Sentiment Analysis on Covid-19 Related Tweets

Wai Phyo
Department of Computer and Information Science, Fordham University
New York City, USA
wphyo@fordham.edu

Saroj Behera
Department of Computer and Information Science, Fordham University
New York City, USA
sbehera@fordham.edu

Abstract—We aim to analyze the Twitter messages, also known as “tweets”, related to Covid-19 posted by millions of users. With the pandemic, the public has been overwhelmed with information whether true or false; from treatment news to masks and from origin of the virus to how it spreads. There is an urgent need of mitigating those false information so that health officials and policy makers can make appropriate decisions. We will identify the public sentiment trends related to the pandemic using sentiment analysis packages. We looked into the sentiment insights leading up to the first peak in the United States, using subjectivity and polarity scores. In addition, we implemented the regression model to predict sentiment scores and fine-tuned the model by using ridge regularization and 10-fold cross validation. We observed the minimization of test error by adding bi-gram to the word features extraction. Furthermore, our experiment on clustering provides insights into topic modeling in states with most cases and topic shift within the dataset.

Keywords—Covid-19, Coronavirus, sentiment analysis, text mining, opinion mining, topical clustering, Twitter

I. INTRODUCTION

With the on-going pandemic, the public has been fueled by information from every possible source and channels. One of the most popular microblogging sites, Twitter, has become a public opinion dataset with Twitter messages, called ‘Tweets’ from worldwide with no more than 280 characters to post (the new limit is expanded in 2017 from 140 characters). Tweets can express anything from opinions, announcements, or in some cases, misinformation. Our work will focus on the public sentiment from tweets related to Covid-19 and analyze the progress of sentiments over time as the pandemic has approached its peak in the United States.

In sentiment analysis, the goal is to determine opinions and emotions from short texts, sentences or reviews. There are different approaches in labeling the sentiments, as in [1] tweet can be expressed as polar which is either positive or negative, otherwise, it’s neutral. Another approach is having multi-dimensional emotions: joy, sadness, anger, fear, etc to express sentiments.

The sentiment analysis is quite ubiquitous in products and other domains because product reviews, movies reviews or restaurant reviews are quite polarized and the rating scores or review stars are often used in building training sets. Our challenges in tweet sentiment analysis are: (i) neutral tweets are as much common as positive and negative tweets. (see Fig. 1) (ii) limited character in tweets makes it harder to grasp the sentiment cues and most of neutral tweets are also concentrated on shorter tweets. (iii) Pre-existing bias on the data about global crisis itself, since the dataset is extracted from Covid-19 related tweets specifically with keywords "COVID", "Coronavirus", "Chinese virus", "school closure", "school closed", "Food scarcity", "reopen business", it implies underlying negative sentiment on the event itself.

Our inspiration is to explore how most impacted states are expressing in terms of sentiments over time and if the sentiments change regarding lockdown, reopening and school closures. Our paper can be summarized into: (i) analysis of the overall US sentiment over the course of March 25 to June 2, with a closer look at trends in California, Florida, Michigan, New Jersey, New York and Texas; (ii) regression model with unigram and bi-gram for feature representations and tune the model by using the ridge regularization and cross-validation method to minimize the test error; (iii) topical clustering using standard Latent Dirichlet Allocation (LDA) clustering and Mallet implementation, using the perplexity and coherence score in both models.

The rest of our paper will focus on the following: section II literature review on related work. Section III will be on general background on sentiment analysis, regression and clustering. Then, section IV will be our approach and methodology, followed by section V, results and section VI, concludes our paper with further discussion on future work and remarks.

II. LITERATURE REVIEW

Most of the literature and research are sentiment analysis on specific product/domains such as movie reviews, restaurant reviews, hotel reviews, etc. Those are predominantly positive or negative and those reviews mostly come with ratings. Therefore, they are easier approaches to train the data based on the ratings.

