Country Classification of Covid-19 Tweets

Miguel Novo Villar
Goergen Institute of Data Science
University of Rochester
Rochester, NY
mnovovil@ur.rochester.edu

Thomas Durkin
Goergen Institute of Data Science
University of Rochester
Rochester, NY
mdurkin7@ur.rochester.edu

***Abstract*—The following work outlines an analysis of our team's process, performance, and evaluation in the ”Country Classification of Covid-19 Tweets Kaggle Competition”. The researchers used natural language processing techniques to clean text, allocate topics and structures in the dataset and then fit machine learning / deep learning models to predict tweets belonging to six different English speaking countries.**

I. Introduction

The aim of this project is to classify the country of origin of text input coming from Twitter. The dataset used in this study was extracted from Twitter and consists of COVID related tweets. The entire data contains 300,000 tweets, which was split into a training set of 240,000 tweets and a test set of 60,000 tweets.

Using a combination of natural language processing techniques to clean the text of the data, and testing a series of machine learning / deep learning models to fit the cleaned data, we were able to predict the country of origin in the test set with an accuracy of 52 percent.

II. DATA

Each country is represented with 40,000 tweets in the training set and 10,000 tweets in the test set. The most common word in the entire training dataset is “covid” with nearly 20,000 repetitions.

It is crucial to highlight that while the Tweets come from English speaking countries, not all of them are written in English. With over 10,000 tweets not written in English and to improve the performance of our model the translate-api based on google translator had to be implemented to create a new ”text translated” column.

The seven attributes in the dataset are ”text”, ”reply to screen name”, ”is quote”, ”is retweet”, ”hashtags”, ”country” ”text translated”. These features help predict the country label, which is the country of origin of the Tweet [Australia, United States, Canada, Ireland, New Zealand and United Kingdom]. For our analysis, we only used the ”text translated” feature, as the other features would not aid in predicting the country.

III. DATA - DESCRIPTIVE ANALYSIS

Figure 1 and Figure 2 display descriptive statistics of the training and testing dataset where the average, median, minimum and maximun number of chracters or words where calculated, both for the training and testing dataset.

In figure 3, the stacked-bar chart indicates that the most used hashtag across all the countries was COVID19, even though five out of top ten hashtags were a form of ”COVID19”. One sub-bar that sticks out the most is StaySafe for Ireland. Tweets from Ireland used this hashtag twice as many times as the next highest country.

Looking at the heatmap in Figure 4, we can see that all the countries are very similar to each other. There were over 60,000 unique hashtags used between the six countries. This high volume of hashtags created a sparse matrix, leading us to not being able to determine a difference between pairs of countries.

IV. METHODS - TOPIC MODELING

Topic modeling is used in Natural Language Processing to uncover hidden structures in a collection of texts - in this case Latent Dirichlet Allocation (LDA). For the purpose of our analysis we created a word-cloud (Figure 5), a histogram (Figure 6) of the most common words, the top 10 most common topics with the LDA algorithm and a dashboard of the most common clusters (Figure 7).

As we previously mentioned, variations of the word 'covid' or 'coronavirus' are the most used in the dataset as it can be seen in the word-cloud and histogram.

The last step in our topic modeling was to allocate topics and structures within the text. The researchers trained an LDA model with the Latent Dirichlet Allocation (LDA) algorithm from the Sk-Learn library and visualized 10 different topics.

The resultant topics from the LDA algorithm were the following;

1- *covid report new zealand day lockdown coronavirus today patient data*

2- *covid pandemic vaccine support new help coronavirus health business impact*

3- *covid case death new day coronavirus number people test positive*

4- *coronavirus covid uk test flu government positive people quarantine trump*

5- *covid dr coronavirus treatment fauci hydroychloroquine study virus immunity say*

6- *covid home stayhome worker school staysafe care family life people*

7- *covid mask people coronavirus spread face wear hand social distancing*

8- *covid coronavirus minister china lockdown say week day medical uk*

9- *covid trump coronavirus american people died america staysafe vote know*

10- *covid pandemic need health people crisis government response economy world*

As it can be seen and according to our previous analysis, the word ”covid” is the main topic in all the found structures. In order to better visualize this topics, a dashboard was created with the pyLDAvis api and a non-negative matrix factorization in Figure 8 (unsupervised learning) for topic analysis. The desired parameters are 10 clusters, were we can allocate the weights of the words in each topic.

