IEOR 243 - Module 1 Report

Group 10

Haolan Mai, Jiaqing Li, Yucong Li, Ruoxin Xu, Makie Maekawa, Bennett Cohen

April 28, 2023

1 Background

1.1 Assumption

We can assume that our main point of contact at the organization has a basic understanding of analytics, including the ability to read plots and interpret simple statistical measures. However, we can also assume that they may not be familiar with the technical details of machine learning methodologies such as unsupervised clustering and supervised multi-label classification.

1.2 Target Audience

As a data analytic service provider for the CDC, our focus is on providing insights into public sentiment, misinformation, and key topics of discussion surrounding COVID-19. By leveraging machine learning methodologies such as unsupervised clustering and supervised multi-label classification, we can categorize COVID-19-related tweets into meaningful topics and identify trends and areas of public concern.

For example, our sentiment classification model can help public health officials create targeted messaging to improve vaccine uptake by understanding people's attitudes toward COVID-19 vaccines. Furthermore, our semantic clustering model can provide retailers with insights into consumers' buying habits and spending patterns during the pandemic, helping them adjust their business models and sales strategies accordingly.

By providing the CDC with these insights, we can assist them in tailoring their communication strategies and public health campaigns more effectively, ultimately leading to increased public awareness, better adherence to safety guidelines, and improved public health outcomes.

1.3 High-level Summary of the Model

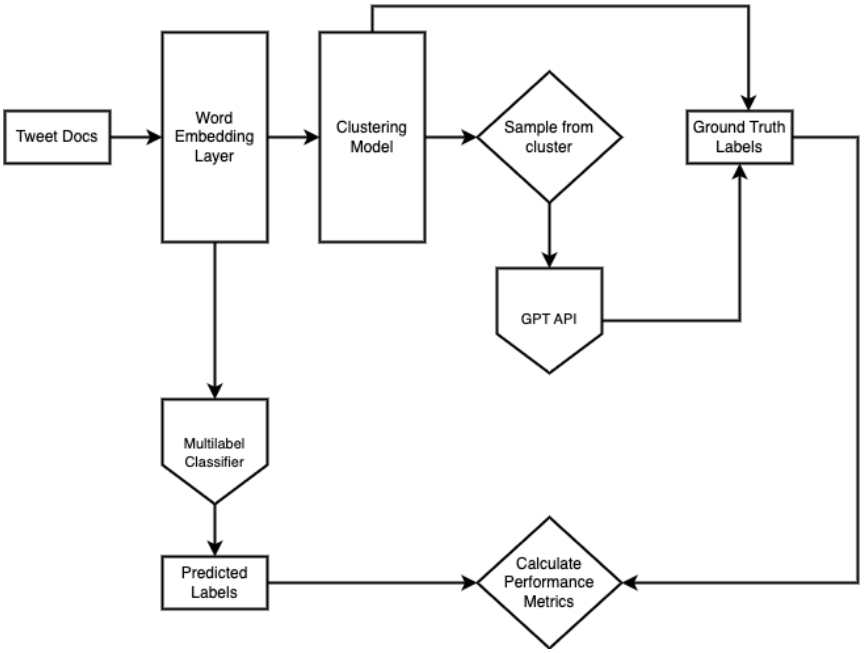

Figure 1: Model Approach Flow Diagram

Figure 1 above shows our modeling approach, which could be summarized into four main steps:

1) Converting all the tweet text into numbers representing different words/tokens. This process is called word embeddings.

2) Using a clustering model to group the tweets into 8 different categories based on the content of the tweet.

3) Using a version of ChatGPT to sample 50-100 tweets from each cluster and describe the tweets using 5-10 keywords. Now, we have a list of 40 labels

4) Training a Multi-output classification model to assign labels to each tweet in our dataset.

We used different training models: logistic regression, linear discriminant analysis, multi-layer perceptrons, random forests, gradient boosting, etc.

1.4 High-Level Metric Summary

The primary metric we use to score models is *Hamming Loss*, which measures the proportion of labels that are incorrectly assigned. For instance, a score of 0.05 means 95 percent of labels were correctly identified. This is an idea similar to accuracy.