Several literature for tweet sentiment analysis has pointed out approaches to build the training sets. As in [2], the paper uses the emoticons or hashtags to identify the polarity of the tweets and class labels. But tweets with both positive and negative emoticons are removed to avoid the ambiguity while
labeling. As in [1], one of the datasets is also obtained from the same approach using the tweets scraper to send a query to positive emotion :) and another query, to negative emotion :( at the same time. The rest of the dataset are either hand-classified or by crowd-sourcing sites like Amazon Mechanical Turk.

There has been extensive research in sentiment classification as in [1-2] on tweets specifically on consumer product research or specific events like the Obama-McCain debate from 2008 or Health Care Reform 2010, etc. Since we are in the midst of Covid pandemic, there has been limited research on Covid related research. The most recent one that was published in June 2020 by J. Samuel et al in [3] used sentiment analysis package in R and analysed the progress of fear sentiment over time as well as compared two classification methods.

A lot of studies do not consider neutral tweets in their research, as in [1-3] and solely focus on binary classification. Research on sentiment regression analysis has been understudied with the exception in [5]. The study by Drake, Ringer and Ventura compare the performance using classification and regression algorithms. Their results show that regression algorithms perform better as the number of sentiment classes grows.

With respect to topical clustering, although LDA has been widely used by researchers due to its modularity and extensibility, there has been limited exploration in terms of its effect on a variety of textual data on generated topics. As in [7], the data consisted of articles from fisheries journals and are purely domain specific. They applied LDA to abstract and full-text data and explored the practical effects on types of documents, taking human topic ranking as gold standard. However, the data normalization was limited and lemmatization wasn’t performed to find the derivational forms of words unlike our approach which aimed at reducing the inflection and analyzing the parts of speech in widely abstract tweets.

III. SENTIMENT ANALYSIS AND METHODS

In this section, we are going to explore some of the general background on sentiment analysis and machine learning methods that we will apply in our experiment with Covid-19 related tweets.

A. Sentiment Analysis

Sentiment analysis is determining if texts or documents contain any contextual polarity and can often be described as labeling sentiments as positive, neutral or negative along with some scores or ratings. Prominent sentiment analysis approaches include Textblob, one of the NLP libraries and VADER, a lexicon and rule-based sentiment analysis tool tailored to sentiments in microblogging sites. Using the two approaches, we can score the subjectivity and polarity scores of tweets.

To understand how the public is reacting to the pandemic on Twitter, we will need to identify if tweets are objective or subjective. Textblob determines the subjectivity of the texts or documents, whether the writer expresses more opinions or fact-based. Textblob scores ranging from 0 being most objective and 1, most subjective. For subjectivity lexicons for adjectives, words from about 2900 lexical features are tagged per sense in different features. While computing the subjectivity score for a single word in Table I, it returns the average scores of all parts of speech tags.

<table>
<thead>
<tr>
<th>Sense</th>
<th>Polarity</th>
<th>Subjectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very good</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Of major significance or importance</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Relatively large in size</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Remarkable or out of ordinary in degree</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Similar to Textblob, as mentioned in [4] from the work of Hutto and Gilbert, VADER relies on a dictionary that maps lexical features to emotion intensities known as sentiment valence. The intensity ratings on each of the 9,000 lexical features were collected from ten independent human raters, totalling 90K+ ratings with validated valence scores to represent both polarity and intensity from a scale of -4 to 4. For instance: “okay” has positive valence of 0.9, and “horrible” has -2.5; the frowning emoticon :( -2.2. It then generates compound score by summing valence scores of words in the lexicon, which then normalized between -1, most negative sentiment and +1 the most positive.

VADER also incorporates word-order sensitive relationships, such as degree modifiers, to increase the sentiment intensity. For example, the sentiment intensity increased from 0.22 to 0.36 in the words ‘good’ to ‘extremely good” while “marginally good” decreased by 0.293 on average. Even with some unique handling such as understanding slang words, word shapes, and degree modifiers, there is limitation on contextual associations in some cases for example, the following tweet is labeled positive while the underlying sentiment is quite negative.

