Columbia University

Peach Speech Analysis via NLP

December 10, 2020

Group Members:
Jung, Jinwoo
Kwon, Hyuk Joon
Lee, Hojin
Lim, Tae Yoon
Mackenzie, Matt

Advisor Groups:
Peter Thomas Coleman
Professor, Psychology and Education
Coleman@exchange.tc.Columbia.edu

Allegra Chen-carrel
Program Manager, The sustaining Peace Project
ac3922@columbia.edu

COLUMBIA UNIVERSITY

Outline

Relationship between peacefulness of countries and languages used in news articles

Hate speech is a very active area of research, however, what about Peace Speech? Some research suggests that peace speech is the DNA of peaceful societies. We wanted to deepen our understanding about this claims through cutting edge data science techniques. We will analyze articles from different countries and study the relationship between peacefulness of countries and languages used in articles.

Img source: https://webstockreview.net/images/peace-clipart-word-wisdom-5.png

COLUMBIA UNIVERSITY

Project Description

Preprocessing & Exploratory Data Analysis

- **Data Engineering**
- **Exploratory Data Analysis**
- **Preprocessing of text data**

Analysis on Initial Hypothesis

- **Testing initial hypothesis:**

  Is there a relationship between peacefulness of country and language used in the articles?

Alt. Hypothesis and Testing

- **Result of the initial hypothesis testing was not promising**
- **Came up with possible reasons for them and tested those ideas**

Conclusion and Solutions

- **Showing result of alternative hypothesis testing**
- **Conclusion of initial hypothesis**
- **Details of the conclusion and suggestions**

Data Overview

Data Source and Structure

News on the Web (NOW)

"11.2 billion words from web-based newspapers and magazines from 2010 to present times"

The data originally came divided into 2 types of files, joined together by an ID:

- Source files: contain metadata like publisher, website, country of origin, etc...
- Text files: contain the raw text for each news article.

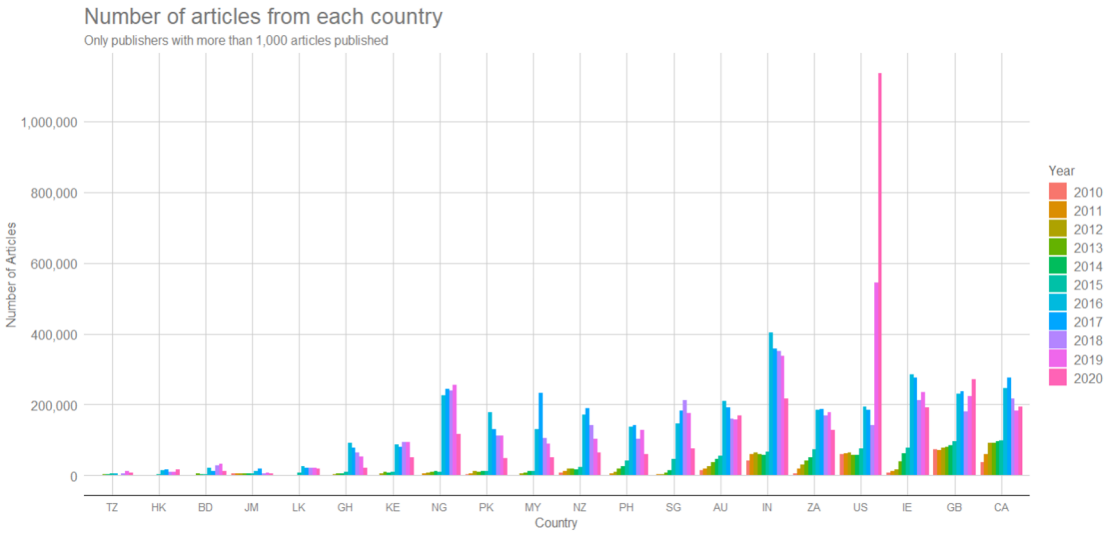
We focus our analysis on the following twenty countries

Peaceful	Non-Peaceful	Other
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)

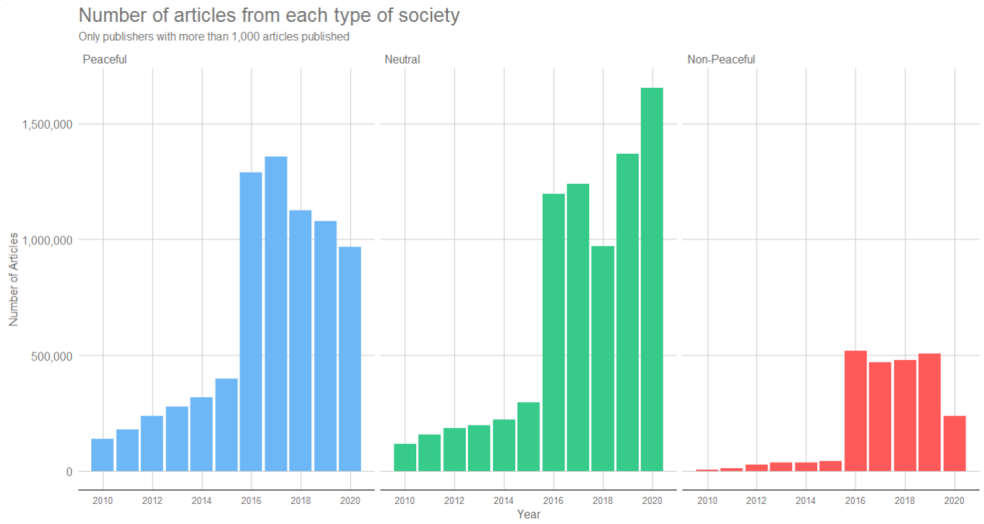
Data Overview

Original Data Size

There is a large imbalance of data between countries and year.



