Predicting the Emotional Valence of COVID-19 Related Tweets

TU, WILSON, The University of British Columbia, Canada
HABIB, SHUAIB, The University of British Columbia, Canada
ADRIAN-HAMAZAKI, ALEXANDER, The University of British Columbia, Canada
MAHAPATRA, TARANG, The University of British Columbia, Canada

1 ABSTRACT

ABSTRACT During the COVID-19 pandemic, social media has been a catalyst in the dissemination of information. This study investigated distribution patterns of news utilizing data from Twitter. Tweets with key words and phrases related to the pandemic were identified, and characterized based on their emotional valence using a machine learning algorithm. Additionally, natural language processing was utilized in order to examine the usage of different words to further characterize the sentiment of a tweet. Using these characterizations, patterns of their dissemination were analyzed. The analysis demonstrated a slight correlation between the sentiment of a 'trending' tweet and its' popularity, however, this does not hold when all tweets are considered. Utilizing our findings, we propose a more precise and refined machine learning model[1]. Targeted investigations into patterns of dissemination may aid in stopping the spread of misinformation, thus supporting the public health measures for the COVID-19 pandemic.

ACM Reference Format:
Tu, Wilson, Habib, Shuaib, Adrian-Hamazaki, Alexander, and Mahapatra, Tarang. 2021. Predicting the Emotional Valence of COVID-19 Related Tweets. 1, 1 (June 2021), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 INTRODUCTION

During the COVID-19 pandemic, social media has played a large role in the spread of information to the general population. This has effectively had an impact on the efficacy of implemented public health measures, as widespread dispersion of unverified claims may impact the responses to these measures from the population. This is demonstrated by the response from the public to false claims such as "hydroxychloroquine can prevent covid", or "injecting disinfectant might kill the virus" headlined across North America. With the sheer amount of false claims being generated and the algorithms embedded in social media designed to keep individuals on an app for longer, creating Echo Chambers, it is highly unlikely to monitor and prevent all misinformation spread. It is imperative that we understand the way these claims disseminate, as it can help lead to further intervention and education, which may lead to more successful public health measures.

In this study, we sought to investigate the dissemination of information by using tweets from Twitter as the primary data source. We hypothesized that the dissemination would be related to the emotional valence (ie. positive, negative, neutral) of the tweet. We expected that tweets that demonstrated high valence (either highly positive or negative) would be more likely to be re-tweeted.

3 METHODS

Our study is aimed at analysing the trend in people's emotions over time. We chose to analyse tweets from Twitter as a result of the widespread use of the Twitter platform. As a result, extracting sentiment information from tweets over time using Natural Language Processing tools would be a good representation of people's emotions over time and place. Our study can have widespread policy implications. A simple example would analysing areas in which people are the most distressed as a result of the pandemic and sending assistance to those areas or analysing which places have the most negative sentiment towards vaccines and devising appropriate measures to curb the spread of the virus in those areas. To proceed with our analysis, we first pre-processed our tweet data to clean it up. We then did data exploration by investigating if there is a difference in the semantics used in tweets across our three emotional valence categories - positive, negative, neutral. Lastly, we built a classifier that given a tweet, would be able to categorise the tweet in one of the three sentiment categories as listed above.

3.1 Server

All of the tweets and their associated information are stored on our private Windows Subsystem for Linux 2 (WSL2) remote machine. The remote machine is then protected to allow only the authors of this article to have Secure Shell Protocol (SSH) access.

3.2 Database System

A MySQL Server is set up in the remote machine to allow the storage and manipulation of the tweets.

3.3 Dataset

Our analysis was performed on a subset of the "Covid-19 Tweets Dataset (SITE). The dataset contains 1,633,553,378 Covid-19 related tweets from across the world and in multiple languages; these tweets have been collected since January 22nd, 2020. However, due to Twitter's distribution policy, these datasets only contain "tweet ids" which can use used as keys to retrieve the full information regarding a tweet. Due to Twitter's limitation of the twitter Application Programming Interface (API) which only allows small subsets of

[1]our github link is github.comTu1026Big_Data_Challenge_Team_46

Authors' addresses: Tu, Wilson, The University of British Columbia, Vancouver, Canada, s31302@gmail.com; Habib, Shuaib, The University of British Columbia, Vancouver, Canada, ~~~~; Adrian-Hamazaki, Alexander, The University of British Columbia, Vancouver, Canada, alexadrian@hotmail.ca; Mahapatra, Tarang, The University of British Columbia, Vancouver, Canada, .

