The Power of Peace Speech

Peace Speech Analysis

Dec 14, 2021

Team Members
Hongling Liu, Haoyue Qi, Xuanhao Wu, Yuxin Zhou, Wenjie Zhu

Advisor Group
Peter Thomas Coleman Professor, Psychology and Education
Allegra Chen-carral Program Manager, Sustaining Peace Project
Larry Liebovitch Professor, Psychology and Physics
Philippe Loustaunau Managing Partner, Vista Consulting LLC

Project Description, Goal, and Data Overview

Peacekeepers are actively conducting research on worldwise hate speech including **news articles and blog posts** in seeking to maintain robust and peaceful communities. But the peace speeches were commonly neglected in past researches.

This year, we continue this research done by previous year's capstone students to fill in this missing side of the story by **finding connections between speech and the peacefulness of a country** with two innovative objectives.

Objective 1 - Contextual Task

- Explore the performance of different models on classifying high-peace and low-peace countries' English article embeddings encoded by **BERT**.

Premise 1: If there exists a high performance model, then there are language differences in the articles from countries with different peace levels.

Objective 2 - Contextless Task

- Study how the words' order affect the previous classification results by evaluating the model and encoder's performance over randomly shuffled articles.

Premise 2: If we see a performance drop after shuffling, then the sentence context is a strong peacefulness indicator.

Datasets studied are news articles from top 10 high and low peace countries ranked by *rule-of-5 (see Appx.1 for definition)* method, and is provided by LexisNexis and is stored on AWS S3 buckets *(see Appx.1 for the full list of countries).*

EDA Methodology:

Step 3
Update our previous conclusion if needed. Terminate the process if no major updates.

Step 1
Take a random subset of real data in a logarithmic way, increasing the subset size by a factor of 10 every time.

Step 2
Measure some metrics, perform analysis, and make conclusions on this subset.

Insights from EDA:

1. Heavy data imbalance between countries in each class.

   Actions taken: Balance by country is less feasible. Balance by class strategy became our top choice for most of the cases.

2. Significant length difference between articles from high peace countries vs. low peace countries.

   Actions taken: Small max length (128 tokens) to confound the length difference between +/- samples.

Data Preparation and Modeling Approaches

Preparation Goal - Minimal preprocessing to **avoid destructions** to natural language structure. Stopwords are kept to **preserve negations and sentence sentiments**.

Cleaning Procedure:

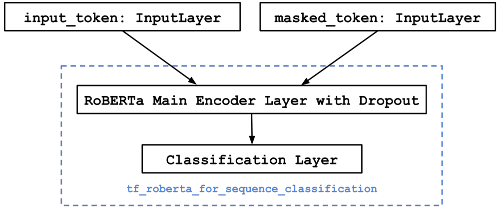
1. URL and Email Pattern identification and Removal
2. Unifying English versions via Dictionary Lookup
3. Named Entity Removal via SpaCy
4. BERT's Paired Preprocessor and Tokenization

Shuffling Procedure:

1. Split cleaned sentence by space and randomly shuffle the order.

BERT Selection Criterion:

1. Pretrained on news articles.
2. Able to interpret sentiments.
3. Reliable model provider.



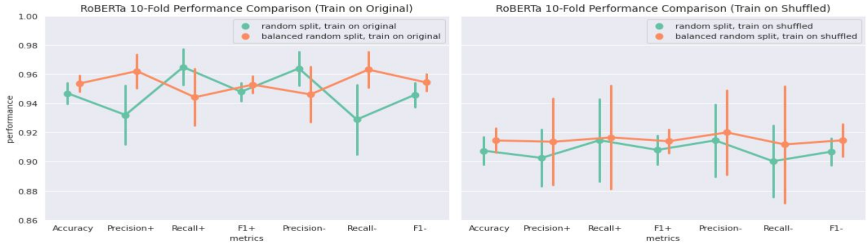
End Classifier Candidates:

- Logistic regression (Main)
- XGBoost
- SVM

Experiment 1 - Random Split: Train on all countries, test on all countries. This **ensured the class balance** in training and validation set, but **ignored the country-wise sample balance** within each peace group in both sets.

Experiment 2 - Split-by-Country: Train on *Data Minorities*, test on *Data Majorities*. This method is designed to bypass the named entity residual issue. It **challenges the model to discover shared language structures** between countries within the same peace group. **Class balance and country-wise sample balance are separately maintained** in training and validation.

