Covid-19 False Information Detection Using Word Vectorization and Passive-Aggressive Classifier

Jerry Guo zemingg2@illinois.edu

Introduction:

Since the Covid-19 outbreak in early 2020, people worldwide living in pandemics have searched for information about the virus online to respond to their queries such as how to protect themselves from being infected. Even though there are reliable health organizations like CDC and WHO, some people are still influenced by the uncertain information they viewed online, which helps the spread of fake news and misinformation. The data from WHO has shown that because of the misleading false information online, a large portion of the population in the UK are anxious about the Covid-19 vaccine, while a small percentage of them adamantly refuse to get vaccinated, which thwarted the progress of the community's immune(WHO, 2021). Therefore, it is vital to detect and stop the spread of false information on the internet as early as possible.

In order to detect false information on the internet, we decided to find an appropriate machine learning algorithm to retrieve and determine whether a social media post contains false information.

Data Collection & Exploration:

The dataset we analyzed in this study comes from Brandwatch. It contains millions of social media posts on Twitter that were about COVID-19 since January 1st, 2020. Because the size of the original dataset is too large to analyze efficiently, we randomly split the dataset into more than 200 parts(each has 30000 to 50000 records), and selected thirty of them at random as our sample data, which has 1341628 entries in total. Then, we filtered out all the non-English posts and dropped all the irrelevant fields from the table. Now, our sample dataset has only seven fields: the user's geographical location, the user's language, the date that the post was created, the number of followers, the retweet number, the content of the post, and which country did the user came from. Here is a line graph showing where the most English Twitter posts about COVID-19 came from. From the graph below, we can see that more information about COVID-19 has been spread by users living in the United States onto the internet than by users from other countries.

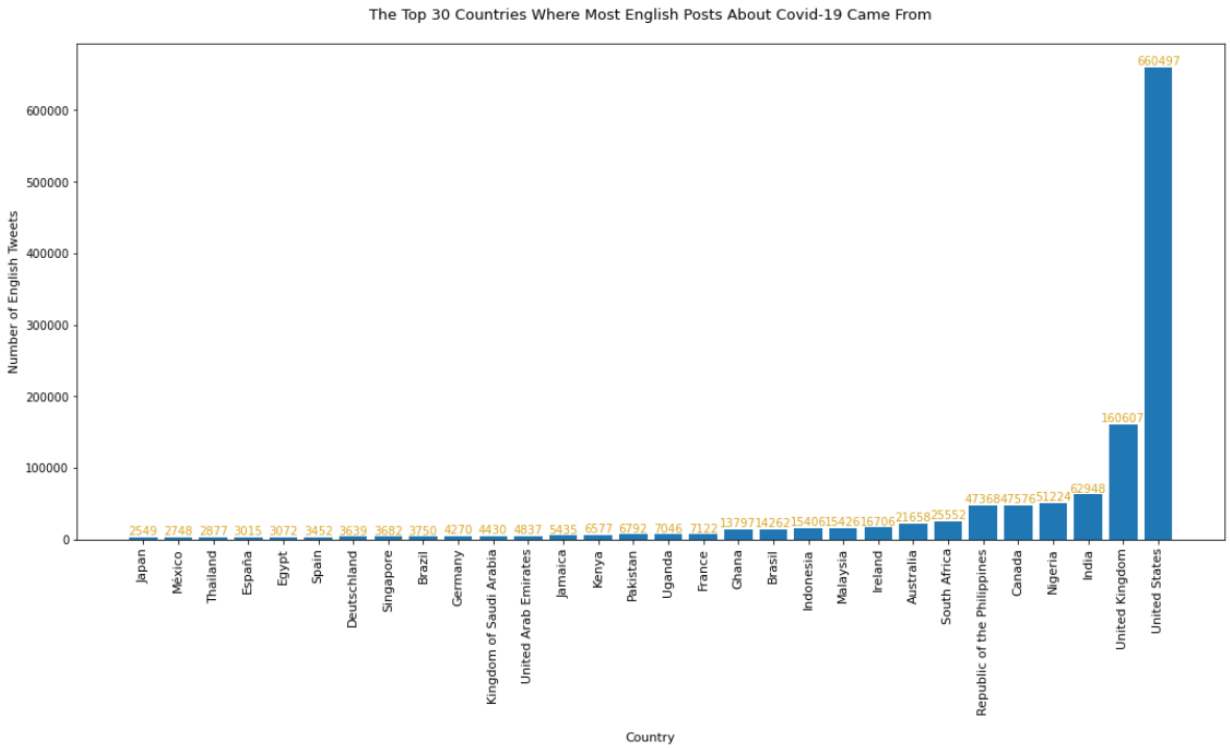

Figure 1: The Top 30 Countries Where Most English Posts About COVID-19 Came From

On the other hand, more information can be found as we look at how the number of Tweets about COVID-19 had been changing in December. We can see that after paying attention to the virus on the second day of the month, people had shifted their attention to preparing for Christmas. The number of posts about COVID-19 hit a trough around Christmas Day but began to bounce back after that, and reached a new peak at the end of the month.