XXXX-XXXX/2021/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

tweet information retrieval in a short amount of time, we were only able to retrieve 5 million tweets in the given time. (However, all of the analysis here is run on a smaller subset of that data set which consists roughly 10,000 tweets due to our limited computing power). Our dataset has 12 distinct columns. Below are the columns are their meaning:

- TweetID[2] - The TweetID of a Tweet
- SentimentLabel[3] - Labels a tweet as having negative, positive, or neutral sentiment
- LogitsNeutral - The neutral score of a tweet
- LogitsPositive - The positive score of a tweet
- LogitsNegative - The negative score of a tweet
- CreatedAt - When was the tweet made
- location - If available, the location of the user
- OriginalTweet - The TweetID of the original tweet if this tweet is a retweet
- followersCount - The number of followers the user has
- UseriD- The unique identifier of the user
- Content - The content of the tweet
- RetweetCount - The number of retweet that the tweet has

The 3 columns, SentimentLabel, LogitsNeutral, LogitsPositive, and LogitsNegative are taken from the results of the paper "An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic" by Christian Lopez and Caleb Gallemore. These results were generated using state-or-the-art Twitter Sentiment algorithm BB_twtr.

3.4 NLP Cleaning

Prior to performing our sentiment analysis the tweets were cleaned of information that might hinder our analysis.To do so, we applied natural language procesing (NLP) techniques, removing extraneous information such as replacing integers with their textual representations, removing non ascii characters, and removing URLs.

3.5 Sentiment Analysis

We employed the following Machine Learning classifier algorithms: Naive Bayes, Random Forest, Logistic regression, SVM, an Ensemble model that regrouped all four of these models as well as two wildly employed and effective deep learning NLP models called the BERT and the LSTM model. We analysed the performance of our algorithm on our data set based on the following metrics: precision, recall, f1-score and support as well as visualising the confusion matrix and chose the best model taht gave us highest accuracy without overfitting. We also made use of Randomized Search CV to tune our hyperparameters for all our ML models excluding the deep learning ones and Random Forest. For the Random Forest algorithm, we made use of SMAC, which is an AutoML package for fast Hyperparameters tuning in Python to speed up computation.

[2]A retweet can have a TweetID just like a original Tweet
[3]The sentiment label from our dataset is our predictor variable and the content of the Tweet is our explanatory variable.

4 RESULTS

4.1 Natural Language Processing Analysis

The following graphs were compiled based strictly on the natural language processing done to the tweets.

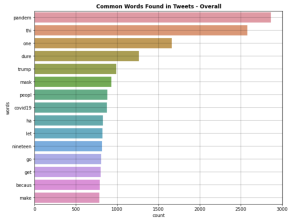

Fig. 1. Most common words found in tweets

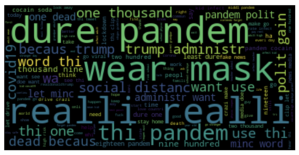

Fig. 2. Generated word cloud of tweets

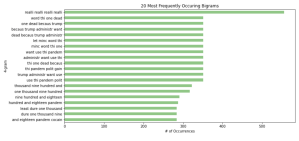

Fig. 3. Most common bigrams found in tweets

4.2 Dataset Valence Categorization

Utilizing the machine learning algorithm, the dataset was categorized based on emotional valence (positive, negative, neutral).

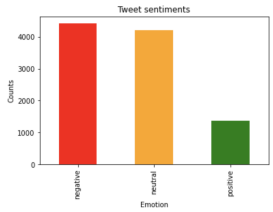

Fig. 4. Number of tweets in dataset categorized by emotional valence

4.3 Natural Language Processing Analysis with Dataset Valence Categorization

Finally, natural language processing analysis was done on the categorized dataset, and characterizations of the tweets based on their emotional valence were elucidated. These include: a) proper nouns, b) readability scores (measure for ease of reading text) c) stop words (common words adding little value to a sentence, ie. the, as, a), d) capital letters e) type-token ratios (TTR - defined as the total number of unique words divided by the total number of words in a given segment of language, which is a primary indicator of 'richness' of a text)

