Sentiment Analysis of COVID-19 Tweets

Xinyi Cai, Sherrie Chen, Sebastian Koh

ABSTRACT

This project uses a Kaggle dataset and aims at performing sentiment analysis to tweets posted with covid-19 hashtags. We use NLP techniques such as Bag of Words and TF-IDF to perform feature transformation to textual inputs and use real value encoding to transform the labels. Then, machine learning models such as logistic regression, Multinomial Naive Bayes, linear SVC and gaussian SVC are performed on the dataset with different metrics. Our experiments have shown that linear SVC produces the highest accuracies and can probably be the most suitable model for stakeholders to assess the general public sentiment situation.

1. INTRODUCTION

On March 11^{th} 2020, the World Health Organization (WHO) declared CoronaVirus Disease of 2019 (COVID-19) as a pandemic. This disease transmits easily and fast and it has caused stress, anxiety and a sense of uncertainty to the general public when a lot of non-essential businesses are shut down and people are asked to shelter in place at home. Social media has played an important role for individuals to showcase their lives and express their feelings. Therefore, learning about posts people made on social media can help the government or public health sectors assess the mental health of the general public.

The objective of this project is to use Natural Language Processing (NLP) and Machine Learning (ML) to perform sentiment analysis and classification to Twitter posts with the #covid19 hashtag. The Kaggle dataset we are using in this project contains raw information on the location (categorical), date of tweet (ordinal), the main tweet message, and the sentiments (ordinal). It has approximately 42000 training samples and 4000 test samples in total.

In this report, we will be revisiting previous COVID-19 sentiment analysis research, our approaches to feature transformation, model selection and regularization. We will also compare and discuss our results and try to come up with directions of future work. We will also reflect on any negative impacts of our project and propose ways to limit these impacts.

2. BACKGROUND

We are not alone in analyzing sentiments from COVID-19 related Twitter posts. A study by Samuel et al. (2020) [1] uses logistic regression and Naive Bayes classifiers to analyze tweet sentiments based on a single keyword tracking and achieves accuracy rates of 74% and 91% respectively. Different from our work, they focus on detecting fear sentiments in tweets with relative shorter lengths. Barkur et al. (2020) [2] try to analyze tweets posted by Indians regarding the government's decision on lockdown using WordCloud and find out that Indians took the strategy of the government positively on imposing the lockdown.

A study done by Chakraborty et al. (2020) [3] uses a fuzzy rule based model to handle the uncertainty that is prevalent in raw sentiments. Often these values are truncated because model training usually rounds up the values when it tries to label different classes. In their research, instead of classifying the sentiments into only five types, they use real values as sentiment scores. They also use a gaussian membership function to characterize the fuzziness of the model. This model can achieve an accuracy rate of up to 81%. This studies has found that even though people have tweeted mostly positively regarding COVID-19, netizens have been re-tweeting the negative tweets and no useful words could be identified by word frequencies.

Aside from categorizing tweets as positive, neutral and negative, a study done by Hung et al. (2020) [4] has identified COVID-19 related tweets into five themes: health care environment, emotional support, business economy, social change, and psychological stress. Latent Dirichlet Allocation (LDA) [5] was used to extract the hidden semantic structures of these tweet posts. It is an unsupervised ML method that groups common words into multiple topics. They have used Valence Aware Dictionary and sEntiment Reasoner (VADER) to determine whether a certain tweet post is positive, neutral, or negative, as well as the degree of each sentiment.

3. DATA PREPROCESSING

3.1 Data Description

Our dataset contains raw information on the location (categorical), date of tweet (ordinal), the main tweet message, and the sentiments (ordinal). It contains approximately 42000 training samples and 4000 test samples.

3.2 Preprocessing

The purpose of preprocessing is to remove information unrelated to semantics so as to ensure consistency for the words across the corpus. Firstly, we dropped duplicate data and checked for missing values. We discovered that 20% of location data are missing, and all the other data fields contain no missing values. Next, we clean each

original tweet by removing https links and urls, hashtags and mentions, punctuation and stop words. Additionally, we converted numerical digits into their equivalent textual representation and set all characters to lowercase. Because some words are in abbreviation, for example, "I'll," we created a dictionary of contractions and mapped abbreviated words into their expanded form.

After running preliminary analyses for the midterm report, we also tried to combine five labels into three labels: extremely negative and negative labels will just all be negative labels; extremely positive and positive labels will just all be positive labels. By combining the labels, we managed to achieve higher accuracy rates for the models in use.

