

Final Report - Peace Speech Project

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

December 18, 2020

1. Problem definition

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 this final report, we will summarize the first and second progress report, which includes data engineering, initial exploratory data analysis, pre-processing, hypothesis testing and modeling. To see more details in earlier progress, please refer to each specific report presented earlier. Also, we will describe new approaches we took after summarizing previous reports, which focuses on how we came up with different models to derive new sets of lexicons.

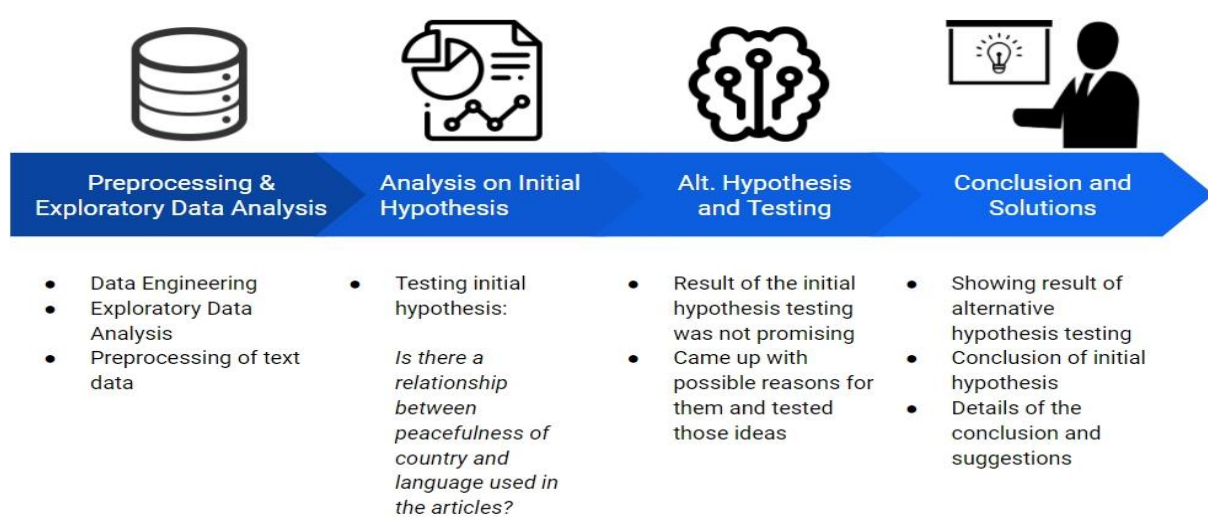


Fig 1-1 Summary of Progress

2. Summary of Progress Report 1

2.1. Data and Data Engineering

Our group worked on the News on the Web (NOW) dataset from corpusdata.org, which is composed of 11.2 billion words from web-based newspapers from 2010 from 20 countries. The dataset is composed of multiple files, divided into the same two types: source files that contain metadata like a publisher, country of origin, and others, and text files that contain the raw texts for each news article. There were about 20 million news articles contained in the full raw dataset of size over 60GB. We were also given predefined lexicons for each peaceful, neutral, conflicting countries, and predefined classifications of the countries into the 3 societies.

The raw datasets we obtained were first organized to meet our project management team's request and to facilitate further analysis. To be able to easily transfer our work to others when this project is finished, they requested each article to be extracted into its own file and placed in a folder such as "[Country]/[Publisher]/[Year]/[text_file].txt". It was not a trivial task since the datasets were more than 60GB, 20 million articles to process where datasets were split into 49 source fields and 129 text folders, with each folder containing at least 20 text files (Fig. 10-1, 10-2, 10-3).

2.2. Exploratory Data Analysis

In order to perform text-preprocessing and further analysis, we wanted to come up with a way to sub-sample the whole dataset. Because the number of articles from publishers that have more than 1000 articles accounts for around 85% of the full data, we decided to use articles from these publishers that have more than 1000 articles for future analysis (Fig 10-4).

We observed that there exists a large imbalance of data between countries and years. Some countries such as the U.S. have significantly more articles than others (Fig 10-6) and there are significantly larger numbers of articles from 2015 and onward. We thus further applied downsampling for years after 2015 so that each year has a similar number of articles. We also made sure to maintain similar distributions of articles from each publisher so the relative representations remain the same. The above procedure allowed us to decrease the dataset's size from 20 million articles of size over 60GB to around 1.5 million articles of size about 1.5GB.

2.3. Text Pre-processing

Inspection of the processed data above showed that there exists unnecessary information that may affect later analysis such as inducing readers to subscribe to their contents or suggesting other unrelated articles. Since these sentences may involve words that are later recognized as the lexicons that represent peaceful or hateful societies, we decided to filter out the noisy information that is not related to the news articles. This text-preprocessing procedure is composed of two parts. The first part is the general procedure, which is to clean the scraped news articles into easily

readable sentences by removing HTML tags and symbols. For the second part, we compared two specific models, the N-Gram model, and the Cosine Similarity Sentence Embedding model.

The main hypothesis behind the N-Gram model is with an assumption that similar noisy patterns are publisher-specific. Particularly, we chose 5 grams since it is uncommon to have phrases of 5 same words appear multiple times. Accordingly, by grouping articles into each publisher and tokenizing each tense, we measured frequencies of 5-gram phrases and removed sentences with the particular phrases that appear above 25% per publisher (Fig 10-7).

Unlike the N-Gram model which works only when we group the data into each publisher, we also thought of performing text preprocessing using Cosine similarity with sentence vectors, which works per each article. After tokenizing each sentence in a document using Sentence-Bert (Reimers & Gurevych, 2019) and HuggingFace transformers (Wolf et al., 2020), we computed the cosine similarity between each sentence and document, then removed sentences found to have a low level of similarity, where we used hyper-parameter of 0.95.

As shown in the below table, both the 5-Gram model and the model with Cosine Similarity each have advantages and disadvantages. However, our group decided to proceed with only the 5-Gram model mainly due to much faster processing time and due to lack of power of our group's local machines (More examples of the comparison on Fig 10-8, 10-9).

	Pros	Cons
N-gram (5-gram)	<ul style="list-style-type: none"> - Safer, which sentence to remove (clean document -> not remove any) - Faster to run (~ 11 sec per 250 articles) 	<ul style="list-style-type: none"> - Fails for recurring phrases of less than 5 words - Processing time increases exponentially (process per doc, depend on # of articles per publisher)
Cosine Similarity	<ul style="list-style-type: none"> - Able to delete phrases with less than 5 words - Time complexity: linear in number of article 	<ul style="list-style-type: none"> - Remove sentences that are not spam - Unable to control (pre-trained, vectorize) - Slower to run (~ 6 mins per 250 articles)

Fig 2-1 Pros and Cons of the N-Gram model and Cosine Similarity Model

3. Summary of Progress Report 2

3.1. Word Frequency Analysis

After preprocessing the data, we decided to look into the preprocessed data to identify if we could find specific patterns using the lexicons. Specifically, we were curious whether lexicons are used proportionally given the classification of countries by peace level and whether the classification of countries into the peace level communities are valid.