After 2015, substantially more articles are captured due to changes in data collection procedures.



Data Overview

Restructuring

Due to the large size of the dataset, we restructured the data to easily access any needed articles.

Original

49 source files, 129 text folders, and each text folder containing at least 20 text files (each country and NAs).

Restructured

Nested directories: **Country/Publisher/Year**; within each folder is one file for each article that belongs in that group.

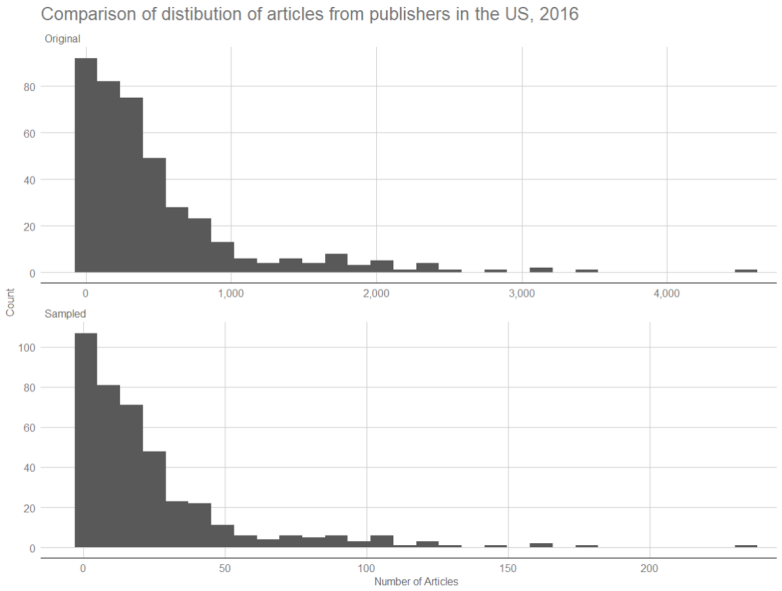
Resampling

We downsampled the number of articles we worked with to make our analysis more manageable.

Original Size: ~60GB, 20 million articles
New Size: ~5GB, 1.5 million articles

Procedure

Filter out articles from publishers that have 1000 or less total articles. Downsample years after 2015 so that each year has a similar number of articles. Within each downsampled year, maintain the distribution of articles from each publisher so their relative representations remains the same.



Preprocessing

Text Pre-processing

Need: unnecessary, noisy information that may affect later analysis

- phrases/sentence that are not related to the article's content
- ex. inducing readers to subscribe to their articles
- ex. suggestions of other articles

2 steps procedure:

1. General text pre-process
2. Model based text pre-process
   - N-Gram
   - Cosine Similarity Sentence Embedding

@@391241 <h> ? 80k cash boost to make netball centre of excellence in Gwynedd <h> ... <p> Invalid e-mail. Thanks for subscribing ! Could not subscribe , try again later <p> Netball in Gwynedd is set to get a cash boost <p> AN ? 80,000 National Lottery cash injection will help turn Bangor into a " hotspot " for netball . <p> … opportunity to spot and support gifted young players . <p> It is forecast that the Dome will allow the creation @ @ @ @ @ @ @ @ @ @ …"

Preprocessing

Text Pre-processing

1. **General text pre-process**
   - Applies to all specific models be used
   - Cleans the scraped news article into easily readable sentences
     - html tags such as <p> and <h>
     - symbols such as {, }, <, >, \, (,), \n, and @
     - convert symbols such as :, ;, ?, ! to periods
     - @ @ @ @ @ @ @ @ @
       - imposed by data provider to prevent violating copyright laws

sample article

@@391241 <h> ? 80k cash boost to make netball centre of excellence in Gwynedd <h> ... <p> Invalid e-mailThanks for subscribing ! Could not subscribe , try again later <p> Netball in Gwynedd is set to get a cash boost <p> AN ? 80,000 National Lottery cash injection will help turn Bangor into a " hotspot " for netball . <p> … opportunity to spot and support gifted young players . <p> It is forecast that the Dome will allow the creation @ @ @ @ @ @ @ @ @ …"

Preprocessing

2. Model based text pre-process

- **methodology used to automatically filters out the noisy information**

N-Gram

- hypothesis : similar noisy patterns exist per publisher
- systemic way to check for particular **publisher-specific** patterns
- measure frequencies of phrases across 5-gram phrases
- remove sentences with particular phrases of > 25% per publisher

Cosine Similarity Sentence Embedding

- tokenize each sentence in a document using Sentence-Bert and HuggingFace
- compute cosine similarity between each sentence and document
- remove sentence found to have low level of similarity (used 0.95)

Preprocessing

Pros / Cons of N-Gram and Cosine Similarity

	Pros	Cons
N-gram (5-gram)	- Safer, which sentence to remove (clean document -> not remove any) - Faster to run (~ 11 sec per 250 articles)	- Fails for recurring phrases of less than 5 words - Processing time increases exponentially (process per doc, depend on # of articles per publisher)
Cosine Similarity	- Able to delete phrases with less than 5 words - Time complexity: linear in number of article	- Remove sentences that are not spam - Unable to control (pre-trained, vectorize) - Slower to run (~ 6 mins per 250 articles)

Preprocessing

Top examples of the removed sentences from 5-Gram [Sample Data]