3.3 Recap on Data Exploration

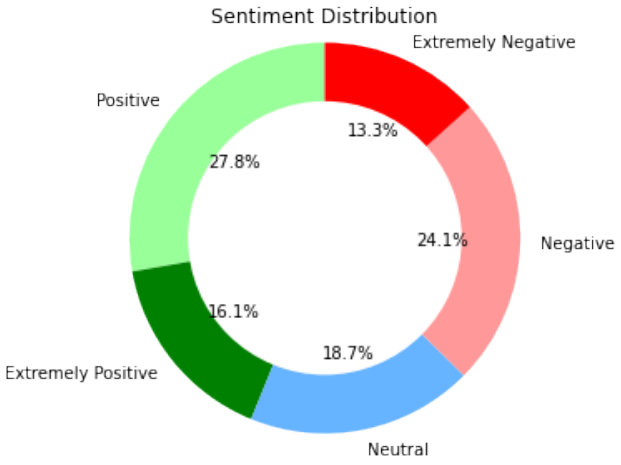

Figure 1: Distribution of Sentiments

The graph above shows an even distribution of samples across different sentiment classes, which favors training of the ML models.

The most frequent words from the tweets generally include the discussion of the virus; "coronavirus" and "pandemic." People also care about their food supplies and the availability of "grocery" or "sanitizer." The frequency shows that the discussion of COVID19 on Twitter mostly involves basic needs in daily life.

The most common bigrams, which are adjacent two words in text, demonstrates that many users are concerned about supplies (i.e. toilet paper, hand sanitizer, grocery). The term "panic" appears both in most common unigram and bigram, which indicates that frequent words related to emotions might be helpful for classifying the tweet sentiment.

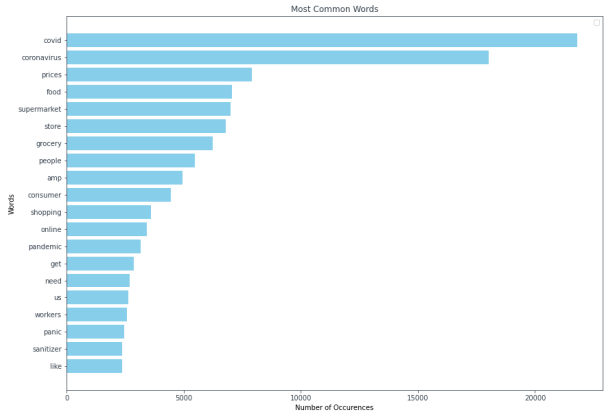

Figure 2: Frequency of the 20 Most Frequent Words from Tweets

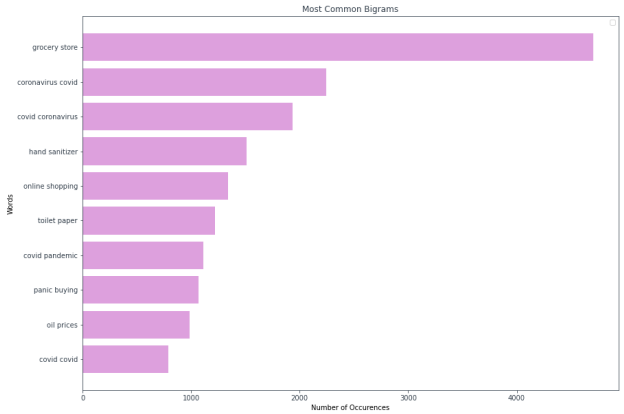

Figure 3: Frequency of the 10 Most Frequent Bigrams from Tweets

3.4 Feature Transformation

Our previous data visualizations showed the usefulness of the data inputs based on the correlation between the available information and our desired output, as well as our sense making hypothesis on the application. These data inputs will undergo different feature transformation techniques according to their suitability for the contexts and data types.

3.4.1 Natural Language Processing

The content within the tweet message will form the most relevant feature for our model. NLP will be performed on the tweet message to extract useful features. For

our project, we explore 3 methods for comparison; (1) Bag of words (BOW) Count, (2) Term Frequency–Inverse Document Frequency (TF-IDF). In addition, we will also extend these 3 models to n-grams where we believe the occurrences of certain words in certain sequences will help in our prediction model. Specifically, our models will have features expanded to a single word and bigram (2 consecutive words).

