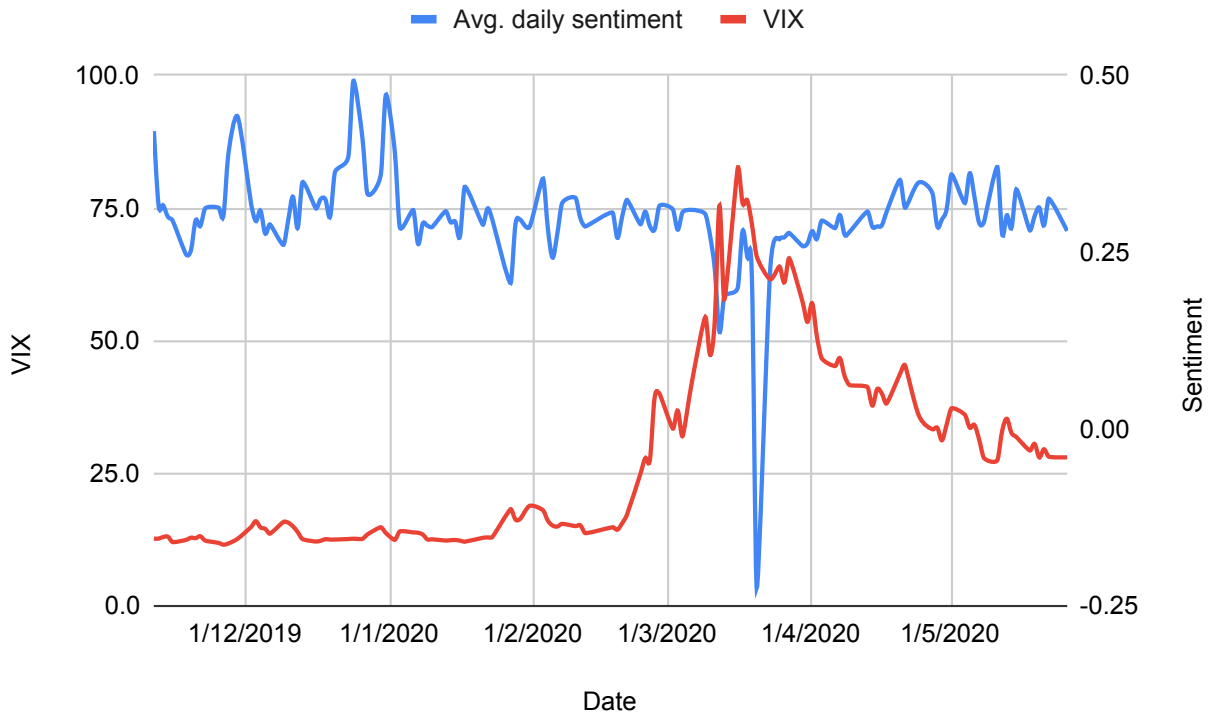

Figure 5: Correlation between VIX and Sentiment score

To investigate further, the dependency of sentiment on news, events and global shocks it was decided first to look into what sentiment analysis was able to detect. The dates with the highest deviation from the average daily sentiment were reviewed. Mainly the USA events were analysed. The dates marked in red on the Figure 6 were with the highest deviation from the average sentiment, which for the whole period is equal to 0.308. After the topic modelling (see in Appendix B) it was possible to identify that the deviation was mainly caused by the following events:

- November 11th, 2019. Veterans day. The sentiment score was 0.42;
- November 28th, 2019. Thanksgiving. The sentiment score was 0.62;

- December 24th - 25th, 2019. Christmas. The sentiment score was 0.56;
- December 31st, 2019 - January 1st, 2020. New Year. The sentiment score was 0.55;
- January 26th, 2020. Kobe Bryant died. The sentiment score was -0.01;
- February 15th, 2020. Valentine's Day. The sentiment score was 0.43;
- March 8th, 2020. International Women's Day. The sentiment score was 0.37;
- March 12th-20th, 2020. Covid-19 is declared as world pandemic and Stock Market drop. The sentiment score was from 0.14 to -0.22;
- April 12th, 2020. Easter. The sentiment score was 0.52;
- May 10th, 2020. Mother's Day. The sentiment score was 0.58;

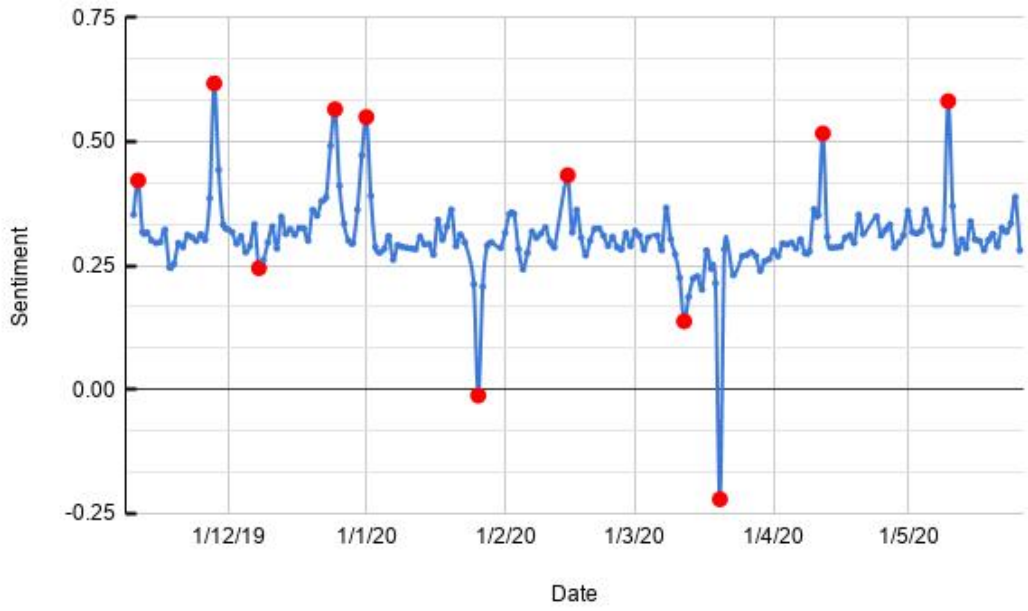

Figure 6: The dates with most sentiment deviation from average

The period from March 12th and March 20th of the year 2020 appeared to be the most negative out of all as seen in Figure 8. During that period high volatility was spotted on the stock markets as well. The reasons for that could be that Covid-19 was declared a world pandemic and negative sentiment due to lockdown and restrictions coming from it.