3.1.a. Word Frequency Analysis: Peace Metric

We defined the peace metric as the percentage of positive words used subtracted by the percentage of negative words used and was used to investigate the originally given lexicons. Since all of the values of the peace metric were positive, the values are then normalized when the graph was drawn (Fig 10-10). The results were not as promising as we expected since results from countries such as Great Britain (GB) and Tanzania (TZ) showed completely opposite of what we expected. Moreover, we drew box plots and tested if the box plots were significantly different from each other using ANOVA, and we found out that the p-value was too big to conclude that there were any significant differences among the peace groups.

3.1.b. Word Frequency Analysis: Word Frequency Count

Our initial assumption was that positive lexicons are used more frequently in peaceful countries and negative lexicons are used more in non-peaceful countries. We performed a simple Bag-of-Words approach, and counted each word from documents. However, unlike our expectation, the articles from peaceful countries do not seem to use words from the peace lexicon more than non-peaceful countries and vice versa (Fig 10-11).

3.2. Word2Vec

Simple word frequency analysis did not provide any meaningful correlations between lexicons and articles. Thus, we decided to train the words with the articles using the Word2Vec model. 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 a word is inserted as input (Mikolov et al., 2013). Then, we put all the words in the lexicon to the model and got a vector of size 300 for each of the words. In order to plot the lexicon words in a 2-dimension space, we transformed the word vector into 2 dimensions by using T-SNE. It was possible to observe a clear distinction between words in the peaceful and non-peaceful lexicons and words between non-peaceful and resilient words but it is hard to differentiate between words in a peace lexicon and a resilient lexicon (Fig 10-12). From this result, we realized that further research is required on initial hypotheses.

3.3. Possible issues for unsuccess of initial approach

The result from the word2vec model showed that 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 the 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.

3.3.a. Potential Solution: Domestic Filter

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 after implementing a domestic article filter to see how the classification of countries by peacefulness and classification of lexicons were related. We utilized the Named Entity Recognition (NER) method to count the names of the countries and filter articles based on the number of times a country is mentioned. However, the test result after the implementation of the filter was similar to previous test results, and it was still difficult to distinguish any meaningful differences among graphs. From this observation, we can conclude that our hypothesis regarding domestic and international articles is not supported (Fig 10-13).

3.3.b. Potential Solution: Lexicon Augmentation

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 combining a number of academic articles and selecting fit words by hand. As we are dealing with news articles and not academic articles the meaningfulness of the lexicons given to us could be problematic, therefore, an effort to create a set of new lexicons was needed. We used term frequency analysis from a bag of words approach combined with WordNet to discover the most frequent verbs, adjectives, and adverbs used in each society. Then, we examined the peace metric once more; firstly with our new lexicon alone, and secondly with the two lexicons combined. 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. This method proved to be successful (Fig 10-14, 10-15).

3.4. Potential Solution: Classification Models

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 the possible reasons 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 the following 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 from peaceful and non-peaceful countries. We have used two

models, a random forest with Doc2Vec embedding and BERT with a linear layer for the model. The Random forest with Doc2Vec achieves moderate results and BERT with linear layer achieved a significant result as most of the accuracy evaluation metrics are above 0.95. Using Classification models we have concluded that there is a boundary between a peaceful and non-peaceful country (Fig 10-16).

4. Defining new lexicons

After finding out that there is a boundary between articles from peaceful and non-peaceful countries, we have decided to find more lexicons residing in the boundary. We have previously used word frequency analysis which counts the frequency of words in both articles and chooses the most frequent and only included in one side of the article to define new lexicons. The new lexicons from the frequency analysis have improved the peace metrics score, hence we believe that the provided lexicons are not enough to define whether the article is peaceful or not. As BERT is a deep learning model, we have decided to create interpretable deep learning models in order to create a bigger and more comprehensive set of lexicons.

4.1. RNN with attention layer

The method is suggested by one of our mentors. We have used the 100,000 sampled articles from both peaceful and non-peaceful nations (equally distributed). The articles have been preprocessed with lemmatization using the NLTK and stopwords have been removed. We have only used the first 360 vocabularies from the texts as it is the mean length of articles.

4.1.a Method

We have retrieved the new set of lexicons with the following procedure. We have trained a classification model with an RNN (BiLSTM) with an attention layer to classify the articles between peaceful and non-peaceful nations. From the trained model, we retrieve the attention weights from the attention layer. Then, we have created a dictionary which matches the attention weights and corresponding tokenized word. We generated these dictionaries from all the sampled articles. We average the weights and sort them to get the top 400 vocabularies from peaceful and non-peaceful. In order to avoid words included in both peaceful and conflict, we have removed the common words.

4.1.b Result

Rank	Peace Lexicon	Conflcit Lexicon
1	south	president
2	press	minister
3	content	south
4	independent	bank
5	please	university
6	alone	news
7	advertisement	star
8	estate	league
9	newspaper	photo
10	cape	town

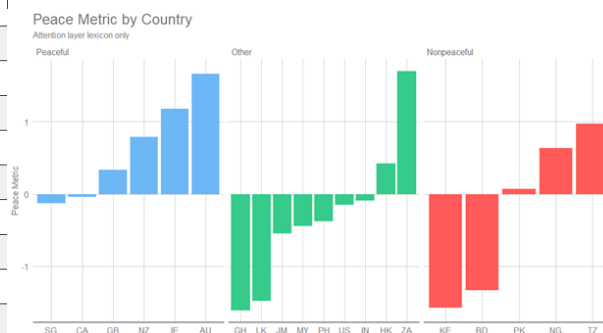


Fig 4-1 Top 10 ranked lexicons and peace metric calculated with the lexicon

As the above result shows, the lexicons from the attention layer are not as insightful as the provided lexicons or lexicons from the word frequency analysis. Since the attention layer weights correspond to the position of the vocabulary instead of vocabulary itself as word frequency analysis or BERT model. The attention weights and model outputs are not correlated, hence the attention weights are not the feature importance (Jain et al., 2019).

4.2. Encoder and Generator to extract rationale

After the failure of RNN with a single attention layer, our group has researched the interpretable NLP model to retrieve the lexicons from the deep learning model. Among a few papers about interpretable models, we have focused on rationalizing Neural Predictions (Lei et al., 2016). We have tried twice with different datasets.

