DSI Capstone Fall 2022 FINAL REPORT

Peace Speech - Revealing Differences in Speech Between High-Peace and Low-Peace Countries

Yibo Chen (yc3837), Hongou Liu (hl3518), Ziheng Ru (zr2206),
Xinfu Su (xs2444), Pinyi Yang (py2290), Yuwen Zhang (yz4157)

Abstract

We are interested in finding the speech differences between high and low-peace countries. We use Twitter and news data for the project. For data processing, we lowercase the words, remove links, country names, line breaks, email addresses, punctuations, and numbers, convert contractions, and apply lemmatization.

We remove country-specific words using TF-IDF, logistic regression, or intersection. We apply logistic regression, XGBoost, and transformers (BERT and XLNet) to the datasets and extract high-peace/low-peace features using coefficients or attention scores. Then we apply KMeans to cluster the features and generalize cluster topics.

With the result, we conclude that 1. High and low-peace countries use different sets of words focusing on different topics 2. High-peace countries talk about more topics than low-peace countries 3. High-peace countries discuss topics in daily life more, such as food, films, and feelings, whereas low-peace countries focus more on political subjects.

Introduction

To date, no technique has been developed to help us understand, measure and track the power of peace speech – peaceable language for building and maintaining more robust, peaceful communities. We can learn much from existing peaceful societies. While research has shown that highly peaceful societies are significantly more stable and have the lowest probability of relapsing into violence, very little is understood about why and how they remain peaceful – as they are rarely studied.

Therefore, the Peace Speech project aims to fill in the blanks in understanding peacefulness. We are interested in finding and proving the power of peace speech. We want to reveal the differences in speech between high-peace and low-peace countries, as peace speech is the DNA of peaceful societies. We have expanded the work from the previous teams, who worked on the LexisNexis dataset, on two more robust datasets and several different metrics with more diverse

approaches. We believe that peace speech is a foundation that helps to construct and maintain peacefulness. It may seem simplistic, but variations in the quality and message of these simple human encounters – which occur millions or billions of times daily in communities – build up to form the norms, taboos, memories and expectations that shape our lives.

Country Selection

To start with the project, we needed to select our set of high-peace and non-peaceful countries. We measured country's peacefulness using five indices, which are the Global Peace Index (GPI), the Positive Peace Index (PPI), the World Happiness Index (WHI), the Human Development Index (HDI), and the Fragile States Index (FSI). First, we took inspiration from the method used in the article *Natural Language Processing and Machine Learning Identified Words in Media Most Associated with Lower and Higher-Peace Countries*[1], and randomly chose 18 countries: Nigeria, Iran, Malaysia, Jamaica, Sri Lanka, Canada, Ireland, U.K., Finland, Norway, Thailand, Greece, France, Uganda, India, Poland, Peru, and Mexico. Then we scaled the data of the 18 columns from 0-100 and divided them into three groups: the highest group marked as green, the medium country marked as yellow, and the lowest marked as red. For any country whose three or more indices are in the highest third and colored as green, we categorized the country as high peace, and if over three of the five indices are in the lower third, we consider the country as low peace (Table 1).

Country	GPI	PPI	WHI	FSI	HDI	
Nigeria	0	0.35	19.16	0	2.29	low
Iran	2.64	10.75	27.47	15.96	57.11	low
Malaysia	87.27	48.63	47.82	49.7	63.76	medium
Jamaica	51.15	39.12	51.26	42.75	42.2	medium
Sri Lanka	49.06	20.22	14.47	21.8	58.94	low
Canada	92.97	95.83	80.32	93.91	94.27	high
Ireland	100	93.24	80.71	93.06	96.33	high
Uk	73.63	83.19	78.29	68.94	92.66	high
Finland	89.49	100	100	100	95.18	high
Norway	88.03	99.38	88.72	99.39	100	high
Thailand	43.63	35.1	52.27	33.13	63.07	medium
Greece	61.73	62.27	53.68	50.43	83.03	medium
France	57.9	87.79	71.96	80.76	86.7	high
Uganda	28.95	0	20.43	6.21	0	low
India	10.23	24.31	0	26.67	24.77	low
poland	81.63	65.87	58.01	66.99	80.5	medium
Peru	44.12	32.51	44.07	33.37	54.36	medium
mexico	7.86	27.91	58.14	32.76	53.44	low

Table 1

Through the technique described above, we found that Canada, Finland, Norway, the UK, Ireland, and France are high-peace countries, while India, Iran, Sri Lanka, Uganda, Nigeria, and

[1] Manuscript_NLP_Peace_2022_08_28b (1).pdf

Mexico are classified as low-peace. However, in the process, we also realized that how we choose the group significantly influences the identification. For example, if we choose all 18 developed countries with similar values for all five indices, then some countries will be defined as low peace even though they are not low peace in our common sense. Thus, we wanted to improve the classification by defining our own peace index based on the five indexes collected from over 137 countries.

To deal with this unsupervised classification job, we naturally used the principal component analysis (PCA) technique, which uses projection and ensures the maximization of variation:

$$w_k = \underset{||w||=1}{argmax} X - \sum_{s=1}^{k-1} X w_s w_s^T$$

Here, we set k=1. It turns out that the most significant eigenvalue is 24 and explains 82.8% of the variance, which is rather satisfying. Then we needed to scale the Index from 0 to 100. The linear combination looks like this:

$$Peace\ Index = Rescaling(0.441 * WHI + 0.444 * HDI + 0.479 * FSI + 0.469 * PPI + 0.398 * GPI)$$

The equation shows that the coefficients of different indexes are pretty close, which means that they contribute to our peace index equally significantly. We thus divided countries into High Peace, Medium Peace, and Low Peace Groups by finding the 33 and 66 percentile (Table 2).

High Peace Country	Low Peace Country
Ireland	Philippines
Denmark	Nicaragua
Finland	Jordan
...	...
Moldova	Democratic Republic of the Congo
Jamaica	Afghanistan

Table 2

From the table, we decided to select eight countries from the High Peace group and eight from the Low Peace group.

High-Peace Countries	Low-Peace Countries
Canada, UK, Finland, Norway, Ireland, France, Australia, Singapore	India, Iran, Nigeria, Uganda, Gambia, Libya, Pakistan, Zimbabwe

Table 3

Data Description

The limitation of the LexisNexis dataset that the previous team has worked on is that it contains primarily financial-related information. Thus, it might not be representative enough of the speech pattern of each country. The other issue of the LexisNexis dataset is that the classification of different countries of each article is frequently inaccurate, and the number of articles from each country is imbalanced. Therefore, our team decided to build up other datasets for the analysis. We constructed two datasets, one Twitter dataset of Tweets extracted using the Twitter API and one news dataset composed of local news scrapped from popular sources in each country. We used two sources because we want to compare if different sources of texts reveal similar information about peacefulness.

I. Twitter

We choose Twitter as one of our sources because it incorporates conversations of ordinary people and can reveal the differences of speech patterns among regular social media users from high-peace/low-peace countries.

To extract Tweets from Twitter, we applied for Academic Access to the Twitter API. We built up a sample dataset of 800k Tweets, consisting of around 50k Tweets from each country from 2017 until now.

When posting Tweets, the user can tag their current location. To classify the country of origin in each Tweet, we utilized the place_country query key, which filters Tweets where the country code associated with a tagged location matches the given ISO alpha-2 character code. We also leveraged the lang query key, which can specify the language of the Tweets we want to extract, in our case English. We also ensured that the Tweets we compiled were not retweets, replies, quotes, or Twitter Ads.