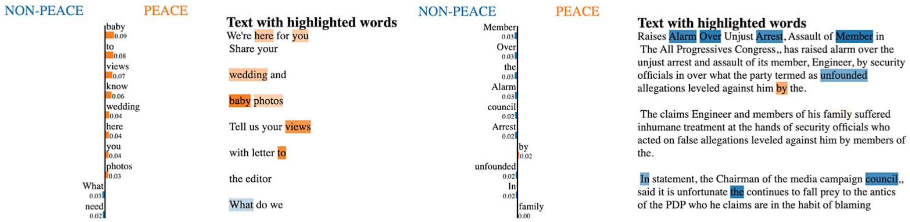
Experiment 3 - Country-Balanced Random Split: Random Split with less data from data majority countries. This helps the model to **put the equal focus on data from different countries**. The **class balance and country-wise sample balance are maintained** in training and validation.



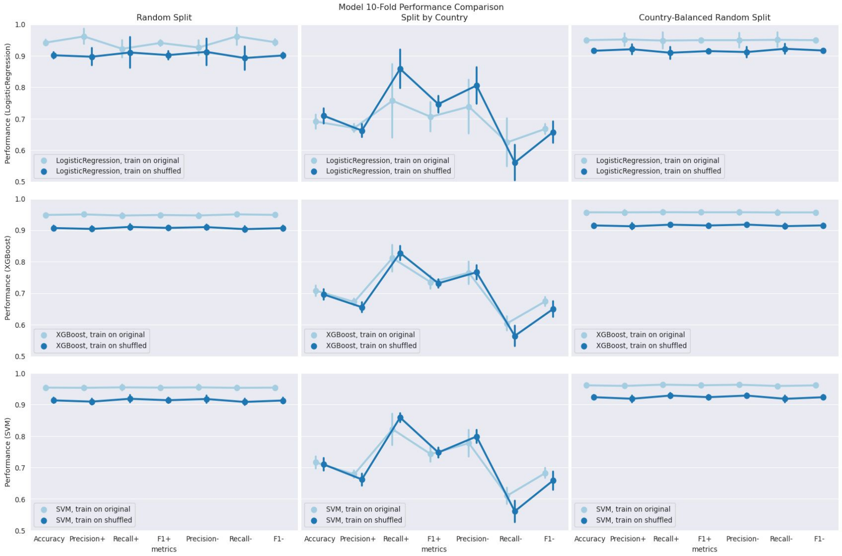
Key Results and Future Improvements

Key Observations and Results:

1. Performance on Random Split (90%+) >> Split-by-Country (70%+).
   ➜ **Linguistic features are a weak predictor** of whether a country is high or low peace across countries. **Countries may have different linguistic features** relating to their level of peacefulness.
2. Split-by-Country gives 20% accuracy improvement compared to binary classification baseline (50%).
   ➜ There **exists a weak link of language structures** in articles from countries in the same peace group.
3. Performance on the original is better (~5%) than on the shuffled articles under same train-test split approach for all experiments.
   ➜ **Sentence context contributes a little** but not significantly to the classification. Comparing to that, **meaning of individual words might be more important.**



- LIME results validates 3



- All Model Performance

Future Improvements

Improve current pipeline - Train our own entity recognizer to better cleanup the sentence.

Explore other models - Try BERT encoders and other classification models to improve the accuracy.

Explain current model - Run current trained model on extreme countries identified by *matched-pair (see Appx.1 for definition)* method and see if the results agrees with manual classification standard.

Additional Comments:

1. RoBERTa is **prone to overfit** and sensitive to class-wise balancing. Meanwhile, more **complex classifiers reduce the prediction performance dispersion**.

Reference

[1] Coleman, P. T., Fisher, J., Fry, D. P., Liebovitch, L. S., Chen-Carrel, A., & Souillac, G. (2020). How to live in peace? Mapping the science of sustaining peace: A progress report. The American psychologist, 10.1037/amp0000745. Advance online publication. https://doi.org/10.1037/amp0000745

[2] “Smart Batching Tutorial - Speed Up BERT Training · Chris McCormick,” Mccormickml.com, Jul. 29, 2020. https://mccormickml.com/2020/07/29/smart-batching-tutorial/.

[3] E. Alzahrani and L. Jololian, “How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors,” arXiv.org, 2021. https://arxiv.org/abs/2109.13890.

[4] J. Camacho-Collados and M. T. Pilehvar, “On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis,” arXiv.org, 2017. https://arxiv.org/abs/1707.01780.

[5] Cathal Horan, “Tokenizers: How machines read,” FloydHub Blog, Jan. 28, 2020. https://blog.floydhub.com/tokenization-nlp/#wordpiece.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv.org, 2018. https://arxiv.org/abs/1810.04805.