V. Methods - Classification

To prepare our model for training, we applied common data cleaning methods to the twitter text. Some of these methods include removing the newline characters, original URLs, stopwords, and extra spacing. These techniques allowed us to remove noise in the dataset, therefore, our model was only trained on the most meaningful part of the text. In order to get more out of our data, we converted each emoji to their text form instead of removing them. Also, each URL within the data was expanded to its original link instead of the shortened twitter link. This approach gave us more data to train our model with because we had access to the original site name and URL path which contained important words used in the article or website title.

As a first attempt, the classification methods we evaluated were Multinomial Naive Bayes, Linear Support Vector Classification, and Logistic Regression. Each of these models were fit using a pipeline that vectorized the data using TF-IDF. The vectorizer used sub-linear scaling to obtain the best results.

Another model we created was a One Dimensional Convolutional Neural Network. We chose to use a 1D CNN because of its efficiency and ability recognize patterns in sentences. Before training this model, the data was tokenized and padded. One-hot encoding was used to convert the labels to numbers. After significant testing, it was deemed using the top 100,000 words and padding each vector to a maximum length of 150 yielded the best results. Our 1D CNN layer consisted of 250 filters and a kernel size of 3 so that the sliding window tracked trigrams. Rectified linear activation function (ReLU) was used for the activation and Adam was used for optimization. Through exhaustive testing, it was decided that the optimal number of epochs was 2. Any amount of epochs higher than this amount lead to overfitting.

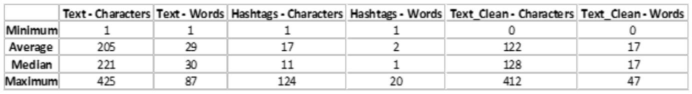
VI. Classification Results

Using our initial models as a baseline, it was determined that Multinomial Naive Bayes gave us the best results with an accuracy of 49.4 percent. Linear SVC was the second best at 48.7 percent with Logistic Regression being our worst model at 45.2 percent. After training our 1D CNN, we were able to achieve our best results with an accuracy of 50.01 percent. Looking at the confusion matrix for Multinomial Naive Bayes (Figure 9) and 1D CNN (Figure 10), there is some disparity between the models for predicting the correct country. Particularly, the 1D CNN is able to more accurately predict which tweets are from the US. By creating an ensemble model from these two models, we are able to see an improvement across all countries (Figure 11). This led us to improving our overall model accuracy to 51.6 percent.

VII. Conclusions and Limitations

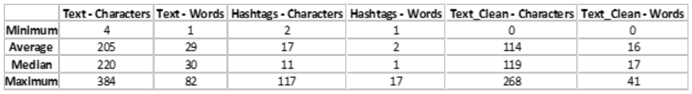
Reflecting on our results, a classification that yields an accuracy around 50 percent is not an accurate prediction and there is a large room for improvement by mastering the techniques of neural networks and transformers. The resulting confusion matrices show that our model labels the tweets from the United States excessively when the tweet was sent from another country. Seeing why these incorrect predictions are being made could drastically improve our model. On the other side, topic modelling algorithms gave reliable results about the underlying structure of the dataset and making it possible to identify topic trends in countries like Ireland. After trying out multiple modeling techniques, it was decided that an ensemble model was our best option. By combining a Multinomial Naive Bayes model and a One Dimensional Convolutional Neural Network we obtained an accuracy of nearly 52 percent.

Fig. 1. Training Dataset Descriptive Statistics



	Text - Characters	Text - Words	Hashtags - Characters	Hashtags - Words	Text_Clean - Characters	Text_Clean - Words
Minimum	1	1	1	1	0	0
Average	205	29	17	2	122	17
Median	221	30	11	1	128	17
Maximum	425	87	124	20	412	47

Fig. 2. Testing Dataset Descriptive Statistics



	Text - Characters	Text - Words	Hashtags - Characters	Hashtags - Words	Text_Clean - Characters	Text_Clean - Words
Minimum	4	1	2	1	0	0
Average	205	29	17	2	114	16
Median	220	30	11	1	119	17
Maximum	384	82	117	17	268	41

