Word Frequencies in Media Articles Predict the Level of Peace in Countries

William Powers
Queens College, CUNY
william.powers81@qmail.cuny.edu

Lin Shi
Queens College, CUNY
lin.shi66@qmail.cuny.edu

Allegra Chen-Carrel
Columbia University, NY
ac3922@columbia.edu

Philippe Loustaunau
Vista Consulting LLC, Arlington VA
ploustaunau@conseil-vista.com

Peter T. Coleman
Columbia University, NY
pc84@tc.columbia.edu

Larry S. Liebovitch
Queens College, CUNY
Columbia University, NY
larry.liebovitch@qc.cuny.edu

ABSTRACT

We found that the words reported by the media in a country can be used to predict the level of peace in that country. We analyzed data collected by the Sustaining Peace Project at Columbia University (http://sustainingpeaceproject.com) from media articles in the News on the Web (NOW) dataset (corpusdata.org) from 20 countries. Those subject matter experts identified which countries they classify as high peace, low peace, or mixed peace. We determined the frequency counts of the words from each country from: 1) all the words within quotations representing the words of local people and 2) the complete set of words in the articles cleaned with stopwords removed. We used those word frequencies to train machine learning models. When trained on only the countries identified as extreme low or high peace, those models had high, 94%, prediction accuracy. We also used those models to analyze articles from more mixed peace countries, that were not in the training set, to determine a peace index for them as a quantitative measure of their level of peace.

CCS CONCEPTS

• **Computing methodologies** → Machine learning; • **Applied computing** → **Sociology**; **Psychology**.

KEYWORDS

Preprocessing; Natural Language Processing; Lexicon Validation; Classification Model; Random Forest Classifier; Logistic Regression Classifier; Peace

ACM Reference Format:
William Powers, Lin Shi, Allegra Chen-Carrel, Philippe Loustaunau, Peter T. Coleman, and Larry S. Liebovitch. 2022. Word Frequencies in Media Articles Predict the Level of Peace in Countries. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Studies of peace have previously focused on the resolution of destructive conflicts, defining peace in a negative way, as the absence of harmful conflict. There is now an increasing emphasis in peace studies on "positive peace", to achieve an understanding of the active social forces that must work together to generate and maintain peace in a society [1] [2] [3] [4] [5]. The Sustaining Peace Project at Columbia University, led by Peter T. Coleman [6], has been studying the conditions that sustain peace. An important aid to such studies would be the ability to classify and measure the degree of peace in a society.

The medium of news article publication in the last decade has largely transitioned from standard print toward an online digital format. With this transition comes greater accessibility and an ability to collect data from a large number of publications from across the world for analysis and comparison. We made use of such data to determine if the degree of peace in a society is reflected in the language used by people in that society and by the language used by the reporting media in that society.

We used machine learning models, logistic regression and random forest, to analyze the words in the articles from those countries. We show that, properly trained, such models, can use word frequencies of local people and reporters to identify countries of low and high peace, as well as to assign a quantitative score of the degree of peace in countries between those extremes. Using the word frequencies and importance methods of the classifiers, we also determined which words were the most important in making the classification. Our work here builds on the previous preliminary work of the students in the 2020 Capstone Project at the Columbia University Data Science Institute [7].

2 DATA COLLECTION AND PRE-PROCESSING

We used the over 60 gigabytes of data that The Sustaining Peace Project had collected from the NOW corpus [8]. The data were news articles published in English, in 20 countries, between January 2010 and September 2020, recorded as text files for each country, for each month. The subject matter experts in that team had also identified the degree of peace in each country as high peace (Australia, Canada, Ireland, New Zealand, Singapore, United Kingdom), low

peace (Bangladesh, Kenya, Nigeria, Pakistan, Tanzania) or countries with mixtures of high peace and low peace characteristics that we call mixed peace (Ghana, Hong Kong, India, Jamaica, Malaysia, South Africa, Philippines, Sri Lanka, United States).