[7] P. Goyal et al., “Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour,” arXiv.org, 2017. https://arxiv.org/abs/1706.02677.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” arXiv.org, 2016. https://arxiv.org/abs/1602.04938.

Appendix 1 - Terminologies and The List of Countries

Rule-of-5	For each country, scores for indices such as Global Peace Index, the Positive Peace Index, etc., are computed. Then countries are placed into low-, mid-, and high-peace categories for each index. If a country has at least 5 indices categorized as low-peace, and no index in high-peace, then the country is low-peace. Same logic follows for high-peace.
Data Majority	Data from India and Australia. They makes up more than 90% and 75% in each category accordingly.
Data Minority	Data from the rest of countries in our dataset.
Matched-pair	Cluster the countries using geography first, then select the highest and lowest peace countries in each geographical region, where the lowest peace country in a region is the one with the lowest number of indices classified as high-peace and highest number of indices classified as low-peace. Same logic follows for high-peace.

A Full List of Countries

lowPeace
Afghanistan
Congo
Guinea
India
Iran
Kenya
Nigeria
Sri Lanka
Uganda
Zimbabwe

highPeace
Austria
Australia
Belgium
Czech Republic
Denmark
Finland
Netherlands
New Zealand
Norway
Sweden

Appendix 2 - *Fine-Tune Statistics under different Train & Validation Split*

C.I and mean obtained from 10 Stratified Folds, 8K data in Train & 2K data in Validation per Fold.

		Accuracy	Precision +	Recall +	F1 +	Precision -	Recall -	F1 -
Original Sentence	Random Split	94.68 (94.13, 95.22)	93.19 (91.70, 94.69)	96.48 (95.56, 97.40)	94.78 (94.31, 95.25)	96.39 (95.53, 97.25)	92.88 (91.10, 94.65)	94.57 (93.94, 95.20)
	Split By Country	71.88 (70.17, 73.58)	67.64 (66.49, 68.79)	84.17 (78.04, 90.29)	74.77 (72.12, 77.43)	80.02 (75.59, 84.45)	59.59 (55.42, 63.76)	67.87 (66.04, 69.69)
	Country Balanced Random Split	95.35 (94.94, 95.77)	96.20 (95.34, 97.07)	94.40 (92.97, 95.85)	95.27 (94.85, 95.70)	94.60 (93.19, 96.01)	96.30 (95.40, 97.21)	95.42 (95.00, 95.84)
Shuffled Sentence	Random Split	90.74 (90.05, 91.43)	90.25 (88.80, 91.70)	91.46 (89.34, 93.58)	90.80 (90.06, 91.53)	91.45 (89.60, 93.30)	90.02 (88.19, 91.84)	90.67 (89.97, 91.37)
	Split By Country	68.72 (67.16, 70.28)	64.77 (63.41, 66.14)	82.47 (77.04, 87.90)	72.38 (70.35, 74.42)	76.94 (71.86, 82.02)	54.97 (50.43, 59.51)	63.57 (61.22, 65.93)
	Country Balanced Random Split	91.45 (90.83, 92.07)	91.36 (89.15, 93.57)	91.65 (89.02, 94.29)	91.39 (90.79, 92.00)	92.00 (89.86, 94.13)	91.17 (88.20, 94.14)	91.46 (90.63, 92.29)

Appendix 2 *- Model Performances Stats (Random Split)*

C.I and mean obtained from 10 Stratified Folds, 8K data in Train & 2K data in Validation per Fold.

		Accuracy	Precision +	Recall +	F1 +	Precision -	Recall -	F1 -
Original Sentence	Logistic Regression	94.19 (93.50, 94.87)	96.13 (94.34, 97.93)	92.20 (90.16, 94.24)	94.06 (93.36, 94.77)	92.61 (90.84, 94.38)	96.17 (94.13, 98.21)	94.29 (93.60, 94.98)
	XGBoost	94.86 (94.59, 95.13)	95.03 (94.66, 95.41)	94.67 (94.09, 95.25)	94.85 (94.57, 95.13)	94.70 (04.17, 95.23)	95.05 (94.65, 95.45)	94.87 (94.61, 95.14)
	SVM	95.35 (94.94, 95.77)	95.29 (94.80, 95.78)	95.43 (94.76, 96.10)	95.36 (94.94, 95.78)	95.43 (94.80, 96.06)	95.28 (94.77, 95.79)	95.35 (94.95, 95.76)
Shuffled Sentence	Logistic Regression	90.15 (89.38, 90.93)	89.68 (87.65, 91.72)	91.03 (87.41, 94.65)	90.20 (89.20, 91.20)	91.22 (88.13, 94.31)	89.28 (86.53, 92.03)	90.08 (89.42, 90.74)
	XGBoost	90.70 (90.12, 91.28)	90.42 (89.79, 91.05)	91.05 (90.40, 91.70)	90.73 (90.16, 91.31)	90..99 (90.35, 91.62)	90.35 (89.69, 91.01)	90.67 (90.08, 91.25)
	SVM	91.30 (90.67, 91.92)	90.90 (90.21, 91.58)	91.80 (90.92, 92.68)	91.34 (90.71, 91.98)	91.73 (90.89, 92.57)	90.80 (90.07, 91.53)	91.26 (90.63, 91.88)