As such the sentiment analysis was able to spot highly negative sentiment in Australia in January (see Figure 7) as the Australian bushfire season was on its peak and was widely covered in the news. In general, the lower sentiment of Australia from some perspective could

be affected by the Australian bushfire season lasting from June 2019- till March 2020. The fires covered many important parts of Australia and had a significant impact on wildlife and the environment. However, it is not a topic of this paper.

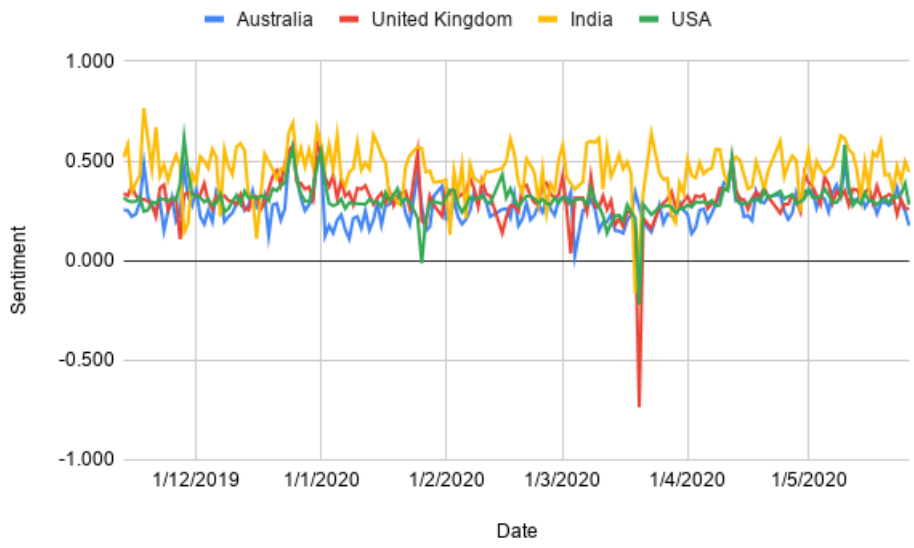

Figure 7: Sentiment of all countries over the full period

As the second step to investigate how the news affected the sentiment of tweets it is decided to run a correlation analysis between the sentiment of news against the sentiment of tweets of all countries. For this purpose, 50 days were picked out of the whole period of study and 10 random news for these days were collected. Then the text of these news was processed and the sentiment was extracted in the same manner as the tweet's sentiment analysis was done. The source for news was MailOnline newspaper archive [18].

As a result of the analysis, the correlation coefficient between the sentiment of tweets of all countries and the sentiment of news was 0.393 as shown in Table 6. The coefficient indicate positive correlation meaning the more positive the news were, the more positive was the sentiment score of tweets. However, the coefficient doesn't show a high correlation between the sentiment of news and tweets. As it is seen in the Figure 8 , both sentiments seem to follow the pattern of each other, however, due to a weak correlation, it is hard to make significant conclusions. Also as seen from Table 6 the correlation between Indian tweets and news is low, which partially could be due to a small number of tweets for India.

Also, the sentiment analysis was able to spot the happiness baseline. Figure 9 shows that on average the sentiment was fluctuating mostly within a range of 0.2 to 0.35. The same is seen in Table 7 where the monthly average sentiment is within the same range. If it were on average below 0, then it would suggest that people are negative in general, which is not the case. However, if it were closer to 1, it would not be so true as well as then it would mean that people are extremely happy most of the time. And also absolute 0 is not possible as then it would mean that people communicate using predominantly only facts and their speech is objective, whereas

Country	News
Australia	0.389
UK	0.337
India	0.138
USA	0.375
Overall	**0.393**

Table 6: Correlation between News' and Tweets' Sentiment scores of 4 countries

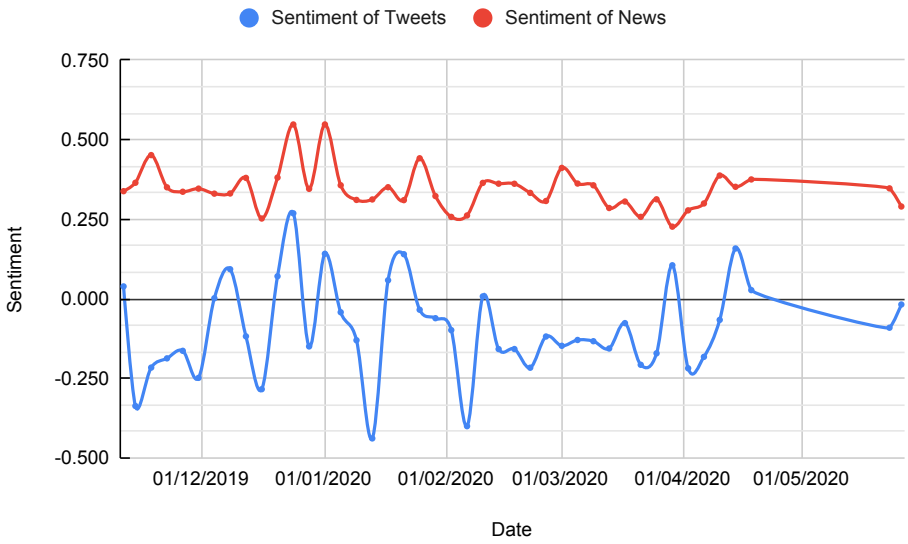

Figure 8: Correlation between News' and Tweets' Sentiment scores

in many cases it bears opinion and thus subjective.

Country	Full period	November 2019	December 2019	January 2020	February 2020	March 2020	April 2020	May 2020
US	0.31	0.33	0.34	0.29	0.31	0.26	0.31	0.33

Table 7: Monthly Baseline Sentiment

As described by S. Lyubomirsky in 2011 [17], there is a phenomenon, known as hedonic adaption. When after experiencing large either negative or positive news or event people return to the previous level of emotional stability. As it can be seen from Figure 7, all countries had their fluctuations due to positive or negative events which either raised or dropped the sentiment. However, the sentiment analysis was able to grasp from data the change in the level of emotions and adjusted back to the baseline range from higher deviations.