Tweet: Cuomo has blood on his hands and is now attempting to blame the nursing homes!

VADER cannot identify the intensity of the phrase ‘blood on his hands’ and is labeled with positive compound score. In this context, ‘blood on his hands’ would have much higher negative scores than the word ‘blood’ alone. VADER has the challenge of sentiment analysis without the context and word association.
B. Linear Regression

Linear regression algorithm is estimating the values of coefficients from a linear relationship between input variables and single output value. It takes the inputs as a matrix and uses linear algebra operation in equation (1) to estimate the optimal values for the coefficients.

$$w = (X^T X)^{-1} X^T y$$

During model training, regularization can be used to minimize the squared error on the training data but also decrease the complexity of the model. One of the two regularization processes is Ridge regression which minimizes the absolute sum of coefficients by adding a penalty term for model complexity.

In addition, K-fold cross-validation works best on small and mid-sized datasets to tune hyperparameters, from which will result in the lowest test error. The data is separated into k subsets, alternatively training on all except kth folds and testing on kth fold. The final error is then calculated with the average of all k trials.

C. Clustering

Topical Clustering is the process of emulating human cognitive ability to discern objects based on the degree of similarity. Latent Dirichlet Allocation (LDA), a generative probabilistic topic model, gives us a way to infer the latent structure behind a collection of documents - in our case is tweets by users. The generative approach defined in [7] as follows:

1) For every topic k = {1, ..., K}, draw a distribution over the vocabulary V, \(\beta_k \sim \text{Dir}(\eta)\)

2) For every document ‘d’,

(a) draw a distribution over topics, \(\theta_d \sim \text{Dir}(\alpha)\) (i.e. per-document topic proportion)

(b) for each word w within document d.

i) draw a topic assignment, \(Z_{d,n} \sim \text{Mult}(\theta_d)\), where \(Z_{d,n} \in \{1, ..., K\}\) (i.e. per-word topic assignment)

ii) draw a word \(W_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})\), where \(W_{d,n} \in \{1, ..., V\}\)

Each topic \(\beta_k\) is a multinomial distribution over the vocabulary V and comes from a Dirichlet distribution \(\beta_k \sim \text{Dir}(\eta)\). Additionally, every document is represented as a distribution over K topics and comes from a Dirichlet distribution \(\theta_d \sim \text{Dir}(\alpha)\). The Dirichlet parameter \(\alpha\) controls the initialization and denotes the smoothing of topics within documents, and \(\eta\) denotes the smoothing of words within topics. The joint distribution of all the hidden variables \(\beta_k\) (topics), \(\theta_d\) (per-document topic proportions), \(z_{d,n}\) (word topic assignments), and observed variables \(w_{d,n}\) (words in documents) is expressed by (2):

$$p(\beta_K, \theta_D, z_{D,W}|\alpha, \eta) = \prod_{k=1}^K \prod_{d=1}^D p(\beta_k|\eta) \prod_{n=1}^N p(\theta_d|\alpha) \prod_{n=1}^N p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta_{z_{d,n}})$$

IV. APPROACH

In our approach, we consider sentiment trend and clustering as unsupervised learning. We will also use machine learning model regression to predict the scores and evaluate the performance metric with RMSE.

A. Dataset

a) Tweets: We use the repository from data collections hosted at TACC. It consists of tweets from Twitter search terms: “COVID”, “Coronavirus”, “Chinese virus”, “school closure”, “school closed”, “Food scarcity”, “Water contamination”, “reopen business”.

The dataset includes only tweet IDs according to Twitter’s content redistribution policy that restricts the sharing of tweets other than tweets IDs and/or user IDs. Thus we have to extract the context of the tweets using the Hydrator tool. With the daily tweets IDs that are provided from mid-March to the beginning of June, we are able to extract about 4 millions of tweets.

b) Daily Infection Data: We use the repository from John Hopkins data aggregated from CDC site.