4.2.a Data

We have used two datasets. The first dataset is the preprocessed articles with lemmatization and stopword removal and sampled 200,000 articles for both training and validation. The second dataset is a further preprocessed dataset from the first dataset. We have excluded digits and a few annotations from spacy to improve the quality of lexicons. We have removed PERSON (People, including fictional), NORP (Nationalities or religious or political groups), and GPE (Countries, cities, states) from the texts. We have only used the first 512 words from the article and padded the sequence if the article is less than 512 words.

4.2.b Model

- ❑ **Embedding:** We have used glove.6B.300d pre-trained word embedding for the representation of a word. The pre-trained dataset contains 400,000 words and we have added one more word for the zero-padding.

- ❑ **Encoder:** Given a training set with a sequence of input (x,y), the encoder predicts the label (peaceful or non-peaceful) of the article (Lei et al., 2016). The encoder can be any neural network (RNN, CNN, RCNN). The paper has suggested the RCNN architecture for the encoder. However, for now, we have used the CNN for text classification to fasten the training time (Kim, 2014).
- ❑ **Generator:** The generator extracts a subset of text from the original input to function as an interpretable summary (Lei et al., 2016). The rationale for a given sequence is defined as the binary variables $\{z_1, z_2, \dots, z_n\}$ where each variable is 0 or 1 indicating the word is selected or not.
- ❑ **Loss/Optimization:** With the above two functions, we generate a single model where the model classifies the article is peaceful or not with encoder and uses input as $\text{enc}(z, x)$ and minimizes the loss with the given label: $\|\text{enc}(z, x) - y\|^2$. The paper introduces a *doubly stochastic gradient* for the optimization, which uses the average gradient of sampled rationale from the generator to calculate the gradient of the encoder to optimize the model.
- ❑ **Lexicon Extraction:** The model outputs the series of words that the generator has chosen for the classification. We have extracted the rationales for every article in both training and validation set, then count the frequencies of the vocabulary in the article to generate a set of vocabulary.

4.2.c Result

Rank	Peace Lexicon	Conflict Lexicon
1	rogers	absa
2	rcmp	enugu
3	rte	naira
4	meath	mombassa
5	limerick	tehreek
6	gardai	awami
7	taranaki	sbp
8	postmedia	odm
9	iwi	vodacom
10	hutt	nakuru

Fig 4-2 Top 10 ranked lexicons from the model using second dataset

	Train Accuracy	Validation Accuracy	Train F1 score	Validation F1 Score
Data 1	0.92	0.90	0.90	0.89
Data 2	0.87	0.84	0.84	0.87

Fig 4-3 Evaluation metrics for the model on the both dataset

The table shows the top 10 most frequent vocabulary extracted from the model from data 2. Some of the vocabs (sdp, odm, etc) are not English vocabs and others do not give any insight about the peacefulness. The top 10 vocabularies from data 1 are mostly the locational names, hence

we have proceeded with the second dataset. Although the model classifies the second dataset well, we can create a hypothesis that the model learns about the locational name for the classification.

4.2.d Possible Reason of Failure

The vocabs generated from the second dataset might be flawed due to the preprocessing. The author of the paper does not mention any preprocess, hence using the raw texts (only n-gram preprocess) for the model might yield better rationales from the texts. We have used CNN for the classification purpose. However, since the text is sequential data, RNN should perform better. As the article suggests, we can use RCNN for encoder architecture.

Since the lexicons from the model do not yield insights and accuracy decreased with the filter, we have not used the lexicons for the evaluation in the next section.

5. Lexicon Evaluation

5.1 Lexicon Overlap

The three lexicon versions, Original, Term Frequency, and Attention Layer were all derived in very different ways. As such, there is extremely little overlap between these sets of words (see appendix 10-17), which raises some concerns. It is likely many of the words obtained from the latter three are not going to be useful and are rather artifacts of the specific sample we chose to work with. Since our experience studying peace has been rather short-lived, we have handed off all the terms we obtained to our advisors, and they will continue to refine and understand those lists.

5.3 Overall Peace Metric

We evaluate our countries once again with the peace metric, but this time we use the words from all three lexicons. As seen below, the scores are even better now than they were from just the term frequency lexicon.

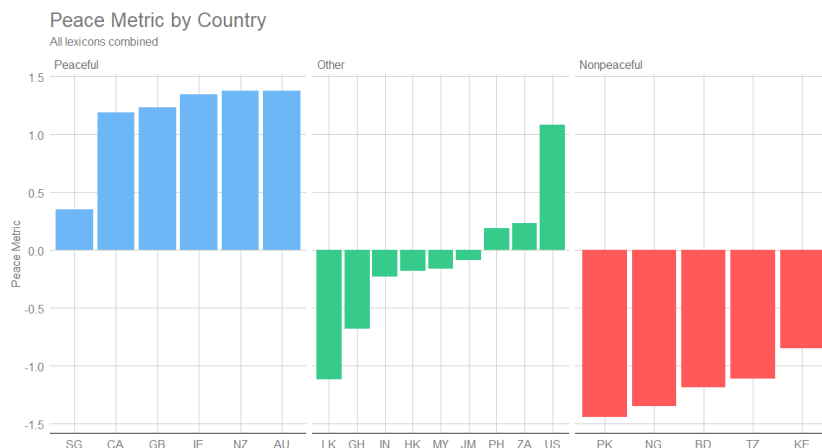


Fig 5-1 FIGURE NAME UPDATE NEEDED

5.3 Overall Predictive Power

With the success of the complex BERT model, our advisors raised the question, “are the lexicon words predictive on their own?” To address this, we used the term frequencies for the words in each lexicon as the predictors in a logistic regression to predict whether a country is peaceful or not (trained with a balanced sample of 150,000 articles split these into an 80/20 train-test split). When we use all the words from the original lexicon and all the unsupervised variations we have created together, the model is able to reach about 81%¹ accuracy.

5.3 Comparing Predictive Power

Since there is so little overlap between the three lexicons, we wanted to compare the predictiveness of each lexicon individually to get a sense if any lexicon is doing significantly better than another. By fitting a different logistic regression with just the words from each lexicon, we can see that the term frequency words do the best, but more importantly, it is clear that each lexicon on its own is not as predictive as all three together.