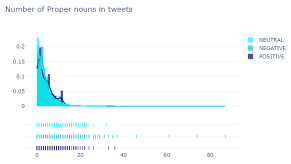

Fig. 5. Number of proper nouns within tweets categorized by emotional valence

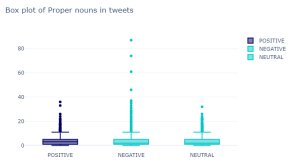

Fig. 6. Boxplot of the number of proper nouns categorized by emotional valence

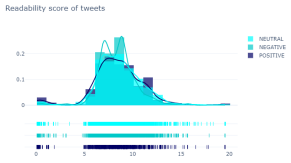

Fig. 7. Readability of tweets categorized by emotional valence

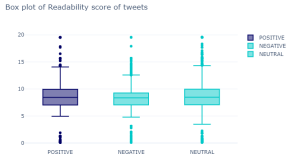

Fig. 8. Boxplot of readability of tweets categorized by emotional valence

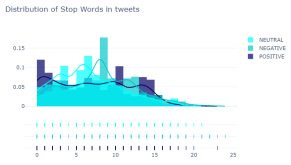

Fig. 9. Distribution of stopwords within tweets categorized by emotional valence

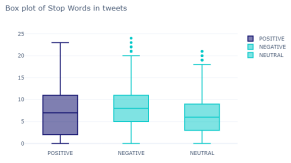

Fig. 10. Boxplot of stopwords within tweets categorized by emotional valence

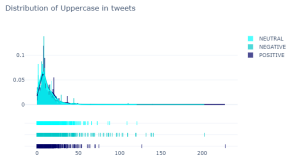

Fig. 11. Distribution of uppercase letters within tweets categorized by emotional valence

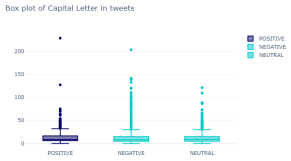

Fig. 12. Boxplot of uppercase letters within tweets categorized by emotional valence

4.4 Models and their performance scores

These were the models we arrived at after training the models on our data set and tuning hyperparamaters: MultinomialNB

RandomForestClassifier with 100 trees, max depth of 6, min samples leaf of 6, min samples split of 50 and class weight being 'balanced'

Linear SVC with C as 1 , max iter as 500

LogisticRegression with C as 2 and max iter as 10.

Ensemble model : VotingClassifier with SVC, Naive Bayes, Logistic, Random Forest

We made use of a 12 layer model for BERT and the following model initialization:

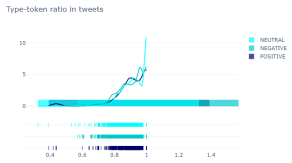

Fig. 13. TTR of tweets categorized by emotional valence

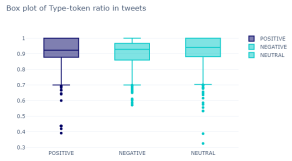

Fig. 14. Boxplot of TTR of tweets categorized by emotional valence

Comparison of all algorithm results

Model	Accuracy
SVM Algorithm	0.83
Naive Bayes Algorithm	0.77
LogisticRegression Algorithm	0.82
Random Forest Algorithm	0.8
Ensemble Modelling	0.82
LSTM Modelling	0.8
BERT Modelling	0.79

Fig. 15. TTR of tweets categorized by emotional valence

$model = BertForSequenceClassification.from_pretrained("bert-base-uncased", numlabels = len(labeldict), outputattentions = False, outputhiddenstates = False$, where label dict represents our three emotional states encoded as $0 = positive, 1 = neutral, 2 = negative$.

For LSTM, this was our model initialization: Sequential model with 0.2 SpatialDropout1D, 100 LSTM layers, 3 dense layers with softmax activation, categorical crossentropy loss, Adam's optimizer, 5 epochs and a batch size of 64. Logistic regression gave us the highest accuracy out of all models with an accuracy score of 0.82.

ACKNOWLEDGMENTS

Acknowledgments to Hebah Hussaina of University of British Columbia for helping us out with the science communication component of our discussion. Acknowledgements to Lopez and Gallemore for their open source datasets.

REFERENCES

[1] Cliche, Mathieu. *BB_twtr at SemEval2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs.* ArXiv:1704.06125 Cs, Stat Apr 2017.
[2] Lopez, Christian: Lopezbec/COVID19_Tweets_Dataset, https://github.com/lopezbec/COVID19_Tweets_Dataset