- Bag of Words (BoW): We simply broke down all texts into words and assigned unique tokens to unique words. Each token was assigned as a feature variable and the value of that variable for each tweet sample would signify the frequency count of the unique word in it.
- TF-IDF: Each tweet was represented by a numerical value that reflects the importance of each word in the tweet to the whole data set. The TF-IDF value increases proportionally to the frequency count of the words in the tweet and is offset by the number of tweets in the training dataset that contains the same words. TF-IDF is more complex than BOW count but it has some element of normalization that re-adjust the feature input value according to how common the same words appear in different samples.

3.4.2 Real Value Encoding

The simplest encoding is to use the real value parameter in its raw form. The length of the tweet message will be perceived from each tweet and input as a feature in its real value form. We believed that the length of a tweet message might have some correlation to the severity of the sentiment. For instance, someone who is extremely unsatisfied or extremely satisfied with the COVID situation might have a higher tendency to express his or her comments in a lengthy manner. This length of the tweet message will be used directly as a feature input.

Besides that, exploration could also be done on extracting real values from the date of the Tweet message as the sentiment could be strongly correlated to when the tweet message was posted. Specifically, we believed that a tweet made nearer to the date when the COVID pandemic happened might contain more element of uncertainty and stronger sentimental value, as compared to a tweet that is made months after the pandemic start date. This portion has yet to be explored in our project but will be a feasible next step for improving the models.

3.4.3 One-Hot Encoding

One-hot encoding allows us to represent categorical variables into a binary vector. As our dataset contains data on location where the Tweet message was made, it is possible to extend all possible unique locations as features and use one-hot encoding to represent the location of each data point. In such case, a sparse matrix would be obtained to ease training of the Machine Learning (ML) model. We believe that the

location could provide some correlation in the general sentiments in that area. This might help in the sentiment prediction of a new tweet if there is strong correlation between location and sentiment value. Alternatively, the location information could also be harnessed via its coordinates and zip code via a library "Geopy". Similarly, this portion has yet to be explored in our project but will be a feasible next step for improving the models.

4. MODELS

4.1 Multinomial Logistic Regression with L2 Regularization

We chose logistic regression because it is a probabilistic model that makes no assumptions of the distributions of labels in the feature space. We employed l2 regularization because it prevents overfitting and does not shrink less important features to zero.

$$\ln\left(L(\beta_i \mid x_i)\right) = \sum_{i=0}^{n} y_i \ln\left(P(x_i)\right) + \sum_{i=0}^{n} (1 - y_i) \ln\left(1 - P(x_i)\right)$$

4.2 Multinomial Naive Bayes

As we only included TF-IDF encoded n-grams as our feature, we could assume that the features are conditionally independent. Since we could easily turn words in tweets into counts, we chose the multinomial distribution to represent the posterior probability.

$$P(C_i \mid x_1, x_2 \ldots, x_n) = \left(\prod_{j=1}^{j=n} P\left(x_j \mid C_i\right) \right) \cdot \frac{P(C_i)}{P(x_1, x_2 \ldots, x_n)} \quad for\ 1 < i < k$$

4.3 Linear Support Vector Classifier (SVC) with Squared Hinge Loss

We chose Linear SVC because it applies a linear kernel function to classification problems and they work well for large datasets. The Linear SVC maximizes the width of the soft margin, and it transforms non-linearly separable data into dimensions that can be separated by a hyperplane. We also included Gaussian SVC because it uses the "one-vs-one" approach for multi-class classification instead of "one-vs-all" in Linear

SVC. We employed squared hinge loss because it is zero when the prediction matches the label and quadratically increasing with the error when the prediction is incorrect.

$$L(y, \hat{y}) = \sum_{i=0}^{N} \left(max(0, 1 - y_i \cdot \hat{y}_i)^2 \right)$$

5. RESULTS AND DISCUSSION

To understand how well our models performed, we employed four metrics: accuracy score, f1 score, precision score and recall score. Accuracy score outputs the ratio of the correctly labeled test samples. Precision score tells us about the ratio of false positive samples to the sum of true positive and false positive samples (fp/(tp+fp)). Recall score tells us about the ratio of true positives to the sum of false negatives and true positives (tp/(fn+tp)). F1 score can be interpreted as a weighted average of precision and recall scores (F1= 2*precision*recall/(precision+recall)). It reaches its best value at 1 and worst value at 0.

