Progress Report II - Peace Speech Project

Jinwoo Jung (jj2762), Hojin Lee (hl3328),
Hyuk Joon Kwon (hk3084), Matt Mackenzie (mbm2228), Tae Yoon Lim (tl2968)

November 23, 2020

1 Problem Definition and Progress Overview

As the world gets more polarized, hate speech has been an active area of research. However, there was not much attention to the other side of the story: peace speech. This project was proposed with the hope that peace speech could play a role in measuring peace and promoting more peaceful societies. Our group's tasks within this project were first to validate the hypothesis by developing techniques to understand, measure, and track the power of peace speech, which will guide us toward building and maintaining more robust and peaceful communities. We looked into news articles from many different countries, analyzed them using natural language processing (NLP) techniques, and studied the relationship between the language used in the articles and the peacefulness of the country.

In the first progress report, we addressed a description, data engineering, initial exploratory data analysis, text pre-processing with provided datasets. This second progress report will discuss frequency analysis, testing hypotheses on whether pre-defined lexicons and classification of countries into peaceful/resilient/conflicting societies are valid, and using other data science techniques to come up with new lexicons that are purely data-driven. Please refer to the graph below, which gives a brief overview of our group's progress on the project.

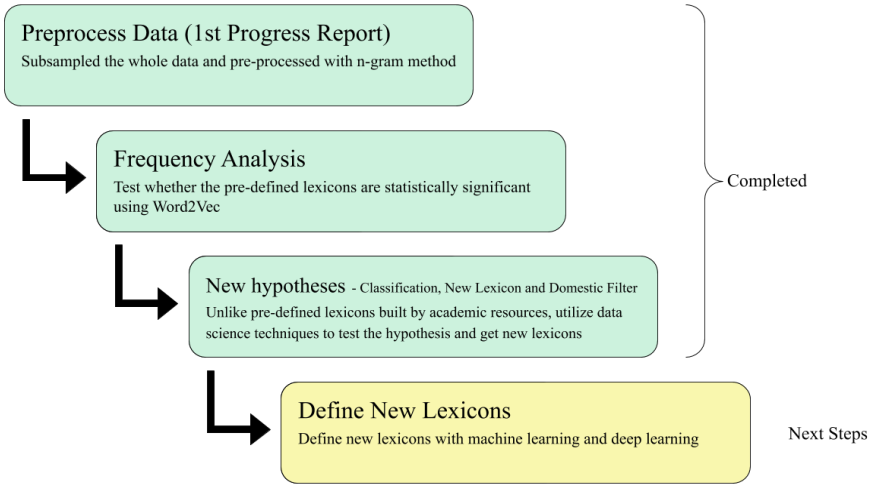

Fig 1-1. Summary of Progress

2 Word Frequency Analysis

2.1 Peace Metric: Bar plot and Box plot.

After cleaning the data we decided to look into the preprocessed data to identify if we could find specific patterns using the lexicons, whether lexicons are used proportionally given among peaceful, resilient, and conflicting countries and whether classification of countries into the 3 communities are valid. The peace metric was assigned to each country and it was simply the percentage of words in the negative lexicon subtracted from the percentage of words in the positive lexicon. This value was normalized across all countries and a bar plot was drawn. As we can see with Figure 2-1 below, it is possible to see that although a few countries such as CA, IN, PK give us expected results, others such as GB, TZ give us completely opposite results. JM has a greater peace score than even CA.

We then grouped the countries and drew a box plot using the Peaceful Score used in Fig 2-1. There seemed to be better results where peaceful had the highest median score and the non-peaceful group had the lowest median score. However, after testing if the box plots were significantly different using ANOVA the scores from each country give a p-value of 0.5, and thus, we conclude that there are no significant differences among the groups. Furthermore, we tested if there is a significant difference between any one of the two peace groups. But even when comparing peaceful countries with non-peaceful countries the p-value was 0.35, which was too high to conclude that the two groups were different.

