Article Classifier Project

Larry S. Liebovitch
William Powers
Lin Shi
Queens College CUNY

ABSTRACT

UPDATED—November 23, 2021. This project involves the analysis of data that was collected and analyzed by the The Power of Peace Speech Project at Columbia University. The data consists of media articles in the NOW database from 18 countries that team had classified as peaceful, non-peaceful, or neutral. We separately analyzed both the the words within quotation marks in the raw data that have the primary source words of actors and the "cleaned " data files with stop-words and the words of the reporters removed. We collected the words and word frequency counts and used them in a machine learning method to predict whether a country is peaceful, non-peaceful, or neutral.

CCS CONCEPTS

• **Computing methodologies** → Machine learning;

KEYWORDS

Preprocessing; Lexicon Validation; Classification Model; Random Forest Classifier

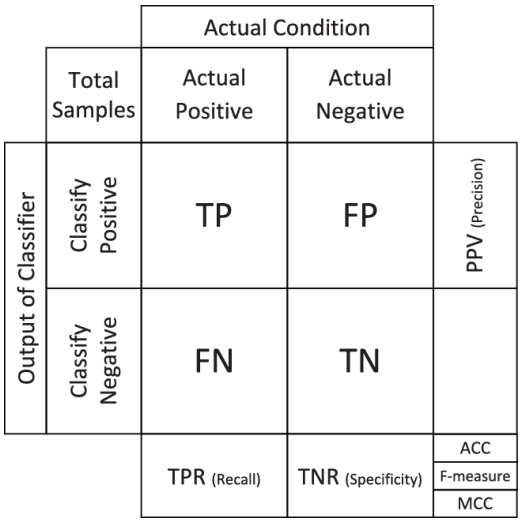

Figure 1: Sample Confusion Matrix

1 INTRODUCTION

The medium of news article publication in the last decade has largely transitioned from standard print toward the online format. With this transition comes greater accessibility and an ability for readers to collect data from a large magnitude of publications from across the world for analysis and comparison. The Power of Peace Speech team collected over 60 gigabytes of raw data from news articles published in 18 countries using the NOW database on corpusdata.org. The lexical information from these articles, such as word frequency, can be used with Machine learning techniques such as the RandomForestClassifier in order to create a prediction model for determining a country's peace status. These models can then be assessed based on their accuracy using a confusion matrix to determine whether the prediction is True Positive, False Positive, False Negative, or True Negative.

Using these tools to determine whether a country is peaceful or not peaceful may give new insight into the significance of language terminology within the articles. It will be necessary to highlight whether there is an associated importance to any particular word used belonging to a peaceful or not peaceful country, and whether the usage of such a word can affect the model's prediction.

1.1 Stop-words

In his article about the use of pronouns and function words, James W. Pennebaker [2] determined that the frequency of these commonly used words, also known as *stop-words*, are potentially meaningful in determining the personality, emotional state, and honesty of the speaker or author. For example, overuse of function words such as *I* and *we* can suggest that a person may be lying, whereas the use of exclusives such as *but* and negations such as *not* may indicate a person is telling the truth.

2 DATA COLLECTION AND PRE-PROCESSING

The data was collected from text files between January 2010 and September 2020. The text files were collected into folders by month, and the files were separated by the countries in which the articles were written. The program, *physicsproject_articleparser.ipynb*, parses through the text files, and searches for all quotes within each article which are led by the word 'said' in order to filter out statements that are connected together by missing quotation marks. If the quote begins and ends in a curly quotation mark (“ ”), they are replaced with straight quotation marks (" ") and stored within a

, ,
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

list inside a pandas dataframes. These lists are separated by their respective countries and year.

The program *physicsproject_article_dataframes_preprocessing.ipynb* filters the quotations, separates word frequencies into lists, and corrects an issue where apostrophes separate contractions into separate words. This information is written into a new dataframe and saved as a pkl file, which enables the dataframes to be accessible outside the program after it has finished running.

3 ARTICLE CLASSIFICATION: QUOTATIONS

In William's version, the program *physicsproject_article_classification.ipynb* reads the updated dataframe files and creates a count of each word that appears within quotations within a year for each country, while excluding punctuation marks. These word counts are then sorted alphanumerically and collected within dataframes with rows representing each country, columns representing each word, according to their respective year. The 11 dataframes representing each year are also combined into a single dataframe which aggregates all word counts and their sums in total. The individual and collective dataframes are assessed according to the ratio of each word's frequency, and compared to a total sum of 1.