We selected the models with the lowest hamming loss value. From *Figure 2* below, our best two models are Linear Discriminant Analysis (LDA) and Logistic Regression (LR).

	train_hamming_loss	test_hamming_loss	train_accuracy	test_accuracy
LR	0.0426	0.0412	0.6829	0.6913
LDA	0.0285	0.0294	0.8071	0.8033
MLP	0.2582	0.2625	0.0069	0.0053
RF	0.1949	0.1966	0.0046	0.0060
GBC	0.1064	0.1046	0.2143	0.2247

Figure 2: Hamming Loss

2 Demo

2.1 Interactive EDA

Before applying any machine learning method, it's often helpful to run specific exploratory data analysis in order to get a general sense of the whole dataset. We want to give the CDC the ability to observe macro-trends from the tweet data for the following concepts.

1) Most common words in tweets *(Bar Plot of Most Common Words in Tweets and Word Cloud of Most Common Words)*

2) Length of tweets *(Distribution of Length of Tweets)*

3) Number of tweets *(Time-series Plot of Tweet Counts)*

Every plot type can be filtered by *Date Range (07/24/2020-08/30/2020)* and *Location (All Countries, United States, Canada, South Africa, Switzerland, India, United Kingdom)*

Because COVID-19 has been a global pandemic, it's important to give our clients the choice to focus on their interests in future analysis.

For instance, if we see different trends in different locations at different times, this can advise the CDC to aim its messaging differently to different locations.

Examples

1) *Figure 3* graphs how many COVID19 tweets between 2020/07/30 and 2020/08/30 are coming from the United States.

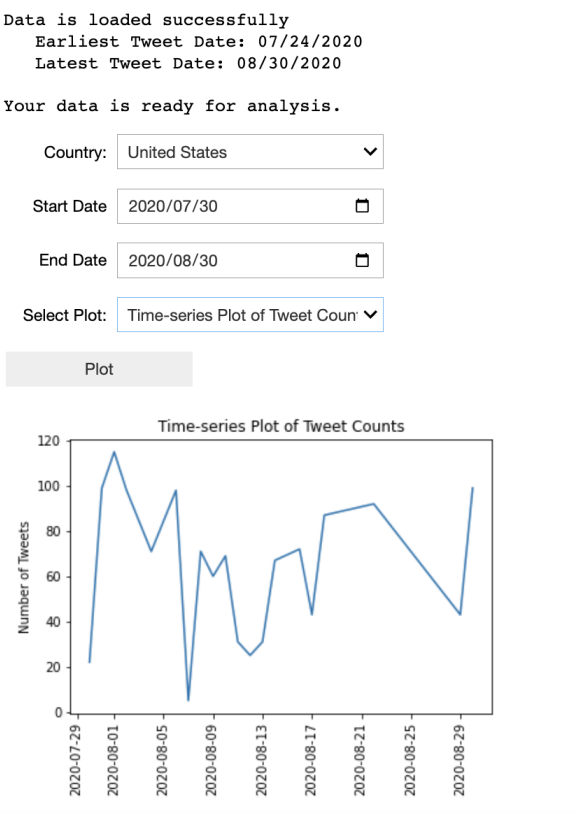

Figure 3: Time-series Plot Tweet Counts

2) *Figure 4* shows how to find the most common Twitter words used in the United Kingdom in July and August 2020 respectively.