II. News

We decided to use news articles as another source for the project because, in contrast to Twitter, which captures speech from ordinary people, news articles are more formal and discuss more focal issues of the target countries. To find the corpus that could better reflect the peacefulness level, we chose each country's local news instead of foreign news. The reason is that, though foreign media may still mention the target countries in their news, we thought words used in that news could not correctly indicate the level of the peacefulness of target countries, unlike local news. Thus, we searched 2-3 local news sources of each country and used official archives of their websites, if available, as our data sources.

We leveraged web scraping to automatically fetch and organize news data from websites of selected sources. We only picked news published between Jan 1st, 2011 to Sep 30th, 2022, to maintain similar and comparable topics among news, thus reducing irrelevant features. After scraping data within the same period from news websites of those ten countries, we eventually got 600k articles, roughly 300k from both high-peace and low-peace countries (Table 4).

Type	Country	Newspaper	News Source URL	Num Articles	Total Articles
High Peace	Canada	The Star	https://www.thestar.com/archive.html	60k+	~300k
	United Kingdom	The Independent	https://lil.nlp.cornell.edu/newsroom/index.html	50k+	
	Norway	News in English	https://www.newsinenglish.no/2022/10/07/	16k	
	Ireland	Sunday World	https://www.sundayworld.com/archive/cnt	18k	
	Finland	Daily Finland	https://www.dailyfinland.fi/archive	29k	
	Singapore	The Straits Times	https://huggingface.co/datasets/teven/pseudo_crawl_en_seeds	50k+	
	Australia	9news	https://lil.nlp.cornell.edu/newsroom/index.html	24k	
	France	France24	https://www.france24.com/en/archives/2021/	60k+	
Low Peace	India	Times of India	https://timesofindia.indiatimes.com/archive.cms	50k	~300k
	Uganda	The Independent	https://www.independent.co.ug/all-news/	28k	
	Iran	Tehran Times	https://www.tehrantimes.com/archive	54k	
	Nigeria	The Nigerian Voice	https://www.thenigerianvoice.com/archive/	32k	
	Zimbabwe	Bulawayo24	https://bulawayo24.com/index-id-archive.html	50k+	
	Gambia	The Point	https://thepoint.gm/africa/gambia/article?start=0	43k	
	Libya	The Libya Observer	https://www.libyaobserver.ly/archive/	24k	
	Pakistan	Dawn	https://www.dawn.com/archive/2022-10-17	28k	

Table 4

Preprocessing

I. Twitter

We started with cleaning the data. Since there is no format restriction to posting on Twitter, the Tweets come in various formats. Therefore, data cleaning is essential. Some steps we took to

clean the data were removing web links in each Tweet, expanding contractions (ex., can't to cannot, I'm to I am), converting emojis to text, and lemmatizing the words using the NLTK package. The other step we took was identifying and removing country names in each Tweet. This step is a bit of a hassle since Twitter users spell country names in various patterns. To accomplish this goal, we found a package in Python called Country Converter[2], which can convert and match different country names in text. We downloaded and edited the source code of the package to meet our project requirements.

After cleaning, Tweets like:

> Congratulations
> Afghanistan 🇦🇫 beat Indonesia 🇮🇩 in a Friendlies International Football 1-0
> 🥅 by Omaid Popalzai ⚽
> #Afghanistan
> #football https://t.co/gozbJek8eT

Will be converted to:

> congratulation beat friendly international football 1-0 goal net omaid popalzai soccer ball football

The pipeline will also record Afghanistan and Indonesia as mentioned countries for this Tweet

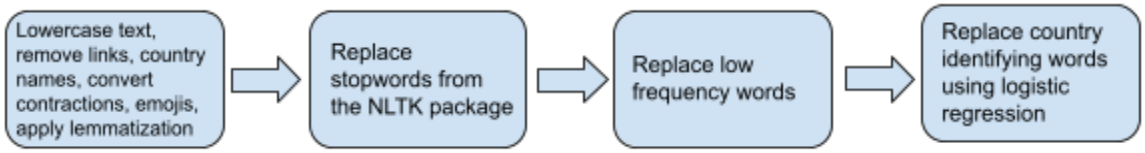
After cleaning, we began to compose a set of stopwords. We started with the standard stopwords from the NLTK library. Then we add words that appear less than four times among all corpus into our collection of stopwords. After removing low-frequency words, we vectorized the text using TF-IDF and applied logistic regression to predict the country based on the given text vector. Then we extract the top 2.5% features that best predict each country (features with the highest coefficients). Our goal is to identify words that uniquely describe a country but do not indicate peacefulness. For example, words like kangaroo will most likely only appear in the text corpus of Australia. Since Australia is a high-peace country, the word kangaroo will likely be categorized as a high-peace indicator. We want to filter out words like kangaroo because we must keep only those that describe peacefulness, not countries. Instead of removing the stopwords, we replace them with general terms based on their part of speech tagging (Table 5). We do that because Tweets are short texts, no longer than 280 characters; simply removing the stopwords may hurt the sentence structure of Tweets too much. Therefore, we replace the stopwords with general terms so that the Tweets will retain more context.

POS Tag	Verb	Noun	Pron	Adj	Adv	Adp	Conj	Det	Num	Prt	X	.
Replacement	DO	TERM	IT	NEUTRAL	NEUTRALLY	IN	AND	A	ZERO	THEM	X	.

Table 5

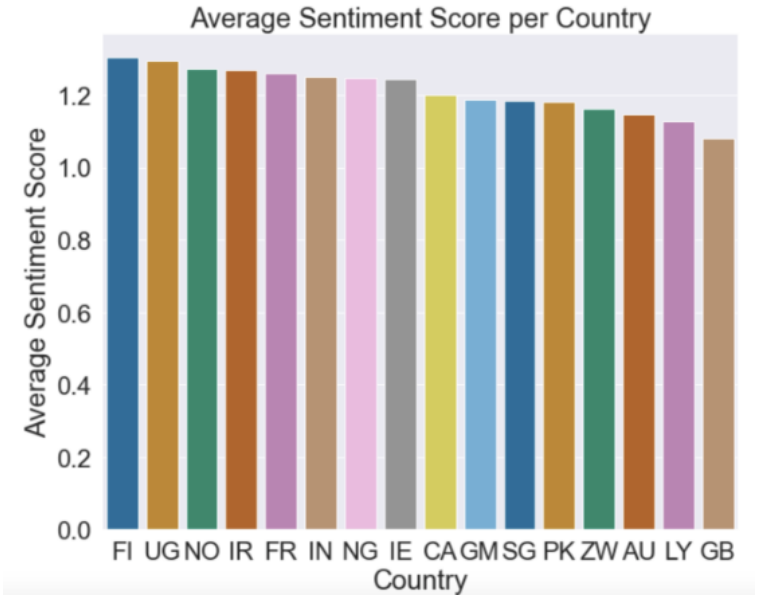
[2] https://github.com/konstantinstadler/country_converter

As a result, if the term *obnubilate* is in the stopwords set and *obnubilate* is a verb. Then the sentence *If a cloud passes in front of the sun, it will obnubilate the ground beneath it* will be converted to *If a cloud passes in front of the sun, it will DO the ground beneath it.*



A. Sentiment Analysis

We initially explored the sentiment scores of each Tweet and the average sentiment score per country due to the concern of comparably more intense sentiment expressions that may appear in social media. After exploratory analysis, we found that all included countries have an average sentiment score above neutral (Graph 1), and using sentiment score alone is not a good predictor of peacefulness. These results suggest that regardless of a country's status in high-peace or low-peace, the expressions of sentiment are, to a certain degree, homogeneous.