Appendix 2 - *Model Performances Stats (Split-by-Country)*

C.I and mean obtained from 10 Stratified Folds, 8K data in Train & 2K data in Validation per Fold.

		Accuracy	Precision +	Recall +	F1 +	Precision -	Recall -	F1 -
Original Sentence	Logistic Regression	69.07 (67.42, 70.72)	67.01 (66.07, 67.96)	75.69 (67.00, 84.38)	70.58 (67.15, 74.02)	73.82 (67.53, 80.11)	62.45 (56.81, 68.10)	66.75 (65.58, 67.91)
	XGBoost	70.81 (69.56, 72.05)	67.22 (66.47, 67.98)	81.19 (78.07, 84.31)	73.51 (71.96, 75.06)	76.49 (73.89, 79.09)	60.42 (58.81, 62.03)	67.43 (66.37, 68.48)
	SVM	71.56 (70.16, 72.96)	67.79 (66.95, 68.64)	82.14 (78.44, 85.84)	74.22 (72.44, 76.00)	77.70 (74.53, 80.86)	60.98 (59.06, 62.91)	68.20 (67.05, 69.35)
Shuffled Sentence	Logistic Regression	70.93 (69.13, 72.72)	66.21 (64.65, 67.76)	85.83 (81.25, 90.41)	74.63 (72.68, 76.57)	80.54 (76.28, 84.80)	56.02 (51.86, 60.19)	65.71 (63.16, 68.27)
	XGBoost	69.61 (68.39, 70.83)	65.56 (64.40, 66.72)	82.78 (81.16, 84.40)	73.15 (72.15, 74.14)	76.67 (75.05, 78.29)	56.44 (54.05, 58.83)	64.96 (63.15, 66.76)
	SVM	70.94 (69.43, 72.45)	66.17 (64.76, 67.58)	85.86 (84.82, 86.90)	74.73 (73.60, 75.85)	79.81 (78.26, 81.37)	56.02 (53.46, 58.58)	65.80 (63.64, 67.95)

Appendix 2 - *Model Performances Stats (Country-Balanced Random Split)*

C.I and mean obtained from 10 Stratified Folds, 8K data in Train & 2K data in Validation per Fold.

		Accuracy	Precision +	Recall +	F1 +	Precision -	Recall -	F1 -
Original Sentence	Logistic Regression	94.95 (94.56, 95.34)	95.17 (93.67, 96.66)	94.83 (92.90, 96.75)	94.94 (94.52, 95.37)	94.94 (93.23, 96.66)	95.06 (93.34, 96.77)	94.95 (94.55, 95.35)
	XGBoost	95.69 (95.49, 95.89)	95.68 (95.15, 96.20)	95.74 (95.24, 96.23)	95.70 (95.53, 95.87)	95.72 (95.21, 96.23)	95.65 (95.09, 96.22)	95.68 (95.44, 95.92)
	SVM	96.08 (95.90, 96.26)	95.90 (95.55, 96.25)	96.29 (95.92, 96.66)	96.09 (95.93, 96.26)	96.26 (95.85, 96.67)	95.87 (95.53, 96.21)	96.07 (95.87, 96.26)
Shuffled Sentence	Logistic Regression	91.57 (91.14, 92.01)	92.07 (90.95, 93.18)	90.95 (89.53, 92.37)	91.48 (90.94, 92.01)	91.17 (89.95, 92.40)	92.20 (91.02, 93.39)	91.66 (91.29, 92.03)
	XGBoost	91.53 (91.10, 91.95)	91.26 (90.51, 92.00)	91.76 (91.45, 92.08)	91.51 (91.02, 92.00)	91.79 (91.55, 92.03)	91.30 (90.65, 91.94)	91.54 (91.15, 91.93)
	SVM	92.31 (91.81, 92.81)	91.81 (91.03, 92.60)	92.82 (92.19, 93.45)	92.31 (91.77, 92.85)	92.82 (92.31, 93.32)	91.79 (90.98, 92.59)	92.30 (91.82, 92.77)