Fig. 3. Top Ten Hashtags by Country

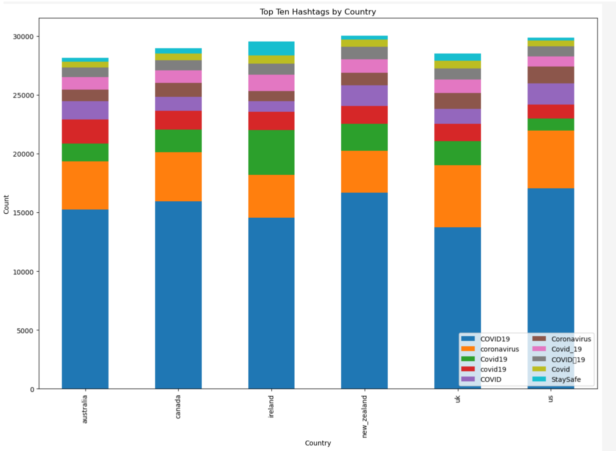


Fig. 4. Country Hashtag Count Heatmap

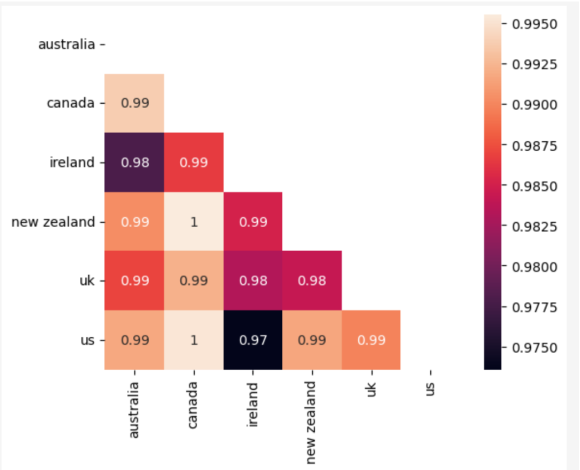


Fig. 5. Word-Cloud of the words in dataset

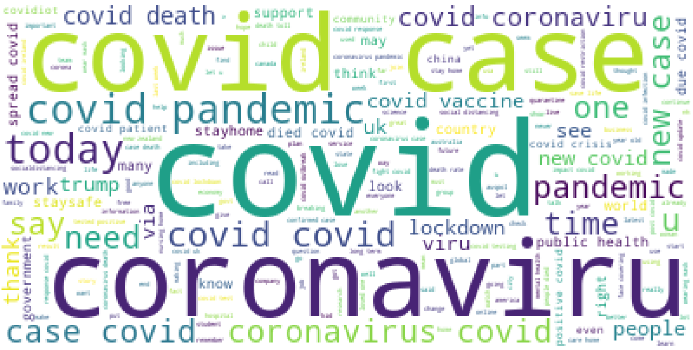


Fig. 6. Histogram of the most common words in dataset

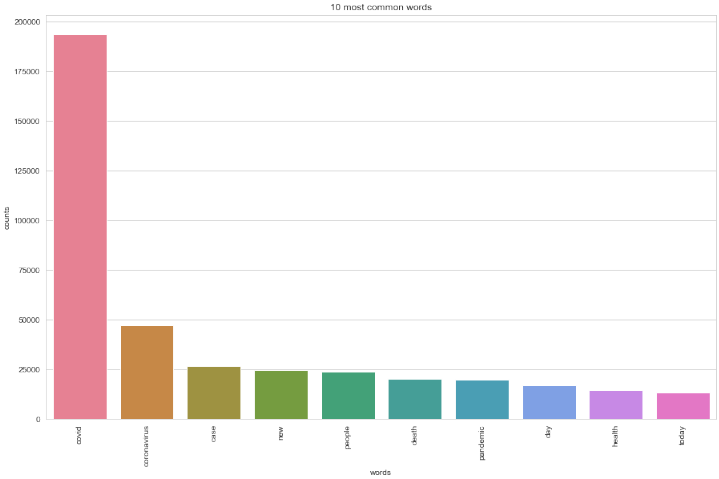
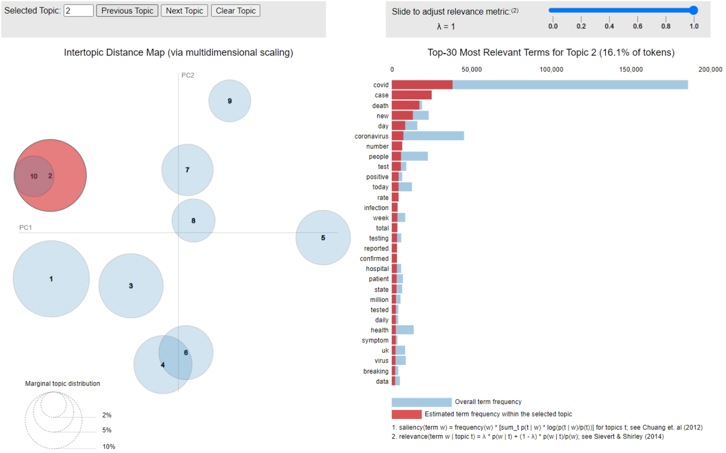