Publisher (num of 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	
	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

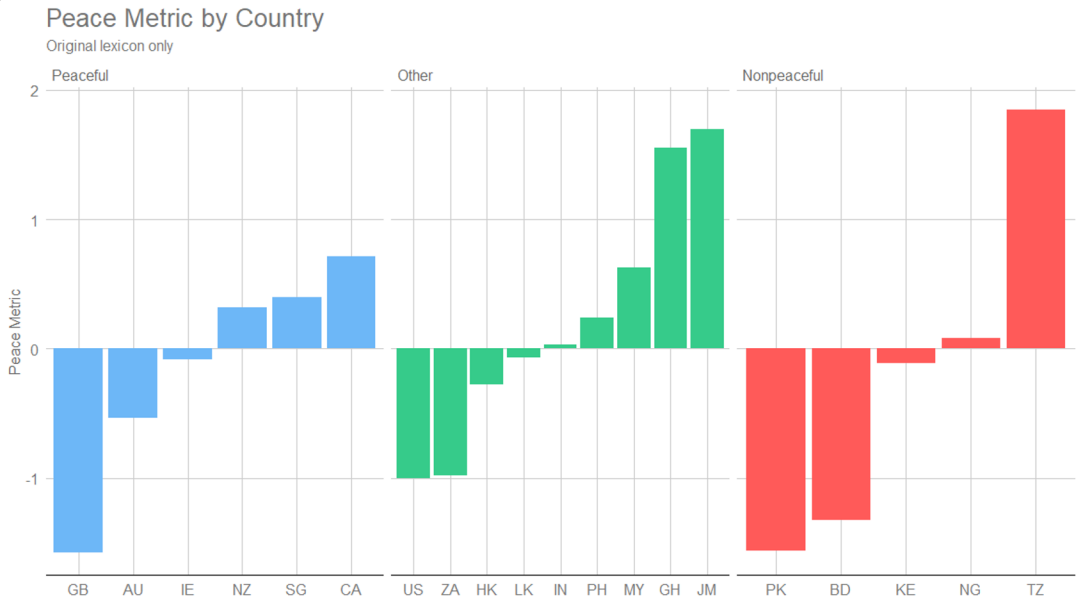
Lexicon Evaluation

Peace Metric

In order to test the validity of the lexicons, the project managers originally developed a metric for comparing the "peacefulness" of different countries. We used this metric as a basis for assessing if the lexicons are working properly at differentiating between peaceful, non-peaceful, and other countries.

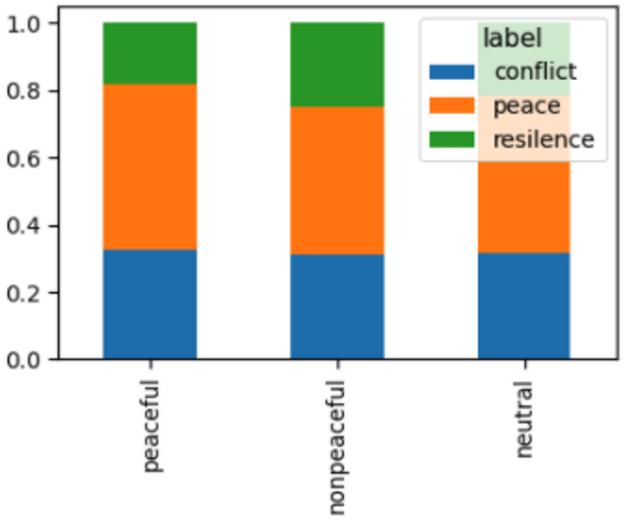
Raw Peace Metric = (% of peaceful terms) - (% of conflict terms)

Once calculated for each country, we can obtain the peace metric by normalizing all the raw peace metric scores to have mean 0 and variance 0 across all the scores.



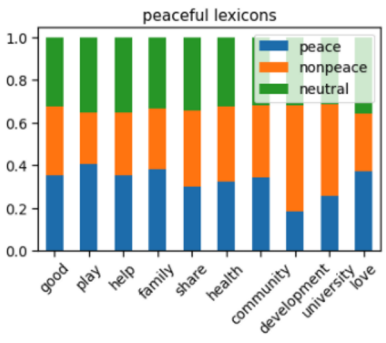
The scores do not work completely as desired.

