How Sentiment Analysis Quantifies the Social Reaction to COVID-19

Jordan Leonard

jleonard21@cmc.edu

A Comprehensive Summary of
Sentiment Analysis Research

CMC

CSCI145 Data Mining
Claremont McKenna College
November 2020

1 Introduction

1.1 What is Sentiment Analysis?

Sentiment analysis is a unique data mining tool that refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to identify, extract, quantify, and study subjective information. It is commonly used to gather information on public opinion by breaking down text to determine whether it contains positive or negative sentiment. Many studies tend to gather their text data from social media platforms due to the large amount of users and content.

1.2 Problem Description

In the following research papers, sentiment analysis is employed to research the social response to COVID-19 on Twitter. One approach develops a model to predict emojis based on the emotional content of the text. Another analyzes the word frequency and overall tweet sentiment between the World Health Organization (WHO) and the general public. The third approach models the evolution of tweet sentiment across the popular coronavirus-related topics mined from the tweet data.

2 Research

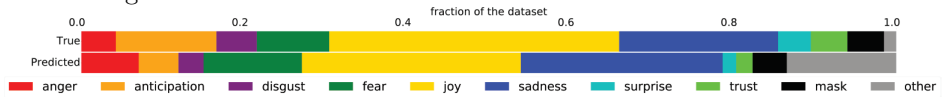
2.1 Multilingual Emoticon Prediction of Tweets about COVID-19

In "Multilingual Emoticon Prediction of Tweets about COVID-19", Mike Izbicki and Stefanos Stoikos develop the first highly multilingual model for emoji prediction, BERTmoticon, using sentiment analysis techniques to analyze the emotional reaction of Twitter users to news about the coronavirus.

Emojis are a text feature used for communicating emotional content in messages, like tweets, so the ability to predict which emoji corresponds to a piece of text can be used as a proxy for measuring its emotional content. The BERTmoticon[10] model was trained on the subset of emojis called emoticons, the original 80 emoji defined in Unicode standard's emoticon code block. The motivation behind this has to do with the fact that they are the most commonly used emoji on Twitter and each emoticon represents an emotion. Since COVID-19 is a worldwide pandemic, it was the perfect opportunity to introduce the BERTmoticon[10] model, which was based on fine-tuning the multilingual BERT model[9], for an emoji prediction analysis.

To produce the BERTmoticon[10] model, a TwitterEmoticon[10] dataset was created by gathering all geolocated tweets sent between January and June 2020 and filtering them to contain one of the 80 target emoticons. Each tweet was preprocessed by removing its emoticon so that each tweet was classfied by a single emoticon. If a tweet contained multiple emoticons, it was duplicated for each emoticon and then granted a classification label. The resulting dataset was then split into training, validation, and test sets to begin the training procedure. The BERT model[9] was fine-tuned on the TwitterEmoticon[10] dataset by training the last layer of the model to generate a model called BERTmoticon-LL[10]. Then, all parameters were trained to generate the BERTmoticon[10] model by warm starting BERTmoticon-LL[10]. Optimal hyperparameters, produced by the validation set, were used to train BERTmoticon-LL[10] and BERTmoticon[10] which produced Macro-F1 scores of 0.159 and 0.210, respectively.

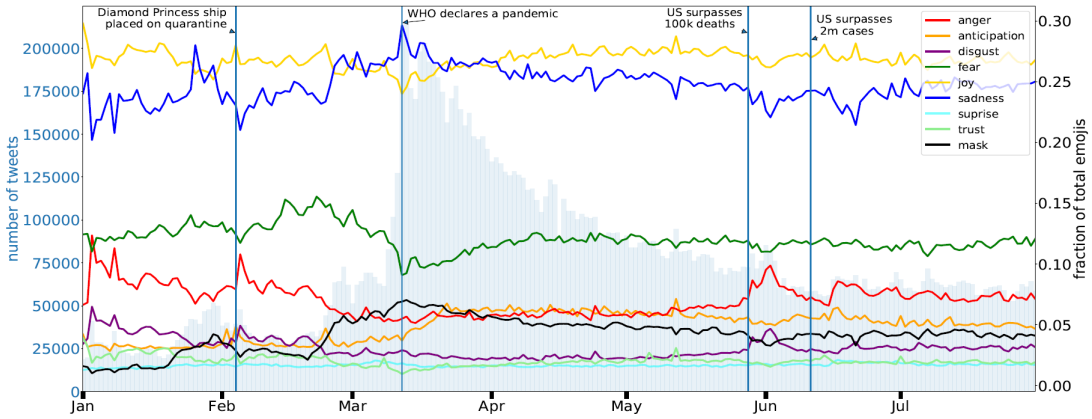
Figure 1: Distribution of Emotions Present vs Predicted in Data