2.1 INPUT 1: Quoted Words of Local People

These articles quote the words of local people such as ordinary citizens, activists, and government figures. This provides a valuable linguistic sample of actual local speech different from that of the reporters and editors. Typically, such text would be pre-processed to remove "stop words", common words that would not be valuable in classification but whose presence would increase computational time. In analyzing this direct speech, we were mindful of the work of James W. Pennebaker [9], who showed that many of these stop words, that are usually removed in NLP analysis, are meaningful in determining the personality, emotional state, and honesty of the speaker or author. For example, overuse of function words such as *I* and *we* can suggest that a person may be lying, whereas the use of exclusives such as *but* and negations such as *not* may indicate a person is telling the truth. For that reason, we did not remove stop words in analyzing the text in these quotations. The Python programs used to retrieve the words in quotations and determine their word frequencies are described in the Appendix [10].

2.2 INPUT 2: Cleaned Full Article Text

Here we used the full text of the articles that had already been cleaned by the students in the 2020 Capstone Project [7] with all the stop words removed as well as phrases and sentences unrelated to the article's content such as ads to subscribe to the publication and suggestions to read other articles. We had access to the cleaned data from all the countries listed above, except Pakistan and South Africa. The Python programs used to determine the word frequencies from these cleaned articles are described in the Appendix [10].

2.3 Word Frequencies

We used two different methods to compare the word frequencies across years and countries. We first collected the most frequent 300 words for each year separately for each country before concatenating the total word count. The advantage of this method is that it shows the details in time, but may not best reflect the overall features since there is an excess number of words not found in all the countries. We avoided that possible issue by only counting the 300 most frequent words among all the articles combined for each country. For that reason we report here only the results of this second method. The total word count of unique words from this method across all the countries was 660 for INPUT 1: Quoted Words and 1, 042 for INPUT 2: Cleaned Articles.

3 MACHINE LEARNING MODELS

We developed three different machine learning models that used the data from INPUT 1: Quoted Words and INPUT 2: Cleaned Articles to predict whether a country is high peace, low peace, or mixed peace. These models differ in the data that they used and how they were trained. They used the classifiers RandomForestClassifier and LogisticRegression from sklearn [11] [12].

(1) ML1: Standard Instance Split. The traditional method would be to divide all the countries (instances) into exclusive train and test sets at a fixed ratio, for example a (80,20) split. Here, since there are only 20 countries, that would lead to only 4 countries in the test set, that would be insufficient for reliable values in the confusion matrix.
(2) ML2: All Others to Predict One. To generate more accurate prediction results we trained the machine learning method on all but one of the 20 countries and evaluated its prediction on that one country excluded from the training set. We repeated this procedure 20 times, each time excluding a different country from the training set. This gave us 20 data points in the confusion matrix, rather than just 4 if we had used ML1 with an (80,20) split. This seamed an obvious way to improve the accuracy, but we have not seen this approach described previously.
(3) ML3; Training on the Extremes. In psychology, it sometimes clarifies the nature of differences to consider only the extreme values. In order to further improve on the results of ML2, we excluded the mixed peace countries from the training set and only trained on the countries that the subject matter experts considered clearly high peace or low peace. As shown below, this significantly improved the prediction accuracy. We also hoped that the machine learning method trained on the extremes, would also be able to properly quantify the degree of peace on the mixed peace countries between those extremes. At least in a different context, we had previously found that a neural network trained on the extreme energy minima of a small molecule also correctly predicted other intermediate, experimentally observed, energy minima [13].

.

4 RESULTS

4.1 INPUT1: Quoted Words

(0) First, as a null hypothesis to compare to the prediction accuracy of the machine learning methods, we predicted the classes 0 (low peace), 1 (high peace), and 2 (mixed peace) of each country at random with a probability proportional to its number of occurrences. This sets the floor of the prediction accuracy. For all the machine learning methods ML1, ML2, and ML3, the z-score, $z = (accuracy - accuracy_{rnd}) / \sqrt{SEM^2 + SEM_{rnd}{}^2} \approx 3.74$ or greater, corresponding to $p < 10^{-4}$, so the improvement in their predictions is statistically significant compared to this null hypothesis.