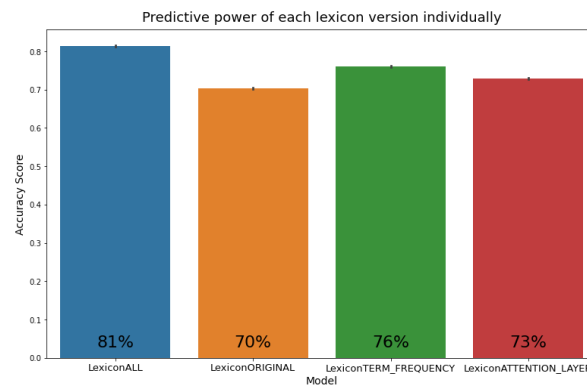


Fig 5-2 Comparison of lexicon performance in logistic regression.

5.4 Driver Analysis

Lastly, we return once more to the logistic regression fit with all 2,500 words from each lexicon. Using the coefficients from this model, we were interested if our lexicon assignments agree with how the model is using each word. Our positive class in the logistic regression was “Peaceful”, so we would expect words from the peace lexicons to have a positive coefficient. This would imply that that word is being used to increase the probability of a nation being peaceful. With the fit logistic regression, we take the top 500 most important terms (by the absolute value of the coefficient) and group the terms by the sign of the coefficient and the lexicon the word belongs to. As we can see, the majority of positive coefficient terms are from the peace lexicon,

¹ All the metrics displayed are a mean of 10-fold cross validation results using the train set.

while the majority of the negative coefficient terms are from the conflict lexicon. This tells us that our assignments are mostly correct.

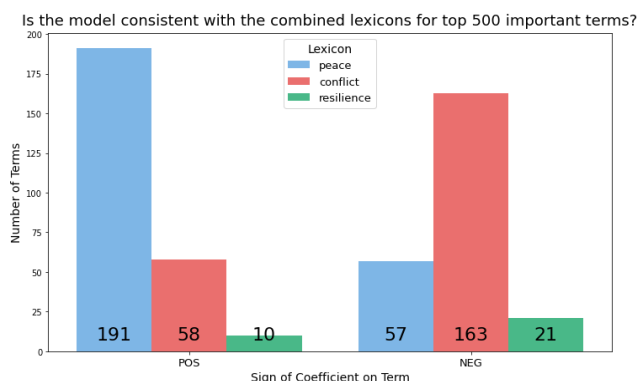


Fig 5-3 Count of terms that fall into each lexicon by the sign of their coefficient in the logistic regression.

6. Future Work

With numerous methods of data science, we have improved the quality of lexicons. In order to retrieve more lexicons from the deep learning model, we suggest using RCNN with n-gram preprocessed articles for the encoder and generator model to extract interpretable rationales from the articles. Further, one can refer to the paper by Bastings in which the model has developed from the encoder and generator model which shows better results than one from Lei (Bastings et al., 2019).

7. Ethical Considerations

Since our project is focused on analyzing words and lexicons that would represent peaceful, resilient, and conflicting countries, and the data we worked on are from publicly available online news media, most of the analysis were free of ethical concerns. However, some parts of our preprocessing steps may introduce small amounts of bias into our analysis.

In the data engineering procedure, due to the large size of the raw dataset, we selected to work only from publishers that have more than 1000 articles, which accounts for about 85% of the full data. This may have led to a larger focus solely on larger publishers, but not from publishers with smaller numbers of articles (Fig 10-4). Also, as Fig 10-5 suggests, there are discrepancies in the number of articles from the communities in that peaceful countries tend to have a larger number of articles than the other groups. Even after the downsampling, this may also have caused our analysis to rely more on articles from peaceful countries compared to resilient and conflicting communities (Fig 10-17).

8. Contribution

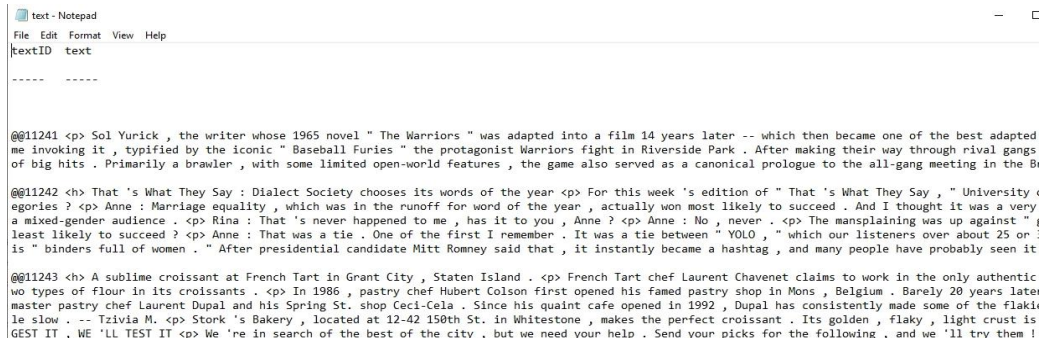
- **Jinwoo Jung** (team captain): Set-up milestones and managed progress. Main contributor to analysis and implementing N-Gram model and domestic filters.
- **Hojin Lee**: Main contributor in word frequency analysis and domestic filters.
- **Hyuk Joon Kwon**: Main contributor in using Word2Vec to cluster lexicons and assisted in creating the random forest model with Doc2Vec.
- **Matt Mackenzie**: Main contributor in development of new lexicon through word frequency analysis and evaluation of lexicons.
- **Tae Yoon Lim**: Main contributor in developing BERT classifier, RNN with Attention layer, and Encoder & Generator to extract rationales.

9. References

- [1] Beltagy, I., Peters, M., & Cohan, A. (2020, April 10). Longformer: The Long-Document Transformer. Retrieved October 22, 2020, from <https://arxiv.org/abs/2004.05150>
- [2] 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>
- [3] Le, Q., & Mikolov, T. (2014, May 22). Distributed Representations of Sentences and Documents. Retrieved November 22, 2020, from <https://arxiv.org/abs/1405.4053>
- [4] 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>
- [5] 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>
- [6] 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>
- [7] 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>

- [8] Reimers, N., & Gurevych, I. (2019, August 27). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Retrieved 2020, from <https://arxiv.org/abs/1908.10084>
- [9] Wei, J., & Zou, K. (2019, August 25). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Retrieved October 22, 2020, from <https://arxiv.org/abs/1901.11196>
- [10] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. (2020, July 14). HuggingFace's Transformers: State-of-the-art Natural Language Processing. Retrieved 2020, from <https://arxiv.org/abs/1910.03771>
- [11] Jain, S., & Wallace, B. (2019, May 08). Attention is not Explanation. Retrieved December 16, 2020, from <https://arxiv.org/abs/1902.10186>
- [12] Lei, T., Barzilay, R., & Jaakkola, T. (2016, November 02). Rationalizing Neural Predictions. Retrieved December 16, 2020, from <https://arxiv.org/abs/1606.04155>
- [13] Kim, Y. (2014, September 03). Convolutional Neural Networks for Sentence Classification. Retrieved December 17, 2020, from <https://arxiv.org/abs/1408.5882>
- [14] Bastings, J., Aziz, W., & Titov, I. (2020, June 19). Interpretable Neural Predictions with Differentiable Binary Variables. Retrieved December 17, 2020, from <https://arxiv.org/abs/1905.08160>