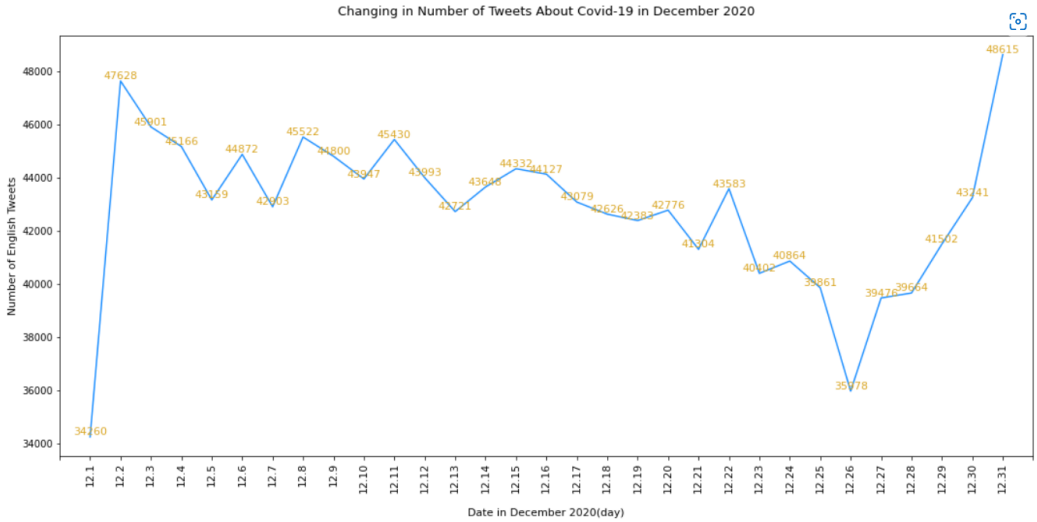

Figure 2: Changing in Number of Tweets About COVID-19 in December 2020

Methodology:

Data Preprocessing:

The next thing we did is data preprocessing, which is about tokenizing and normalizing the content of the posts to make our following predictions more accurate. We first split all the text into words and removed spaces, numbers, and all marks. Then, we wanted to let the computer ignore all the stopwords when it is making predictions. Stopwords are words like “is” and “a” that do not have significant meaning but appear frequently in sentences. So we imported the stopwords from the nltk module and took them out of the normalized tokens.

TF-IDF Vectorizer:

Because machine learning algorithms usually do not interpret raw textual data, we have to represent our text by numbers. In other words, we need to vectorize our tokens. The word vectorization technique we choose to use is called TF-IDF(term frequency-inverse document frequency) vectorizer. It is a statistical technique that measures how a term is relevant to a document in a collection of documents. TF-IDF can be calculated by multiplying TF(term frequency of a word in a document) by IDF(inverse document frequency of the word across a set of documents).

$$TF(t)=freq(t)n$$

Where t is the term, n is the total number of terms in the document. And

$$IDF(t)=log(1+n1+df(t))$$

Where n is the total number of documents, df(t) is the number of documents has the term in them. A term’s TF-IDF increase as it appears more times in each document, but decreases as it appears more times in the scope of all documents.

$$TF-IDF(t)=TF(t)*IDF(t)$$

Compared with TF, TF-IDF adjusts the influence of common stop words(like “a” and “the”), making the outcome more accurate.

Passive Aggressive Classifier:

Passive-Aggressive Algorithm is one of the most useful online-learning algorithms. Not like the batch algorithms that only compute the data at once, it computes and updates its prediction dynamically as new data is collected. This makes it a better fit for our task, which is computing a

very large scale of data from social media platforms. The name Passive-Aggressive Algorithm can be explained by how it works. It is passive when the lately collected data does not affect the prediction, but it will turn aggressive and makes changes to the model if the lately collected data has a significant impact on the prediction. It will adjust its weight vector for each misclassified result and try to correct it.

Model Training and Validating Predicted Result:

After building out the model, we trained and tested it with additional labeled datasets of fake and real news from Kaggle. Each of these datasets contains two fields: the social media post, and whether it is true or false. They have 1788 entries in total. We divided these records into two groups: the group(75%) to be used for training and the group(25%) to be used for testing the model. And below is the result of the model testing:

	precision	recall	f1-score	support
	0.90	0.98	0.94	357
Fake	0.00	0.00	0.00	5
Real	0.85	0.59	0.69	85
accuracy			0.89	447
macro avg	0.58	0.52	0.54	447
weighted avg	0.88	0.89	0.88	447

Figure 3: The Model Testing Result

the Precision of a model represents the rate of getting the relevant elements as the result. And the Recall of a model indicates what proportion of all relevant elements was recognized as relevant. The F1-Score in the third column balances the Precision and the Recall and measures the accuracy more comprehensively. Here is how F1-Score is calculated:

$$F(\beta) - Score = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

Because we want our model to predict the correctness of users' posts, we value Precision more than Recall. Therefore we consider the model with an overall accuracy of 88% acceptable. In the end, we input our vectorized data into the model and merged the result column into the original table. Now, all the suspicious false information is labeled with "Fake". Below is a statistic of the countries where most false information about COVID-19 came from in December 2020.