Accuracy Mean	Sample SD	SEM
0.332	0.107	0.0239

(1) ML1: Standard Instance Split. When trained on all 20 of the high peace, low peace, and mixed peace countries, the random forest classifier had a mean accuracy of only 0.550. Each

run of the random forest classifier yields slightly different results. The prediction accuracy averaged over 20 runs is shown below.

Accuracy Mean	Sample SD	SEM
0.550	0.238	0.0532

(2) ML2: All Others to Predict One. Training and testing this method on all 20 countries did not much change the result. A sample confusion matrix is shown in Figure 1.

Accuracy Mean	Sample SD	SEM
0.530	0.0801	0.0179

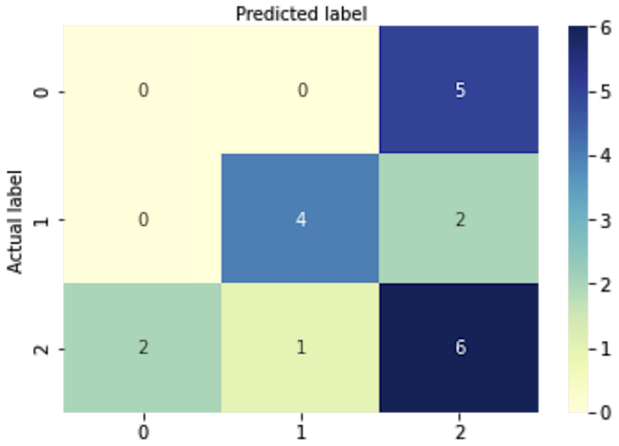

Figure 1: Sample confusion matrix from INPUT 1: Quoted Words with all 20 countries using ML2: All Others to Predict One, with approximately 50% prediction accuracy. 0=low peace, 1=high peace, and 2=mixed peace.

(3) ML3: Training on the Extremes. In 20 consecutive experiments using ML2, we never found that a low peace country was predicted to be high peace and that a high peace country was predicted to be low peace. This suggested that the prediction method is better at distinguishing extreme situations and that including the mixed peace countries in the training set was actually reducing the prediction accuracy. To prove this hypotheses, all mixed peace countries were eliminated and only high peace and low peace countries were used in the training set. Now, the ML2 All Others to Predict One method yielded much higher accuracies of approximately, 94%. A sample confusion matrix is shown in Figure 2.

Accuracy Mean	Sample SD	SEM
0.941	0.0445	0.0100

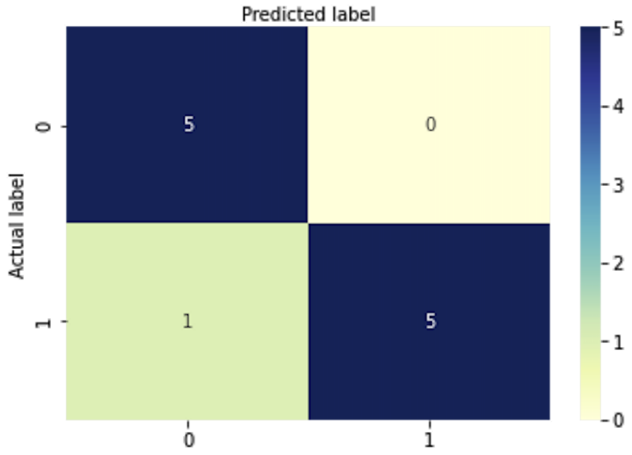

Figure 2: Sample confusion matrix from INPUT 1: Quoted Words with only the 11 most high peace or low peace countries using ML3: Training on the Extremes using all others to predict one, with approximately 91% prediction accuracy. 0=low peace and 1=high peace.

4.2 INPUT2: Cleaned Articles

(1) As was true from INPUT1: Quoted Words, the most accurate predictions from the random forest classifier from INPUT2: Cleaned Articles was achieved using ML3: Training on the Extremes.