B. Preprocessing

Tweet mentions with @ while retweeting, URLs, stop words, special characters were removed. Short words less than 3 or equal 3 letters are removed. Lemmatization is performed to avoid data sparsity. We also filtered user location with “USA”, “United States” or “US” and randomly selected about 6000+ records per day from March 25 to June 2, which total over 400,000 tweets. Additionally, for clustering, we filtered the parts of speech using the Spacy library and kept
nouns, pronouns, verbs, adverbs, and adjectives as part of normalization.

C. Feature extraction

For feature representation in the regression model, we used TF-IDF (term frequency - inverse document frequency) features extraction. TF-IDF is based on the frequency of word occurrences not just in a single document (tweet) but in the whole corpus. It works by assigning lower scores to the common words while giving more importance to the rare words in the corpus but with a good number of occurrences. If a word is common and occurs in multiple documents, it will approach 0. Otherwise, it will be close to 1. In mathematical terms, TF-IDF can be calculated with word t in document d from document set D as equation (3):

\[ TFIDF(t, d, D) = TF(t, d) \times IDF(t, D) \]  

(3)

where:

\[ TF(t, d) = \log (1 + freq(t, d)) \]

\[ IDF(t, D) = \log (N / count(d \in D : t \in d)) \]

With this approach, we are able to extract unigram and bigram features for our regression model.

D. Methodology

First we extract subjectivity score using Textblob to look at the trend of tweet subjectivity over time. The Textblob subjectivity ranges from 0 being objective to 1 being subjective. The polarity scores will be generated from VADER, ranging from -1 to 1. With subjectivity and polarity scores, we will analyze the overall average scores trend in the US. Then, using the same scores we examined the state level sentiment trends with some external factors such as lockdown period, reopening timeline, etc.

We apply a regression model for the tweets sentiments prediction. After the initial model building and evaluation with RMSE, we decided to tune our model with ridge regularization. It is used to optimize the model parameters by adding sum of squares of coefficients with a hyperparameter called alpha (lambda) which determines the strength of the regularization. With 8 different alpha values ranging from 0.001 to 1000, we apply 10 fold cross-validation to tune the hyperparameter and examine how it affects the training and test error. Lastly, we pick the model that would evaluate the test set with the model hyperparameters to predict the sentiment score.

Finally, for clustering, we used both the standard LDA along with Machine Learning for Language Toolkit (Mallet) implementation (via Gensim package). Our extraction process was aimed at gaining insight from the tweets by deriving semantic coherence. For our base model, we created a dictionary and corpus required as input for Topic Modeling and kept the model hyperparameter ‘k’ (number of topics) as 10 for each dataset as a coarse-grained clustering approach. As alpha being the hyperparameter for Dirichlet prior influences how the topic distribution is done, we kept it > 0.5 for a uniform distribution. The Gensim wrapper for LDA passes the alpha value as ‘auto’, symmetric and asymmetric. We used the coherence measure as ‘C_v’, which is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity. We obtained the optimal number topics using the elbow method of the coherence score.

E. Evaluation Metrics

For regression analysis, we first split the dataset into a 60% training set, 20% validation set and 20% for the test set in our simple validation . We evaluate the model using the root mean square error (RMSE), or the standard deviation of the residuals, i.e. prediction errors. We use the formula for RMSE in equation (4),

\[ \sqrt{\frac{1}{n} \sum (actual\ score - prediction\ score)^2} \]  

(4)

The equation will compute the error metrics for both training and test dataset with each dataset having n records. RMSE can interpret the error into similar units. In our case, it’s the sentiment score unit.

V. RESULTS

Results will be grouped into (i) Sentiment trends results from US overall and state level, (ii) Regression performance metrics with unigrams and combination of unigrams and bigrams as features and (iii) Clustering.

A. Sentiment Trends

![Fig. 2. Daily case increase with line representing moving average.](image)

According to the infection data from John Hopkins dataset, there are more than 2.7 million people infected as of July 2, 2020 and at least 121,000 have died. As Dr. Anthony S. Fauci warns in early June that the Coronavirus pandemic is far from over in early June, that “we are still at the beginning of it”. The data also support his claim that it’s starting to rise from mid-June and hitting a new daily tally starting July with over 52K on July 2. (see Fig. 2).