In Lin's version, the program *project_article_classification.ipynb* reads 20 pkl files to generate 20 series. The program then concatenates all series and produces a single dataframe. This new dataframe contains 20 columns representing 20 countries and 122 rows representing all articles from corresponding countries. For each country, every word from the articles of this country is counted and results are stored as the form of tuples. By sorting these tuples, the 300 most frequently observed words are extracted in terms of each country, and these words are stored into specific lists.

Imagine two scenarios in which the most 300 frequently observed words from all countries are either identical or distinct. With regard to the situation in which words are identical, 300 distinct words will be collected in order to be analyzed at the next step. On the other hand, in a situation that words are totally different 5400 distinct words can be used at the next step. In this case, 2000 distinct words in total are tallied from William's version, and 660 distinct words in total are tallied from Lin's version.

This information is used in order to create predictions according to whether a country is peaceful, non-peaceful, or neutral using the RandomForesetClassifier sklearn prediction method. These predictions are made using

- the traditional method that divided all countries into train groups and test groups at a fixed ratio on the collective dataframe with all years included
- another method in which 19 rows of countries are used to predict the class of the excluded country with a different country excluded on each of the 20 iterations
- in addition, predictions were also made by modifying the collective dataframe to exclude neutral countries to use the most extreme peaceful and nonpeaceful countries to make the predictions

3.1 Difference in Word Count

When collecting the total word count of unique words across all countries, the word count collected by Lin was 660, whereas the word count collected by William was 2000. This occurred because William's version collected a word count of each year separately before concatenating the total word count, whereas Lin's version collected all the articles from all years together before establishing a single aggregate word count. The advantage of the method used in William's version is showing more comprehensive features from different months, but may change the trend of the overall features of a country at a certain degree since an excess number of distinct words are to be counted. Lin's version avoided this issue, and only counted the top words among all the articles combined. Therefore, for the purposes of article classification, Lin's version of 660 words was utilized.

Not Peaceful	Count	Peaceful	Count
state	197326	year	638740
nigeria	164717	new	486784
government	163651	time	416247
people	137769	people	379067
country	137002	work	310320

Figure 2: Top five most frequent words featured from countries that are 'Peaceful' and countries that are 'Not Peaceful'.

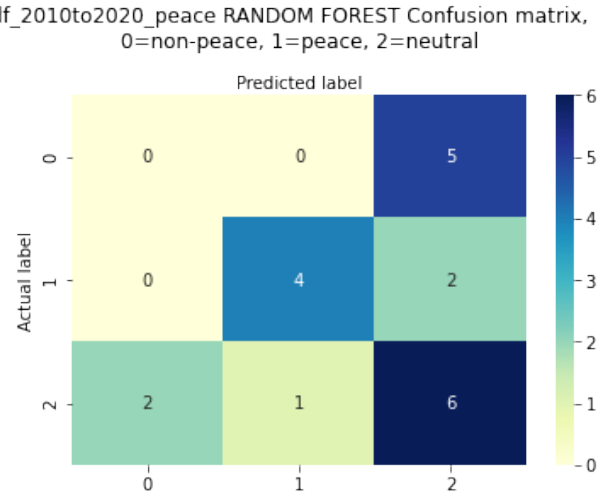

Figure 3: Sample Confusion Matrix making prediction using all countries, producing approximately .50 percent accuracy

3.2 Results

A brief summary of the RandomForestClassifier results of the words in the quotations from the raw data that have the primary source

words of local actors is given here. All results below keep 3 significant figures:

- data from the peaceful and non-peaceful, and neutral countries: using the traditional method with the collective dataframe produced:

Accuracy Mean	Population SD	Sample SD	SEM
.550	.232	.238	.0532

- data from the peaceful, non-peaceful, and neutral countries: using the 19 rows of countries method produced:

Accuracy Mean	Population SD	Sample SD	SEM
.530	.0781	.0801	.0179

- control null hypothesis: using randomly chosen values proportional to the number of peaceful, non-peaceful and neutral countries to predict the class of a country produced:

Accuracy Mean	Population SD	Sample SD	SEM
.332	.104	.107	.0239

  This sets the floor of the prediction accuracy and suggest that the prediction results of the two methods above are statistically significant $z = (x - x_{rnd})/stddev \approx (2.0 - 2.5)$
- data from only the disjoint peaceful and non-peaceful countries using the 19 rows of countries method produced:

Accuracy Mean	Population SD	Sample SD	SEM
.941	.0434	.0445	.0100

3.3 Observations

In the results of 20 consecutive experiments, two situations were observed. The first was that no non-peaceful country was predicted to become a peaceful country. The other was that no peaceful country was predicted to become a non-peaceful country. This suggests that the prediction method can be better utilized for distinguishing extreme situations. With that in mind, a hypothesis can be made that the possibility of a higher accuracy was constrained by the disturbance caused by the presence of neutral cities. To prove this hypotheses, all neutral countries were eliminated and only peaceful and non-peaceful countries were left before another 20 consecutive experiments were redone.