Accuracy Mean	Sample SD	SEM
0.940	0.0821	0.0184

(2) We also determined which words were the most important in making the prediction by the random forest classifier using the *feature_importances_* method. With this method, the words in file *p_n_notp_words.xlsx* recognized as the most important were highlighted with yellow background color. The highest frequency words were more likely the words of the highest feature importance, but interestingly, many words of lower frequency were also important in predicting whether a country was high peace or low peace, as shown in Figure 3. The words of the highest feature importance, with their size scaled to their frequency of occurrence, for the high peace and low peace countries are shown in Figure 4 and Figure 5.

(3) Peace Index. We used logistic regression on ML3: Training on the Extremes to then compute $p = prob(highpeace)$ of the mixed peace countries from their data. This provides a quantitative measure of the degree of peace in those countries, using $100p$ as shown in Figure 6. In Figure 7 we compare our Peace Index to five peace indices from different sources. Note that our Peace Index values for the mixed peace countries really do fall in-between those of the extreme high peace and low peace countries.

	High Peace	Count		High Peace	Count		Low peace	Count		Low peace	Count
1	year	6.4E+05	51	lead	1.2E+05	1	state	2.0E+05	51	federal	4.2E+04
2	new	4.9E+05	52	public	1.2E+05	2	nigeria	1.6E+05	52	africa	4.2E+04
3	time	4.2E+05	53	child	1.1E+05	3	governr	1.6E+05	53	right	4.2E+04
4	people	3.8E+05	54	set	1.1E+05	4	people	1.4E+05	54	woman	4.2E+04
5	work	3.1E+05	55	australi	1.1E+05	5	country	1.4E+05	55	security	4.2E+04
6	use	3.0E+05	56	woman	1.1E+05	6	year	1.4E+05	56	power	4.2E+04
7	well	2.8E+05	57	share	1.1E+05	7	preside	9.4E+04	57	child	4.1E+04
8	day	2.4E+05	58	run	1.1E+05	8	time	9.3E+04	58	leader	4.1E+04
9	include	2.0E+05	59	issue	1.1E+05	9	nationa	8.5E+04	59	support	4.1E+04
10	world	1.9E+05	60	lot	1.1E+05	10	well	7.4E+04	60	house	4.1E+04
11	compar	1.9E+05	61	canada	1.0E+05	11	new	7.4E+04	61	start	4.0E+04
12	governr	1.8E+05	62	case	1.0E+05	12	work	7.0E+04	62	universi	4.0E+04
13	game	1.7E+05	63	state	1.0E+05	13	day	6.9E+04	63	local	4.0E+04
14	think	1.7E+05	64	move	1.0E+05	14	use	6.9E+04	64	place	3.9E+04
15	good	1.7E+05	65	comme	1.0E+05	15	kenya	6.3E+04	65	sector	3.9E+04
16	week	1.7E+05	66	area	1.0E+05	16	nigeriar	6.2E+04	66	internat	3.8E+04