With the surging numbers across the US, there is no doubt it’s affecting people’s physical health and causing over 121,000 lives in the United States alone. The devastating reality and overwhelming state of information world would
add threats to the emotional stability. As the research by Johnston and Davey has shown in [6], “negatively valenced news conveys that negativity to the consumer and makes them feel more anxious, stressed or even depressed”. Thus, public sentiments will play a huge role in understanding life during the pandemic and social distancing.

With Textblob and VADER, we are able to assess how subjective or objective the tweets are. According to Fig. 4, with the development of Covid-19 from early stage in March to peak months in the US, the expression in tweets becomes more negative (from about -0.02 to -0.08) on average. With a quarter of the dataset being neutral tweets, the scores are less subjective.

![Fig. 4. Average polarity scores (in red) vs. subjectivity scores (in blue)](image)

We want to further analyze if the overall trend stays consistent with some of the states. For the timeframe from end of March to beginning of June, we visualized how the sentiment trends compared to the daily cases increase, see Fig. 3. We picked six states, namely California, Florida, Michigan, New York, New Jersey, and Texas, grouped them into two groups: Peak states vs. Non-Peak states. In peak states: NY, NJ and MI, the cases were at peaks and decreasing in the time period. In non-peak states: CA, FL and TX, the cases were not at peaks but they are either on the upward trend or plateauing at the time period.

In addition, we examine three significant periods, namely: beginning of lockdown, reopening and BLM movement. As a result, trends stay consistent in all the states in those three periods. Table II represents the sum of average sentiment change within 5 days of each significant period. There is positive sentiment change around reopening in all six states while negative sentiment change during lockdown and BLM movement. When the Black Lives Matter movement started, average sentiment scores went down drastically and hit lowest in some states towards the end of May. Out of all six states, New York has a sentiment shift after the reopening metrics was announced on May 11 followed by three regions starting to reopen. Daily average sentiment scores increase from -0.03 on May 12 to 0.01 on May 22.

<table>
<thead>
<tr>
<th></th>
<th>Sentiment Δ within 5 days of lockdown</th>
<th>Sentiment Δ within 5 days of reopening</th>
<th>Sentiment Δ within 5 days of BLM movement</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York</td>
<td>-0.010</td>
<td>0.028</td>
<td>-0.024</td>
</tr>
<tr>
<td>New Jersey</td>
<td>-0.056</td>
<td>0.043</td>
<td>-0.071</td>
</tr>
<tr>
<td>Michigan</td>
<td>-0.115</td>
<td>0.035</td>
<td>-0.069</td>
</tr>
<tr>
<td>California</td>
<td>-0.062</td>
<td>0.011</td>
<td>-0.036</td>
</tr>
<tr>
<td>Florida</td>
<td>-0.016</td>
<td>0.053</td>
<td>-0.063</td>
</tr>
<tr>
<td>Texas</td>
<td>-0.002</td>
<td>0.027</td>
<td>-0.022</td>
</tr>
</tbody>
</table>

B. Results from Regression

With the sentiment scores generated from NLTK-VADER, about a quarter of the dataset is neutral. By excluding the
neutral values that are concentrated between -0.05 and 0.05 (see Fig. 5), we could predict the binary classification of positive and negative sentiment tweets. Instead, we apply a linear regression model to predict sentiment score and explore the usage of unigrams vs. unigrams and bigrams as features. The RMSE results are summarized in TABLE III and IV.

With 10-fold cross validation, the best hyperparameter gives the least validation error. With the alpha value with Ridge regularization, we build a fresh new model to mitigate the slight overfitting as well as the least test error.

**Unigrams**: simplest way to extract features from the tweet. The test error is 0.2351. After model tuning with regularization, the test error is 0.2229.