To apply the BERTmoticon[10] model, a TwitterCOVID[10] dataset was created in a 4-step process. First, a list of 29 English-language search terms related to the coronavirus was generated from the terms used in a dataset generated by Chen et al.[5]. Twitterisms like "kungflu" which do not translate well into non-English languages were removed while generic terms like "china" and "trump" were included since these terms were likely to reference the coronavirus in the given timeline. Second, each term was translated into

the 72 languages supported by Bing translate. Third, Python's spaCy library was accessed[8] to tokenize and lemmatize each of the 400 million geolocated tweets posted between January and June 2020. Finally, the TwitterCOVID[10] dataset was filtered to be the set of all tweets whose lemmatized text contains any of the search terms from the tweet's language or English. In total, 16.2 million tweets met the criteria to be included within the TwitterCOVID[10] dataset.

Figure 2: Distribution of Emojis Present vs Predicted in Data



Given that 15.11% of tweets in the TwitterCOVID[10] dataset contained an emoticon, the BERTmoticon[10] model was used to label the remaining tweets. Figure 1[1][10] displays the distribution of emoji present in the dataset versus those predicted. The anticipation, disgust, joy, surprise, and trust emoticons appear less frequently in the predicted dataset, while, the anger, sadness, and fear emotions appear more often. It is hypothesized that this difference is caused by more-formal Twitter accounts being less likely to use an emoji in their tweets and having the tendency to tweet about different topics more than informal accounts of the general public. Figure 2[2][10] represents the change in emotional content and the fraction of tweets in a day between January and June. Follow up studies intend to analyze how different countries and language communities reacted to the events within the COVID-19 timeline. The authors also hope that the BERTmoticon[10] model will prove useful for analyzing the emotions of text in other contexts outside of COVID-19.

2.2 Word Frequency Sentiment Analysis of Twitter Messages During Coronavirus Pandemic

In "Word Frequency and Sentiment Analysis of Twitter Messages During Coronavirus Pandemic", by Bhavya Ahuja Grover, Nikhil Kumar Rajput and Vipin Kumar Rathi, the textual content within Twitter tweets is analyzed using word frequencies and sentiment analysis based on the massive influx of tweets from the coronavirus. Determining word frequencies in a document gives a strong idea about the patterns of words used, meanwhile, sentiment analysis helps to understand attitudes and infer emotional content about a subject.

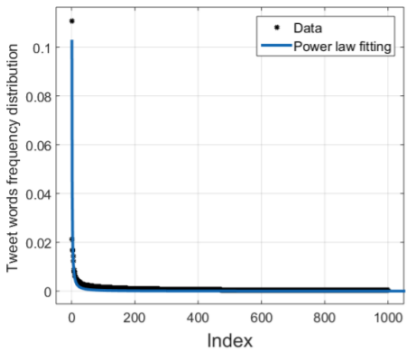
In order to model the word frequency distributions of twitter messages, the authors use the power law, $f(x) = ax^b$, where a is a constant and b is the scaling or exponential parameter. For the statistical analysis, four datasets were created using Twitter data collected by Banda et al.[6] which contained all users tweeting about COVID-19 between January and March. One dataset involved Twitter Id evolution while the others contained the unigram, bigram, and trigram word frequency distributions. The Twitter Id dataset simply focused on the number of users tweeting about COVID-19, but the word frequency distributions were a little

[1]Figure 1: The distribution of emotions in the subset of TwitterCOVID that contain emojis (top). The distribution of emotions in the subset of TwitterCOVID that did not contain emojis, and that we used BERTmoticon[10] to assign predictions for (bottom).