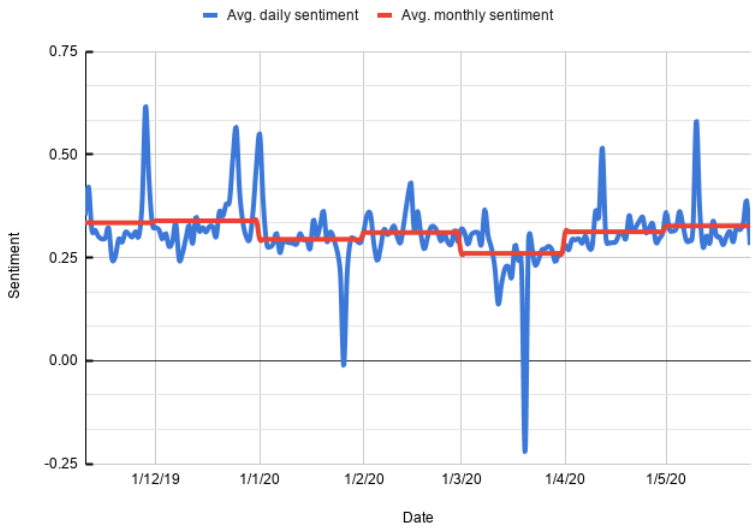

Figure 9: Monthly Baseline Sentiment

6.3 Discussion

The aim of this work was to understand if sentiment analysis can be used as a tool to estimate how happy people are and which socio-economic factors affect it the most. In order to achieve this goal, the Twitter posts were collected for a period from 10.11.2019 to 26.05.2020 and their sentiment was extracted. Also, data on socio-economic ranking of the USA states were collected and correlation analysis was done against tweets' sentiment scores grouped by the USA states.

It was expected that the socio-economic factors would correlate to a high degree with a sentiment. In simple terms, the more socio-economically developed the state is the more positive it would be. However, it was found that tweet's sentiment and overall socio-economic factor has a correlation coefficient of -0.307, whereas the highest correlation was with health care quality, which was -0.468. The coefficients do not suggest that there is a high dependency between sentiment and socio-economic factors. Also, the negative coefficient implies that the more socio-economically developed the states are, the less happy it is. One explanation for such a low correlation coefficient could be the idea that positivity or happiness is not the same as wellbeing, or how developed the environment is. There is Human development index (HDI) which consists of socio-economic factors, however, it does not necessarily reflect if people in the country are happy or sad.

As part of further analysis, it was tested if the sentiment was affected by the volatility of the stock market. Correlation analysis with VIX showed that the coefficient is 0.452 which is higher than for socio-economic factors. The assumption was made that sentiment can be affected by the news. It is noticed with topic modelling that on days with high deviations from the norm, certain news was discussed a lot more, which implies that news may cause an increase or decrease in sentiment. The same finding was noticed in the paper of Ahmet Onur Durahim et al (2015) [7],

where the peaks of sentiment score were on dates with big events. However, when analysing how tweet's sentiment correlates with the sentiment of multiple news on different days the results are not so obvious. The correlation coefficient showed a low correlation of 0.37. It could lead to a conclusion that high fluctuations could be caused by news on a corresponding day, but not all news have a significantly high effect on the sentiment of tweets.

There were a few limitations that need to be addressed, possibly limiting the generalization of this study. One of the One was that the number of analysed news articles was not as large compared to the total number of news articles existing on a specific date. Also, the number of dates which were under analysis for news does not cover the whole period of study. It could lead to not fully representative news sentiment scores and thus not reliable correlation coefficient for news-to-tweets analysis.

Another limitation is that the English language is not the only language spoken in the areas of study. Thus the sentiment scores of tweets, as well as news, could be not consistent if added tweets in other languages. Another limitation is that Twitter is not representative of the population, the share people who use Twitter is still not high. The number of users of Twitter in the USA is about 62 million people. Lastly, it is important to note that it was complicated to collect similar socio-economic data for different countries therefore it has limited the correlation analysis to be conducted on for the state of the USA.

Future studies could address the limitations of this work. For a deeper analysis of the effect of the news on sentiment, it is recommended to expand the number of articles for analysis as well as several sources. It is interesting to see if there is a correlation between what news articles publish about and what Twitter users mention in their posts over a long period.

In conclusion, this work contributes to the field of sentiment analysis, by providing evidence that socio-economic factors might not correlate with sentiment extracted from tweets and thus might not explain the happiness of Twitter users. Also, this work presents evidence that some news can influence the positivity of tweets.

References

[1] O. Abdelwahab, M. Bahgat, C. J. Lowrance, and A. Elmaghraby. Effect of training set size on svm and naive bayes for twitter sentiment analysis. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 46–51. IEEE, 2015.

[2] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, pages 1–6. IEEE, 2013.

[3] K. Ahmad, D. Cheng, and Y. Almas. Multi-lingual sentiment analysis of financial news streams. In *1st International Workshop on Grid Technology for Financial Modeling and Simulation*, volume 26, page 001. SISSA Medialab, 2007.

[4] K. Ahmed, N. El Tazi, and A. H. Hossny. Sentiment analysis over social networks: an overview. In *2015 IEEE international conference on systems, man, and cybernetics*, pages 2174–2179. IEEE, 2015.

[5] M. Al-Smadi, ·. B. Talafha, M. Al-Ayyoub, and ·. Y. Jararweh. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics*, 10(3):2163–2175, 2019. doi: 10.1007/s13042-018-0799-4. URL `https://doi.org/10.1007/s13042-018-0799-4`.

[6] J. C. De Albornoz, L. Plaza, P. Gervás, and A. Díaz. A joint model of feature mining and sentiment analysis for product review rating. In *European conference on information retrieval*, pages 55–66. Springer, 2011.

[7] A. O. Durahim and M. Coşkun. # iamhappybecause: Gross national happiness through twitter analysis and big data. *Technological Forecasting and Social Change*, 99:92–105, 2015.