**Unigrams and Bi-grams**: Bigrams alone can be too sparse, which could lead to serious overfitting. Thus, both unigrams and bigrams are extracted from text features. Compared to unigrams features, the test error is improved with 0.2291 in unigrams and bigrams. With slight improvement in test error after regularization, the model also improved the overfitting issues.

From Table III and IV, each of the models outperformed the simple baseline. The best performance was from the unigram and bigram combination with the hyperparameter (alpha = 1.1) with an RMSE of 0.2261.

<table>
<thead>
<tr>
<th>Features</th>
<th>Train error</th>
<th>Test error</th>
<th>Test vs Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.2023</td>
<td>0.2678</td>
<td>1.3237</td>
</tr>
<tr>
<td>Unigram</td>
<td>0.2165</td>
<td>0.2351</td>
<td>1.0859</td>
</tr>
<tr>
<td>Uni+Bigram</td>
<td>0.2211</td>
<td><strong>0.2291</strong></td>
<td><strong>1.0852</strong></td>
</tr>
</tbody>
</table>

**TABLE IV**: Error metrics after tuning hyperparameter

<table>
<thead>
<tr>
<th>Features</th>
<th>Train error</th>
<th>Test error</th>
<th>Test vs Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>0.2192</td>
<td>0.2311</td>
<td>1.0543</td>
</tr>
<tr>
<td>Uni+Bigram</td>
<td>0.2133</td>
<td><strong>0.2261</strong></td>
<td><strong>1.0598</strong></td>
</tr>
</tbody>
</table>

C. Results from Clustering

With the same 400K US dataset, we focused on the most affected states like NY and CA in particular to pick up marrow of the tweets. We addressed the following key cases: (i) Topics from majorly impacted states – NY, NJ, CA, FL, GA (ii) Most Dominant words from each tweet and (iii) Topic drift with datasets of 50k multiples for our entire 400K total tweets.

i) Tweets by Users in most affected states: We filtered the tweets based on the user location which consisted of little over 20k tweets for each state. For the first run, we randomly picked 10 topics. Next we ran our model against a range of Topics from (1..50) and derived the coherence score for each Topic (Fig. 6-7). Next we picked the optimal topic number based on the results i.e. where the model appeared to have flattened (elbow method). We noticed the coherence score spikes up to 20 & 25 topics respectively but then degrades or growth is minimal.

Further, we ran our model against the statewise datasets and derived the perplexity and coherence score. Table V, highlights the significant improvement in coherence score with Mallet against standard LDA with hyperparameter ‘k’ (Optimal Topics). i.e. (the average/median of the pairwise word-similarity scores of the words in the topic) for both the models which proved that Mallet in fact provides better topic segregation.
TABLE V. State wise coherence score with optimal topic selected

<table>
<thead>
<tr>
<th>State</th>
<th>LDA</th>
<th>LDA Mallet</th>
<th>Perplexity</th>
<th>Optimal Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>NY</td>
<td>0.26</td>
<td>0.32</td>
<td>-9.74</td>
<td>21</td>
</tr>
<tr>
<td>NJ</td>
<td>0.23</td>
<td>0.33</td>
<td>-9.37</td>
<td>14</td>
</tr>
<tr>
<td>CA</td>
<td>0.22</td>
<td>0.39</td>
<td>-9.33</td>
<td>25</td>
</tr>
<tr>
<td>FL</td>
<td>0.22</td>
<td>0.23</td>
<td>-9.18</td>
<td>17</td>
</tr>
<tr>
<td>GA</td>
<td>0.21</td>
<td>0.26</td>
<td>-9.31</td>
<td>25</td>
</tr>
</tbody>
</table>

ii) Dominant topics for each tweet: Next we attempted to fetch the dominant topic in each tweet out of multiple topics as it is expected in LDA models, each document comprises of multiple topics. Table VI shows the dominant topic for each sentence and shows the weight of the topic and the keywords in a nicely formatted output. This way, we know which document belongs predominantly to which topic and the keywords associated with the topic number. For example, Topic 7 can be categorized as “Economy” with keywords like bill, money, relief, etc whereas Topic 15 focuses on death rate, etc. Similarly, words like mask, worker, etc that contribute to Topic 13 can be grouped into ‘Healthcare’ as coarse grain approach.