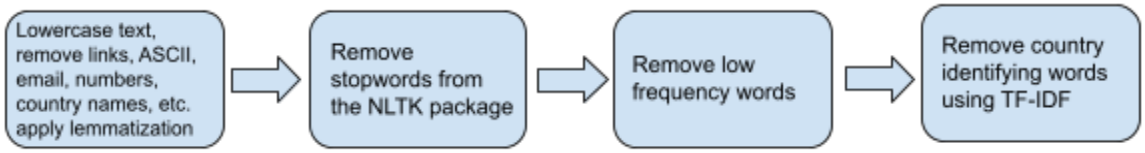


Graph 1

II. News

Like what we did to the Twitter dataset, we applied a similar approach to clean the news data. We first removed ASCII, punctuations, links, emails, numbers, country names, and other unique patterns. We converted contractions, lowercased the text, and applied lemmatization.

Since news articles' language is much more formal and standardized, we chose not to use logistic regression to remove country-specific words. Instead, we decided only to use TF-IDF for the process. We first concatenated all articles from one country into a single document to get one document for each country. Then, we ran the TF-IDF algorithm over the documents and ranked words within each document (country) by the TF-IDF value. Then we selected the top words of the TF-IDF matrix as the stopwords we wanted to filter. This method works because the high TF-IDF value of one word in one document means that word frequently occurs in that document but hardly exists in the others. This process will help us filter out the country-identifying words as expected. Unlike Tweets, news articles are much longer, so simply removing the stopwords, instead of replacing them with general terms, will not hurt the context of the articles too much. Therefore, we removed the stopwords for the text.



Method

I. Models

A. Twitter

We applied several models to the Twitter data. We first applied TF-IDF transformation to the dataset and performed a simple logistic regression as our baseline model. After that, we moved forward to more advanced models, BERT. The two models that we decided to use are BertForSequenceClassification and BERTweet. BertForSequenceClassification[3] is a BERT transformer model with a general sequence classification/regression head on top. BERTweet[4] is a pre-trained language model for English Tweets. We decided to use two different BERT models because we want to compare their performance and see if there is any difference in features picked up by the two models. We tokenized the Tweets using BertTokenizer and performed an 80:20 train-test split to the Twitter data.

---

[3] https://huggingface.co/docs/transformers/v4.25.1/en/model_doc/bert#transformers.BertForSequenceClassification

[4] https://huggingface.co/docs/transformers/model_doc/bertweet

B. News

For our baseline model, we applied a leave-one-out approach to make the model more robust. We ignored news articles from one specific country and rebalanced our dataset (high-peace v.s. low-peace). We repeated this process 16 times. We then collected words in high-frequency from high-peace and low-peace countries separately and took the intersection of the words. This process helped us filter out extremely unique or unfamiliar words in one country versus another. Afterwards, we used TF-IDF to encode the articles with the words we got from the second step. We applied logistic regression to the encoded dataset.

We also utilized the XGBoost classifier to classify the peacefulness of a news article based on its TF-IDF matrix. Because the news media of different countries may focus on different topics based on periods and regions they are in, the corpus extracted from the news is also time-sensitive and region-sensitive. To avoid overfitting and ensure the robustness of the XGBoost model, we applied 5-fold cross-validation.

For more advanced models, we used tokenized the dataset using BertTokenizer and trained a DistilBERT[5] model, a small, fast, cheap and light Transformer model trained by distilling the BERT base.

We also applied the XLNet transformer model to the dataset. The XLNet transformer has similar functionality as BERT but is Auto-Regressive based, which is different from the Autoencoding based scheme of BERT. We used this model in news data because XLNet is more apt at longer textual lengths[6]. Specifically, we used XLNetTokenizer to tokenize the text and XLNetForSequenceClassification[7] to train the data. Like the Twitter dataset, we performed an 80:20 train-test split to the news data.

II. Feature Selection

We used different approaches to extract the top features from each model. For logistic regression, we simile utilized the model coefficients. For the XGBoost model, we took out keywords from each model trained during the 5-fold cross-validation and got 2500 words each for high-peace and low-peace corpora. After removing duplicated and location-related words, the high-peace corpus has 1704 words, and the low-peace corpus has 1182 words. We utilized the GloVe embedding (mentioned below) to convert words to vectors and applied the support vector machine (SVM) to classify words in high-peace and low-peace corpora. Words with longer distances to the decision boundary of SVM are classified as intending to demonstrate higher or

[5] https://huggingface.co/docs/transformers/model_doc/distilbert
[6] https://medium.com/dataseries/why-does-xlnet-outperform-bert-da98a8503d5b
[7] https://huggingface.co/docs/transformers/model_doc/xlnet

lower peace. We selected words with the top 500 longest distances to the decision boundary as the top high-peace and low-peace words.

For transformers, we selected the keywords based on the integrated gradient method. In predicting each article's peacefulness, the integrated gradient method would generate an attention score associated with each word, where positiveness indicates the word is more relevant to high peace and vice versa. We computed the algorithmic average of the scores of each word obtained from all articles to be the global importance weight of words. The higher (more positive) the importance weight of a word, the higher peace the word is associated with. Based on the importance weight, we could easily select top-k keywords from both high-peace and low-peace countries.

III. Feature Interpretation

Having obtained the top-k keywords, we focused on how to explain the result. We utilize pre-trained word-to-vec models and unsupervised learning techniques to extract useful information from the keywords. We first used a pre-trained GloVe model[8] to generate vector representations of keywords. With the set of word vectors, we utilized the K-means algorithm with the Elbow Method to cluster the keywords. We then used the GloVe methods to retrieve the most representative words of each cluster and manually assign topics to each cluster. We also applied K-means clustering on the vectorized dataset to generalize text topics and identify target features in each topic.

Result

I. Twitter

The base case logistic regression model yielded roughly 72% percent of accuracy. Some important features that contribute to the low-peace classification include politically-related words, such as *independence*, *vote*, *chairman*, *president*, *nation*, *crown*, and *advocate*. Additionally, words related to religion are relatively frequent among these features, including *mosque*, *prophet*, *almighty*, *prayer*, *worship*, *amen*, and *blessed*. Important features that contribute to the classification of high-peace cover a range of topics, including topics focusing on food, such as *ham*, *pie*, *champagne*, and *coke*.

The pretrained BERTweet model is fine-tuned on a cleaned Twitter dataset, and the validation accuracy is around 82% with the following confusion matrix (Table 6). We also experimented

[8] https://nlp.stanford.edu/projects/glove/

with this neural network approach on shuffled Tweets. However, there is no significant difference in accuracy between the shuffled and unshuffled datasets. Therefore, we focused our analysis on the unshuffled dataset.

	High Peace	Low Peace
High Peace	68607	11538
Low Peace	17168	62687

Table 6

Identical to the approach taken by explaining the logistic regression model, we have also sought to explain the global explainability of the BERTweet model.

For the low-peace classification, consistent with the findings from logistic regression, politically related words are still a comparably important category that contributes to the low-peace classification. These words include but are not limited to *capital*, *crown*, *freedom*, *government*, *king*, *people*, *leader*, *military*, *minister*, *nation*, *president* and *queen*. There are also several religious-related words, including *holy*, *worship*, *soul*, and *god*.

