COVID-19 Tweets Classification

Kopal Garg

*Biomedical Engineering*

*University of Waterloo*

*Waterloo, Canada*

kopal.garg@uwaterloo.ca

# Abstract

*Around March 2020, the outbreak of a global pandemic was declared by the World Health Organization. Naturally, the pandemic caused global outrage – people shared their opinions on public health responses along with their sentiments in form of fear, and panic that was fueled by incomplete or inaccurate information. The main objective of this paper is to explore these behaviors of societies at large by analyzing this informational crisis at a time when COVID-19 approached peak levels across the globe. More specifically, this entails using COVID-19 related tweets for textual preprocessing, data visualization and analytics as well as a comparison of 4 supervised sentiment classification models: Linear Support Vector Machines, Naïve Bayes, Decision Trees and K-Nearest Neighbor. The impact of data preprocessing and feature extraction methods like Bag-of-Words, term frequency-inverse document frequency and hashing were explored on each model. The data used was a collection of ~45,000 Tweets found in the Coronavirus Tweets NLP Dataset on Kaggle. Among all approaches, \_\_ proved to have the best prediction accuracy of \_\_%.*

**Keywords**: COVID-19, Sentiment Analysis, Twitter, Textual Analytics, Linear Support Vector Machines, Naïve Bayes, Decision Trees, K-Nearest Neighbor, Bag-of-Words, TF-IDF, NLP

# Introduction

As the world faces the Coronavirus disease (COVID-19), Twitter has become a significant resource that helps relay important information to users in real-time. As the global conversation continues around the spread of the pandemic, tens of millions of Tweets on this topic have been shared. Given the sensitive and evolving nature of this topic, people continue to feel strongly about public health policies, and announcements. They rightfully feel the need to debunk misleading or unverified claims and typically do so by sharing or responding to sentimental Tweets. Some Tweets are positive, while some incite panic. In order to protect public conversation, Twitter is creating new rules surrounding behaviors that are considered acceptable on the platform [1]. It has a zero-tolerance policy for any attempts of abuse or malicious behaviors [1]. Given the volume of Tweets that are shared per second, it is largely impossible to manually identify those that are positive or negative. As expected during times of crises, negative emotions are dominant. The main motivation behind this paper is to identify Tweets that are excessively negative and may potentially lead to panic and public outrage. To maintain the public’s mental well-being, these Tweets should be counterbalanced with strategic public health communication. This paper presents a method of predicting whether a Tweet indicates a positive, negative or neutral sentiment.

Since the sentiment analysis focuses on textual data, this method involves a fair bit of text pre-processing.

Twitter has become a considerable source for streaming data that is specifically useful in prediction, knowledge extraction and analytical tasks in machine learning (ML).

* Pandemic – outrage, sharing of sentiments to public health policies
* Type of data being used – and how tweets have replaced traditional news outlets
* Purpose of the analysis

# Background Review

Similar methods have been implemented in the past. Samuel et al., published a method of identifying public sentiments associated with the pandemic [2]. They demonstrated the use of two machine learning (ML) classification methods for textual analytics, and compared the effectiveness of each in classifying tweets of varying lengths [2]. They reported a 91% classification accuracy of Naïve Bayes for short tweets, a 74% classification accuracy of Logistic Regression for shorter tweets and relatively weaker accuracies of both methods for longer tweets [2]. Rustam et al., evaluated the performance of various ML classifiers on a training dataset formed by different feature extraction techniques like bag-of-words, hashing and term frequency-inverse document frequency. They proposed a method that involves concatenating the two feature extraction techniques. They trained five ML models, namely, Random Forest, XGBoost, Support Vector Classifier, Extra Trees Classifier and Decision Tree, and used measures like accuracy, precision, recall, and F1 score to evaluate the methods performance.

* Research on current methods of understanding mass behavior
  + Overview of current work in the field
  + Who has worked on this problem?
  + What method have they used
* Literature review
  + Textual analytics
  + Twitter analytics
  + Classification methods
    - Linear regression model
    - Naïve Bayes classifier
    - Logistic regression
    - K-nearest neighbor
  + Benchmarks

# Application/Dataset

The data is obtained from the “Coronavirus Tweets NLP - Text Classification” dataset on Kaggle [3]. It contains 2 .csv files - *Corona\_NLP\_test.csv*, *Corona\_NLP\_train.csv*. Columns include geographic locations, dates on which the Tweets were published, the original/raw Tweets and sentiment labels for a combined ~45,000 Tweets from March to April in 2020. The usernames are coded to avoid privacy concerns. Table 1., contains a high-level summary of the training data.