[8] X. Fang and J. Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5, 2015.

[9] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. *Icwsm*, 7(21):219–222, 2007.

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

[11] S. Jaswanth Sai Chowdary, S. Waseem Farooq, S. Abubakar Siddique, and S. Dinesh. Sentiment Analysis of movie reviews using Microsoft Text Analytics and Google Cloud Natural Language API. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2019 IJSRCSEIT* |, 5(1):2456–3307, 2019. doi: 10.32628/CSEIT195130. URL https://doi.org/10.32628/CSEIT195130.

[12] N. S. Joshi and S. A. Itkat. A survey on feature level sentiment analysis. *International Journal of Computer Science and Information Technologies*, 5(4):5422–5425, 2014.

[13] S. Kolkur, G. Dantal, and R. Mahe. Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology*, 5(2):768–770, 2015.

[14] A. D. Kramer. An unobtrusive behavioral model of" gross national happiness". In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 287–290, 2010.

[15] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23, 2014.

[16] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[17] S. Lyubomirsky. *Hedonic adaptation to positive and negative experiences.* Oxford University Press, 2011.

[18] MailOnline. Mailonline archive. https://www.dailymail.co.uk/home/sitemaparchive/index.html.

[19] R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144, 2006.

[20] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE*, 8(5):64417, 2013. doi: 10.1371/journal.pone.0064417. URL www.plosone.org.

[21] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu. Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 205–208. IEEE, 2019.

[22] C. Musto, G. Semeraro, and M. Polignano. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. In *DART@ AI* IA*, pages 59–68, 2014.

[23] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[24] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[25] U. N. D. Programme. Human development report. `http://hdr.undp.org/en/2019-report`, 2019.

[26] D. Quercia, J. Ellis, L. Capra, and J. Crowcroft. Tracking" gross community happiness" from tweets. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 965–968, 2012.

[27] U. N. . W. REPORT. Best states, overall index. `https://www.usnews.com/news/best-states/rankings`, 2019.

[28] A. Richter and M. Koch. Functions of social networking services. *From CSCW to Web 2.0: European Developments in Collaborative Design Selected Papers from COOP08*, 2008.

[29] R. G. Rodrigues, R. M. das Dores, C. G. Camilo-Junior, and T. C. Rosa. Sentihealth-cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. *International journal of medical informatics*, 85(1):80–95, 2016.

[30] M. d. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodriguez-Garcia, and R. Valencia-Garcia. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, 2017, 2017.

[31] M. Salathé and S. Khandelwal. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology*, 7 (10), 2011.

[32] V. K. Singh, R. Piryani, A. Uddin, and P. Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pages 712–717. IEEE, 2013.

[33] V. Subramaniyaswamy, R. Logesh, M. Abejith, S. Umasankar, and A. Umamakeswari. Sentiment analysis of tweets for estimating criticality and security of events. In *Improving the Safety and Efficiency of Emergency Services: Emerging Tools and Technologies for First Responders*, pages 293–319. IGI Global, 2020.

[34] T. T. Thet, J.-C. Na, and C. S. Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848, 2010.

[35] G. Vinodhini and R. Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

[36] K. Wu, B. Zhang, J. Zheng, and H. Yao. Sentiment classification for topical chinese microblog based on sentences' relations. In *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing*, pages 2221–2225. IEEE, 2013.

[37] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Nikolai Rol**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Twitter sentiment analysis to estimate happiness level,

   supervised by Rajesh Sharma.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Nikolai Rol
10/08/2020

A Appendix A

The table of sentiment of news and tweets per country for 50 dates.

Date	Sentiment of news	Sentiment of tweets in Australia	Sentiment of tweets in the UK	Sentiment of tweets in India	Sentiment of tweets in the USA	Sentiment of tweets for four countries
11/11/2019	0.039	0.158	0.311	0.464	0.421	0.338
14/11/2019	-0.337	0.252	0.326	0.581	0.300	0.365
18/11/2019	-0.216	0.487	0.308	0.765	0.246	0.452
22/11/2019	-0.187	0.301	0.364	0.425	0.312	0.350
26/11/2019	-0.163	0.197	0.317	0.530	0.302	0.337
30/11/2019	-0.248	0.277	0.338	0.439	0.332	0.346
04/12/2019	0.002	0.252	0.297	0.464	0.309	0.331
08/12/2019	0.093	0.195	0.326	0.559	0.245	0.331
12/12/2019	-0.118	0.317	0.331	0.588	0.284	0.380
16/12/2019	-0.284	0.267	0.320	0.111	0.312	0.252
20/12/2019	0.071	0.281	0.423	0.459	0.362	0.381
24/12/2019	0.269	0.504	0.553	0.644	0.491	0.548
28/12/2019	-0.150	0.251	0.360	0.472	0.300	0.346
01/01/2020	0.142	0.531	0.528	0.583	0.549	0.548
05/01/2020	-0.042	0.202	0.312	0.630	0.284	0.357
09/01/2020	-0.130	0.212	0.282	0.462	0.288	0.311
13/01/2020	-0.439	0.154	0.329	0.459	0.308	0.312
17/01/2020	0.058	0.276	0.293	0.493	0.342	0.351
21/01/2020	0.140	0.340	0.336	0.275	0.289	0.310
25/01/2020	-0.034	0.439	0.544	0.572	0.212	0.442
29/01/2020	-0.061	0.324	0.279	0.393	0.296	0.323
02/02/2020	-0.098	0.212	0.334	0.130	0.354	0.258
06/02/2020	-0.400	0.208	0.351	0.214	0.275	0.262
10/02/2020	0.008	0.344	0.400	0.387	0.326	0.364
14/02/2020	-0.158	0.332	0.288	0.520	0.308	0.362
18/02/2020	-0.158	0.331	0.275	0.535	0.306	0.362
22/02/2020	-0.216	0.205	0.323	0.480	0.324	0.333
26/02/2020	-0.118	0.310	0.391	0.241	0.287	0.307
01/03/2020	-0.148	0.318	0.428	0.580	0.319	0.411
05/03/2020	-0.129	0.265	0.288	0.591	0.305	0.362
09/03/2020	-0.133	0.259	0.272	0.593	0.303	0.357
13/03/2020	-0.156	0.231	0.302	0.422	0.186	0.285
17/03/2020	-0.076	0.202	0.246	0.495	0.280	0.306
21/03/2020	-0.207	0.193	0.213	0.342	0.283	0.258
25/03/2020	-0.171	0.279	0.264	0.439	0.269	0.313
29/03/2020	0.106	0.209	0.269	0.190	0.239	0.227
02/04/2020	-0.218	0.136	0.279	0.428	0.269	0.278
06/04/2020	-0.182	0.196	0.262	0.456	0.284	0.300
10/04/2020	-0.066	0.391	0.354	0.442	0.363	0.388
14/04/2020	0.158	0.313	0.305	0.504	0.286	0.352
18/04/2020	0.027	0.349	0.344	0.497	0.311	0.375
23/05/2020	-0.091	0.368	0.231	0.472	0.317	0.347
26/05/2020	-0.018	0.176	0.259	0.446	0.280	0.290
Total Period	**-0.095**	**0.279**	**0.329**	**0.460**	**0.308**	**0.344**