17	team	1.7E+05	67	health	1.0E+05	17	party	6.0E+04	67	money	3.8E+04
18	high	1.6E+05	68	base	9.9E+04	18	bank	5.8E+04	68	provide	3.7E+04
19	right	1.6E+05	69	nationa	9.9E+04	19	world	5.7E+04	69	help	3.6E+04
20	help	1.5E+05	70	player	9.9E+04	20	police	5.6E+04	70	office	3.6E+04
21	change	1.5E+05	71	local	9.8E+04	21	service	5.5E+04	71	nation	3.6E+04
22	busines	1.5E+05	72	add	9.7E+04	22	high	5.4E+04	72	ensure	3.5E+04
23	life	1.5E+05	73	follow	9.6E+04	23	include	5.4E+04	73	number	3.5E+04
24	start	1.5E+05	74	site	9.5E+04	24	call	5.4E+04	74	billion	3.5E+04
25	service	1.4E+05	75	police	9.5E+04	25	compar	5.4E+04	75	former	3.5E+04
26	call	1.4E+05	76	plan	9.3E+04	26	court	5.4E+04	76	health	3.4E+04
27	thing	1.4E+05	77	news	9.3E+04	27	public	5.4E+04	77	road	3.4E+04
28	month	1.4E+05	78	win	9.1E+04	28	governc	5.3E+04	78	month	3.4E+04
29	family	1.4E+05	79	season	9.0E+04	29	develop	5.3E+04	79	commu	3.4E+04
30	place	1.4E+05	80	young	9.0E+04	30	issue	5.3E+04	80	set	3.4E+04
31	country	1.4E+05	81	find	8.9E+04	31	membe	5.2E+04	81	chief	3.3E+04
32	city	1.4E+05	82	story	8.9E+04	32	busines	5.2E+04	82	cent	3.3E+04
33	report	1.4E+05	83	cost	8.9E+04	33	electior	5.1E+04	83	order	3.3E+04
34	play	1.4E+05	84	increase	8.8E+04	34	report	5.1E+04	84	lagos	3.3E+04
35	big	1.4E+05	85	membe	8.8E+04	35	area	4.9E+04	85	increase	3.3E+04
36	market	1.3E+05	86	event	8.8E+04	36	good	4.8E+04	86	educati	3.2E+04
37	informa	1.3E+05	87	court	8.7E+04	37	school	4.8E+04	87	team	3.2E+04
38	great	1.3E+05	88	house	8.7E+04	38	banglac	4.7E+04	88	accordi	3.2E+04
39	point	1.3E+05	89	top	8.6E+04	39	life	4.7E+04	89	econom	3.2E+04
40	provide	1.2E+05	90	man	8.6E+04	40	million	4.6E+04	90	follow	3.2E+04
41	long	1.2E+05	91	open	8.5E+04	41	case	4.5E+04	91	system	3.1E+04
42	group	1.2E+05	92	today	8.5E+04	42	general	4.4E+04	92	thing	3.1E+04
43	zealand	1.2E+05	93	ministe	8.5E+04	43	market	4.4E+04	93	director	3.1E+04
44	support	1.2E+05	94	party	8.5E+04	44	lead	4.4E+04	94	nigeriar	3.1E+04
45	commu	1.2E+05	95	system	8.5E+04	45	political	4.4E+04	95	news	3.1E+04
46	best	1.2E+05	96	believe	8.4E+04	46	project	4.4E+04	96	officer	3.1E+04
47	school	1.2E+05	97	result	8.4E+04	47	ministe	4.3E+04	97	change	3.1E+04
48	number	1.2E+05	98	level	8.4E+04	48	group	4.3E+04	98	point	3.0E+04
49	million	1.2E+05	99	continu	8.3E+04	49	law	4.3E+04	99	industry	3.0E+04
50	cent	1.2E+05	100	ireland	8.3E+04	50	add	4.2E+04	100	family	3.0E+04