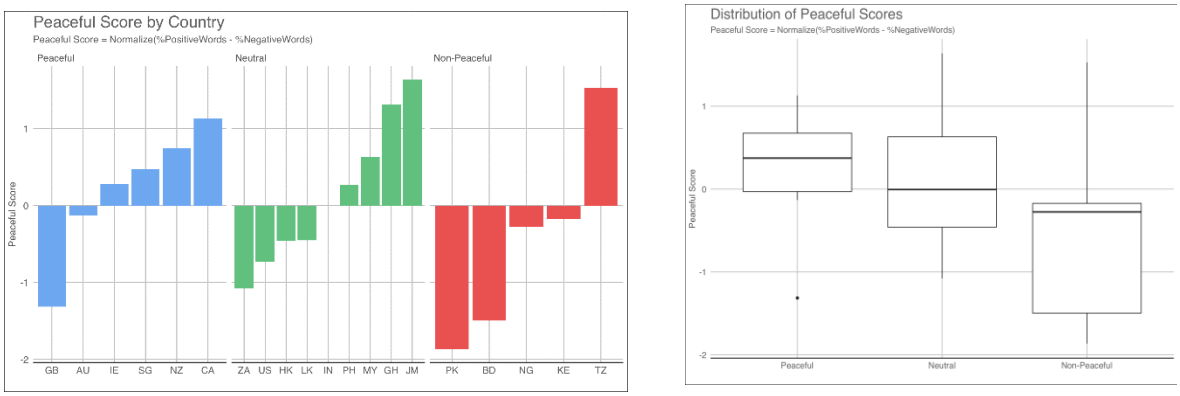

Fig 2-1. Peaceful Score by Country and Breakdown of news articles each type of country

2.2 Word Frequency Count

The results above were not as promising as expected, and we decided to look at the actual lexicons that were given to us at the start of this project. Initially, we had the assumption that words in positive lexicon are used more in peaceful countries and words in the negative lexicon are used more in non-peaceful countries. There, however, does not seem to be any clear patterns in the three

graphs drawn below. The articles from peaceful countries do not seem to use words from the peace lexicon any more than non-peaceful countries. Also, articles from non-peaceful countries do not seem to use words from the conflict lexicons any more than the articles from the other two groups.

Additionally, to see if there are patterns between the three groups we drew a graph where each bar represented each peace group. Each bar was colored based on the percentage of conflict/peace/resilience terms used in the articles of each group. The graphs don't seem to give us any insightful information. The peace group has the largest amount of peace lexicons used but there is no real difference between the number of conflict terms used within the three groups.

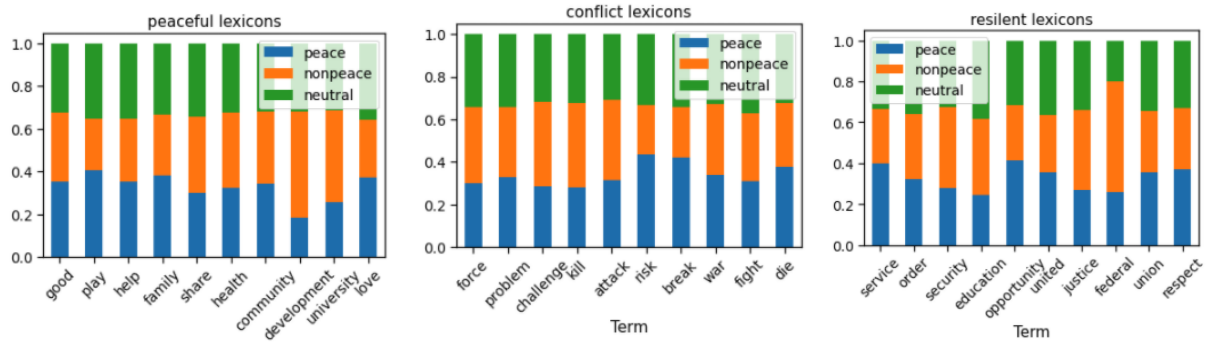

Fig 2-2. Percentage of articles that use a specific term from the lexicon.