Perhaps our most important result is that the most significant classification success was achieved using this modified dataframe which removes the neutral values, that broke the constraints and increased accuracy mean by 77.5% compared with the accuracy mean produced by using the same method of 19 rows of countries.

4 ARTICLE CLASSIFICATION: CLEAN DATA, PEACEFUL VS. NON-PEACEFUL COUNTRIES

In this section, the program *project_article_classification.ipynb* is created in order to read and analyze the complete "cleaned" data files. This data came from the *article_text_Ngram_stopword_lemmatize* column of the respective data files from each peaceful or non-peaceful country, while omitting neutral countries. In this program the word frequencies were collected using the entire articles, as opposed to being limited by what was within quotation marks. These word counts are then sorted alphanumerically and collected within dataframes with rows representing each country and columns representing each word for all years combined. The dataframes are assessed according to the ratio of each word's frequency, and compared to a total sum of 1.

This information is used in order to create predictions according to whether a country is peaceful or non-peaceful using the RandomForesetClassifier sklearn prediction method. These predictions are made using a method in which 9 rows of countries are used to predict the class of the excluded country with a different country excluded on each of the 20 iterations.

Two lists were also produced featuring the top 1,000 most frequently featured words in an added collection of all the peaceful countries, and an added collection of all the non-peaceful countries. This can be found within the *list_peace_nonpeace.xlsx* excel file.

Similarly, using the most 300 frequently observed words from each peaceful and not-peaceful country, another excel file named *p_n_notp_words.xlsx* was produced featuring all peaceful and non-peaceful words with total number of occurrences of these words. The *feature_importances_* method will also be used to determine which of the features (that is, which words) are the most important in the RandomForestClassifier making its classification prediction. With this method, the words in file *p_n_notp_words.xlsx* that were recognized as the most important words were highlighted with yellow back-ground color. The reason for doing this is to verify whether the results produced by RandomForestClassifier method are more dependent on the more frequently occurring words.

4.1 *feature_importances_* using NumPy

When testing the RandomForestClassifier on the data, the rows containing the collected data for each country are first converted into NumPy arrays. The *feature_importances_* function of the RandomForestClassifier is specifically meant for testing data within a Pandas dataframe. Therefore, converting the rows of data into NumPy arrays for use within the RandomForestClassifier model made it difficult to test for the importance of specific words. As a result, a separate two-dimensional array was created containing the list of important words and their importance values. The important words are sorted to display by the descending order of their corresponding importance value.

4.2 Results

A brief summary of the RandomForestClassifier results of the words in the cleaned raw data that have the primary source words is given here:

- Data from the peaceful and non-peaceful countries using the 9 rows of countries method produced:

Accuracy Mean	Population SD	Sample SD	SEM
.940	.0800	.0821	.0184

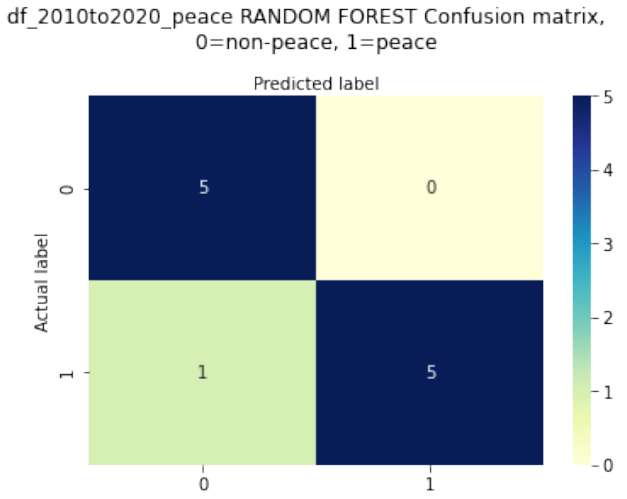

Figure 4: Sample Peaceful vs. Non-peaceful Confusion Matrix producing approximately .91 percent accuracy

Word	Importance
without	0.057026
not	0.042462
give	0.040000
country	0.039257
development	0.037701

Figure 5: Top five words listed in order of importance, collected using importance_features_ method.

- Of the 667 features within the training set, only 24 of the terms hold an importance value greater than 0. The most important term, *without*, has an importance value of 0.057.