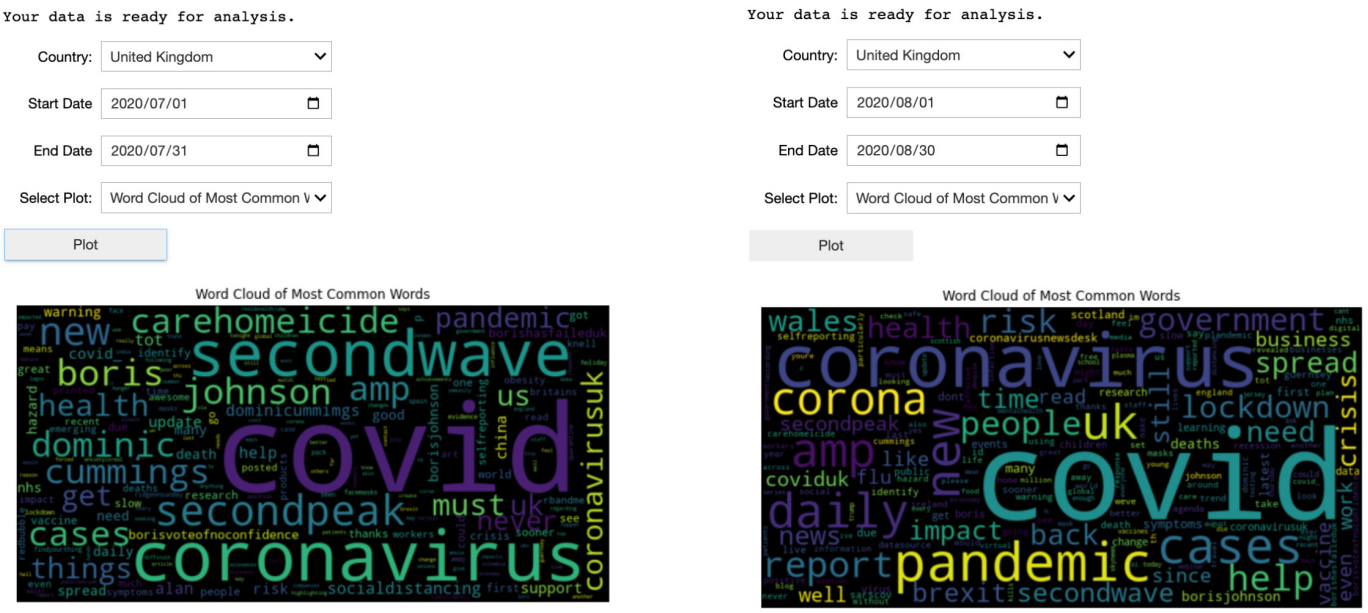

Figure 4: Most Common Words in UK in July 2020 and August 2020

Decisions Effect

These interactions, provided to the CDC, could have a significant impact on their communication effectiveness and public health campaigns. By understanding the public sentiment and identifying key topics of discussion regarding COVID-19, the CDC can tailor their messaging to address specific concerns and promote governmental decisions such as vaccination uptake. Additionally, by analyzing consumers' buying habits and spending patterns during the pandemic, retailers can adjust their business models and sales strategies accordingly, ultimately leading to improved economic outcomes.

The decisions to provide the CDC with exploratory data analysis capabilities can have a positive impact on the organization's ability to understand macro-trends in tweet data related to COVID-19. By allowing the CDC to observe the most common words in tweets, the length of tweets, and the number of tweets over time, they can gain insights into public sentiment and concerns related to COVID-19. Additionally, by providing filtering options for date range and location, the CDC can focus on their specific areas of interest and potentially adjust their messaging based on location-specific trends.

Communication with the Audience

In terms of communicating these effects to the audience, it would be important to provide clear and concise explanations of how the exploratory data analysis tools work and how they can be used by the CDC to gain insights into public sentiment related to COVID-19. Additionally, it would be important to provide visual representations of the data to help the audience understand the trends and insights that are being uncovered. It may also be helpful to provide examples of how the CDC has used these insights to adjust their messaging or response to the COVID-19 pandemic.

2.2 Real-Time Inference

Our model was trained on a dataset from 07/24/2020 to 08/30/2020. After our models are done, it would be helpful to let our clients use them to get real-time labels and categorize recent data. Our dataset was not crawled recently, so many more #COVID19 tweets have been created since then with new protocols, vaccine information, regulation, etc. Thus, we built the *Real-Time Inference* demo to let our clients explore new Twitter trends regarding #COVID19 topic by applying our models.

Decisions Effect

Our demo grants real organizations the ability to choose the model they want to apply when exploring trends. Even if there are similar performances between the two best models, they can give slightly different results. Our clients could compare these two models to balance the decisions that they make. In short, our *Real-Time Inference* demo provide the framework to analyze data properly.
Here are some sample tweets to try:

- "COVID-19 cases are on the rise again. Let's all do our part to stop the spread: get vaccinated, wear a mask, and social distance. Together, we can beat this virus."

- ”Indian municipalities have been a great failure in controlling #COVID19, barring a few.”
- ”Coronavirus infections top half a million in South Africa...#SouthAfrica #Gauteng #Pretoria #COVID.”