Figure 3: 100 most frequency words for the high peace and low peace countries countries from INPUT2: Cleaned Articles. Yellow indicates words of highest feature importance in making the classification by the random forest feature importance method.

Figure 4: Word cloud of the words of highest feature importance, with their size scaled to their frequency of occurrence, for high peace countries generated from INPUT 2: Cleaned Articles.

Figure 5: Word cloud of the words of highest feature importance, with their size scaled to their frequency of occurrence, for low peace countries generated from INPUT2: Cleaned Articles.

INPUT2: CLEANED FULL TEXT	
Country	Peace Index
Nigeria	6
Bangladesh	7
Tanzania	8
Kenya	9
Ghana	21
Sri Lanka	24
Jamaica	51
India	55
Malaysia	57
Hong Kong	67
Philippines	68
United States	92
Singapore	94
Canada	94
United Kingdom	94
Australia	95
Ireland	96
New Zealand	97

Figure 6: Peace Index measured from the logistic regression p value using data INPUT2: Cleaned Articles. The low peace countries in red and the high peace countries in green were used in the training set to compute the Peace Index of the other mixed peace countries.

	Peace Index	GPI	PPI	WHI	FSI	HDI
Nigeria	6	2.8	3.9	5.2	101	0.52
Bangladesh	7	2.1	3.6	4.7	92	0.59
Tanzania	8	1.8	3.4	3.6	81	0.51
Kenya	9	2.4	3.6	4.4	98	0.56
Ghana	21	1.8	2.9	4.7	69	0.58
Sri Lanka	24	2.2	3.2	4.3	90	0.77
Jamaica	51	2.1	2.5	5.6	65	0.72
India	55	2.6	3.3	4.5	78	0.62
Malaysia	57	1.6	2.5	5.9	66	0.79
Hong Kong	67			5.5		0.93
Philippines	68	2.5	3.3	5.2	85	0.70
United States	92	2.3	1.8	7.0	35	0.92
Singapore	94	1.4	1.7	6.5	33	0.93
Canada	94	1.4	1.5	7.4	25	0.91
United Kingdom	94	1.8	1.6	6.9	34	0.92
Australia	95	1.4	1.5	7.3	25	0.93
Ireland	96	1.4	1.4	7.0	24	0.93
New Zealand	97	1.2	1.5	7.3	23	0.91

Figure 7: Our Peace Index compared to five peace indices from different sources. The tertiles of each index for these countries are colored: low peace=red, mixed peace=yellow, and high peace=green. GPI=Global Peace Index [14], PPI=Positive Peace Index [15], WHI=World Happiness Index [16], FSI=Fragile States Index [17], and HDI=Human Development Index [18].

5 CONCLUSIONS

(1) We found that the word frequencies from the quotes of local people and the reporters words in media articles can be used to accurately classify countries as high peace or low peace.
(2) The prediction accuracy was the same, 94.0% vs. 94.1%, whether the complete text within the quotations or the clean data (with stop words removed) from the whole article was used as the input data.
(3) The prediction accuracy was much higher 94% vs. 53% when the training and test sets were limited to only the most high peace and low peace countries, and did not include mixed peace countries in between those extremes. In fact, different peace indices report different numerical measures of the degree of peace in those mixed peace countries [14] [15] [16] [17] [18]. The machine learning models used here are not good at learning how to classify countries given input about mixed peace countries whose degree of peace is itself ambiguous to human classifiers.
(4) However, training on only the extreme high peace and low peace countries made it possible to use the p values from the logistic regression classifier to assign a quantitive peace index of the degree of peace in such mixed peace countries.
(5) It is also important to keep in mind the limitations of this study. The NOW corpus adds approximately 10K articles/day, which is considerably smaller than the total number of media articles published in English just in the USA alone and words are excised at random so that it can remain open source without violating copyright protections. NLP and machine learning can also reflect cultural or racial bias inherent in the original data due to historical and current forces of the countries where the data was available. Many African countries are considered to be low peace countries according to several peace indices. We found that the word "African" was often used in low peace contexts. This association does not suggest that "African" is a non-peaceful word, rather its inclusion here points to deeper historical patterns of conflict, imperialism, colonialism, and racism inherent in the structure of global peace indices and a reminder of the vital importance of how data is used and interpreted. To help mitigate these biases in future research, studies may consider training using longitudinal comparisons of the same countries, or comparisons of countries within the same region.

6 VALUE OF THIS ANALYSIS

6.1 Measuring Peace

Understanding the level of peace in a country is a key starting point to assess how well a society is functioning, how happy its citizens are, and how effectively that country is interacting with its neighboring countries and the rest of the world. Quantitative peace indices that measure the levels of peace in countries are frequently reported in global media. These indices influence the goals of governments and other agencies and therefore how they spend their resources, so they have important consequences on people's lives.

Previous peace indices have been based on the assumptions of their authors about what factors they think best represents a high

peace society. Using those assumptions, those authors assembled data and the opinions of experts to construct quantitative measures of peace. As far as we know, our work here is the first data driven, rather than assumption driven, approach to determining measures of peace. Our hypothesis was that the degree of peace in a country would be reflected in the language of that country. We did use subject matter experts to identify the most extreme high peace and low peace countries. But we did not make any assumptions about what linguistic features are more representative of high peace or low peace countries. Rather, our algorithms learned that information on their own from the text data in media articles. We then used those algorithms to provide quantitative measures of the level of peace in mixed peace counties that were not at the extremes of high peace or low peace countries. This data driven approach, using less assumptions, provides a new valuable window into assessing the level of peace in countries.