10. Appendix



```

text - Notepad
File Edit Format View Help
textID text

-----

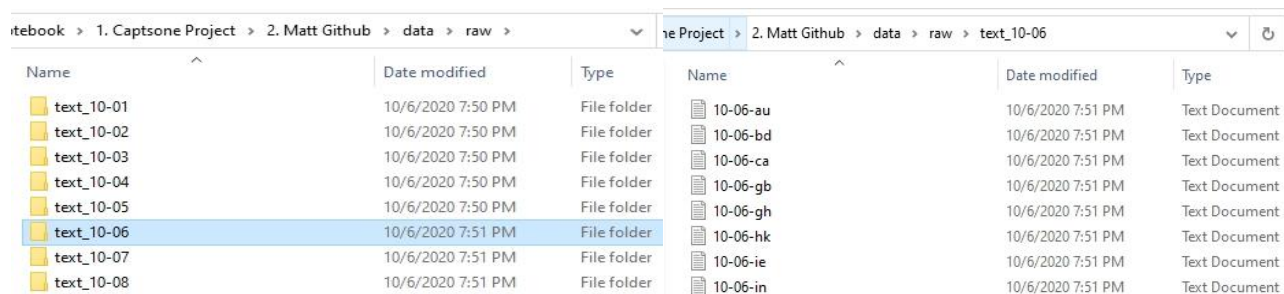
@@11241 <p> Sol Yurick , the writer whose 1965 novel " The Warriors " was adapted into a film 14 years later -- which then became one of the best adapted
me invoking it , typified by the iconic " Baseball Furies " the protagonist Warriors fight in Riverside Park . After making their way through rival gangs
of big hits . Primarily a brawler , with some limited open-world features , the game also served as a canonical prologue to the all-gang meeting in the B

@@11242 <h> That 's What They Say : Dialect Society chooses its words of the year <p> For this week 's edition of " That 's What They Say , " University c
egories ? <p> Anne : Marriage equality , which was in the runoff for word of the year , actually won most likely to succeed . And I thought it was a very
a mixed-gender audience . <p> Rina : That 's never happened to me , has it to you , Anne ? <p> Anne : No , never . <p> The mansplaining was up against " I
least likely to succeed ? <p> Anne : That was a tie . One of the first I remember . It was a tie between " YOLO , " which our listeners over about 25 or 1
is " binders full of women . " After presidential candidate Mitt Romney said that , it instantly became a hashtag , and many people have probably seen it

@@11243 <h> A sublime croissant at French Tart in Grant City , Staten Island . <p> French Tart chef Laurent Chavenet claims to work in the only authentic
wo types of flour in its croissants . <p> In 1986 , pastry chef Hubert Colson first opened his famed pastry shop in Mons , Belgium . Barely 20 years later
master pastry chef Laurent Dupal and his Spring St. shop Ceci-Cela . Since his quaint cafe opened in 1992 , Dupal has consistently made some of the flaki
le slow . -- Tzivia M. <p> Stork 's Bakery , located at 12-42 150th St. in Whitestone , makes the perfect croissant . Its golden , flaky , light crust is
GEST IT , WE 'LL TEST IT <p> We 're in search of the best of the city , but we need your help . Send your picks for the following , and we 'll try them !

```

Fig 10-1 Screenshot of Sample Text File



Name	Date modified	Type
text_10-01	10/6/2020 7:50 PM	File folder
text_10-02	10/6/2020 7:50 PM	File folder
text_10-03	10/6/2020 7:50 PM	File folder
text_10-04	10/6/2020 7:50 PM	File folder
text_10-05	10/6/2020 7:50 PM	File folder
text_10-06	10/6/2020 7:51 PM	File folder
text_10-07	10/6/2020 7:51 PM	File folder
text_10-08	10/6/2020 7:51 PM	File folder

Name	Date modified	Type
10-06-au	10/6/2020 7:51 PM	Text Document
10-06-bd	10/6/2020 7:51 PM	Text Document
10-06-ca	10/6/2020 7:51 PM	Text Document
10-06-gb	10/6/2020 7:51 PM	Text Document
10-06-gh	10/6/2020 7:51 PM	Text Document
10-06-hk	10/6/2020 7:51 PM	Text Document
10-06-ie	10/6/2020 7:51 PM	Text Document
10-06-in	10/6/2020 7:51 PM	Text Document

Fig 10-2 Screenshot of full data structure

Peaceful	Non-Peaceful (Conflicting)	Neutral
Australia (AU)	Bangladesh (BD)	Ghana (GH)
Canada (CA)	Kenya (KE)	Hong Kong (HK)
Ireland (IE)	Nigeria (NG)	India (IN)
New Zealand (NZ)	Pakistan (PK)	Jamaica (JM)
Singapore (SG)	Tanzania (TZ)	Malaysia (MY)
United Kingdom (UK, GB)		Philippines (PH)
		South Africa (ZA)
		Sri Lanka (LK)
		United States (US)

Fig 10-3 Breakdown of the predefined countries by peace level

Number of Articles Published	Publishers		Articles	
	Count	%	Count	%
1	6,762	18.1%	6,762	0.03%
2	3,256	8.7%	6,512	0.03%
3	1,968	5.3%	5,904	0.03%
4-5	2,540	6.8%	11,250	0.06%
6-10	3,312	8.9%	25,581	0.13%
11-25	4,171	11.2%	69,968	0.35%
26-100	5,487	14.7%	292,704	1.45%
101-500	5,321	14.2%	1,266,238	6.27%
501-1000	1,586	4.2%	1,124,211	5.57%
1000+	2,994	8.0%	17,381,823	86.09%