There are also some but fewer politically related words for the high-peace classification, which include *election*, *vote*, *ministry*, and *democracy*. Additionally, there are more expressive and descriptive terms in high-peace countries, such as *lovely*, *boring*, *confused*, *excited*, *exhausted*, *blessed*, *starving* and *fantastic*. Lastly, an interesting fact is that more curse words, internet slangs, and internet trending topics have been seen as contributing to high-peace classification, such as *wtf*, *hoe*, *fuck*, *slay*, *bts*, and *crypto*.

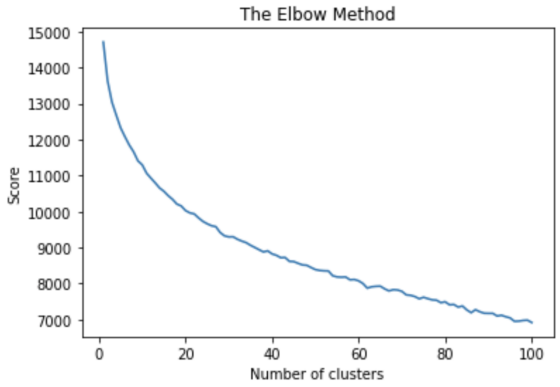
We also fitted the dataset on BertForSequenceClassification. Like the BERTweet model, BertForSequenceClassification model also yielded an accuracy of around 82%, with the confusion matrix below (Table 7).

	High Peace	Low Peace
High Peace	66334	15824
Low Peace	13725	64089

Table 7

When going through the features learned by the BertForSequenceClassification model, we found that the words for high-peace and low-peace countries also roughly follow the general trend as shown in both the logistic regression and BERTweet models. In order to better understand the features, we applied GloVe embedding to vectorize and performed K-means clustering on the top

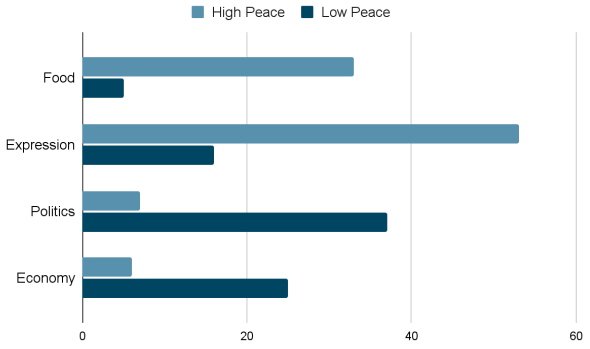
high-peace and low-peace features. To determine the number of clusters, we applied the Elbow method, and from the graph (Graph 2), we decided to use 15 clusters.



Graph 2

With the K-means clustering, we found results consistent with the findings from both the logistic regression and BERTweet models. Firstly, high-peace countries tend to talk more about food or food-related subjects, such as *wine*, *beer*, *fish*, *taste*, *ham*, *cheese*, and *drinking*. Expressive and descriptive terms are also more significant in high-peace classification, such as *beautiful*, *favorite*, *glamor*, *lonely*, and *sexy*.

For low-peace classification, political topics still appear more frequently, including *independence*, *election*, *political*, *chairman*, *official*, *congress*, and *parliament*. Topics about the economy are also more significant in low-peace classification, with words like *bank*, *budget*, *tax*, *money*, *financial*, *fund*, and *cash*.

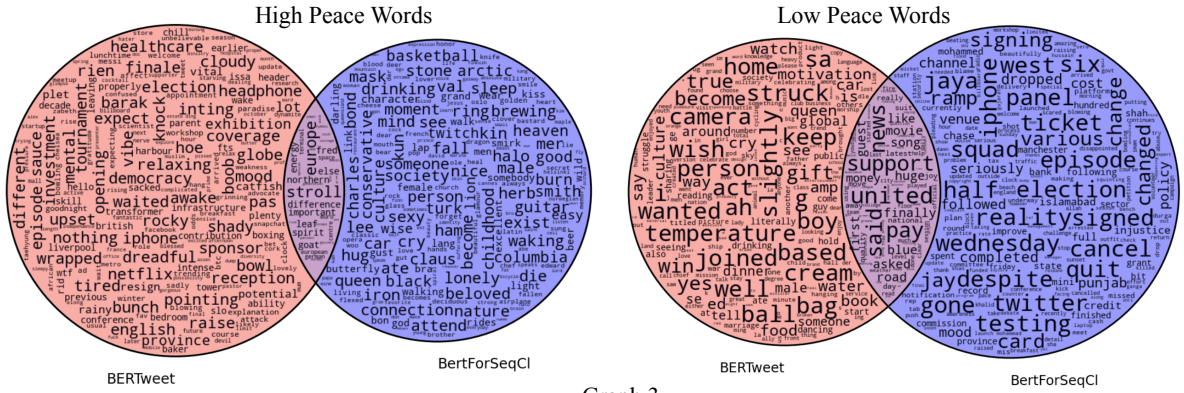


Number of Keywords in High/Low Peace Countries Based on Different Topics

The full word clusters generated by the BertForSequenceClassification can be found at Appendix - II. BertForSequenceClassification Clusters.

However, discrepancies among the models also exist. For example, logistic regression tends to classify sports-related features as high-peace topics, whereas the opposite is true for the other two models.

In order to compare the findings from different approaches, we have created Venn diagrams to observe the important words extracted from the two BERT models (Graph 3). An interesting finding from the graph is that despite common thematic similarities across all models, such as high-peace countries' focus on expressive topics and low-peace countries' focus on political topics, the key features picked up by the models are not necessarily identical. For example, *lonely* and *relaxing* are both classified as high-peace expressive terms, but the former is learned by the BERTweet model, whereas the latter is learned by the BertForSequenceClassification model.



Graph 3

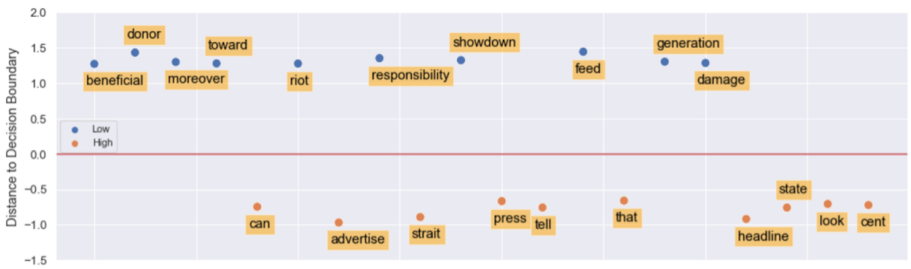
II. News

The baseline logistic regression model with leave-one-out approach yielded 87% accuracy. Building on that result, the XGBoost model received an average train accuracy of 90.12% and average test accuracy of 88.37% with the below confusion matrix (Table 8).

	High Peace	Low Peace
High Peace	28000	1700
Low Peace	1500	26000

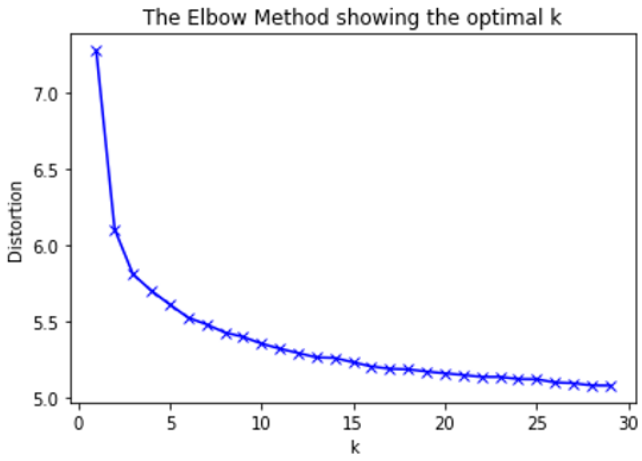
Table 8

Some important features that contribute to the low-peace classification include *donor*, *riot*, and *damage*, and key features for high-peace classification include *advertise*, *strait*, and *headline*.