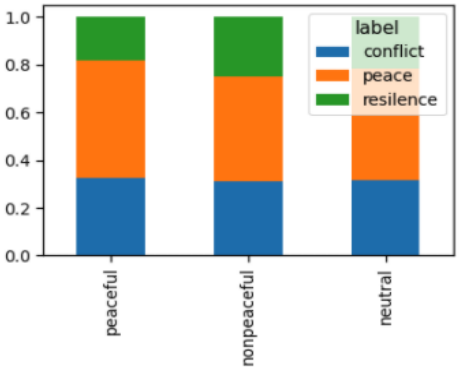

Fig 2-3. Percentage of lexicon terms used from each peace group.

2.3 Word2Vec

As section 2.2 did not provide any meaningful correlations between lexicons and articles we decided to plot the words of the lexicons and color them by their respective lexicon groups. With this plot, we can see if there are clusters between different lexicon groups. In order to plot the lexicon words in a 2 dimension space, we have to first vectorize the words and then transform the word vector into 2 dimensions. Initially, we got all the articles from our dataset and split them up by the spaces in between the words. Using this, we trained a Word2Vec model that outputs a vector of size 300 when provided a word as input (Mikolov et al., 2013). We put all the words in

the lexicon to the model and got a vector of size 300 for each of the words. Finally, by using T-SNE we transformed the vectors into 2 dimensions and used them to plot the graph. As seen in figure 2-3 we can clearly observe that there is a distinction between words in the peaceful and non-peaceful lexicons and between non-peaceful and resilient words but it is hard to differentiate between words in a peace lexicon and a resilient lexicon.

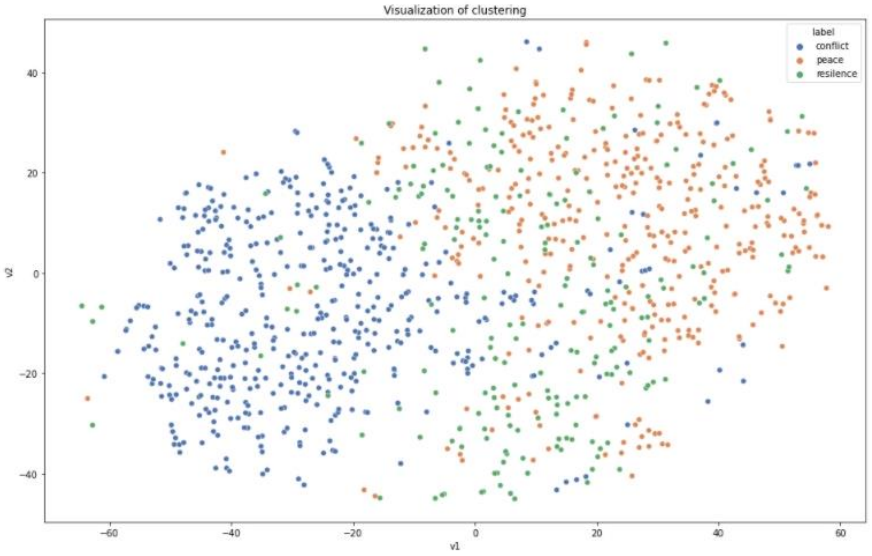

Fig 2-4. Plotting lexicon words using Word2Vec

2.4 Possible reasons of failure

The word2vec model did show some results where there seemed to be a distinction between the positive lexicons and the conflict lexicons. However, holistically, it was difficult to claim that there exists a relationship between peaceful countries with peaceful lexicons and non-peaceful countries with conflict lexicons. With careful discussion with mentors, we have narrowed down to the following three reasons: sets of pre-defined lexicon are flawed, the classification of countries is flawed, or an article from a peaceful country writing about a non-peaceful country has an impact.