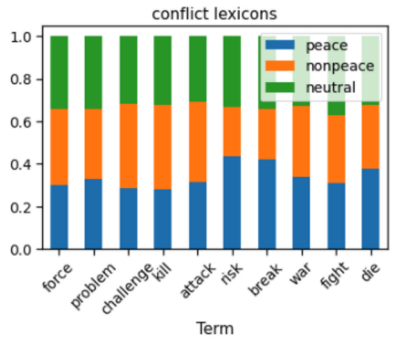
Word Frequency Analysis

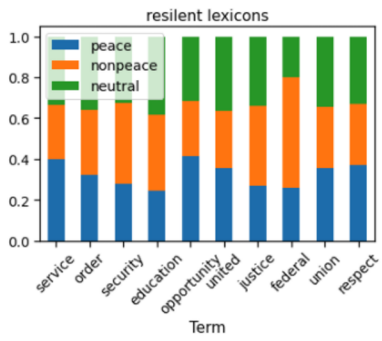


Comparison of frequency of vocab. used in each categories for each peace group

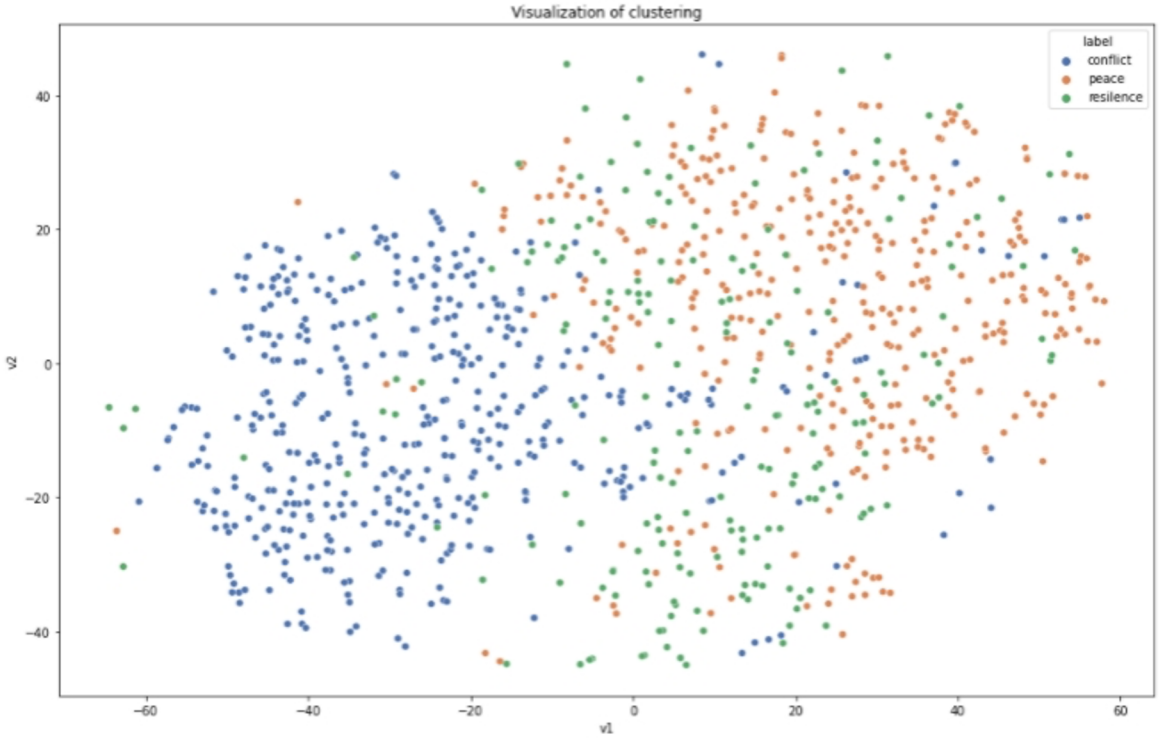
- Initial hypothesis: there should be a correlation between the two
- Truth is that it is hard to tell if there exists any significant differences
- Performed ANOVA and p-value suggested that there is no statistically significant differences among groups







Word2Vec



- Converted pre-determined vocab. to vectors by training word2vec model using given articles
- Implemented T-SNE for dimension reduction to 2D
- Somewhat able to distinguish vocab. by categories with limitation

Potential Reasons for Failure of Initial Approach

International articles affect the result of the analysis

The unclear distinction among different categories of lexicons could be due to a mix of domestic and international articles

- **Implement Domestic Filter using NER (Named Entity Recognition) method**
- **Run same word frequency and word2vec analysis after implementing the filter**

Flaws in predefined set of lexicons

Predefined set of lexicons were derived from dictionary and non-data scientific way

- **Perform word count analysis to come up with new set of lexicons**
- **Validate new analysis through running peace score metrics**

Classification by peace level for countries is flawed

There is a possibility that classification of peace level per country might be flawed.

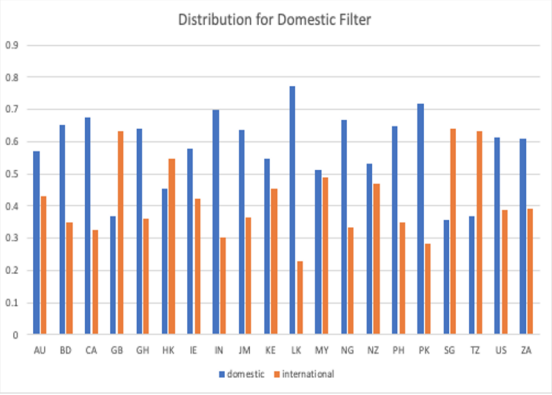
- **Implement text classification model and clustering models to compare articles and countries**
- **Extract important features out of models and reflect them on final conclusion**

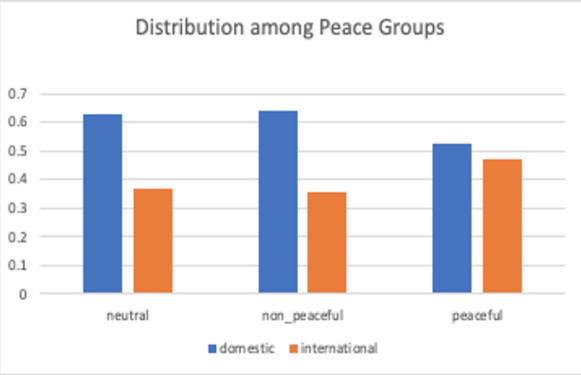
Domestic Filter

General Methodology:

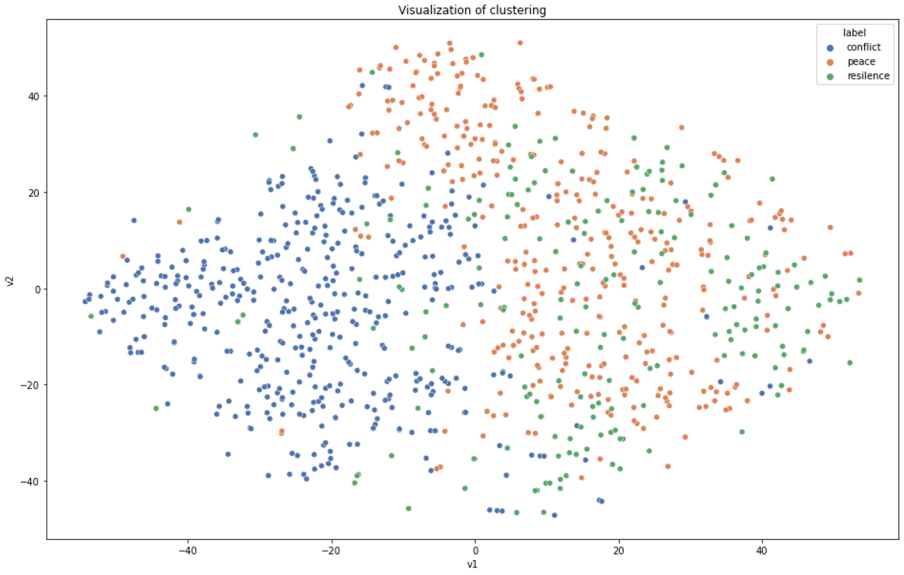
- Performed Named Entity Recognition (NER) analysis to extract the name of places that are mentioned in the article.
  - If the particular country where the article is from is mentioned at least once, then classified as Domestic article
  - If none of the places were recognized, classify as Domestic article

- Result of the Filter:
  - Countries that are classified as peaceful had slightly higher proportion of International Article than other groups



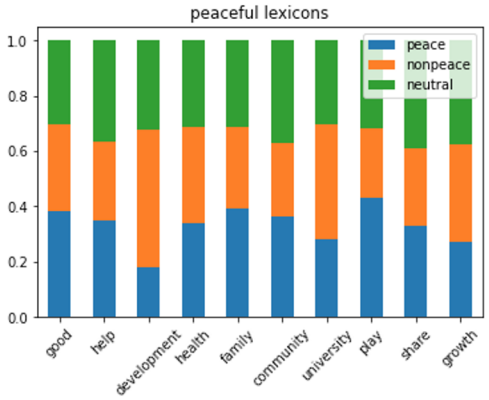


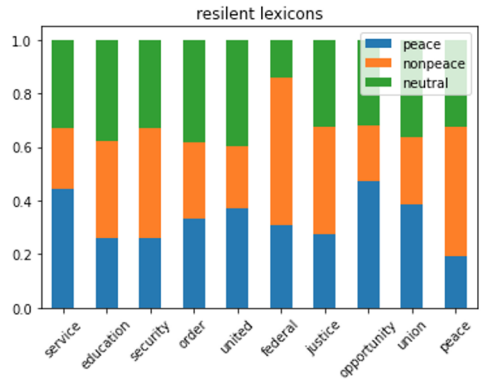
Domestic Filter - Result

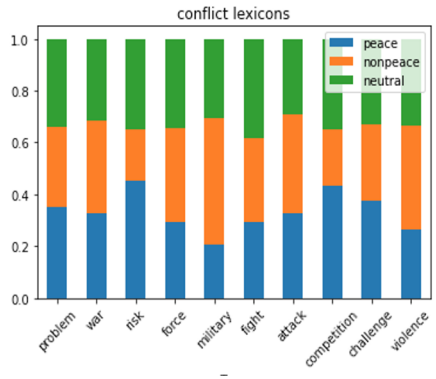


Performed Word2Vec & Word Frequency Analysis

- Result did not change as much as we expected
- Concluded that the international articles do not have much influence on analysis





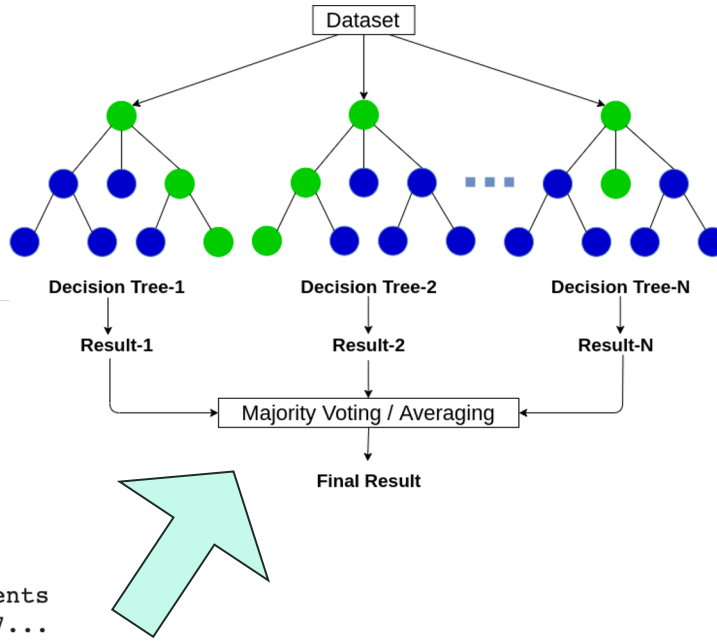


Classification Models - Doc2Vec

Methodology:

- Doc2Vec
  - With Gensim's Doc2Vec model, convert each document into a vector
  - Use Random Forest for the classification