TABLE VI. Dominant words for each topic produced by LDA

<table>
<thead>
<tr>
<th>Topic</th>
<th>Keywords</th>
<th>Original Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>mask, worker, face, public, medical, wear, con...</td>
<td>NIH-approved 3-D printed COVID-19 protective mask for health care workers. #coronaviruskenya #COVID19 Let's get printing!</td>
</tr>
<tr>
<td>15</td>
<td>death, case, report, toll, number, high, rate,...</td>
<td>The mystery of the true coronavirus death rate \nhttps://t.co/nQVzVtA3mz</td>
</tr>
<tr>
<td>7</td>
<td>president, government, americans, vaccine, america, bill, money, relief, treatment, system</td>
<td>Watch live: Senate votes on coronavirus stimulus package</td>
</tr>
<tr>
<td>9</td>
<td>trump, president, americans, america, vote, bl...</td>
<td>While Trump was holding hate rallies and playing round after round of golf. 'nAll your fuc*g excuses are bullshit <a href="https://t.co/DH4ADf6Wf">https://t.co/DH4ADf6Wf</a>..</td>
</tr>
</tbody>
</table>

iii) Topic drift over time: We noticed evidence of change in content of the tweets by users signifying topic drift if data sets are too far displaced in time or if data sets are too large and spread out across too much time. Topic drift is summarized in Table VII. For instance, although masks, vaccine, number, deaths continued to dominate the discussion but towards the end of May and beginning of June, much of the focus shifted to BLM protests and reopening of educational institutions, religious places and businesses.

TABLE VII. Topics distribution across tranches from overall dataset.

<table>
<thead>
<tr>
<th>Dataset Size</th>
<th>From Mar’25 till</th>
<th>Most discussed tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>50k</td>
<td>April 2</td>
<td>State, China, Chinese, spread, mask, response, death, news</td>
</tr>
<tr>
<td>100k</td>
<td>April 10</td>
<td>Mask, hospital, patient, drug, death, case, workers</td>
</tr>
<tr>
<td>150k</td>
<td>April 19</td>
<td>Country, governor, money, relief, drug, doctor</td>
</tr>
<tr>
<td>200k</td>
<td>April 28</td>
<td>Vaccine, mask, testing, symptom, crisis, economy, report</td>
</tr>
<tr>
<td>250k</td>
<td>May 6</td>
<td>Vaccine, supply, community, democrats, infection</td>
</tr>
<tr>
<td>300k</td>
<td>May 14</td>
<td>Fauci, open, vaccine, outbreak, mask, government</td>
</tr>
<tr>
<td>350k</td>
<td>May 23</td>
<td>Church, school, nursing, risk, economy, drug</td>
</tr>
<tr>
<td>400k</td>
<td>June 2</td>
<td>Riot, protester, fire, black, George, racism</td>
</tr>
</tbody>
</table>

VI. CONCLUSION AND FUTURE WORK

We have presented how overall sentiment progresses over time from March 25 to June 2 and how significant events like lockdown, reopening and protest are affecting the sentiment in majorly impacted states. We also demonstrated how majorly impacted states shape the topical clustering along with the topic drifting towards the end of May. The regression model also demonstrates promising performance with the combination of unigrams and bigrams. As model building is largely affected by training sets, we can further incorporate emoticons and count of positive and negative lexicons to enrich the feature sets. Although the dataset is from Twitter, we believe our study can also apply to other social media platforms.

With limited exploration on Covid related data in sentiment analysis, there is much more research to be done not just from social media, but also from other channels including the traditional news platforms and blogs. We hope that the foundation implemented in this paper will provide much insight into communication strategies for health officials or government officials for the recovery pathway from the pandemic and other researchers to implement further complex and extensive solutions to address the pandemic relief plan.
REFERENCES