Table 8: The sentiment of news and tweets per country

B Appendix B

Topic modelling of dates with high fluctuation in sentiment

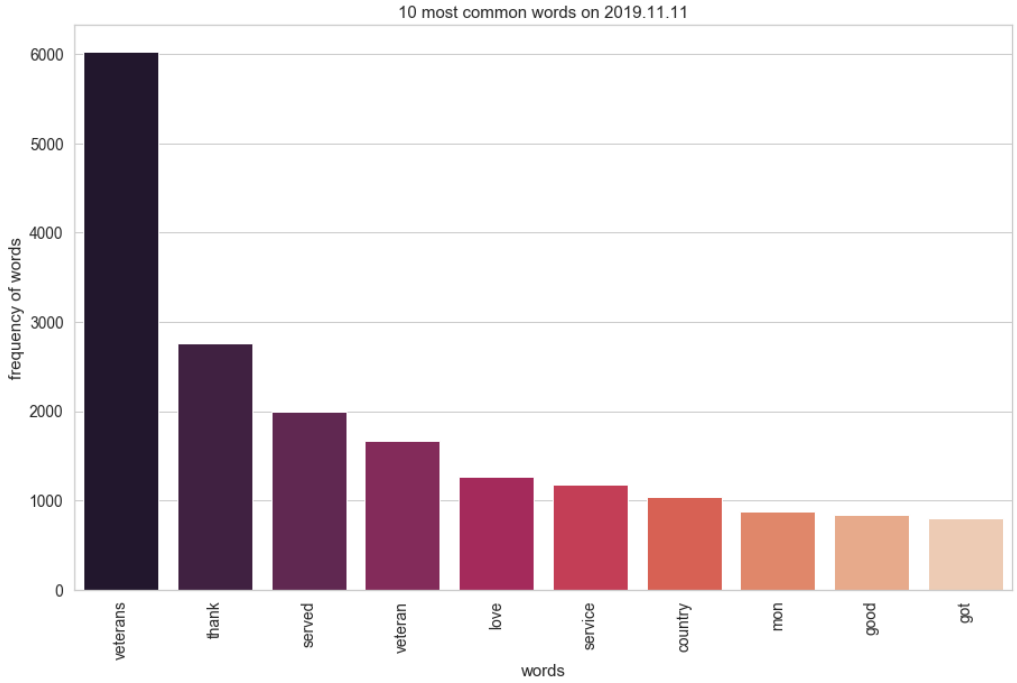

Figure 10: Frequent words on 11.11.2019

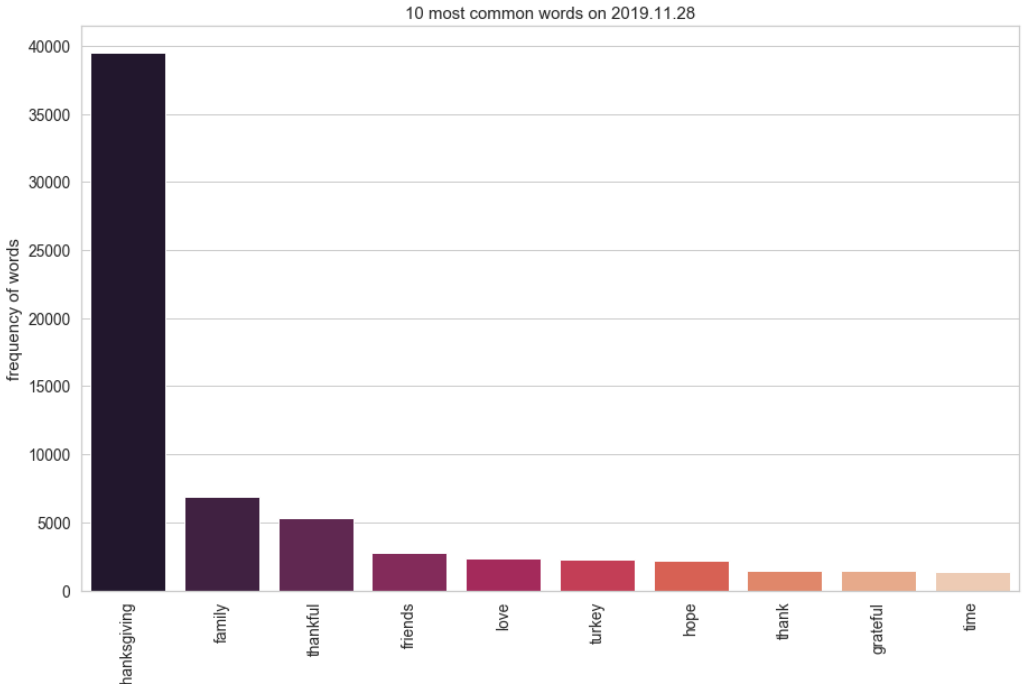