[2]Figure 2: The shaded bar plot in the background shows the total number of tweets in the TwitterCOVID dataset sent on a particular day (left y-axis scale), and the colored line charts show the fraction of tweets in a particular day that correspond to each emotion on the Plutchik wheel or the mask emoji (right y-axis scale).

more complex. Unigrams, bigrams, and trigrams are three forms of tokens that represent the frequencies of one word, two words and three words paired together, respectively. Each of the datasets provide either the top 1000 unigrams, bigrams, or trigrams in which coronavirus was the most frequent word among unigrams. Figures 3-4[2] depict plots of the index versus frequency distributions for unigrams, bigrams and trigrams to demonstrate that the pattern of the data closely follows the power law.

Figure 3: Unigram Frequency vs. Rank



The authors used the Python package TextBlob[11] to perform sentiment analysis of the tweets of their Twitter data [6]. The analysis was conducted on two filtered datasets: tweets made by WHO and tweets that have been retweeted more than 1000 times. The sentiment polarity values of individual tweets were also computed such that the polarities range between -1 to 1, where -1 is the most negative and 1 is the most positive. Table 1[2] shows the positive, neutral, and negative sentiment polarity percentages and Figure 5[2] display the histograms of the sentiment polarities of tweets from the general public versus WHO. It can be seen that the majority of the tweets have a neutral or positive sentiment rather than negative.

Figure 4: Bigram and Trigram Frequencies vs. Rank

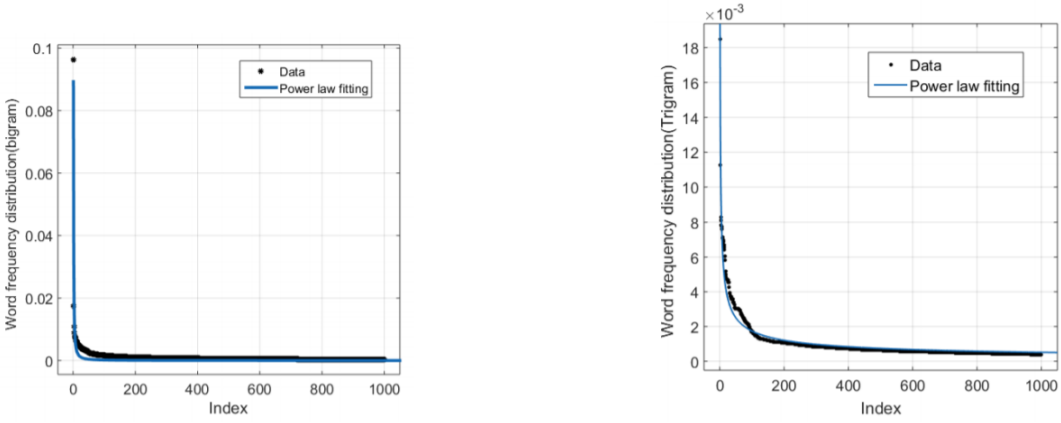
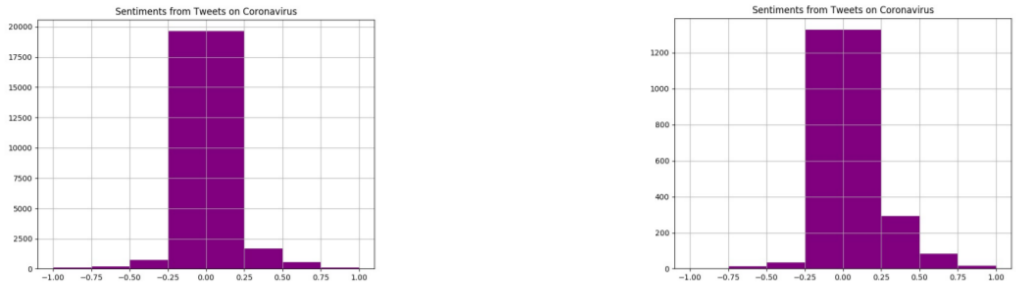


Table 1: Sentiment Polarity Percentages

Sentiment Polarity	Positive	Neutral	Negative
Tweets by WHO	60.27%	24.05%	15.68%
Tweets by General Public	29.33%	54.92%	15.75%

From the analyses, the authors were able to determine the amount that coronavirus-related tweets fluctuated relative to the number of Twitter Ids from January to December. The power law also proved to be an effective model when comparing the word frequency distributions of the unigram, bigram, and trigram tokens to the tweet data. Although the severity of the pandemic increased, overall tweet sentiment of WHO and the general public was determined to either be fairly positive or neutral.