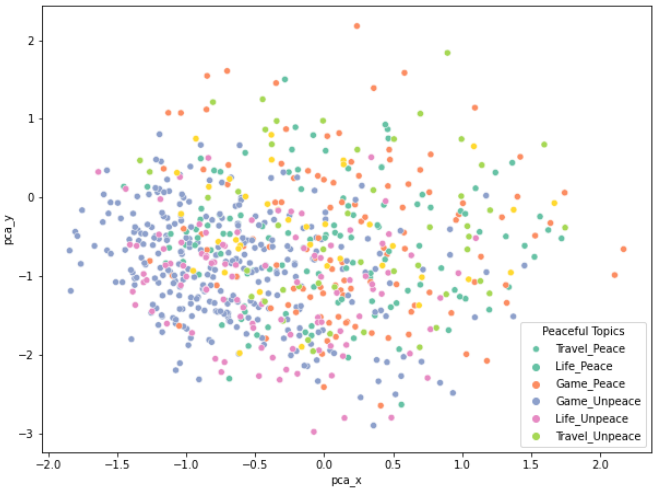
High-Peace and Low-Peace Words with Top 10 Largest Distances to the Decision Boundary

To clarify if high-peace and low-peace speeches have different tendencies on specific topics, we performed K-means clustering on word vectors. We chose 8 clusters based on the distortion calculated by the Elbow method (Graph 4).

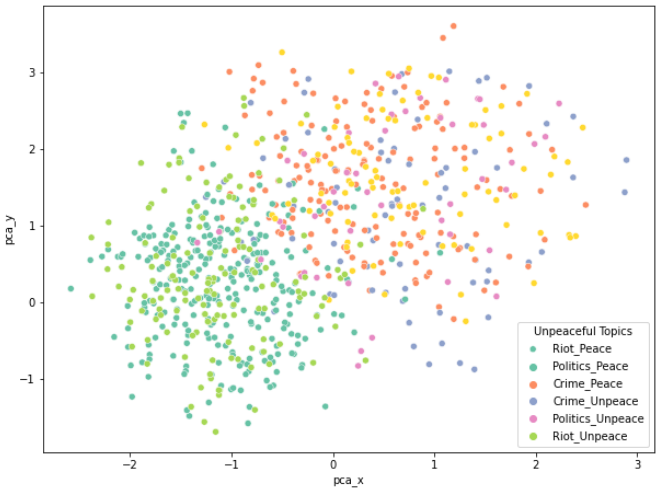


Graph 4

From those 8 clusters, we found 6 clusters that have a bias in the distribution of word sources. Thus, high-peace and low-peace speeches do have different preferences on topics. Clusters with more words coming from high-peace speeches mention travel, daily life, and game. On the other hand, clusters consisting of more low-peace words focus on riots, politics, and crime. Using dimension reduction, we can visualize the clusters of high-peace and low-peace countries (Graph 5). These findings are consistent with the findings from the Twitter dataset.



Clusters with More Words from High Peace Speeches

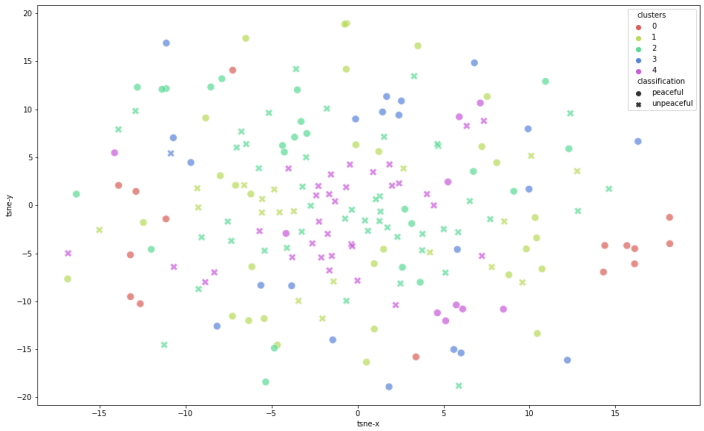


Clusters with More Words from Low Peace Speeches

Graph 5

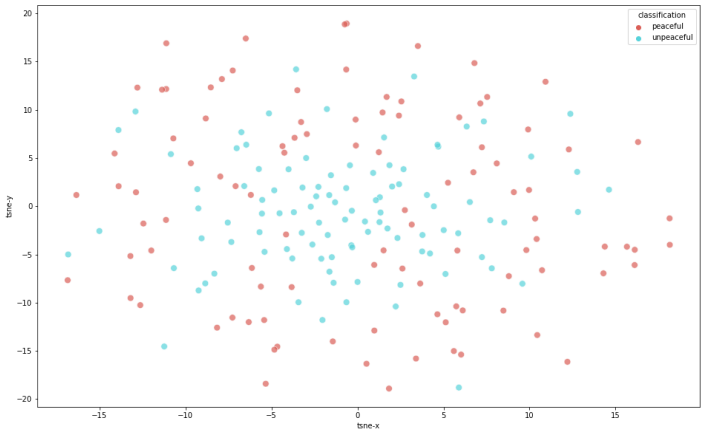
Similar to the XGBoost model, the BERT model also demonstrated different interests in topics from high-peace and low-peace countries. High-peace countries tend to focus more on business,

job market and economy, while low-peace countries are more likely to talk about government and authorities. The evidence can be found in the visualizations of clusters with reduced dimension (Graph 6), where selected keywords have distinct distributions in 5 clusters of different topics (Cluster 3 represents business topic, Cluster 4 represents government topic).



Graph 6

The other interesting finding shown by the BERT model is that words used in news texts from high-peace countries tend to be more diverse than those in low-peace countries, which indicates that high-peace countries talk about a wider variety of topics than low-peace countries. For example, in the t-sne visualization of dimension-reduced word embeddings of representative words in high-peace and low-peace countries (Graph 7), words from high-peace countries are more scattered in the graph. Moreover, the F-norm (representing variance) of the top 100 words from high-peace countries is 20.53, nearly twice the value of low-peace countries, 10.80. This finding is absent from the Twitter dataset, possibly because Twitter users consist of ordinary people with higher tendencies to talk about more random topics, regardless of whether they are from a high-peace or low-peace country. In contrast, such randomness does not exist in news outlets, which report on events in a much more formal and structured manner.



Graph 7

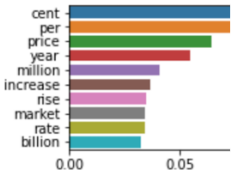
Similar findings can also be seen in the XLNet model. We fine-tuned the XLNetForSequenceClassification model with the XLNetTokenizer when training with the data. The accuracy on the testing set is 0.9176, with the confusion matrix below (Table 9).

	High Peace	Low Peace
High Peace	56694	4857
Low Peace	5331	56831

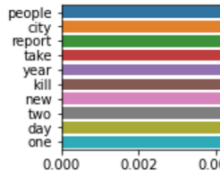
Table 9

After extracting the top high-peace and low-peace words with the XLNet model and vectorizing the words using the GloVe embedding, we proceeded to find the topics of keywords. However, instead of finding topics among top words, we wanted to find topics among articles. Therefore, instead of clustering on the top words, we clustered the documents using K-means. We performed TF-IDF transformation to the documents and then conducted K-means clustering according to the TF-IDF scores, with K=10. We looked at the top 10 terms of each cluster and named each cluster accordingly (Graph 8).