Figure 11: Frequent words on 28.11.2019

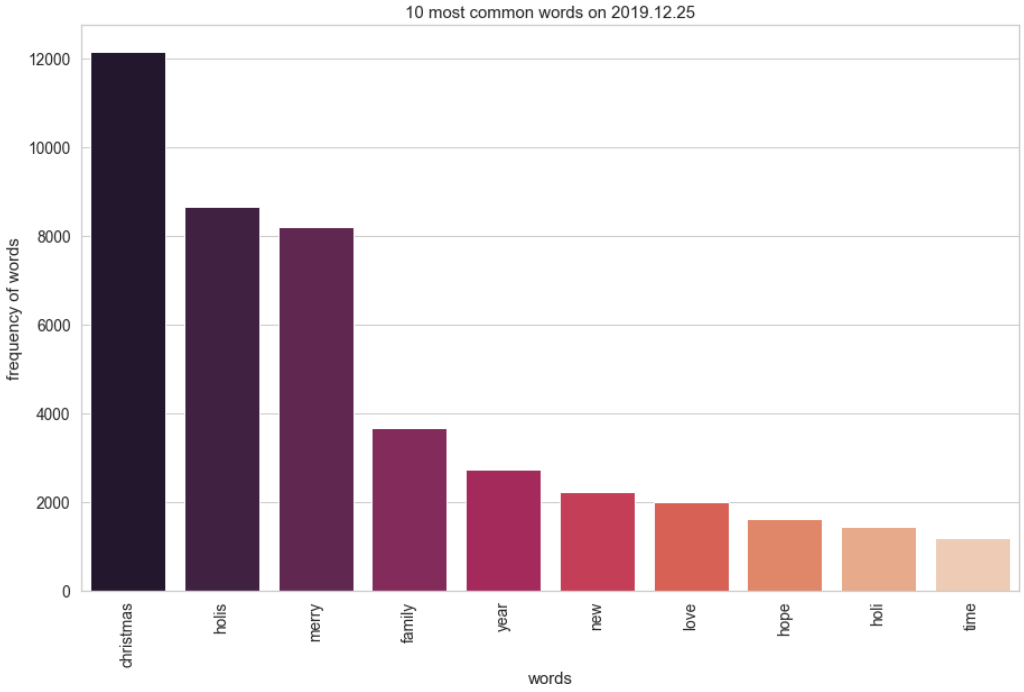

Figure 12: Frequent words on 25.12.2019

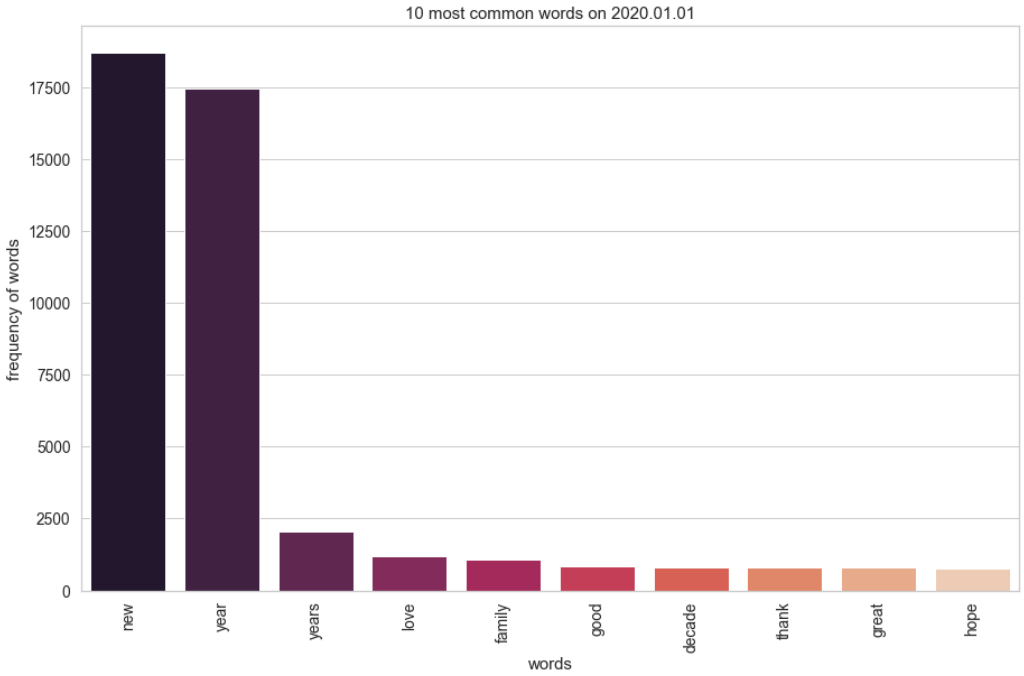

Figure 13: Frequent words on 01.01.2020