***Table 1.*** *Training Data Summary*

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Summary** | **Attribute** | **Summary** |
| # Unique Tweets | 41,156 | Timeframe | 03-16-20 to 04-14-20 |
| # Unique locations | 12,221 | # Unique sentiments | 5 |

******Visualization and analytical techniques were implemented to conduct a preliminary exploration of the data. Figure 1.A. represents the frequency at which Tweets were published from March 16 to April 14, 2020. Figures 1.B. and E. show word clouds consisting of the top-50 frequent unigrams and bigrams found in positive and negative Tweets. In Figure 1.B., words, like, “thank”, “good”, “help”, etc. are seen, while in Figure 1.E. words and phrases with negative connotations, like, “panic”, “crisis”, and “oil price”, etc. are frequently observed. Figure 1.C. represents a circular bar plot showing the top-5 locations from which Tweets were published. Figure 1.D. represents the sentiment distribution in Tweets. It is observed that the number of Tweets with a positive polarity is fairly larger in the training dataset.

**B**

**A**

**D**

**E**



**C**

***Figure 1. A.*** *Tweet Frequency.* ***B.*** *Word Cloud – Top-50 Positive Words.* ***C.*** *Top-5 Locations.*

***D.*** *Sentiment Distribution.* ***E.*** *Word Cloud – Top-50 Negative Words.*

# Proposed Scheme/Algorithms

A preprocessing pipeline was written to create a data set ready for further analysis. This involved eliminating textual or non-textual variables in the data that don’t necessarily contribute to sentiment analysis. Twitter handles, URLs, non-ASCII or special characters, punctuation and numeric values were filtered out and the Tweets were lower cased. Using a corpus of stop-words from the Natural Language Toolkit (NLTK) suite, words with little lexical content (e.g. *the*, *a*, *also*, etc.) were removed. Using the emot package, all emojis and emoticons were replaced with their Unicode CLDR short name. The pyspellchecker package was used to return the most probably results for misspelled words. Segments of text were further processed with tokenization and lemmatization, both of which were performed using packages in the NLTK suite. Tokenization converts text to analysis relevant word tokens while lemmatization transforms words to a simpler form, returning the word’s lemma, which is a canonical form of all its inflectional forms (e.g. *go* represents its inflected forms of *goes*, *going*, *went*, *gone*).

*Input:*‘Coronavirus Australia: Woolworths to give elderly disabled dedicated shoppeng hours amid outbreak ☺ <https://t.co/bInCA9Vp8P>’

*Output:*[‘coronavirus australia woolworths give elder disable dedicate shop hour amid outbreak happy face smiley’]

Three techniques for encoding text data into numerical vectors were implemented to extract suitable features from the processed text.

*Feature-set I*: Bag-of-Words (BoW) coverts variable-length texts into fixed-length vectors, without considering the semantic relation between words. Since the dataset is large, it can contain a vocabulary of a few thousand words. Preprocessing text before employing this technique is therefore useful. Using unigrams or words as features, both of the sentences, although differing in sentiment, will be given the same score. Hence unigram features may be insufficient in discriminating the two sentiment classes.

*Feature-set II*: The Bigram BoW (BBoW) also represents a text document as a weakly ordered collection of contiguous sequences of 2 items. However, it allows for the preservation of more word locality information. Using bigrams, both sentences would generate different sets of features, and varying subjectivity scores may be assigned. It is therefore hypothesized that bigrams can substantially raise the quality of the feature set.

*Feature-set III*:

*Classification Methods:*

*Validation Methods:*

# Experiments and Results

* What was found
  + Include tables, figures and analysis
  + Comparison of all classification algorithms

# References

[1] <https://blog.twitter.com/en_us/topics/company/2020/covid-19.html>

[2] <https://www.mdpi.com/2078-2489/11/6/314/htm>

[3] Coronavirus tweets NLP - Text Classification” dataset