3 Domestic Filter

3.1 Motivation

As one of the possible reasons that we discussed for the unclear distinction among different categories of lexicons was due to a mix of domestic and international articles, we decided to implement a domestic filter, which will separate articles into domestic and international articles. Our hypothesis is that if we filter articles based on content and perform our analysis on articles that were classified as domestic, we might be able to get more accurate results. In detail, we planned to perform word count analysis and word2vec model analysis to see how the classification of countries by peacefulness and classification of lexicons were related.

3.2 Methodology

Implementation of domestic filters was not as trivial as we expected when we first started to develop the filter. We utilized the Named Entity Recognition (NER) method to count the names of the countries. Specifically, we used a python package called locationtagger, the NER package based on NLTK. However, the biggest problem was that the accuracy for recognizing a country name was significantly low when the text is already uncapitalized and does not have punctuation. In order for us to increase the accuracy of the NER task, we had to reproduce the preprocess method in order to maintain capitalized character and punctuations.

Separating domestic articles from international articles is not as clear as we expected in the first place. For example, there was an article from one of the publishers in Australia that discussed multiple accidents that Australian tourists went through in South America. It was difficult to tell whether this article should fall into domestic or international. On the other hand, there were articles that did not have any country names in it. To reconcile these issues, we came up with 2 assumptions. Firstly, if the country name of the particular publisher is from shows up, then we categorized this article as domestic. Secondly, we assumed the articles that do not have any country names to be domestic as well. Using these assumptions, the algorithm will first count all the country names appearing in the article. Then, if either no country name appears or the name of the country that particular publisher is from appears at least once, the algorithm will assign the article as domestic, and otherwise, international.

3.3 Results

The algorithm above rather created a similar distribution of articles between domestic and international across different countries. Below is the bar graph that compares distribution among countries. We also grouped these countries by their peace level, and were able to find an interesting pattern that peaceful countries tend to focus more on international articles compared to neutral and non-peaceful countries. Because peaceful countries tend to be developed countries, while neutral and non-peaceful countries tend to be developing countries - we can come to the conclusion that developed nations have a higher interest in international articles.

One possible explanation for this may be that developed nations, which generally have more educated populations/readers, are more aware of the interconnectivity of the global community. Another explanation could be that developed nations, being more peaceful, have less “breaking” newsworthy events in comparison to developing, politically unstable countries, and tend to report about these “breaking” news events in other nations.

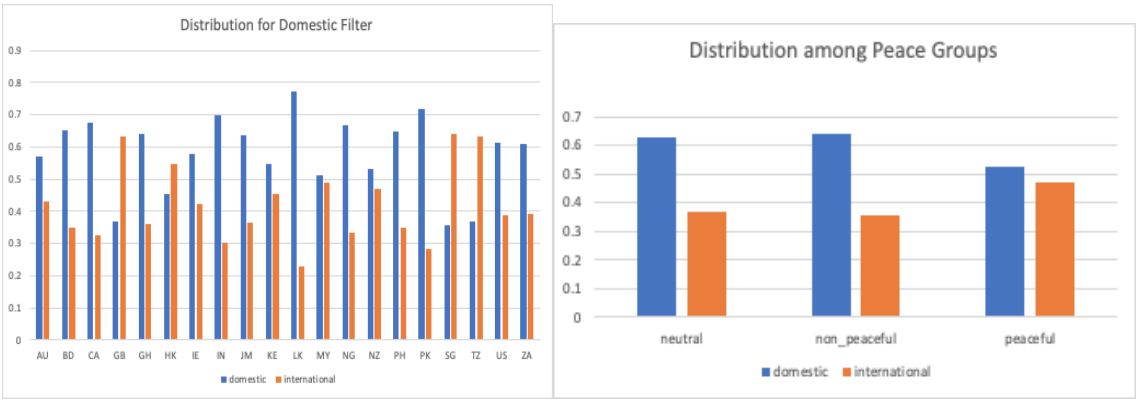

Fig 3-1. Distributions of Domestic and International articles

3.4 Conclusion

Below are the graphs for word count analysis for the top 10 words from peaceful, resilient, and non-peaceful lexicons. If our null hypothesis were to be supported, we should see bigger blue bars for peaceful lexicons, green bars for resilient lexicons, and orange bars for non-peaceful lexicons. However, it was hard to distinguish any meaningful differences among bar graphs. Rather, it was very similar to a word count analysis graph using both domestic and international articles. From this observation, we can conclude that our hypothesis regarding domestic and international articles are not supported.