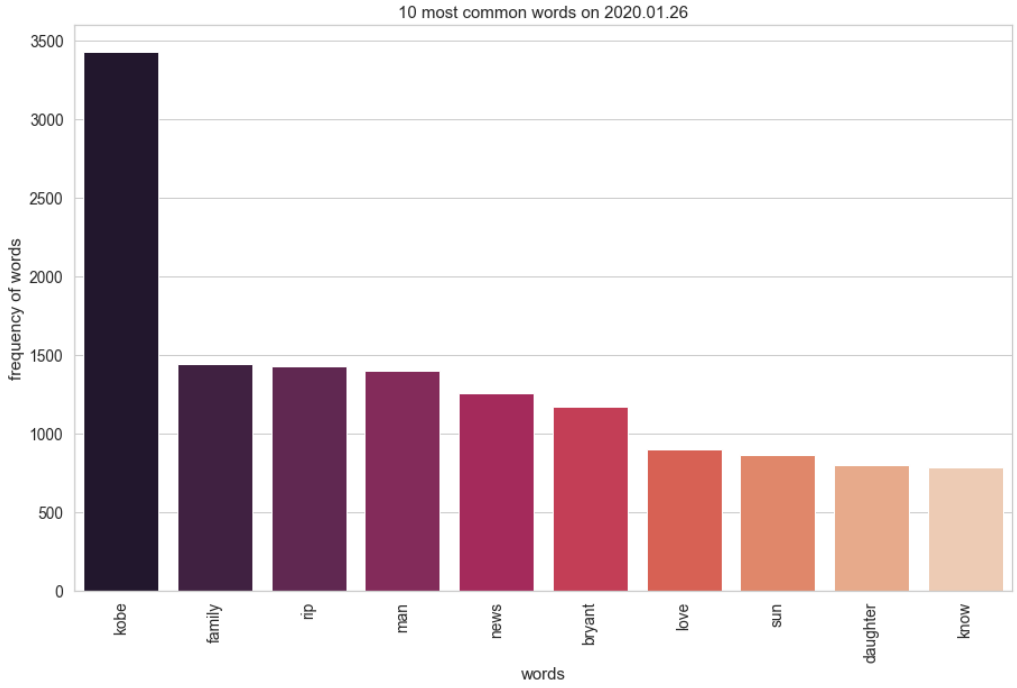

Figure 14: Frequent words on 26.01.2020

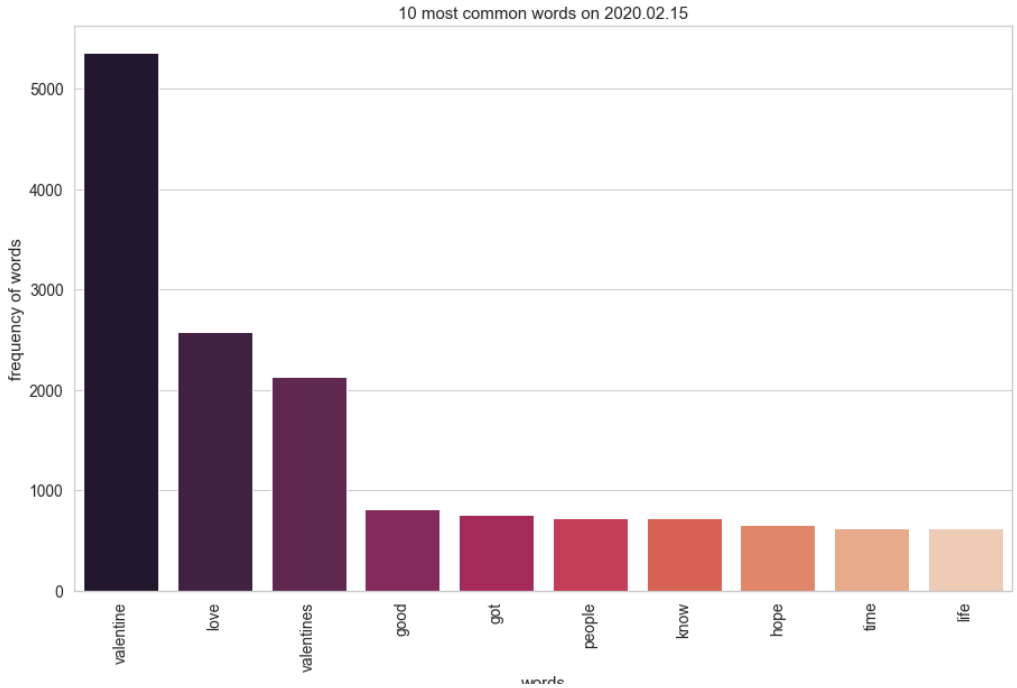

Figure 15: Frequent words on 15.02.2020

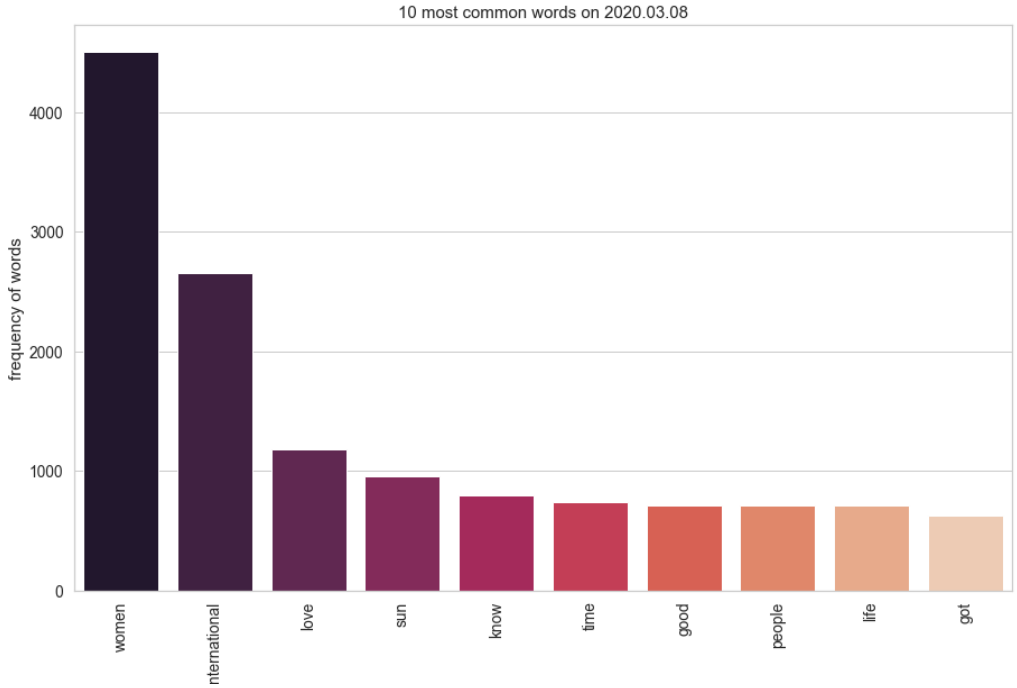

Figure 16: Frequent words on 08.03.2020