Our data driven approach will also make it possible to use longitudinal data over time to develop a real-time dashboard that can monitor the peace status of a country, uncover long-term trends, detect unusual events, and evaluate if the changes made in a society are increasing or decreasing peace.

6.2 Revealing Social Factors that Support Peace

Data driven linguistic processing, such as ours, can also yield important insights into the constructs that people use to see the world and interact with it. For example, Michele J. Gelfand and her collaborators [19] [20] used linguistic analysis to identify measures of “tightness" or “looseness" in societies, finding “ ’tight’ cultural groups with relatively strong norms and little tolerance for deviance and ’loose’ groups with weaker norms and more tolerance for dissent." [19]. Some social challenges may be more successfully addressed by a tight society and others by a loose society. A society with the flexibility to shift in time from one to other may be most effective, using its tightness to resist aggression in a war, or its looseness in developing creative solutions to a changing climate.

Our work here also reveals evidence of tightness and looseness in the words that we have found important in classifying high peace and low peace countries. Hgh peace countries tend to use more words with a generally positive connotation, such as *new*, *well*, *day*, and *great*. Low peace countries tend to use more words associated with an executive and administrative association, such as *government*, *police*, *national*, and *president*. These words are not exclusive to either high peace or low peace countries, there are some overlaps between the two groups.

6.3 Words Matter

The words that reporters choose to use in a story are influenced by, and also influence, their society. Reporters may not even be aware of the effects of their choice of words. Our work here identifies words more associated with high peace or low peace societies. Those words may be a cause and a consequence of enhancing peace or reducing it. Identifying those words may help governmental, organizational, and press sources choose words that enhance peace rather than reduce it.

REFERENCES

[1] Morton Deutsch and Peter T Coleman. The psychological components of a sustainable peace: An introduction. In *Handbook on Sustainability Transition and Sustainable Peace*, pages 139–148. Springer, 2016.

[2] Paul F Diehl. Exploring peace: Looking beyond war and negative peace. *International Studies Quarterly*, 60(1):1–10, 2016.

[3] Douglas P Fry. *The human potential for peace: An anthropological challenge to assumptions about war and violence.* Oxford University Press, USA, 2006.

[4] Gary Goertz, Paul Francis Diehl, and Alexandru Balas. *The puzzle of peace: The evolution of peace in the international system.* Oxford University Press, 2016.

[5] Youssef Manmoud and Anupah Makoond. *Sustaining peace: What does it mean in practice?* International Peace Institute, 2017.

[6] AC4. Columbia University Sustainable Peace Project, 2018. http://sustainingpeaceproject.com. Accessed: December 2, 2021.

[7] Jinwoo Jung, Hojin Lee, Hyuk Joon Kwon, Matt Mackenzie, and Tae Yoon Lim. power-of-peace-speech, 2021. https://github.com/mbmackenzie/power-of-peace-speech/. Accessed: December 2, 2021.

[8] NOW. News on the Web corpus, 2021. corpusdata.org. Accessed: December 2, 2021.

[9] James W 211.. Pennebaker. Your use of pronouns reveals your personality. *Harvard Business Review*, 89(12):32–33, 2011.

[10] ArticleClassifier, GitHub, 2011. https://github.com/wpqc21/ArticleClassifier. Accessed: December 2, 2021.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. sklearn.ensemble.RandomForestClassifier - Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python
sklearn.linear model.logisticregression. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] L. S. Liebovitch, N. D Arnold, and L. Y Selector. Neural networks to compute molecular dynamics. *Journal of Biological systems*, 2(02):193–228, 1994.

[14] GPI Global Peace Index, 2021. https://www.visionofhumanity.org/wp-content/uploads/2021/06/GPI-2021-web-1.pdf. Accessed: December 2, 2021.

[15] PPI Positive Peace Index, 2021. https://www.visionofhumanity.org/wp-content/uploads/2021/04/PPR-2020web.pdf. Accessed: December 2, 2021.

[16] World Happiness Index, 2021. https://countryeconomy.com/demography/world-happiness-index. Accessed: December 2, 2021.