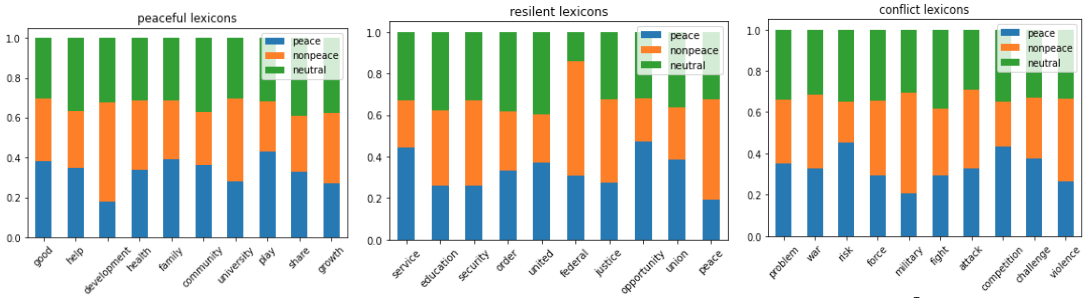

Fig 3-2. Distributions of lexicons by peace level for domestic articles only

4 Lexicon Augmentation

4.1 Background

The key hypothesis of this project is that language in media can distinguish peaceful from non-peaceful nations. Prior to our investigation, the project managers had derived three lexicons: peace, conflict, and resilience. Focusing on the first two, these lexicons contain the words most commonly found in peaceful and in-conflict countries, each containing 814 and 750 words,

respectively. These words were curated by combing through a number of academic articles and selecting fit words by hand.

Peace	Conflict
give	force
good	problem
play	challenge
help	kill
family	attack

Fig 4-1. Top 5 most frequent words from original Peace and Conflict lexicon. Obtained from a sample of 1 million articles.

4.2 Finding New Words through Frequency Analysis

The most basic method of finding new words specific to our data is through word frequency analysis. Using a bag of words approach, we can count the occurrence of each word in our sample, and compare what words are appearing more frequently in peaceful nations as compared to non-peaceful nations. Selecting what words to be included in our "new" lexicon is slightly arbitrary at this point, as we lack the domain-specific knowledge to make well-informed decisions about what words make sense to be included in the "new" lexicon.

4.2.1 General Selection Algorithm

To select new words for our lexicons, we follow the following algorithm. Given a set of words with their respective frequencies for each society type:

1. Select the top *N* words by frequency in each society.
2. From those, select the unique words in each society and assign them to that lexicon.
3. For the words that appear in both societies, pick a difference *D*. Compare the ranks of the common words, and if the difference in rank is at least *D*, append the word to the lexicon where the rank is higher.

For example, below is a demonstration of if we performed the frequency analysis and selected $N = 5$ and $D = 3$. These are not the actual words we obtained.

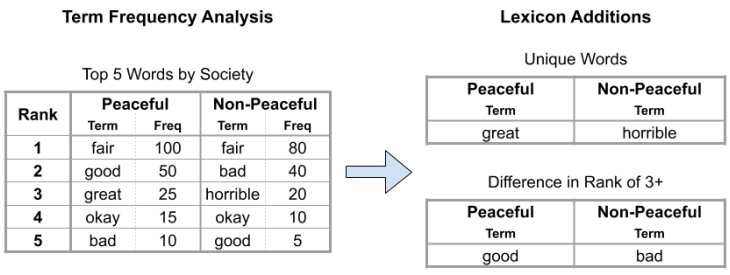

Fig 4-2. Example of creating new lexicons from word frequency rankings ($N = 5$, $D = 3$).

