Larry S. Liebovitch, Yihan Wang

In today’s world, there are lots of diversities among various perspectives. Globalization has rapidly influenced the links between countries. Peace has been a noteworthy mark of some countries with obvious peace contexts. In this study, we collected the text data of speeches from 11 countries with distinct peaceful/non-peaceful characteristics and conducted Machine Learning models to do classifications.

We used 8 countries’ datasets for the training process, and to test the trained models, we selected the rest 3 countries’ datasets as our “future datasets” for predictions. The 8 countries for training were Canada, US, Singapore, Philippines, New Zealand, Nigeria, Ireland, and Bangladesh, and the 3 countries for predictions were Australia, Kenya, and the United Kingdom. From those first 8 countries, we grouped them in 4 pairs by their demographic region similarity.

|  |  |
| --- | --- |
| Peaceful Country | Non-Peaceful Country |
| Canada | US |
| Singapore | Philippines |
| New Zealand | Nigeria |
| Ireland | Bangladesh |

**Processing**

We removed named entities such as person’s names, places, and countries’ names through applying the function “en\_core\_web\_sm” in package Spacy to the lemmatized texts of each country. After removing stop words, some symbols, and single letters, we added one column to each dataset to specify their peace level (0 = non-peaceful, 1 = peaceful).

**Training**

To compare the differences between pairs of countries from different regions, we created TF-IDF features and trained three models for each pair. The three models were Logistic Regression, Random Forest, and XGB. Logistic Regression was to classify the peace level using logistic function, Random Forest and XGB were two extends of Decision Trees, where Random Forest conducted bagging methods for Decision Trees, and XGB used boosting Decision Trees. Furthermore, we combined the 8 countries together to do the same training process. For each of the 5 pairs, we split data to training and testing set (80/20 ratio).

For the three trained models for each pair, we loaded the top 200 important features (words) and their feature importance and used to trained models to predict Australia, Kenya, and the United Kingdom’s datasets.

**Results**

The Canada/US pair showed higher accuracy score for predicting non-peaceful countries’ data, and the New Zealand/Nigeria pair and Ireland/Bangladesh pair showed higher score for predicting peaceful countries’ data. However, the Singapore/Philippines pair tended to predict each text to be peaceful whether it was from peaceful countries or not. Moreover, the combination of eight countries showed higher score for predicting Australia’s data (accuracy = 0.85 Logistic Regression, accuracy = 0.93 Random Forest) and United Kingdom’s data (accuracy = 0.82 Logistic Regression, accuracy = 0.90 Random Forest). The score for Kenya was lower (accuracy = 0.60 Logistic Regression, accuracy = 0.46 Random Forest). Overall, the five choices for country pairs had higher score for predicting peaceful countries’ data than predicting non-peaceful countries’ data. From the word list of Singapore/Philippines, we noticed that all words from Philippines were missing the letter “o”, and that was one of the reasons why this pair has such low predicting accuracy. *Table 1* showed the results from the training process.

**Table 1. Results From 5 Choices of Country Pairs**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | | | | Random Forest | | | | XGB |
|  | Accuracy | Australia (acc) | Kenya (acc) | UK (acc) | Accuracy | Australia (acc) | Kenya (acc) | UK (acc) | Accuracy |
| CA vs. US | 0.88 | 0.48 | 0.66 | 0.40 | 0.85 | 0.31 | 0.90 | 0.26 | 0.83 |
| SG vs. PH | 1.00 | 1 | 0 | 1 | 1.00 | 0 | 1 | 1 | 1.00 |
| NZ vs. NG | 0.98 | 0.98 | 0.53 | 0.94 | 0.96 | 0.94 | 0.43 | 0.92 | 0.93 |
| IE vs. BD | 0.98 | 0.89 | 0.61 | 0.89 | 0.94 | 0.78 | 0.69 | 0.78 | 0.91 |
| Eight | 0.92 | 0.85 | 0.60 | 0.82 | 0.87 | 0.93 | 0.46 | 0.90 | 0.83 |

**Discussion**

Since Random Forest had higher accuracy overall, we chose the top 200 important words and their feature importance of each trained pairs and joined their peaceful level (1 or 0). The Word Cloud showed the appearances of these words.

From the Word Cloud of non-peaceful countries, we noticed that there were more words that were related to politics, such as “president”, “government”, and “chairman”, where the words from peaceful countries were related to daily life, such as “contact”, “comment”, and “say”.

Text

Description automatically generated

Figure 1*.* Word Cloud for Non-Peaceful Countries (Combination of 4 Pairs)

Text

Description automatically generated

Figure 2. Word Cloud for Peaceful Countries (Combination of 4 Pairs)

Text

Description automatically generatedFigure 3. Word Cloud for Non-Peaceful Countries (Eight Countries)

Text

Description automatically generated

Figure 4. Word Cloud for Peaceful Countries (Eight Countries)