Fig 10-4. Breakdown of full dataset by size of publisher

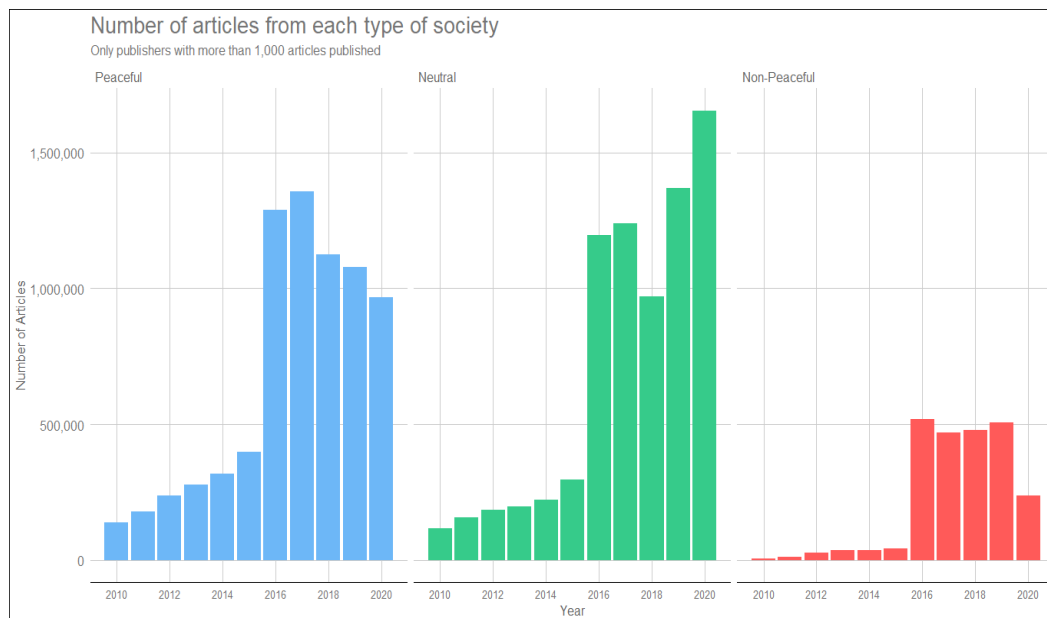


Fig 10-5 Breakdown of news articles each type of country

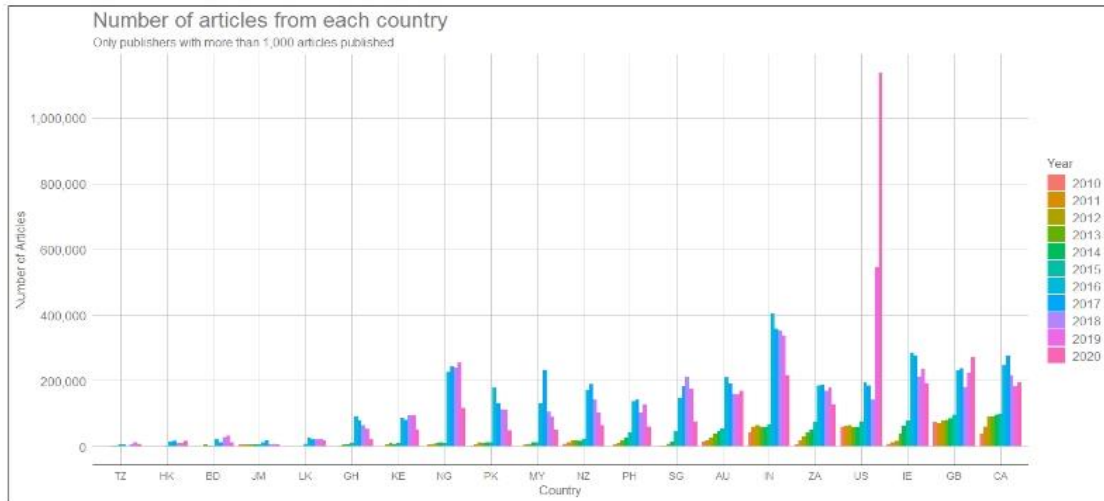


Fig 10-6 Number of Articles per Country per Year

Publisher (# articles)	Top 5 Frequency of phrase	Frequency	Sentences associated with phrase
Times of India (91)	from the times of india	156	more from the times of india / what better than donating blood and saving lives gupta added from the times of india / sex ratio has improved from 1991 to 2001 and till now more from the times of india
	More from the times of	154	improved from 1991 to 2001 and till now more from the times of india
	guidelines by marking them offensive	82	help us delete comments that do not follow these guidelines by marking them offensive / refrain from posting comments that are obscene defamatory or inflammatory and do not indulge in personal attacks us delete comments that do not follow these guidelines by marking them offensive (and other variations of this sentence)
	that do not follow these	81	
	follow these guidelines by marking	81	
Telegraph (52)	N/A	N/A	N/A
Independent Online (49)	addresses all users on independent	20	verified email addresses all users on independent email address before being allowed to comment on articles(and other variations of this)
	for more information please read	20	for more information please read our comment guidelines / for more information please read our
	hover your mouse over the	20	hover your mouse over the comment and wait until a small triangle appears on the right hand side
	our moderators will take action	20	our moderators will take action if need be
	and select flag as inappropriate	20	click triangle and select flag as inappropriate

Fig 10-7 Top examples of the removed sentences from 5-gram from top 3 publishers of the largest number of articles published in the sample dataset

treat obesity through modification of the gastroin
testinal tract and reduction of nutrient intake .
related. from around the web.
more from the times of india. recommended by colom
bia. from around the web. more from the times of i
ndia. recommended by colombia. comments. character
s remaining 3000. or proceed without registration.
share on twitter. sign in with. facebookgoogleemai
l. refrain from posting comments that are obscene
defamatory or inflammatory and inciting hatred aga
inst any community . help us delete comments that
do not follow these guidelines by marking them off
ensive . let's work together to keep the conversat
ion civil .

treat obesity through modification of the gastroin
testinal tract and reduction of nutrient intake .
from around the web. characters remaining 300
0. or

Fig 10-8 Comparison between initially preprocessed news article vs pre-processed version using N-Gram, 2010.01.02 Times of India's article, ID: 1335637

on of nutrient intake . related. from around the w
eb. more from the times of india.
recommended by colombia.
from around the web. more from the times of india.
recommended by colombia. comments. characters rema
ining 3000. or proceed without registration. share
on twitter. sign in with. facebookgoogleemail. ref
rain from posting comments that are obscene defama
tory or inflammatory and inciting hatred against a
ny community .
help us delete comments that do not follow these g
uidelines by marking them offensive
. let's work together to keep the conversation civ
il
.

on of nutrient intake . related. from around the w
1 eb. more from the times of india.
from around the web. more from the times of india.
sign in with.
help us delete comments that do not follow these g
uidelines by marking them offensive
.