4.2.2 Filtering with WordNet

WordNet is a lexical database that provides groupings for nouns, verbs, adjectives, and adverbs into cognitive synonym sets (synsets). For example, the word "educate" has the synset "educate" and "train." The NLTK implementation of wordnet in python also tells us the part of speech of each of these words, in this case, they are both verbs. Another example is "Nigeria." This is a common word in the non-peaceful countries, and it is synset is just itself, "nigeria," which is a noun. One last example is the nonsense word "aaaa"; this sysnet is empty. From these examples, we can perform powerful filtering when we search for words to add to the lexicons. One way we do this is by removing words that have empty synsets. These words are likely nonsense, and therefore can be excluded. Another way we use wordnet is by filtering for words that are of a specific part of speech. Nouns can be problematic in this analysis so ignoring nouns altogether is one possible approach. Using synsets, we can determine the most likely part of speech for a word, so if we want to ignore words that are nouns, we can ignore words where the synset only contains nouns. Similarly, if we wanted to only consider verbs, which are very useful in this analysis, we could consider only words where there is at least one verb in the synset.

When we combine this filtering with the general selection algorithm, we can discover the most frequent verbs, adjectives, and adverbs used in each society. To form our "new" lexicon, we created two sets of words using the noun and verb filters discussed prior and applied the general selection algorithm (with N = 250 and D = 30) to obtain a list of 414 unique words that we can append to the existing lexicon.

Peace	Conflict
look	state
think	minister
play	court
home	accord
open	project

Fig 4-3. Top 5 most frequent words from new Peace and Conflict lexicons. Obtained from a sample of 1 million articles.

4.3 Lexicon Evaluation

Now that we have the term frequencies for all the words, we can investigate more thoroughly why the peace metric (discussed in section 2 and 3) did not perform well originally, and how our "new" lexicon can remedy it. When we select just the words from our original lexicon and visualize their distribution, we see that peaceful and non-peaceful societies are using a very similar proportion of peace and conflict terms. This explains why the scores are not dramatically different. However, when we plot the same thing for our new lexicon words, we clearly see that the non-peaceful countries are using conflict terms more, and peaceful countries are using peace terms more. Now, this is rather self-fulfilling, since these new lexicons were derived from the word

counts in each country. However, this serves as evidence that language does contain the power to differentiate between peaceful and non-peaceful nations.

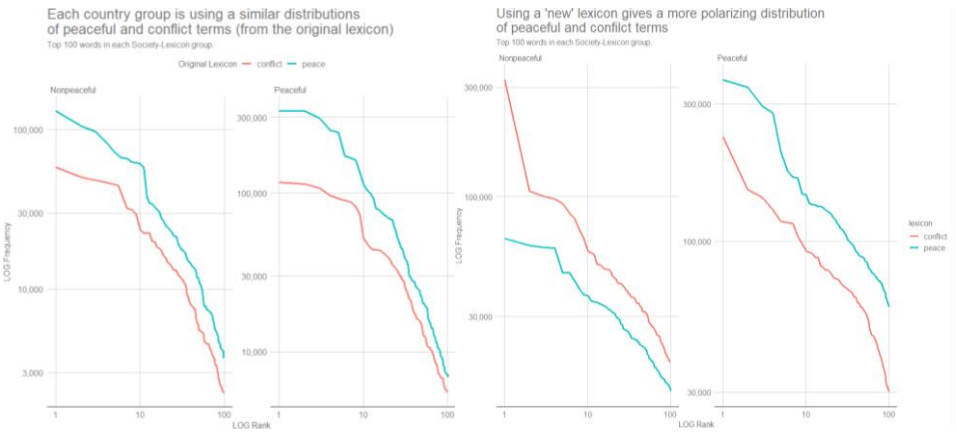

Fig 4-4. Distribution of the top 100 words for each society-lexicon group.