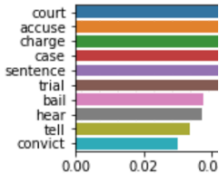
Class 1: Economy/Market



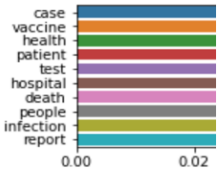
Class 2: Report



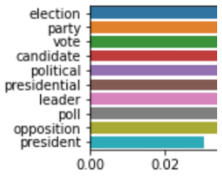
Class 3: Litigation



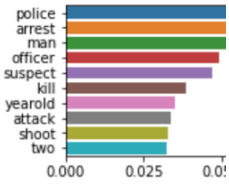
Class 4: Health



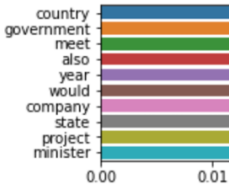
Class 5: Election



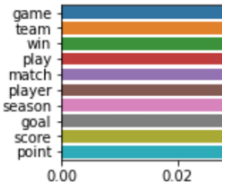
Class 6: Violence/Crime



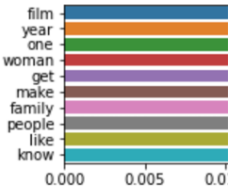
Class 7: Government Views



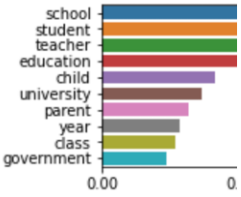
Class 8: Sports



Class 9: Film



Class 10: Education

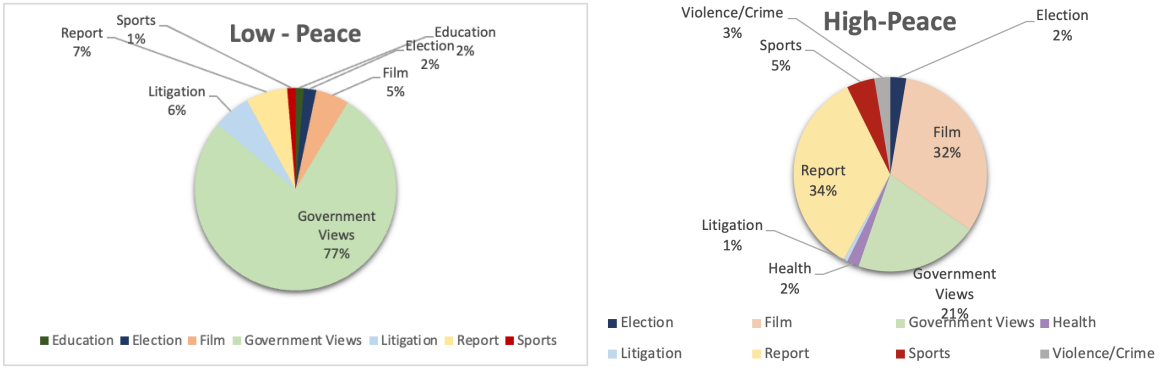


Graph 8

Based on the result of the K-means clustering of documents, we categorize the top words based on their appearances in each cluster. Whichever cluster contains the highest frequency of the target keywords will be classified as the topic of the cluster (Graph 9).

The result also indicates similar conclusions from the previous models. We can show that the topics of news articles in high-peace counties are more diverse than those in low-peace countries,

since words from various topics, such as film and report, take up significant proportions in the pie chart, whereas governmental views are the only dominant news topic in low-peace countries, with a percentage of 78%.



Graph 9

An important thing to note is that we did not perform the same clustering method on the Twitter data because Tweets are overall much more random than news articles and are much more difficult to generalize.

The full word clusters generated by the XLNet can be found at Appendix - III. XLNet Clusters.

Conclusion

Results from both the Twitter and News datasets affirm our initial hypothesis that there is a difference in speech between high and low peace countries. Despite having different nuances, both the Twitter and News datasets confirm that low-peace countries talk more about political issues, whereas high-peace countries talk more about daily activities or expressions. High-peace countries also tend to discuss a broader range of topics than low-peace countries. We think this difference in speech is consistent with the social background of high-peace and low-peace countries. Speech is an embodiment of reality: low-peace countries talk more about politics because they are experiencing certain political circumstances that lead to unpeaceful conditions, making them naturally discuss more about these conditions, and because high-peace countries are unbothered by those political uncertainties that low-peace countries are facing, they are granted with more freedom to explore and, as a result, talk about more and different daily activities and expressions, resulting in such discrepancies in speech.

Ethical Considerations

There are some ethical considerations in the data collection process. For the news articles collected, we were only able to gather data from news sources that were available to us. As for Twitter data, we have applied for Twitter API for Academic Research. By doing so, we have answered a list of application questions and promised that the use of data is non-commercial.

We are aware of the bias in the research project, including the data collection process, the fact that we are only looking at English data, etc. And these biases could extend to plausible future applications. Being aware of the biases and limitations of the research process, we're hoping that future studies could improve upon this.

References

Ma, E. (Sep 23, 2019). Why does XLNet outperform BERT? *Medium*. https://medium.com/dataseries/why-does-xlnet-outperform-bert-da98a8503d5b

Natural language processing and machine learning identified words in media most associated with lower and higher peace countries. (Aug, 2022). https://drive.google.com/file/d/1emJpgW504KMuTjM6z4b4A0RTUfdCiYWD/view

Pennington, J., Socher, R., & Manning C.D. (2014). GloVe: Global Vectors for Word Representation. *Stanford University*. https://nlp.stanford.edu/pubs/glove.pdf

Stadler, K. (2017). The country converter coco - a Python package for converting country names between different classification schemes. *The Journal of Open Source Software*. doi: 10.21105/joss.00332

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., et al.. 2020. Transformers: State-of-the-Art Natural Language Processing. *Association for Computational Linguistics*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45.

Appendix

I. Task Distribution

Country Selection: Pinyi Yang, Yibo Chen
Data Description - Twitter: Yuwen Zhang, Ziheng Ru
Data Description - News: Pinyi Yang, Yibo Chen, Xinfu Su, Hongou Liu
Preprocessing - Twitter: Yuwen Zhang
Preprocessing - Twitter - Sentiment Analysis: Ziheng Ru
Preprocessing - News: Pinyi Yang, Yibo Chen, Xinfu Su, Hongou Liu
Method - Models - Twitter: Ziheng Ru, Yuwen Zhang
Method - Models - News: Pinyi Yang, Yibo Chen, Xinfu Su, Hongou Liu
Method - Feature Selection: Ziheng Ru, Yuwen Zhang, Pinyi Yang, Yibo Chen, Xinfu Su, Hongou Liu
Method - Feature Interpretation: Ziheng Ru, Yuwen Zhang, Pinyi Yang, Yibo Chen, Xinfu Su, Hongou Liu
Result - Twitter: Ziheng Ru, Yuwen Zhang
Result - News: Pinyi Yang, Yibo Chen, Xinfu Su, Hongou Liu