Fig 10-9 Comparison between initially preprocessed news article vs pre-processed version using Sentence Embedding Cosine Similarity, 2010.01.02 Times of India's article, ID: 1335637

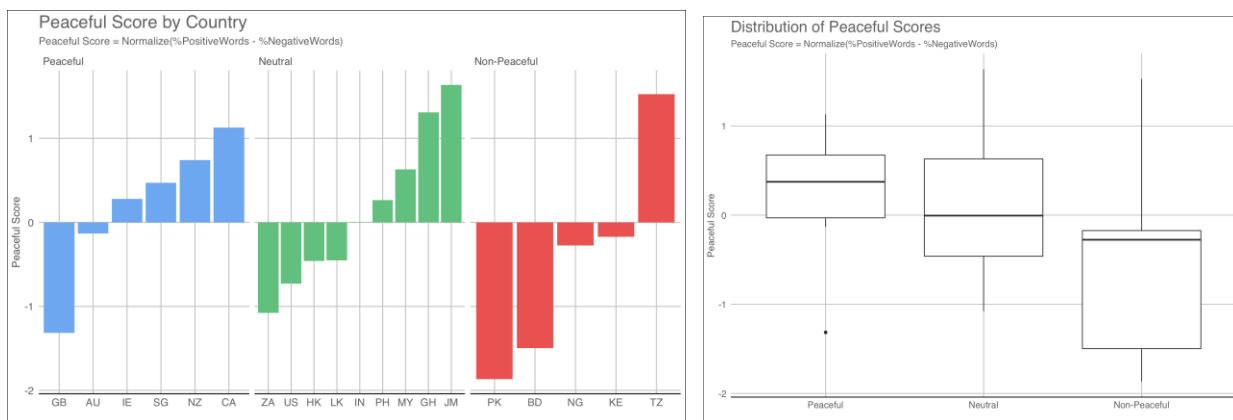


Fig 10-10 Peaceful Score by Country and Breakdown of news articles each type of country

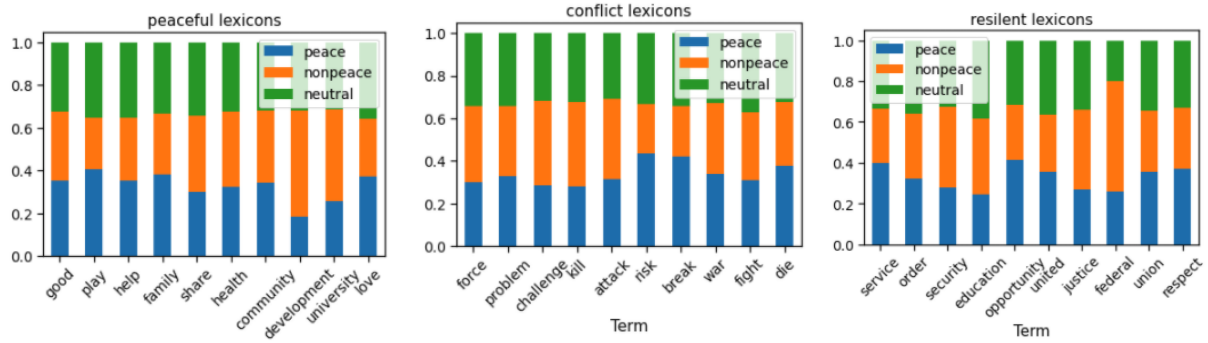


Fig 10-11 Percentage of articles that use a specific term from the lexicon.

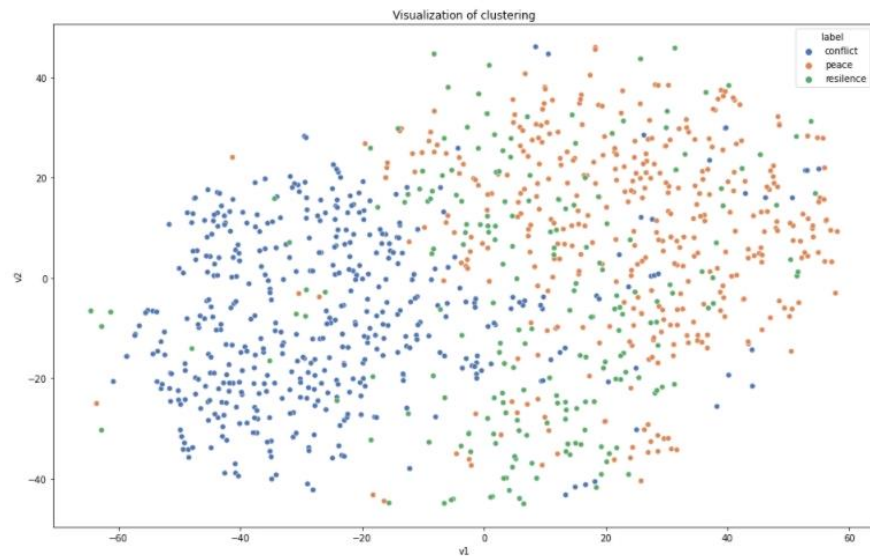


Fig 10-12 Plotting lexicon words using Word2Vec

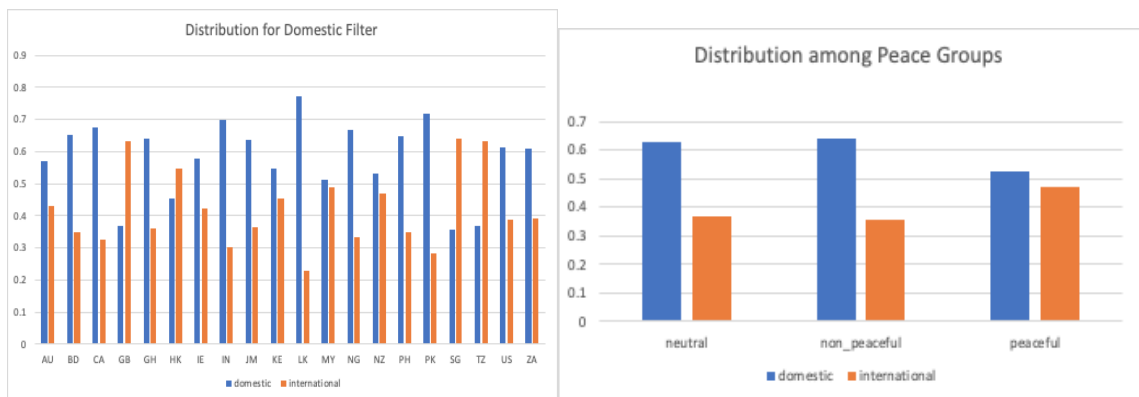


Fig 10-13 Distributions of Domestic and International articles

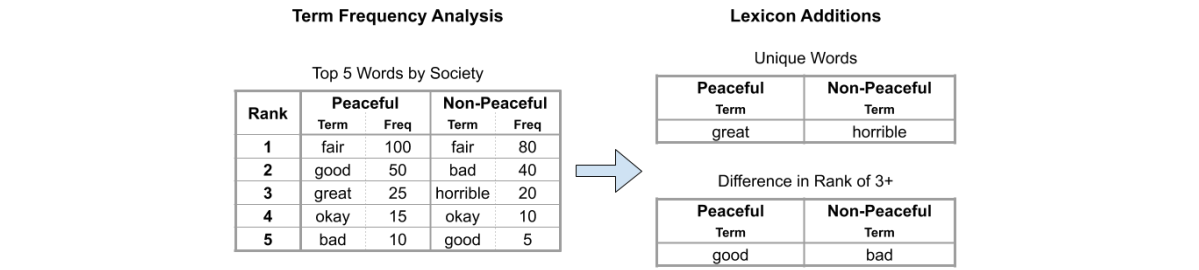


Fig 10-14 Example of creating new lexicons from word frequency rankings ($N = 5$, $D = 3$).