Lastly, we examine the peace metric once more. First with our new lexicon alone, and secondly with the two lexicons combined. We can see that using our new lexicon alone gives a peace metric that is clearly divisive among peaceful and non-peaceful countries while being mostly around 0 for other countries. When we use the original lexicon with our new lexicon, the separation remains clear.

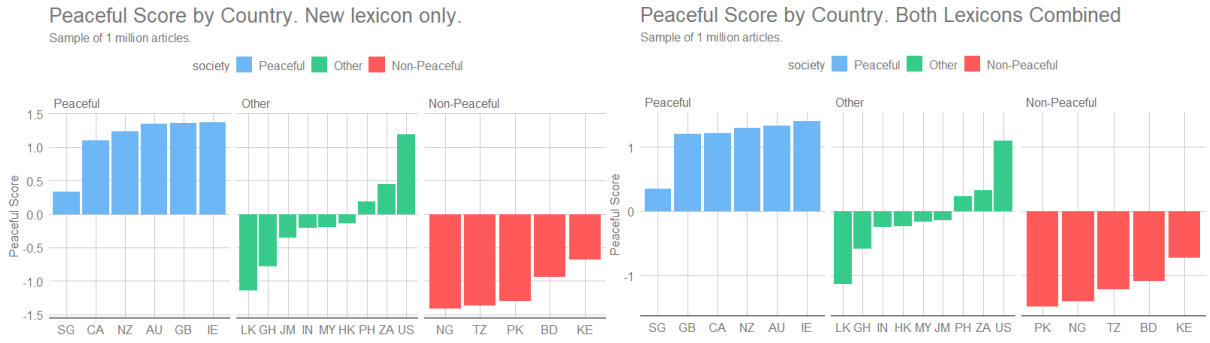

Fig 4-5. Comparison of peace metric when using our new lexicon alone and combined with the original lexicon.

5 Classification Model

5.1 Motivation and Method

As the word frequency analysis in section 3 does not yield much insight, we came up with the possible reasons for the result. One of them is that the categorization of neutral, peace and non-peaceful nations is flawed or the articles from peaceful and non-peaceful countries are similar in the matter of linguistic features. We want to test a hypothesis: is there a boundary between articles from peaceful or non-peaceful countries? In order to test the hypothesis, we have decided to build

a classification model and check whether the model can find the difference between the articles. We have used random forest with Doc2Vec embedding and BERT with a linear layer for the model.

5.2 Data

We want to find whether there is a boundary between peaceful and non-peaceful nations. Therefore, we have used only peaceful and non-peaceful nations, which is in a total of 874,535 articles from 11 countries. Since the purpose of the classification is to test the hypothesis and BERT's fine-tuning process takes a long time, we have subsampled 100,000 articles with equal distribution of peaceful and non-peaceful for the train & validation, and 50,000 articles for the test. Moreover, we have used n-gram preprocessed text data (no lemmatization and stopword removal) for both models.

5.3 Random Forest with Doc2Vec

In order to use a random forest, we need to transform our news article into a vector. We have used a gensim's Doc2Vec embedding for the vectorization of a text (Le et al., 2014). We have trained the Doc2Vec model for the embedding on the training dataset and used it for training Random Forest. We have tuned parameters of Random Forest Classifier with grid search, and test it on the test data.

5.4 BERT

As BERT's performance on classification tasks is well-known, we have tried to leverage the knowledge of pre-training on the classification task. For the preprocessing of text, we have used BERT's tokenizer for both tokenization and encoding. Since the BERT can only take 512 tokens (including [CLS] and [SEP]) and our articles have a longer length than 512 tokens, we have only used the 512 tokens from the article for the input data (Devlin et al., 2019). For the model, we have used the BERT-base-uncased from huggingface.co. With the BERT model, we have passed the first token ([CLS]) to the linear layer for the classification purpose.