Figure 5: Sentiment Polarities

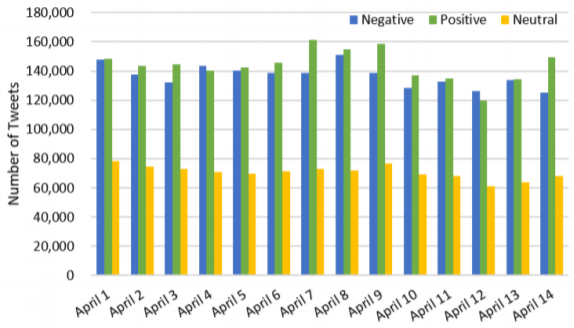


2.3 Detecting Topic and Sentiment Dynamics Due to Covid-19 Pandemic Using Social Media

In "Detecting Topic and Sentiment Dynamics Due to Covid-19 Pandemic Using Social Media" by Hui Yin, Shuiqioa Yang, and Jianxin Li, the popular Twitter topics surrounding the coronavirus and their corresponding sentiment polarity are investigated to understand the mental health of Twitter users. To do so, the Dynamic Topic Model[4] is implemented to generate daily topics such that the sentiment polarity of each topic and tweet can be calculated by the lexicon tool, VADER[3].

The methodology behind the tweet data collection began with obtaining tweet IDs from the coronavirus Twitter dataset collected by Chen et al.[5] based on keywords such as "Coronavirus", "Covid", "Covid19", and "Wuhanlockdown". Account names such as "CDCemergency", "CDCgov", "WHO", and "HHSGov" were also considered as a measure to track tweets. After data filtration, preprocessing, and phrase extraction, the dataset contained a total of 4,919,471 tweets extracted from April 1-14 with 269,391 unique tokens.

Figure 6: Tweet Polarity over Time



Topic modeling is the process of learning, recognizing, and extracting high-level semantic topics across a body of text. A popular technique involves Latent Dirichlet Allocation proposed by Blei et al.[4], which assumes each body of text to be a mixture of topics where each word's presence is associated with a topic.

Once topics are presented, the Dynamic Topic Model[4] is proposed to mine topic evolution over time by extending LDA to allow for topic mining over fixed time intervals. To set up the DTM, LDA was trained on the first day of the dataset to learn the best topic number. At this point, coherence[1] measures the degree of semantic similarity between high scoring words in topics as an indicator to choose the best topic number. The coherence score is resourceful because it helps distinguish between human understandable topics and artifacts of statistical inference.

