Kaggle Project Report

Jacob Freiheit

I. ABSTRACT

Classification models can have a variety of different uses. In any problem that involves the separation of data, classification models can be implemented to attempt to predict what classification different variables would have given some other information. In this lab, we're using training data and our own custom-designed classification models to determine the political favor of people tweeting from different European countries.

II. INTRODUCTION

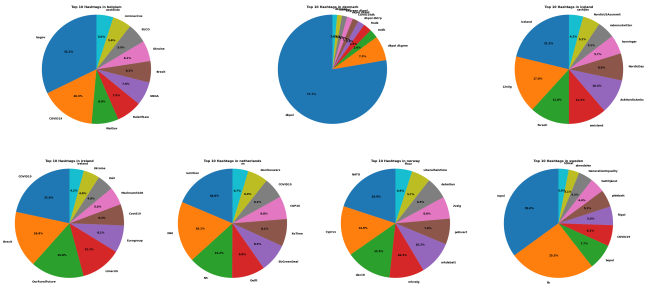
The training data is a subset of tweets taken, and contains information to contextualize each tweet. There is hashtags (the list of hashtags used in the tweet), full text (the text of the tweet), in reply to screen name (the Twitter screen name of the person the tweet is replying to), country user (country of the owner of the tweet), pol spec user (the political association of the user, found only in the training dataset), and Id (an index number associated with tweets, found only in the test dataset). To accurately describe the people who are sending tweets from these locations, visualization techniques were first used to show the kinds of people that are tweeting in the first place. I managed to achieve an accuracy score of 60.2 percent on the test data, though

III. DATA

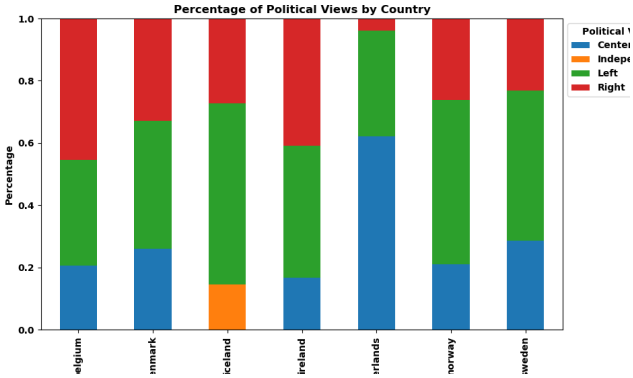
	Metric	Tweet Length (Characters)	Tweet Length (Words)	\
0	Minimum	4.000000	1.00000	
1	Average	167.304121	1.18023	
2	Median	156.000000	1.00000	
3	Maximum	2994.000000	16.00000	

	Hashtag Length (Characters)	Hashtag Length (Words)
0	1.000000	1.00000
1	6.459694	1.18023
2	3.000000	1.00000
3	145.000000	16.00000

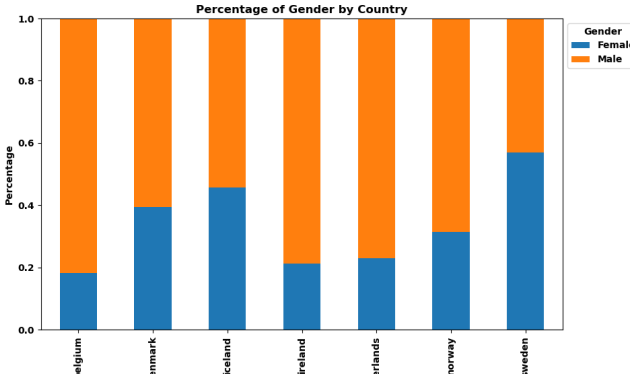
The first thing to show was the average length of the tweets, as well as the average lengths of the hashtags used in them.



The second thing to show was the 10 most frequent hashtags in each of the countries through pie chart. This revealed which topics were getting the most coverage on twitter, and usually discussed topics relevant to the respective country. Ireland's top 10 included Brexit, as an example.



This next part breaks down the percentage of tweets for each political viewpoint from each country. I was surprised to learn about the popular nature of centrism in many of these countries, as in the United States there seems to be an aversion to centrist viewpoints and as a result they are far less popular than more partisan statements.



Finally, this last chart shows gender distribution among tweeters in each of these countries. Across the board, it's clear that women aren't tweeting nearly as much as men. the effects of these can be seen to have some correlation in overall political views in each country. Iceland, which has the highest percetange of Left-wing tweets, also has the second highest percentage of women who are doing the Tweeting. Conversely, Belgium, which has by far the lowest percentage of female tweeters, has the highest percentage of Right-wing tweets out of every country analyzed. It seems women contribute greatly to a country's number of Left-wing tweets.

IV. METHODS

The approach I used involves using a pipeline with TF-IDF vectorization and a logistic regression model. The datasets are first loaded from their associated csv files and placed into dataframe objects, where their labels are encoded.

Label encoding is the process of converting categorical labels into numeric values. In the context of machine learning, many algorithms (including logistic regression) work with numerical inputs, making this process necessary for training most machine learning models.

In the preprocessing phase, the text clean column of the dataset is created. This will be the column on which the machine learning algorithms will be applied, as the text cleaning process both removes extra characters, and lemmatizes the tweet's content. The lemmatization process maps certain words to their general meaning, which reduces the dimensionality of the information being processed. This, in turn, makes the machine learning algorithms more accurate, as connections between certain words will be branched with the lemmatization process.

Unfortunately, The lemmatization only worked for the tweets in the English language. Many of the tweets were in French and German, and working to translate the tweets to English quickly proved futile because of the size of the dataset.

The dataset is then split into training and testing sets using train test split: one for training the model, and the other for evaluating its performance. Predictions are made on the test set, and the model's accuracy and confusion matrix are calculated.

Finally, the labels are decoded by using the inverse transform function, returning them to what they were before they were converted into numerical format. These predicted labels are then placed into a csv file along with the associated ID of the tweet,

V. RESULTS

Despite the numerous different methods and fine tuning I tried, I was only ever able to achieve an accuracy score of 61.94 percent on the training data. This figure was slightly reduced for the smaller test set (60.2 percent).

I expected that this amount of time would allow me to find a higher accuracy score than what I got. I primarily tried to use the things that we used earlier in the lab itself, but I was lead astray by some designs that took too long to run.

If I were to do this lab again, I would have cut my losses on some implementations earlier, and focused on new solutions once it became clear that optimizing the solutions I became attached to wasn't getting me any further.

The problem that I most would like to solve moving forward is how to work around the translation problem moving forward. I ended up developing my classifier without translating, but it'd have been far better if lemmatization were to be applied to all the tweets rather than just the English ones.