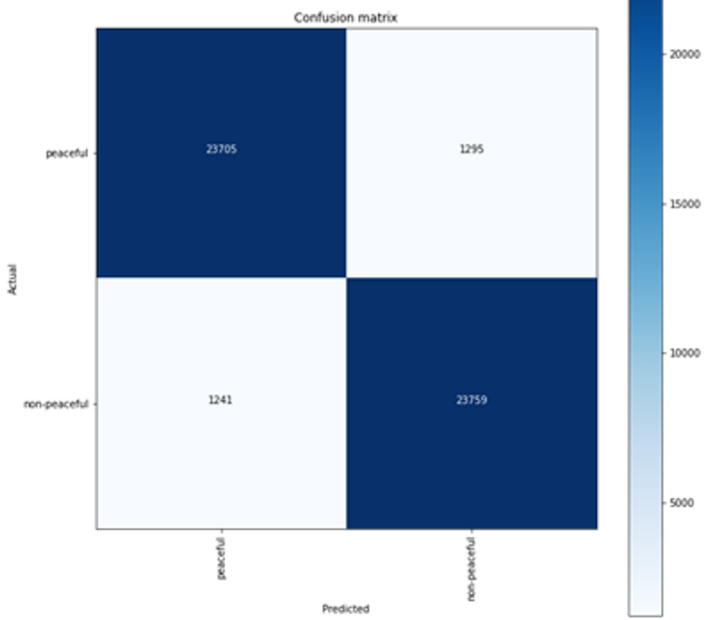
```
   peacefulness                                               text
0  non-peaceful  milestone achieve set world track future econo...
1  non-peaceful  lagos state commissioner police mr umaru manko...
2  non-peaceful  young mahin create history pakistan pakistan p...
3  non-peaceful  otieno otieno diminutive striker great night n...
4  non-peaceful  femi fani kayode way eagle lion king permit be...
(100000, 2)
```




```
   peacefulness  ...                                vectorized_comments
0  non-peaceful  ...  [-1.9986039, 0.9381497, 1.1435105, -0.06345977...
1  non-peaceful  ...  [-4.4682717, -0.8725656, -1.9141116, 0.5902374...
```

Classification Models - Doc2Vec - Results

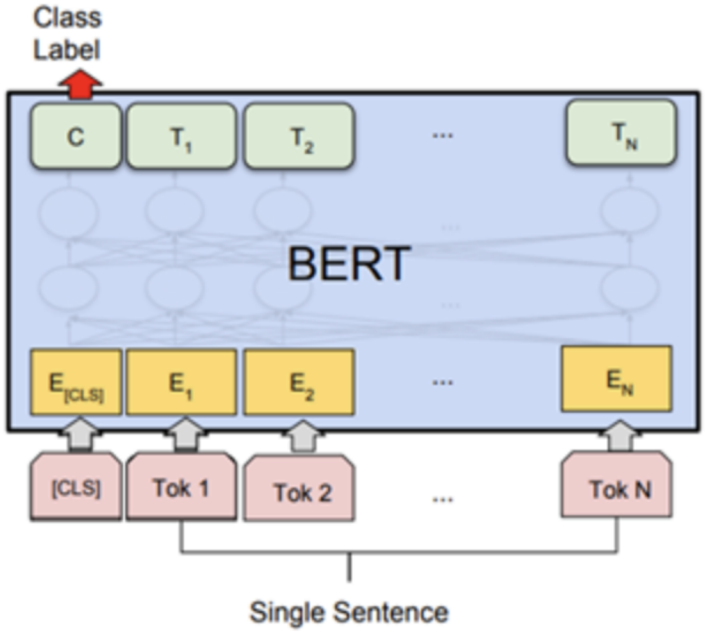
	precision	recall	fl-score	support
non-peaceful	0.78	0.72	0.75	1021
peaceful	0.73	0.79	0.76	979
accuracy			0.76	2000
macro avg	0.76	0.76	0.76	2000
weighted avg	0.76	0.76	0.76	2000



Classification Models - BERT

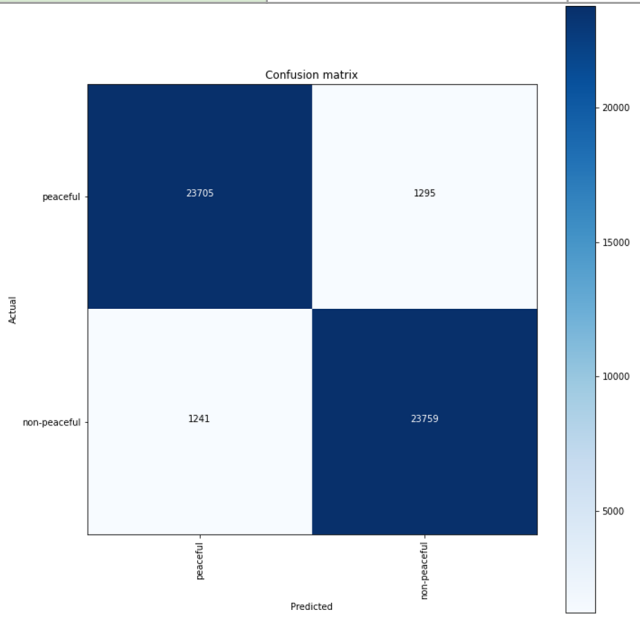
Methodology:

- BERT classifier
  - Import Pre-trained BERT model (bert-base-uncased)
  - Add a Fully connected linear layer for the classification, only takes the first token from the BERT model
  - Fine-Tune (train) the model with 100,000 randomly sampled data
  - Test the model with randomly sampled 50,000 data



Classification Models - BERT - Results

	precision	recall	fl-score	support
non-peaceful	0.95	0.95	0.95	25000
peaceful	0.95	0.95	0.95	25000
accuracy			0.95	50000
macro avg	0.95	0.95	0.95	50000
weighted avg	0.95	0.95	0.95	50000



New Lexicon

New Lexicon from Term Frequency Analysis

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 Words

Select the top **N** words by frequency in each society.