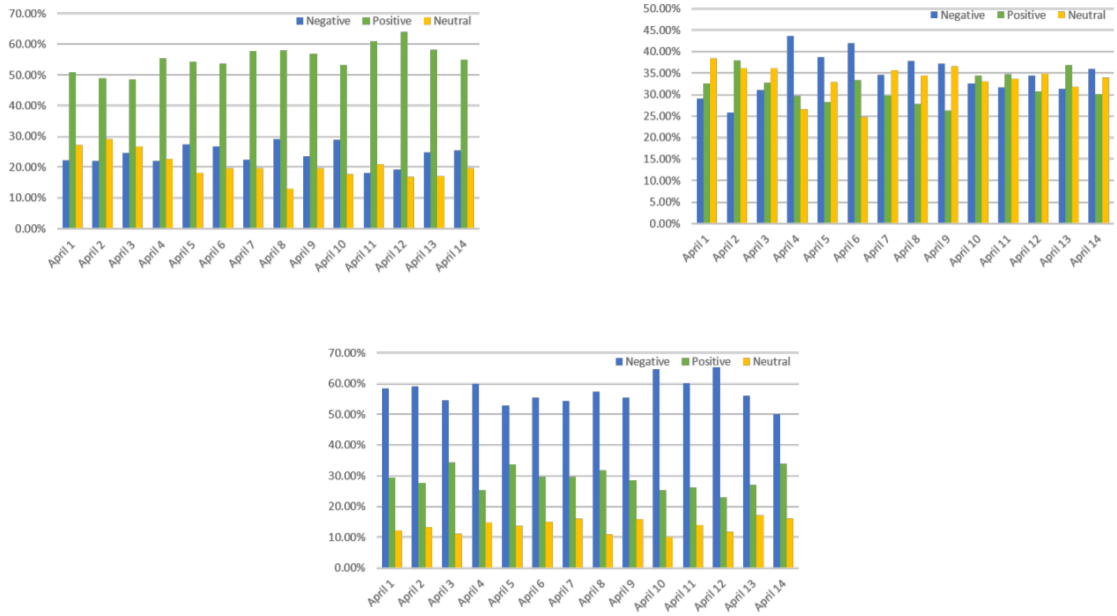
Figure 7: Top 3 COVID-19 Topics

(a) Topic 11 (b) Topic 49 (c) Topic 64

VADER[3] is utilized to classify the sentiment polarities into positive, neutral, and negative categories in the range of -1 to 1. Sentiment polarities calculated as greater than 0.05 are considered positive, less than -0.05 are considered negative, and between 0.05 and -0.05 are considered neutral. Topic sentiment was measured such that the tweets were clustered into their corresponding topics and marked with sentiment polarity. In this case, VADER[3] produced the sentiment polarity of each tweet and combined it with the topic mining result from DTM[4] to analyze the topic-level sentiment. To estimate the overall distribution, the topic sentiment was aggregated whereas the distribution for topics per day was estimated by summing the number of positive, negative and neutral tweets in the topic. Thus, each topic was associated with three sentiment counts.

Figure 6[7] presents the overall sentiment distribution on Twitter during the two-week period of April 1-14. There were approximately 350,000 tweets per day that referred to COVID-19 in which the daily number of positive/negative tweets was similar, but still greater than the number of neutral tweets. Despite the spread of COVID-19, the tweet data showed an overall positive sentiment during the study period. Figure 7[7] displays the most significant words associated with the top three topics-topics 11, 49, and 64-that reflect the common concerns discussed by people. The top three topics consist of staying at home to ensure safety, the latest case reports, and deaths due to the disease. By analyzing the proportion of sentiment polarity of tweets under these topics, Figure 8[7] exhibits changes in user's sentiment as the pandemic spreads. Similar to the prior research paper, sentiment polarity maintained an overall positive attitude despite concerns expressed within every topic. The authors hope that the study can assist in predicting the mental health of Twitter users to future news events, with past reactions as a baseline.

Figure 8: Sentiment Polarities for Topics 11, 49 & 69



3 Conclusion

The coronavirus proved to be a resourceful topic to study given the impact that it continues to have on the world and the ability to gather data on it. The research papers above demonstrated the versatility of sentiment analysis and its ability to quantify the social sentiment of Twitter users in three different approaches. All three analyses provide great insight on the evolution of sentiment analysis and motivation to apply these techniques in future studies on COVID-19 related data or other topic areas.

4 References

[1] A. Both, A. Hinneburg, M. Röder: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining. pp. 399–408 (2015)

[2] Bhavya Ahuja Grover, Nikhil Kumar Rajput, and Vipin Kumar Rathi. Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. arXiv preprint arXiv:2004.03925, 2020.

[3] C.J., Hutto, 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)

[4] D.M. Blei, J.D. Lafferty: Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. pp. 113–120 (2006)

[5] Emily Chen, Emilio Ferrara, and Kristina Lerman. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance, 6(2):e19273, 2020.

[6] Gerardo Chowell, Guanyu Wang, Jingyuan Yu, Juan M. Banda, Ramya Tekumalla, Tuo Liu, and Yuning Ding. A Twitter Dataset of 100+ million tweets related to COVID-19, March 2020. This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates.

[7] Hui Yin, Jianxin Li, and Shuiqiao Yang. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. arXiv preprint arXiv:2007.02304, 2020.

[8] Ines Montani and Matthew Honnibal. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[9] Jacob Devlin, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[10] Mike Izbicki, Stefanos Stoikos, "Multilingual Emoticon Prediction of Tweets about COVID-19", (2020), Bertmoticon, https://github.com/Stefanos-stk/Bertmoticon/blob/master/paper/main.pdf

[11] Steven Loria. textblob documentation. Release 0.15, 2, 2018.