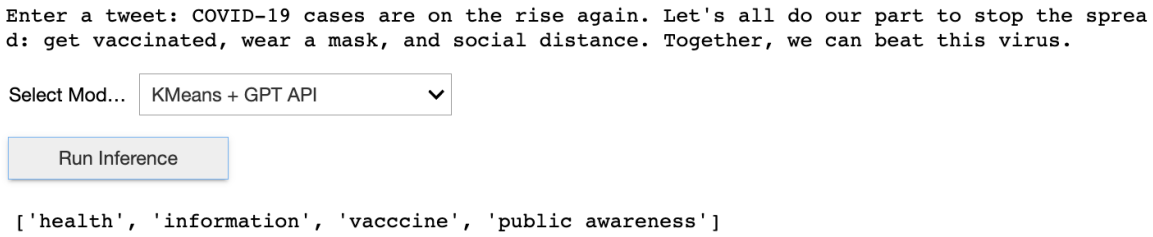

Figure 5: Sample Keyword Generated from KMeans + GPT API

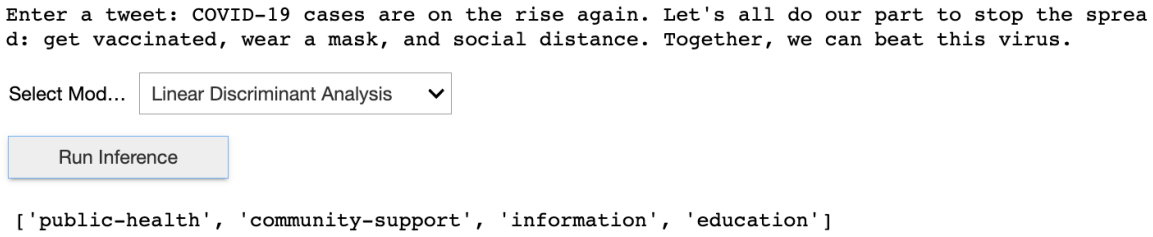

Figure 6: Sample Keyword Generated from LDA

Figure 5 and 6 above show the keyword results gotten from our models when inputting the first sample tweet given above using the *KMeans + GPT API* method versus using the *Linear Discriminant Analysis* method. We could see that these keywords are similar but not exactly the same.

Communication with the Audience

By using this *Real-Time Inference* demo, our client could enter COVID-19-related tweets that they are interested in, select different models, and get a group of keywords according to the input tweets. Organizations such as CDC could then compare outputs from different models and better understand public sentiments before they make regulatory decisions. For example, if they would like to make new governmental decisions on vaccination, they could enter vaccination-related tweets under hashtags such as #vaccine or #vaccination, to get public sentiment keywords.

Moreover, coming out of the pandemic, many people’s mental health might be at the state where they need to have professionals to talk to. Psychologists and marketers could also use our demo to get the specific topics under COVID-19 that people are most concerned about when doing their research.

2.3 User Tweet Micro-Trends

This part demonstrates a demo of an inference tool based on pre-trained models and machine learning models. Users can input a Twitter username and select a model type, then click the "Run Inference" button to get the predicted label distribution of the user's tweets. The organization can choose three different types of models: "KMeans + GPT API", "Logistic Regression", and "Linear Discriminant Analysis" to understand the micro-trends in specific users.

Health organizations are inextricably linked with politics (for some reason). The purpose of this demo is to help organizations visualize the micro-trends related to COVID-19 by analyzing subtle trends of highly influential Twitter users and gaining insights from them. The focus of the demo is to assist organizations in understanding the opinions and attitudes of specific users towards COVID-19 topics, and to conduct comparative analyses as needed to further deepen understanding.

For example, *Figure7* shows that we can analyze the subtle trends of COVID-19 topics of Shankar Prasad (Minister of Electronics and Information Technology) and gain insights from them.