Metrics	Multinomial Naive Bayes	LVC	Logistic Regression
train accuracy	0.659	0.786	N/A
test accuracy	0.668	0.793	0.375
f1 score	0.574	0.772	N/A
precision score	0.694	0.774	N/A
recall score	0.576	0.769	N/A

The results for the three models shown above showed that linear SVC yielded the highest test accuracy and will be the most suitable model for predicting the general sentiments of a cluster of tweet messages.

However, we noted that NLP converts a text into data inputs for many word features where the features are the unique n-gram words found in our whole training dataset. Extending the features to larger n-grams may result in more accurate NLP analysis, but at the expense of much more features and computational power. For our project, we explored up to Bi-grams, which entails a total of 10449 features.

While our test and training data accuracy seems promising, we are not confident enough to use the model for production due to the low training dataset size to features ratio (42000 data points to 10500 features). We assessed that for NLP feature transformation, a larger training dataset is required. In general, our model should suffice to provide some preliminary assessment on the general sentiments on the public's view towards COVID.

While our model may serve as a prediction model for the public to gain awareness on the general sentiment trend of the tweeters, the prediction results could influence how stakeholders perceive the COVID situation and in turn, steer the new tweets towards similar general sentiment. This potentially causes a Weapon of Math Destruction (WMD) tool where our predictions can create a vicious cycle of negative consequences in continuous feedback loops. Such new data points (tweets) will be biased and detrimental to the management of COVID situation.

Hence, our model should only be made available to stakeholders who have non-profit interests in improving public sentiments on the COVID situation. For instance, the government could use our model as a distress alarm when the general sentiment trend approaches extremely negative and counter the low public confidence with strategic campaigns. Our model also intuitively encourages fairness as the sentiments in the training data were classified based solely on the text in their tweet instead of the demographics of the tweeters. This allows an unbiased fair prediction of any new tweet using our model.

6. CONCLUSIONS AND FUTURE WORK

We first performed data visualization on our Tweet training dataset to verify the feasibility of the data for training ML models and explored the common words (unigram, bi-gram) used in the tweets to assess the possibility of performing NLP on the models, as shown in the midterm report. Following that, we processed and cleaned the data before performing feature transformation techniques (e.g. one-hot encoding, TF-IDF) to prepare the data set for model fitting. Finally, we used the selected features and fit the processed training data into 3 different models (Logistic Regression, Multinomial Naive Bayes, Support Vector Classifier). Our results showed that linear SVC produced the highest accuracy and could probably be the most suitable model for stakeholders to assess the general public sentiment situation.

However, the model could be further improved in the future with other word embedding techniques such as Word2Vec. Furthermore, sentence embedding, such as Infersend and Doc2Vec, can add the representation of word vectors and allow us to detect the intent of the tweet. Other than alternating the vectorization of texts, one could also explore the possibility of incorporating features relating to the tweet location and date. Ultimately, a classification model product with higher accuracy would be desired by stakeholders to make correct judgement and the associated policies to address the general population's needs during pandemics or public health crises.

7. REFERENCES

1. Samuel J., Ali G.G.M.N., Rahman M.M., Esawi E., Samuel Y.COVID-19 public sentiment insights and machine learning for tweets classification Information, 11 (2020), p. 314 (2020). https://doi.org/10.1016/j.asoc.2020.106754

2. N2-Barkur G., Vibha, Kamath G.B. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India Asian J. Psychiatry, 51 (2020), Article 102089 Advance online publication. https://doi.org/10.1016/j.ajp.2020.102089

3. Koyel C., Surbhi B., Siddhartha B., Jan P., Rajib B., Aboul E. H. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Applied Soft Computing Volume 97, Part A, December 2020, 106754

4. Man H., Evelyn L., Eric S. H., Wendy C. B., Julie X., Sharon S., Shirley D. H., Jungweon P., Peter D., Martin S. L. Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence. Journal of Medical Internet Research, Vol 22, No 8 (2020). https://www.jmir.org/2020/8/e22590/

5. Rodriguez-Morales AJ, Cardona-Ospina JA, Gutiérrez-Ocampo E, Villamizar-Peña R, Holguin-Rivera Y, Escalera-Antezana JP, Latin American Network of Coronavirus Disease 2019-COVID-19 Research (LANCOVID-19). Electronic address: https://www.lancovid.org. Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis. Travel Med Infect Dis 2020 Mar;34:101623