[17] Fragile States Index, 2021. https://fragilestatesindex.org. Accessed: December 2, 2021.

[18] HDI Human Development Index, 2021. http://hdr.undp.org/en/content/human-development-index-hdi. Accessed: December 2, 2021.

[19] Joshua Conrad Jackson, Michele Gelfand, Soham De, and Amber Fox. The loosening of american culture over 200 years is associated with a creativity–order trade-off. *Nature human behaviour*, 3(3):244–250, 2019.

[20] Michele J Gelfand, Joshua Conrad Jackson, Xinyue Pan, Dana Nau, Dylan Pieper, Emmy Denison, Munqith Dagher, Paul AM Van Lange, Chi-Yue Chiu, and Mo Wang. The relationship between cultural tightness–looseness and covid-19 cases and deaths: a global analysis. *The Lancet Planetary Health*, 5(3):e135–e144, 2021.

A PRE-PROCESSING METHODS

These are the Python programs used to retrieve the words in quotations and determine the word frequencies [10].

A.1 INPUT1: Quoted Words

The program,*INPUT1-Quotes-Article-Preprocessing-1.ipynb*, parses through the text files, and searches for all quotes within each article which are led by the word ’said’ in order to filter out statements that are connected together by missing quotation marks. If the quote begins and ends in a curly quotation mark (“ ”), they are replaced with straight quotation marks (" ") and stored within a list inside a pandas dataframes. These lists are separated by their respective countries and year.

The program *INPUT1-Quotes-Article-Preprocessing-2.ipynb* filters

the quotations, separates word frequencies into lists, and corrects an issue where apostrophes separate contractions into separate words. This information is written into a new dataframe and saved as a pkl file, which enables the dataframes to be accessible outside the program after it has finished running.

The program *INPUT1-Quoted-Words-Classification.ipynb* reads the updated dataframe files and creates a count of each word that appears within quotations within a year for each country, while excluding punctuation marks. These word counts are then sorted alphanumerically and collected within dataframes with rows representing each country, columns representing each word, according to their respective year. The dataframes from each country representing each year are also combined into a single dataframe which aggregates all word counts and their sums in total. The word frequencies are normalized by dividing the frequency of each word by the sum of all the word frequencies.

The program *INPUT1-Quoted-Words-Classification.ipynb* also reads 20 pkl files to generate 20 series. The program then concatenates all series and produces a single dataframe. This new dataframe contains 20 columns representing 20 countries and 122 rows representing all articles from corresponding countries. For each country, every word from the articles of this country is counted and results are stored as the form of tuples. By sorting these tuples, the 300 most frequently observed words are extracted in terms of each country, and these words are stored into specific lists.

A.2 INPUT2: Cleaned Articles

The program *INPUT2-Cleaned-Articles-Classification.ipynb* reads and analyzes the data from the *article_text_Ngram_stopword_lemmatize* column of the respective data files from each country. In this program the word frequencies were collected using the entire articles, as opposed to being limited by what was within quotation marks. These word counts are then sorted alphanumerically and collected within dataframes with rows representing each country and columns representing each word for all years combined. The word frequencies are normalized by dividing the frequency of each word by the sum of all the word frequencies.

Two lists were also produced featuring the top 1,000 most frequently featured words in an added collection of all the high peace countries, and an added collection of all the low peace countries. This can be found within the *list_peace_nonpeace.xlsx* excel file. Similarly, using the most 300 frequently observed words from each high peace and low peace country, another excel file named *p_n_notp_words.xlsx* was produced featuring all high peace and low peace words with total number of occurrences of these words.

When testing the RandomForestClassifier on the data, the rows containing the collected data for each country are first converted into NumPy arrays. The *feature_importances_* function of the RandomForestClassifier is specifically meant for testing data within a Pandas dataframe. Therefore, converting the rows of data into NumPy arrays for use within the RandomForestClassifier model made it difficult to test for the importance of specific words. As a result, a separate two-dimensional array was created containing the list of important words and their importance values. The important words are sorted to display by the descending order of their corresponding importance value.