II. BertForSequenceClassification Clusters

Note: Cluster names, if existed, are only tentative

<u>Cluster 1 (Food)</u>

High-Peace: 33 Words Total
'wine', 'skin', 'leaf', 'grow', 'beer', 'goat', 'herb', 'drinking', 'white', 'natural', 'hair', 'glass', 'orchard', 'taste', 'fish', 'ate', 'rich', 'wild', 'ale', 'pure', 'arctic', 'brewing', 'iron', 'bread', 'meat', 'pig', 'palm', 'cheese', 'maple', 'deer', 'fat', 'blood', 'deciduous'

Low-Peace: 5 Words Total
'yellow', 'hot', 'add', 'breakfast', 'green'

<u>Cluster 2 (Date/Time)</u>

High-Peace: 8 Words Total
'french', 'place', 'autumn', 'mark', 'attend', 'today', 'german', 'fall'

Low-Peace: 72 Words Total

'news', 'morning', 'started', 'session', 'friday', 'half', 'saturday', 'back', 'last', 'soon', 'strike', 'week', 'day', 'delivery', 'afternoon', 'finally', 'weekend', 'completed', 'series', 'conference', 'thursday', 'launched', 'took', 'upcoming', 'next', 'dropped', 'run', 'meeting', 'starting', 'month', 'out', 'date', 'latest', 'meet', 'monday', 'dropping', 'set', 'wednesday', 'schedule', 'hour', 'passed', 'tuesday', 'september', 'worst', 'three', 'six', 'return', 'following', 'ahead', 'summer', 'arrived', 'added', 'despite', 'top', 'past', 'cancelled', 'recent', 'cancel', 'recently', 'booked', 'spent', 'made', 'underway', 'daily', 'pace', 'signed', 'followed', 'first', 'held', 'sunday', 'incident', 'signing'

Cluster 3 (Religion)

High-Peace: 31 Words Total
'heart', 'life', 'person', 'die', 'jesus', 'saint', 'young', 'patient', 'born', 'daughter', 'woman', 'son', 'family', 'man', 'mate', 'lover', 'husband', 'death', 'christ', 'living', 'childhood', 'church', 'female', 'male', 'suffering', 'blind', 'friend', 'marie', 'roman', 'men', 'brother'

Low-Peace: 5 Words Total
'pleading', 'arrested', 'killed', 'abuse', 'couple'

Cluster 4 (Art/Creativity/Expression)

High-Peace: 53 Words Total
'soul', 'wolf', 'beautiful', 'don', 'museum', 'sun', 'queen', 'latin', 'art', 'folk', 'bear', 'angel', 'chef', 'dark', 'image', 'heaven', 'grand', 'sport', 'culture', 'disneyland', 'beauty', 'crown', 'stone', 'gallery', 'wear', 'symbol', 'ring', 'kiss', 'figure', 'favorite', 'artist', 'butterfly', 'guitar', 'cowboy', 'rock', 'glamour', 'beast', 'star', 'dragon', 'cannes', 'classic', 'name', 'jewel', 'golden', 'honor', 'black', 'sexy', 'beloved', 'castle', 'character', 'pop', 'opera', 'lion'

Low-Peace: 16 Words Total
'miss', 'suit', 'award', 'theme', 'episode', 'tribute', 'show', 'beautifully', 'song', 'guest', 'jay', 'movie', 'photography', 'workshop', 'outfit', 'blue'

Cluster 5

High-Peace: 53 Words Total
'lie', 'god', 'easy', 'simple', 'smile', 'sometimes', 'grim', 'dear', 'expression', 'cry', 'laughing', 'relationship', 'someone', 'always', 'moment', 'reason', 'forever', 'absolute', 'everyday', 'nice', 'calm', 'wise', 'sense', 'self', 'people', 'listen', 'nature', 'else', 'let', 'simply', 'voice', 'excellent', 'learn', 'becomes', 'love', 'wonder', 'speak', 'lonely', 'happiness', 'spirit', 'see', 'hands', 'mean', 'somebody', 'pretty', 'difference', 'meaning', 'good', 'ruin', 'forget', 'speaks', 'mind', 'truly'

Low-Peace: 43 Words Total
'note', 'done', 'relieved', 'happening', 'sure', 'mood', 'getting', 'going', 'got', 'actually', 'much', 'reality', 'though', 'detail', 'thank', 'reply', 'welcome', 'happened', 'wrong', 'amazing', 'question', 'joy', 'disappointed', 'awesome', 'everyone', 'really', 'like', 'excited', 'writing', 'complain', 'yet', 'happy', 'gone', 'blame', 'even', 'seriously', 'sorry', 'idea', 'read', 'bit', 'injustice', 'scared', 'damn'

Cluster 6 (Transportation)

High-Peace: 19 Words Total
'restaurant', 'light', 'mountain', 'rides', 'fly', 'space', 'walking', 'park', 'stroll', 'bicycle', 'wind', 'cross', 'walk', 'small', 'car', 'automobile', 'airplane', 'square', 'terminal'

Low-Peace: 34 Words Total
'fire', 'block', 'weather', 'ground', 'sunset', 'dump', 'running', 'flood', 'hundred', 'track', 'venue', 'rolling', 'outside', 'stopped', 'ramp', 'bar', 'floor', 'chase', 'traffic', 'train', 'station', 'area', 'along', 'collision', 'air', 'shoot', 'away', 'bomb', 'road', 'bus', 'tip', 'blocked', 'zone', 'beach'

Cluster 7 (Social Media)

High-Peace: 6 Words Total
'connection', 'identity', 'secret', 'memory', 'link', 'cell'

Low-Peace: 36 Words Total
'posted', 'update', 'available', 'check', 'video', 'quiz', 'online', 'booking', 'iphone', 'released', 'info', 'app', 'code', 'paper', 'register', 'release', 'fake', 'nationwide', 'twitter', 'billboard', 'copyright', 'updated', 'mini', 'youtube', 'card', 'notification', 'radio', 'teaser', 'photo', 'registration', 'platform', 'using', 'laptop', 'print', 'channel', 'page'

Cluster 8 (Economy)

High-Peace: 6 Words Total
'rose', 'growth', 'fallen', 'bond', 'energy', 'euro'

Low-Peace: 25 Words Total
'bank', 'paid', 'raising', 'pay', 'ticket', 'registered', 'drop', 'sale', 'budget', 'sector', 'huge', 'tax', 'biggest', 'rate', 'credit', 'raised', 'cost', 'management', 'money', 'paying', 'financial', 'million', 'fund', 'funded', 'cash'

Cluster 9 (Politics)

High-Peace: 7 Words Total
'edward', 'david', 'fred', 'smith', 'conservative', 'partner', 'charles'

Low-Peace: 37 Words Total
'report', 'election', 'said', 'review', 'board', 'state', 'interview', 'government', 'told', 'bill', 'justice', 'office', 'staff', 'chairman', 'letter', 'petition', 'secretary', 'official', 'asked', 'panel', 'manager', 'congress', 'announced', 'officer', 'court', 'department', 'comment', 'saying', 'quit', 'volunteer', 'parliament', 'grant', 'expert', 'national', 'request', 'committee', 'commission'

Cluster 10

High-Peace: 39 Words Total
'bro', 'sin', 'bergen', 'est', 'mer', 'woo', 'marina', 'por', 'mono', 'mon', 'bon', 'kam', 'finn', 'lille', 'nord', 'sur', 'lil', 'gen', 'tory', 'kun', 'ole', 'der', 'cock', 'lad', 'darling', 'ave', 'bio', 'ham', 'claus', 'val', 'lyon', 'fra', 'clover', 'bra', 'pussy', 'bastard', 'turk', 'lang', 'papa'