4.3 Observations

The accuracy mean from clean data using 9 rows of countries method is very close to the accuracy mean from quotation data using 19 rows of countries. This implies that the accuracy mean produced by this method is not related to the size of the collected data.

In addition, it is interesting to observe that highlighted words in excel file *p_n_notp_words.xlsx* have a tendency to appear in the front of both lists. This verifies the prediction mentioned above that the results produced by RandomForestClassifier method are more dependent on the more frequently occurring words.

5 DISCUSSION AND ANALYSIS

The results of the project demonstrates that using the RandomForestClassifier model with all the countries together using the traditional method does not produce a prediction model that is much more statistically advantageous than that of the randomly generated null hypothesis (.550 vs. .332) However, removing the neutral countries does significantly increase the accuracy of the RandomForestClassifier prediciton model (.941). The results of the confusion matrices similarly display this improvement when removing the neutral countries from the prediction model. This also allows for tools to be used such as the *feature_importances_* method in order to isolate those words that play a role in affecting whether a country is predicted to be peaceful or not peaceful.

Examining the most important words between the Peaceful and Non-Peaceful countries separately does demonstrate some context and associative differences. Peaceful countries tend to use words with a generally positive connotation, such as *new*, *well*, *day*, and *great*. Non-peaceful countries tend to use words which can be generally associated with an executive and administrative association, such as *government*, *police*, *national*, and *president*. However, these words are not exclusive to either peace status, and do show some overlap between the two groups.

It was also critical to analyze the prediction models as a function of multiple attempts. Testing the RandomForestClassifier sometimes produced varying accuracy results, and required finding the average mean of multiple attempts combined. Creating arrays with all results of the accuracy, standard deviation, and standard error of mean allowed for a more detailed representation of whether the prediction model was dependable after multiple trial runs as opposed to being successful after a single attempt.

Another consideration is the removal of stop-words from the cleaned dataset. When using the "cleaned" dataset, several stop-words such as *the* and *at* were still present in the dataset, and had to be removed manually. As a result, a list was created to check whether each word was a non-meaningful stop-word, and removed from the word frequency tally if it matched. However, it may be necessary to clarify which stop-words are meaningful, such as *I* and *you*, which were left in the dataset during the collection of frequencies.

6 NEXT STEPS

We are particularly interested in determining whether the words that are the most frequent are, or are not, the most important words in making the classification prediction. A very preliminary result suggests that it is not the most frequent words, but the modestly frequent words, rank order 50 - 100, that are the most important in making the classification prediction.

Another tool that is considered is the use of a logistic regression model in order to test whether the neutral countries are more closely related to a peaceful or non-peaceful country. However, logistic regression is more traditionally applied toward single-feature

Figure 6: A Word Cloud created by Allegra Chen-Carrel of the most important words within Peaceful countries

Figure 7: A Word Cloud created by Allegra Chen-Carrel of the most important words within Non-Peaceful countries

analysis. It may be necessary to compare the efficacy of the RandomForestClassifier with other predictive analysis tools in order to see if there is a perhaps stronger method for testing whether a country may be peaceful or non-peaceful.

The presence and frequency of certain stop-words, such as those in the first person perspective or with a negative connotation, may also be useful in determining the truthfulness of a given country. Testing whether a country's truth and peace statuses are correlative values using the frequency of these stop-words can serve to reinforce James Pennebaker's theory on function words as well as reveal the significance of these stop-words within a country's written documents.

The unequal sample size of each country's data is another important consideration with regard to the results of the prediction models. Some countries had much fewer articles collected than others, which can potentially skew results unfavorably. Also, some countries may not have quotations as frequently as other countries, which may have significantly restricted the available data for analysis. An improvement to future analysis may involve using a larger dataset that is more evenly distributed among countries. However, this may require a different method for data pre-processsing, as the current method of using Jupyter Notebook on a physical machine for pre-processing 60 gigabytes of data necessitated a run-time over several days. Using a larger dataset may require an online compiler over cloud services, such as *Google Collab* or *Amazon Web Services*.

REFERENCES

[1] Jinwoo Jung, Hojin Lee, Hyuk Joon Kwon, Matt Mackenzie, Tae Yoon Lim (2020) *The Power of Peace Speech.* Columbia Universtiy. https://github.com/mbmackenzie/power-of-peace-speech

[2] James W. Pennebaker (2011) *Your Use of Pronouns Reveals Your Personality.* Harvard Business Review. https://hbr.org/2011/12/your-use-of-pronouns-reveals-your-personality