Select the unique words in each society and assign them to that lexicon.

For 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.

Example with N = 5, D = 3 (not real data)

Term Frequency Analysis

Top 5 Words by Society

Rank	Peaceful		Non-Peaceful	
	Term	Freq	Term	Freq
1	fair	100	fair	80
2	good	50	bad	40
3	great	25	horrible	20
4	okay	15	okay	10
5	bad	10	good	5

Lexicon Additions

Unique Words

Peaceful Term	Non-Peaceful Term
great	horrible

Difference in Rank of 3+

Peaceful Term	Non-Peaceful Term
good	bad

New Lexicon

New Lexicon from Term Frequency Analysis

WordNet

WordNet is a lexical database that provides groupings for nouns, verbs, adjectives, and adverbs into cognitive synonym sets (synsets). For a BoW approach, with no context we can infer the part of speech (PoS) from a word's synset.

For example, consider the synsets for the following terms.

Term from BoW	Synset - "word" (PoS)	PoS Implication
"educate"	"educate" (verb) and "train" (verb)	Verb
"Nigeria"	"nigeria" (noun)	Noun
"research"	"research" (noun), "inquiry" (noun), and "research" (verb)	Noun or Verb
"aaaa"	Empty	Nonsense

Final Selection Process

Using WordNet, we use the selection method explained on the previous slide (N = 250, D = 30) on the following sets of words:

- Non-nouns (filter out words where the synsets only contain nouns)
- Verbs only (filter for words where the synsets only contain verbs)

We perform a union of the results to obtain a new lexicon of 414 words.

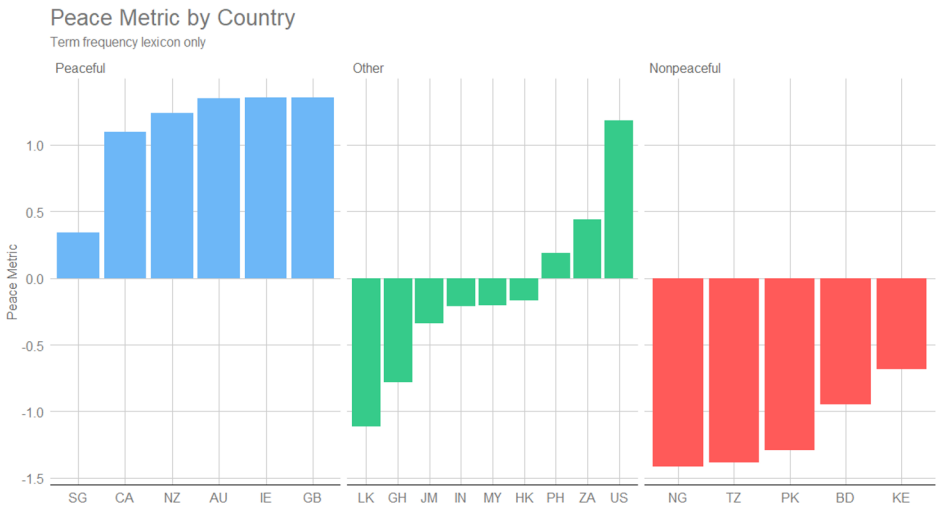
New Lexicon - Result

New Lexicon from Term Frequency Analysis

To the right are the top 10 most frequent terms from the lexicon generated from the term frequency analysis

Rank	Peace Lexicon	Conflict Lexicon
	Term	Term
1	look	state
2	think	minister
3	play	court
4	home	accord
5	open	project
6	really	bank
7	keep	order
8	season	force
9	mean	fund
10	offer	political

We can see that using our new lexicon with the original gives a peace metric that is clearly divisive among peaceful and non-peaceful countries while being mostly around zero for other countries.



New Lexicon - Result

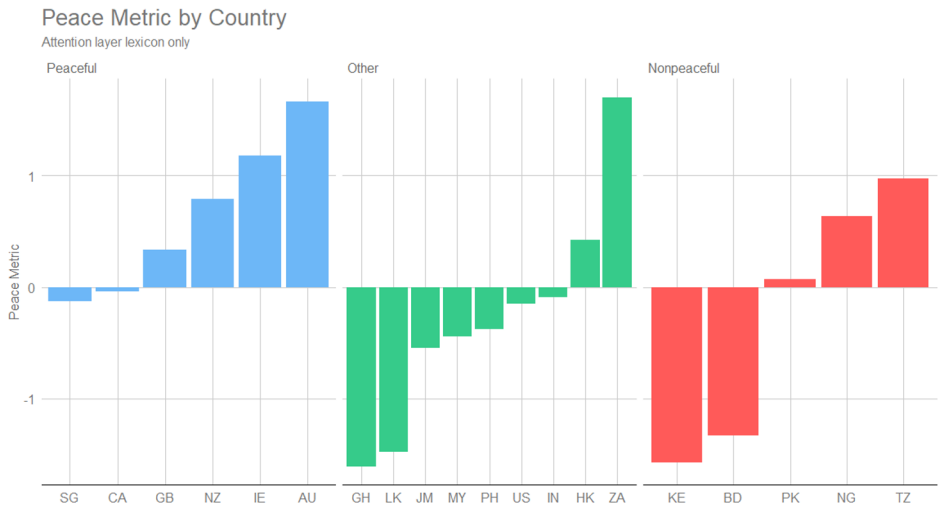
New Lexicon with RNN & Attention Weight

Build RNN with Attention Layer for classification

Retrieve Attention weight, match weights with vocab to extract lexicons which has highest average weight

Rank	Peace Lexicon	Conflict Lexicon
	Term	Term
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

Since Attention Layer corresponds to the position of vocab, it does not reveal a great performance on the peace metrics.