Low-Peace: 12 Words Total
'ganga', 'cong', 'rep', 'ref', 'hai', 'etc', 'sha', 'bye', 'mis', 'hara', 'una', 'pak'

Cluster 11 (Location)

High-Peace: 14 Words Total
'helsinki', 'oslo', 'ontario', 'columbia', 'finnish', 'ottawa', 'trondheim', 'alberta', 'norwegian', 'northern', 'galway', 'nsw', 'centre', 'queensland'

Low-Peace: 26 Words Total
'gambia', 'lahore', 'karachi', 'mining', 'tripoli', 'islamabad', 'independence', 'province', 'city', 'govt', 'district', 'sindh', 'maharashtra', 'organised', 'jaya', 'town', 'punjab', 'community', 'muslim', 'islamic', 'west', 'college', 'kashmir', 'university', 'hyderabad', 'gujarat'

Cluster 12 (International Relations)

High-Peace: 17 Words Total
'european', 'strong', 'america', 'europe', 'generation', 'become', 'protect', 'normal', 'longer', 'military', 'exist', 'becoming', 'example', 'important', 'risk', 'society', 'nato'

Low-Peace: 94 Words Total
'issue', 'launch', 'zero', 'making', 'follow', 'training', 'course', 'ready', 'sub', 'full', 'program', 'impact', 'level', 'new', 'needed', 'project', 'campaign', 'working', 'push', 'stop', 'taking', 'clear', 'ban', 'challenge', 'explore', 'problem', 'able', 'programme', 'phase', 'effort', 'due', 'counter', 'taken',

'development', 'changed', 'political', 'awareness', 'support', 'required', 'however', 'progress', 'complete', 'warning', 'move', 'sending', 'change', 'international', 'deal', 'thanks', 'help', 'relief', 'effect', 'practice', 'policy', 'extra', 'promised', 'process', 'illegal', 'highlight', 'facing', 'limited', 'towards', 'successful', 'try', 'various', 'fully', 'special', 'slow', 'continue', 'plan', 'testing', 'serious', 'united', 'action', 'supporting', 'situation', 'emergency', 'currently', 'job', 'putting', 'free', 'debate', 'continues', 'force', 'quality', 'launching', 'managed', 'take', 'delay', 'tried', 'improve', 'security', 'ongoing', 'promotion'

Cluster 13 (Biology)

High-Peace: 28 Words Total
'skull', 'truss', 'gust', 'smirk', 'optical', 'brain', 'biceps', 'mouth', 'doe', 'horn', 'voltage', 'halo', 'knife', 'hug', 'heal', 'mask', 'gut', 'robot', 'woke', 'burn', 'waking', 'ego', 'sleep', 'twitch', 'spiral', 'flexed', 'homo', 'loft'

Low-Peace: 13 Words Total
'blowing', 'tear', 'squinting', 'thumb', 'flash', 'utc', 'mug', 'tab', 'salute', 'waving', 'clapping', 'earthquake', 'clock'

Cluster 14

High-Peace: 10 Words Total
'lord', 'sung', 'chang', 'kim', 'abe', 'prince', 'soo', 'lee', 'kin', 'kang'

Low-Peace: 16 Words Total
'durga', 'khan', 'hussain', 'shi', 'allah', 'shri', 'imam', 'asha', 'mohammed', 'khalid', 'singh', 'gandhi', 'maha', 'mohammad', 'shah', 'muhammad'

Cluster 15 (Sports)

High-Peace: 2 Words Total
'lap', 'basketball'

Low-Peace: 33 Words Total
'super', 'final', 'team', 'hit', 'league', 'match', 'cricket', 'played', 'wicket', 'test', 'record', 'goal', 'shot', 'leg', 'missed', 'field', 'finished', 'bowling', 'england', 'season', 'forward', 'squad', 'kick', 'second', 'manchester', 'bench', 'losing', 'batting', 'coach', 'beating', 'hitting', 'inning', 'round'

III. XLNet Clusters

Note: Cluster names, if existed, are only tentative

Cluster 1 (Economy/Market)

High-Peace: 0 Word Total

Low-Peace: 0 Word Total

Cluster 2 (Report)

High-Peace: 52 Words Total
headline, advertise, french, restaurant, metre, deadly, weather, photo, mile, page, north, island, airline, nearby, scientist, rebel, researcher, temperature, troop, update, cancel, aircraft, crew, store, crash, winter, user, drug, bomb, spark, demonstration, plane, capture, display, surround, refugee, alert, location, ship, storm, inquiry, wind, sea, fly, air, warm, drive, wound, flood, mayor, tho, explosion

Low-Peace: 10 Words Total
exhibition, historical, kidnap, commander, registration, section, personnel, certificate, exhibit, collection

Cluster 3(Litigation)

High-Peace: 1 Word Total
conviction

Low-Peace: 9 Words Total
adjourn, petition, prosecution, pending, counsel, grant, application, bail, possession

Cluster 4 (Health)

High-Peace: 3 Words Total
surge, healthcare, toll

Low-Peace: 0 Word Total

Cluster 5 (Election)

High-Peace: 4 Words Total
conservative, referendum, coalition, ballot

Low-Peace: 3 Words Total
constituency, electoral, primary

Cluster 6 (Violence/Crime)

High-Peace: 4 Words Total
footage, violent, gun, stab

Low-Peace: 0 Word Total

Cluster 7 (Government Views)

High-Peace: 31 Words Total
increasingly, euro, profit, climate, employer, ease, territory, worst, giant, nearly, lower, tackle, initial, professor, overall, estate, react, pose, largely, union, decrease, slow, reaction, combine, fall, curb, breach, safe, closer, figure, ally

Low-Peace: 116 Words Total
facilitate, reiterate, implementation, commend, assure, organize, delegation, donate, allocate, bilateral, neighbor, governance, municipality, constitutional, enhance, stability, shall, farmer, trillion, establishment, commissioner, poverty, cultural, participation, unity, calendar, export, ambassador, mechanism, guideline, youth, owe, electricity, peaceful, sustainable, participant, awareness, provision, appointment, currency, occupy, partnership, principle, dialogue, representative, thus, contribution, provincial, collaboration, mandate, relevant, knowledge, recognize, stress, agriculture, forum, purpose, basic, construct, rural, procedure, objective, import, strategic, enemy, duty, resume, religious, effective, intervention, resource, obligation, construction, employment, distribute, embassy, destination, maximum, sustain, preparation, exercise, workshop, loan, skill, freedom, quality, function, structure, continent, indeed, medicine, source, association, assistance, manner, essential, arrangement, adopt, commission, civil, plot, direct, phase, expose, desire, immediate, sponsor, settle, vice, private, industrial, absence, appropriate, transaction, commercial, proper

Cluster 8 (Sports)

High-Peace: 7 Words Total
appearance, tough, half, fifth, ball, captain, squad

Low-Peace: 2 Words Total
encounter, division

Cluster 9 (Film)

High-Peace: 48 Words Total
favourite, email, version, singer, bite, pair, device, race, sex, television, image, sexual, onto, certainly, kid, guy, trouble, felt, holiday, really, perhaps, star, worse, alongside, athlete, throw, dress, extra, camera, novel, heart, mistake, almost, eat, eventually, listen, choice, original, anything, powerful, pick, little, print, actually, usually, probably, something, front

Low-Peace: 8 Words Total
pray, category, select, festival, recall, theme, teach, prayer

Cluster 10 (Education)

High-Peace: 0 Word Total

Low-Peace: 2 Words Total
examination, academic