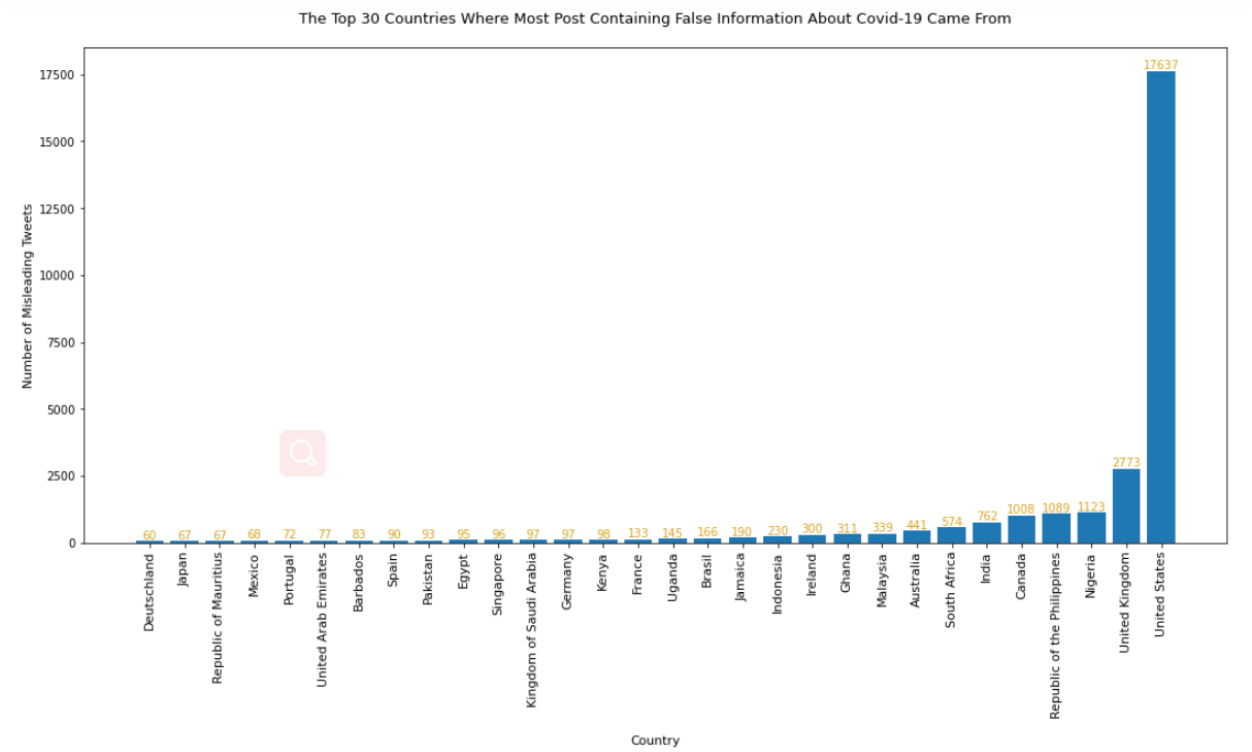

Figure 4: The Top 30 Countries Where Most Post Containing False Information About Covid-19 Came From

Keyword Extraction:

The last thing we did is vectorizing the “Fake” information again, and sorting them by their weight. We took the top 500 most weighted vectors as the keywords and found all the unique words represented by them. These keywords are the ones that we believe are “currently” trending as false information in the online community.

Figure 5: The Keywords Currently Trending In False Information About COVID-19

Some of these words do not provide enough information individually. However, if we think about these words in context with the recent news, we will probably find out what topic the misleading posts are now focusing on(which is also what the public now most cares about). For example, in December 2020, there were arguments about whether cats and dogs could get the virus and infect people. On the other hand, drinking bleach was once believed to be able to kill

the virus in the human body. We hope this result could provide useful hints to scientists and stop the spread of false information before they get really trending online.

Conclusion:

In this study, we tried to detect false information about COVID-19 using word vectorization and Passive-Aggressive Classifier. We first filtered our dataset and normalized the contents of the posts. Then, we vectorize the textual data with the TF-IDF vectorizer. After that, we built the Passive-Aggressive model and trained as well as tested it with another labeled dataset. In the end, we made predictions with this model, and extract the most common keywords in the false information with word vectorization again.

Because of the characteristics of the Passive-Aggressive Algorithm as an online-learning algorithm, the model would be still useful for finding out false information in the current online community. The accuracy of the prediction will mostly depend on how latest the data is.

Future Work:

Although individual words already tell a lot about the topics of misleading posts, we believe there may be other ways except word vectorization to extract key information. Linear regression has also been used for extracting keywords or key phrases, it is one of the new directions we will try in the future. Moreover, we will continue looking for ways to improve the accuracy of our model to over 90%, including adding other algorithms to the model.

References:

WHO. (2021, April 27). Fighting misinformation in the time of covid-19, one click at a time. World Health Organization. Retrieved May 29, 2022, from https://www.who.int/news-room/feature-stories/detail/fighting-misinformation-in-the-time-of-covid-19-one-click-at-a-time