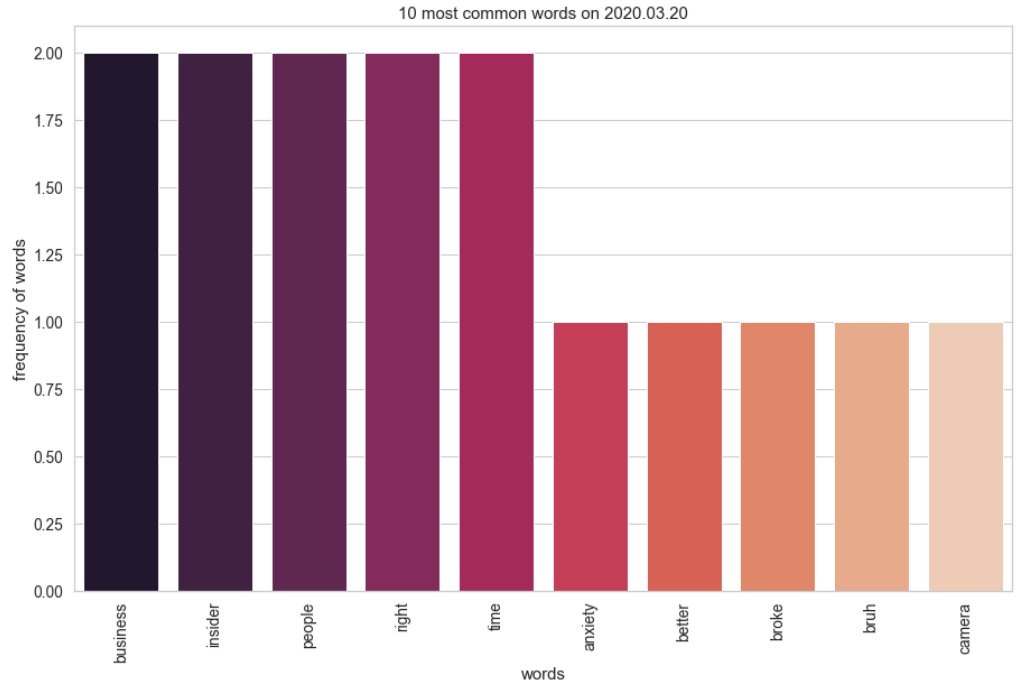

Figure 17: Frequent words on 20.03.2020

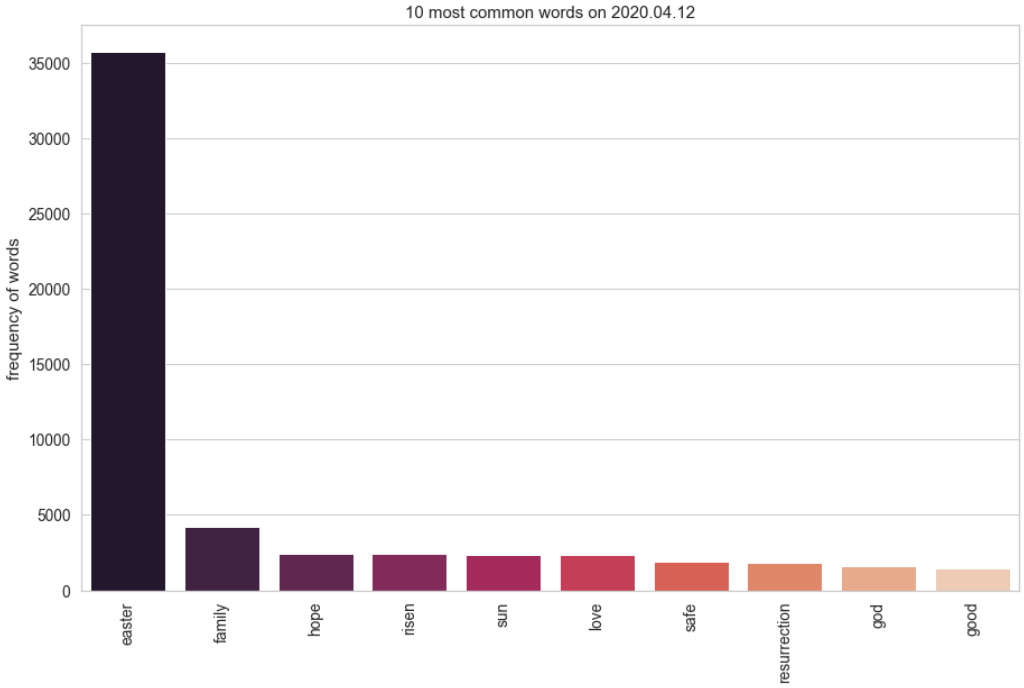

Figure 18: Frequent words on 12.04.2020

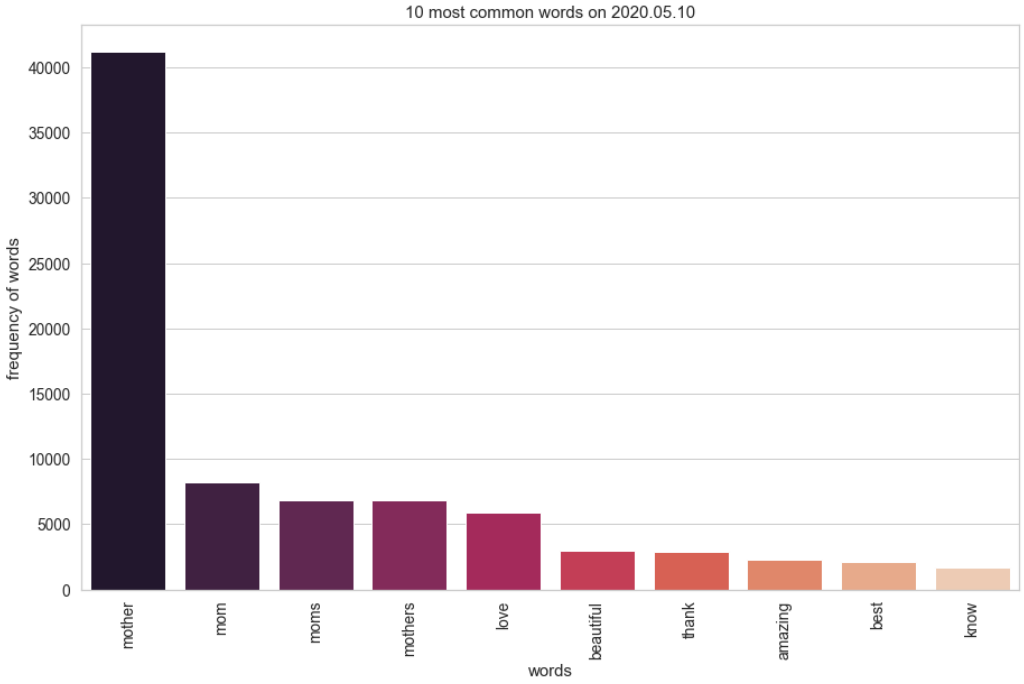

Figure 19: Frequent words on 10.05.2020