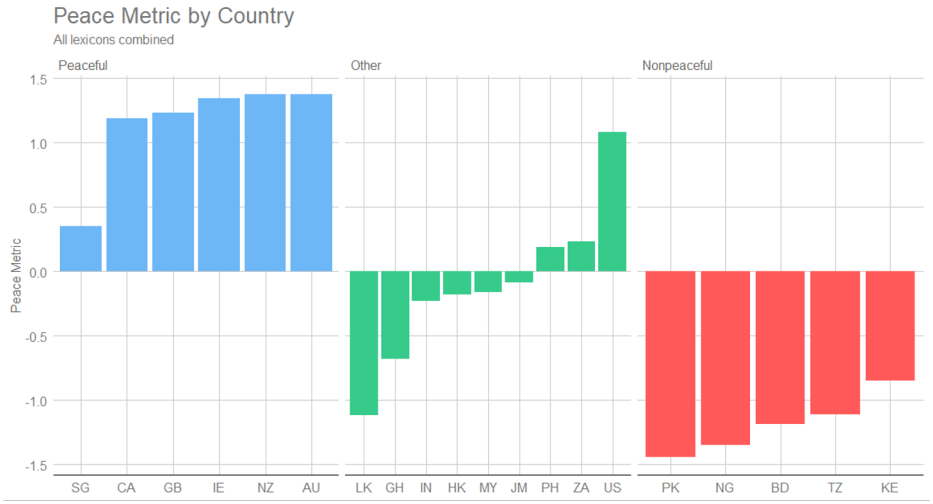
Conclusion

Augmented Lexicon

To the right are the top 10 most frequent terms from each lexicon when we combine the three lexicon versions.

Rank	Peace Lexicon		Conflict Lexicon	
	Term	Version	Term	Version
1	give	Original	state	Term Freq
2	good	Original	president	Atten Layer
3	look	Term Freq	minister	Term Freq
4	think	Term Freq	court	Term Freq
5	play	Original	accord	Term Freq
6	help	Original	project	Term Freq
7	home	Term Freq	south	Atten Layer
8	family	Original	bank	Term Freq
9	share	Original	university	Atten Layer
10	health	Original	order	Term Freq

Using all three lexicons, we get the best separation yet!



Conclusions

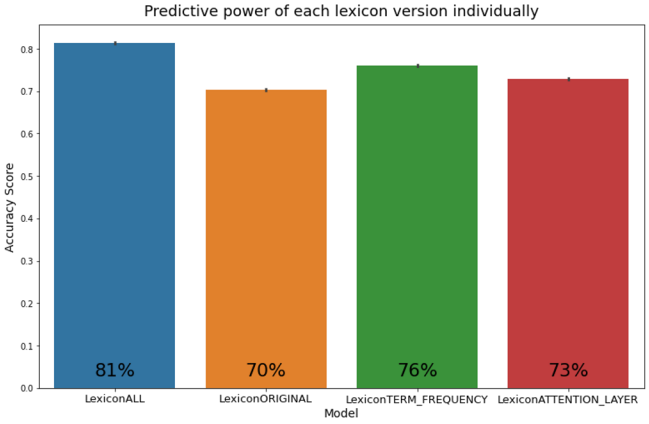
Logistic Regression Models

For each lexicon and all the lexicons combined, we train a logistic regression model to predict peaceful vs non peaceful countries using just the frequencies of the lexicon words.

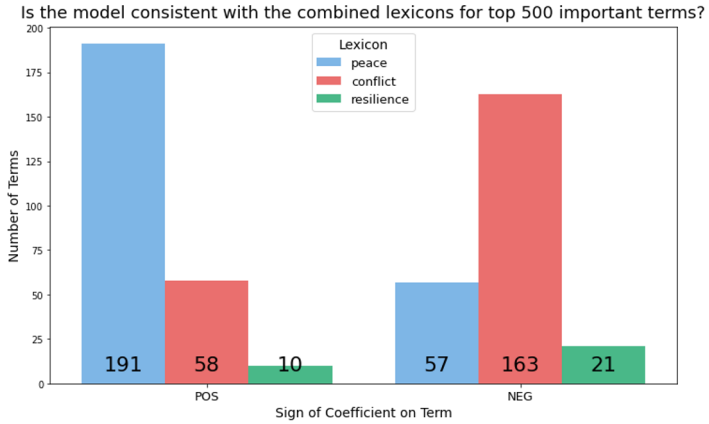
How much overlap is there between lexicons?

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

On its own, each lexicon performs worse than all the lexicons combined.



Peace lexicon terms tend to drive the probability of being a peaceful nation up.



Relationship between peacefulness of countries and languages used in news articles

sustainingpeaceproject.com

Img source: https://webstockreview.net/images/peace-clipart-word-wisdom-5.png

COLUMBIA UNIVERSITY

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. 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

10. 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