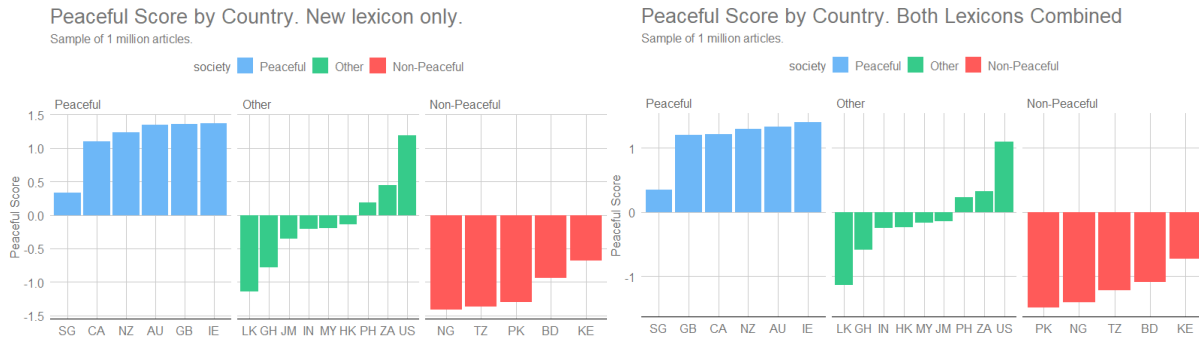


Fig 10-15 Comparison of peace metric when using our new lexicon alone and combined with the original lexicon.

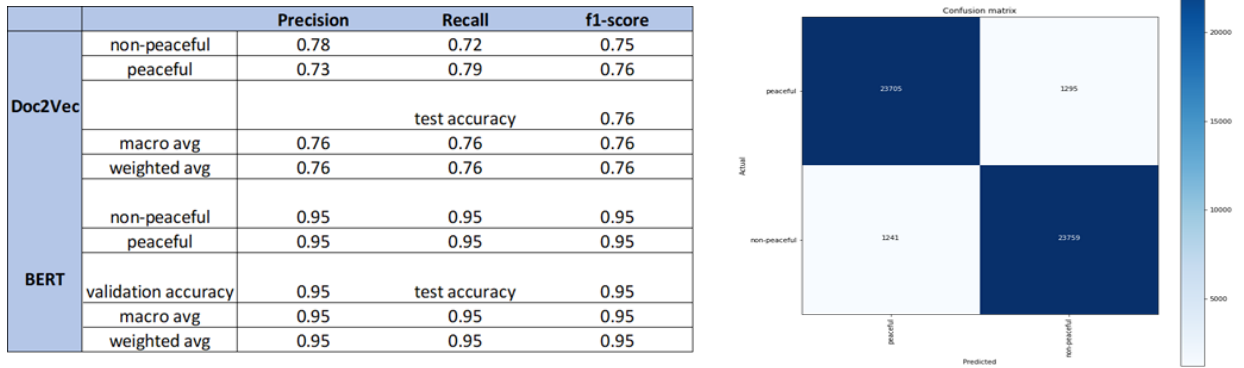


Fig 10-16 Result Table on the classification models and Confusion matrix from BERT classifier

Is the word present in this lexicon?			Freq
Original	Term Freq	Atten Layer	
No	No	No	0
No	No	Yes	453
No	Yes	No	390
No	Yes	Yes	6
Yes	No	No	1931
Yes	No	Yes	5
Yes	Yes	No	18
Yes	Yes	Yes	0

Fig 10-17 Comparisons of how many words appear in each lexicon.

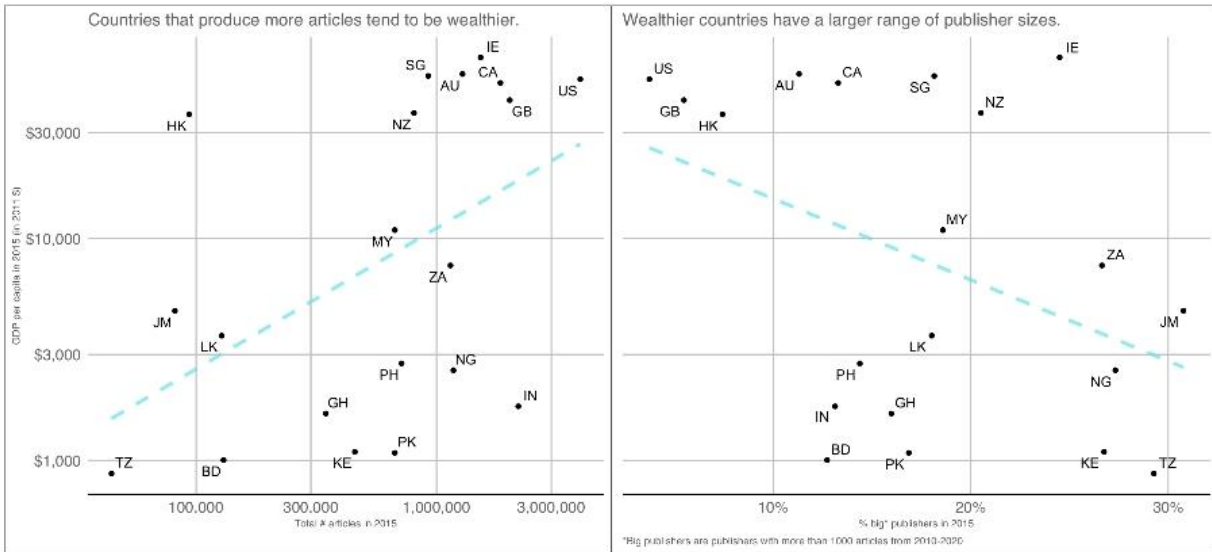


Fig 10-18 Number of articles of countries per GDP per capita