5.5 Result & Analysis

As the table shows, Random forest with Doc2Vec achieves moderate results and BERT with linear layer achieves a great result. As most of the evaluation metrics are above .95, and the confusion matrix shows that it can classify whether an article is from a peaceful country or a non-peaceful country. Hence, we have concluded that there is a boundary between a peaceful and non-peaceful country.

		Precision	Recall	f1-score
Doc2Vec	non-peaceful	0.78	0.72	0.75
	peaceful	0.73	0.79	0.76
			test accuracy	0.76
	macro avg	0.76	0.76	0.76
	weighted avg	0.76	0.76	0.76
BERT	non-peaceful	0.95	0.95	0.95
	peaceful	0.95	0.95	0.95
	validation accuracy	0.95	test accuracy	0.95
	macro avg	0.95	0.95	0.95
	weighted avg	0.95	0.95	0.95

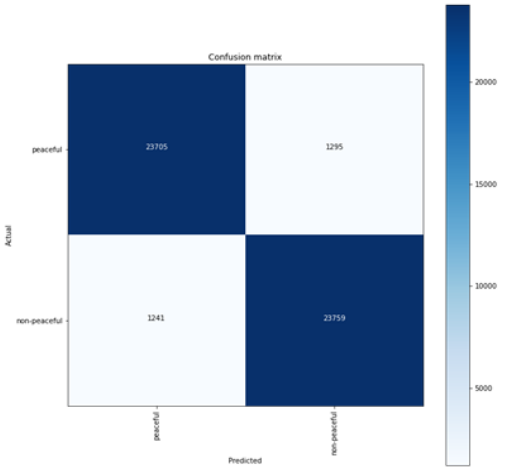

Fig 5-1. Result Table on the classification models and Confusion matrix from BERT classifier

6 Next Goals

Since we can infer that pre-defined lexicons are flawed from newly found lexicons and there is a boundary between articles from peaceful and non-peaceful nations, we will try to focus on finding the boundary between them. We want to investigate the difference between articles in peaceful and non-peaceful nations so that we can create an index or score of an article that contributes to peace.

The first method we will try is to build an interpretable deep learning model using a single layer of RNN with an attention layer as a classification and interpret the attention weights for finding a new set of lexicons or contexts that have an impact on the boundary. In such a way, we will be able to create a new set of lexicons which are added by a deep learning model. If we cannot interpret the attention weight, we will try to validate the power of selected lexicons by filtering them from the news articles and see the results of classification without the impactful lexicons.

7 Contribution

- **Jinwoo Jung** (team captain): Set-up milestones and managed progress. Main contributor to analysis and implementing domestic filters.
- **Hojin Lee**: Main contributor in word frequency analysis and domestic filters.
- **Hyuk Joon Kwon**: Main contributor in using Word2Vec to cluster lexicons and assisted in creating the random forest model with Doc2Vec.
- **Matt Mackenzie**: Main contributor in development of new lexicon through word frequency analysis.
- **Tae Yoon Lim**: Main contributor in developing classification model using Doc2Vec and BERT with a linear layer

8 References

[1] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019, May 24). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Retrieved 2020, from https://arxiv.org/abs/1810.04805

[2] Le, Q., & Mikolov, T. (2014, May 22). Distributed Representations of Sentences and Documents. Retrieved November 22, 2020, from https://arxiv.org/abs/1405.4053

[3] Lin, Y., Michel, J., Aiden, E., Orwant, J., Brockman, W., & Petrov, S. (2012, July 01). Syntactic annotations for the Google Books Ngram Corpus. Retrieved 2020, from https://dl.acm.org/doi/10.5555/2390470.2390499

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013, October 16). Distributed Representations of Words and Phrases and their Compositionality. Retrieved November 24, 2020, from https://arxiv.org/abs/1310.4546

[5] Miller, G. A. (1995). WordNet: A lexical database for English [Abstract]. *Association for Computing Machinery, 38*(11), 39-41. https://dl.acm.org/doi/10.1145/219717.219748

[6] Piantadosi S. T. (2014). Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6