Fig. 7. Topic Visualization Dashboard



Selected Topic: 2

Slide to adjust relevance metric:(2)

λ = 1

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 2 (16.1% of tokens)

Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 8. Non-Negative Matrix Factorization - Frobernius Norm

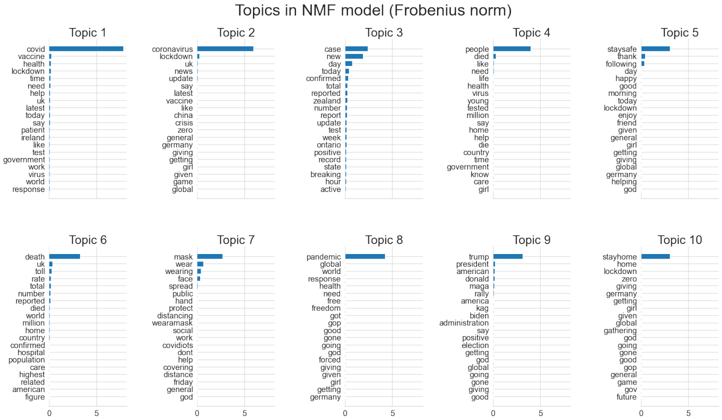


Fig. 9. Multinomial Naive Bayes Confusion Matrix

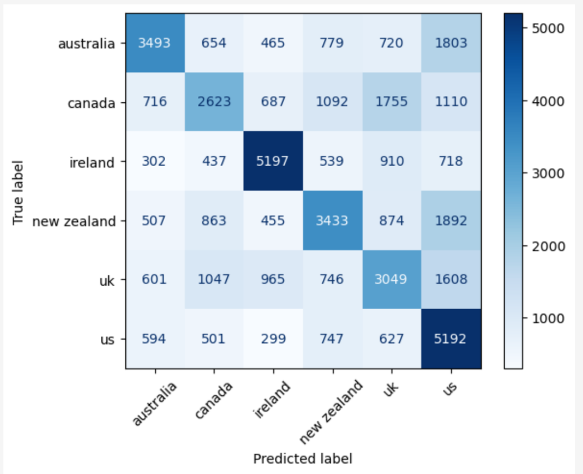


Fig. 10. 1D CNN Confusion Matrix

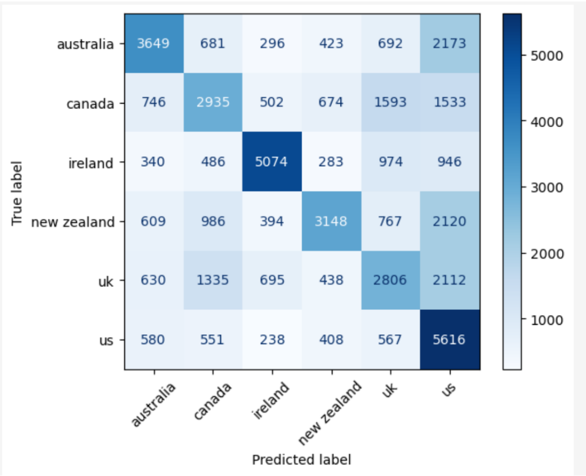
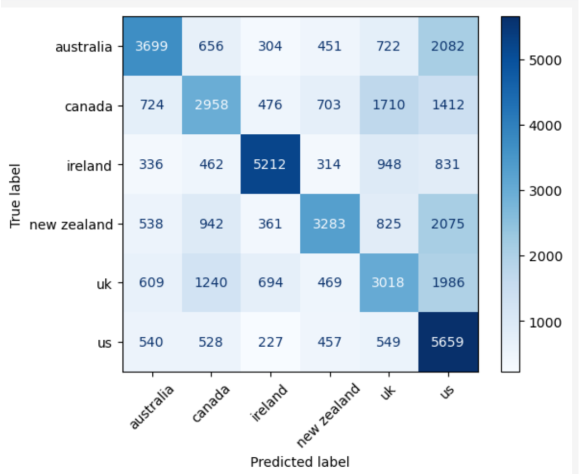


Fig. 11. Ensemble Model Confusion Matrix


australia 3699 656 304 451 722 2082
canada 724 2958 476 703 1710 1412
ireland 336 462 5212 314 948 831
new zealand 538 942 361 3283 825 2075
uk 609 1240 694 469 3018 1986
us 540 528 227 457 549 5659
True label
australia
canada
ireland
new zealand
uk
us
Predicted label