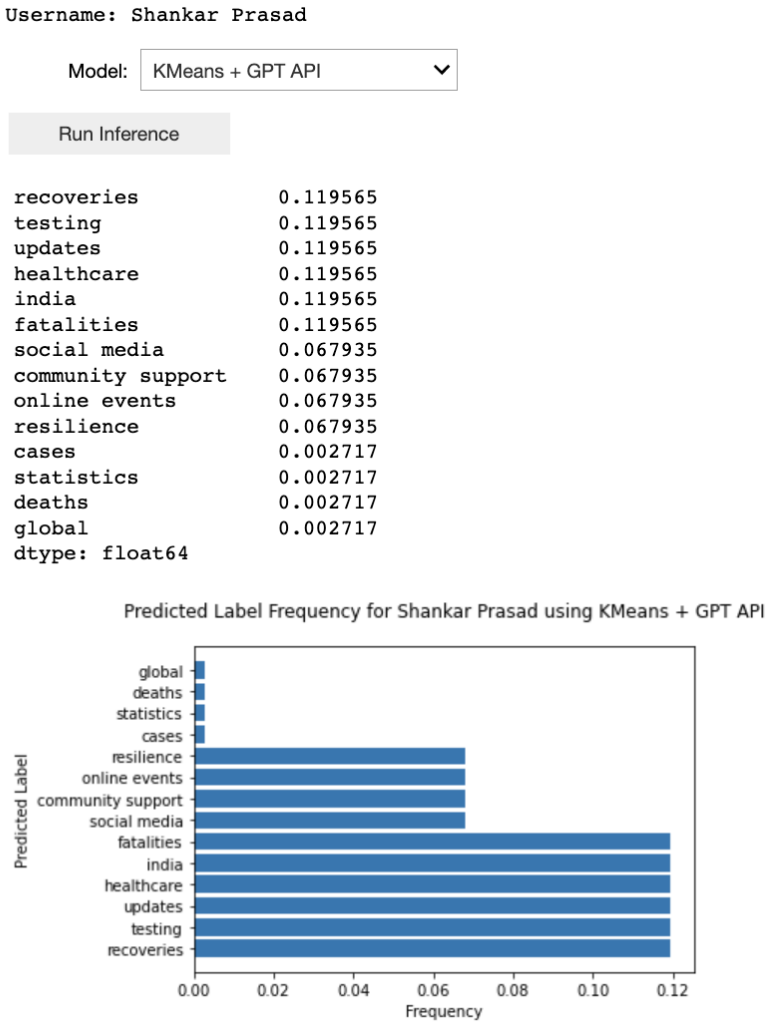

Figure 7: Example of Predicted Label Frequency for Shankar Prasad using KMeans + GPT API

Decisions Effect

The tweets published by high influential users, such as Joe Biden, Shankar Prasad, are official and authoritative. By analyzing the micro-trends related to Covid-19 of high influential users can help the organization make some quarantine plans in advance based on the content of the tweets and identify misinformation and address it in a timely manner. By correcting the misinformation, the organization can prevent it from spreading further. To some extent, the credibility of the organization can be enhanced. When the organization consider these influential voices and amplify them, its perceived authority and trustworthiness of its information will be solidified.

Communication with the Audience

By using the User Tweet Micro-Trends demo, the organization can type in the name of celebrity they are interested in and select one of the three models. Our demo will predict this user's high frequency words in terms of numbers and bar plot. CDC could have a better understanding of the discussions of this people related to the Covid-19 and make relevent prediction and political guidance by visualizing these frequency words. Take the *Figure 7* as an example. CDC could obtain that the words "recoveries", "testing", "updates", and etc. would be discussed by Shankar Prasad more frequently. Combining these words with the professional background of Shankar Prasad, Minister of Electronics and Information Technology, CDC would receive effective policy guidance and adjust its decision accordingly.

3 Conclusion

In general, our approach as a CDC data analytic service provider is to use machine learning methods, such as unsupervised clustering and supervised multi-tag categorization, to categorize COVID-19 related tweets into meaningful topics and identify public trends and domain concerns. Through analysis, I choose the best two models are linear discriminant analysis (LDA) and logistic regression (LR). Three demos have been established, which are Interactive EDA, Real-Time Inference, and User Tweet Micro-Trends.

In the first demo, by observing the most common words in tweets, the length of tweets, and the number of tweets over a period of time, we can understand the national mood and concerns related to COVID-19. In addition, by providing filtering options for date ranges and locations, CDC can focus on their specific areas of interest and use these insights to tailor their messaging to address specific issues and facilitate government decision-making.

We also built a Real-Time Inference demo that allowed customers to explore new Twitter trends on COVID-19 topics through applied models. CDC can choose which models they want to apply when exploring trends and use them to get real-time tags to categorize the most recent data. Ultimately, our clients can compare the results to balance the decisions they make.

Finally, we demonstrated an inference tool (User Tweet Micro-Trends) based on a pre-training model and a machine learning model. The goal of this presentation is to help organizations visualize micro-

trends related to COVID-19 by analyzing the subtle trends of highly influential Twitter users and gaining insights from them.

In summary, the CDC has adopted machine learning methods and the latest technological tools, such as interactive EDA, Real-Time inference, and User Tweet Micro-Trends, to conduct large-scale data analysis on COVID-19 and gain meaningful insights related to the disease. These tools not only help the CDC better understand the public's attitudes and concerns about COVID-19, but also provide real-time data analysis and predictions to assist in making better decisions on public health issues.