Words that Represent Peace

Tushar Prasad
Graduate Student, DSI Scholar, Data Science Institute, Columbia University,
Peter T. Coleman
Columbia University Director, The Morton Deutsch International Center for Cooperation and Conflict Resolution, Executive Director, AC^4, Climate School
Larry S. Liebovitch
Columbia University, Adjunct Senior Research Scholar, AC^4, Climate School
Melissa Mannis
Columbia University, Project Manager, AC^4, Climate School
Zach Stone and Natalie Zadrozna
Columbia University, Consultants, AC^4, Climate School

SUMMARY

- We identified the most important words in news used by machine learning models to classify a country as peaceful or non-peaceful.
- These words help us understand the underlying social processes that are happening in these countries.
- Using clustering and GPT-4 to group those words, the themes most found in peaceful countries were Finance and Personal Activities; and those in non-peaceful were Politics and Legal Processes.

Sustaining Peace Project

Objectives

Redefining peace by going beyond the traditional notion of the absence of war or conflict. To measure positive peace by identifying the social, political, and economic systems that lead to peaceful societies.

Data

Using a Lexis-Nexis's extensive database to scrutinize over 2 million media articles across 20 countries from 2010 to 2020.

Peace Speech through Machine Learning

Using advanced machine learning we find the words and phrases that are more prevalent in peaceful societies.

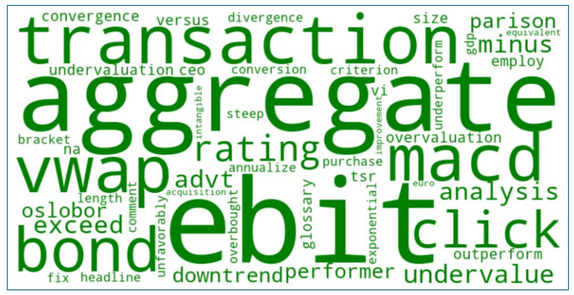
Impact

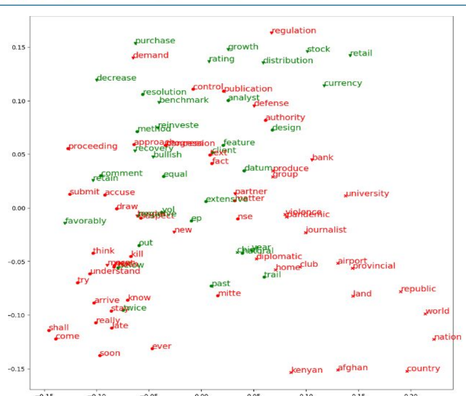
By focusing on 'positive peace,' we hope to identify the structures that create a sustainable and peaceful society.

Training Data and Methodology

- From the Lexis-Nexis dataset, we had 10,000 words from peaceful and non-peaceful countries. We defined the peacefulness of a country using UNICEF Human Development Indexes.
- Peaceful Countries list : Australia, New Zealand, Sweden, Austria, Belgium, Denmark, Norway, FInland, Netherlands, Czech Republic.
- Non-peaceful Countries list: India, Iran, Nigeria, Kenya, Congo, Zimbabwe, Sri Lanka, Uganda, Afghanistan, Guinea
- We removed very commonly occuring stop words. We also decided to take the top most 1000 words by occurrences, since we wanted to avoid overfitting on words which were very rare but particular to a single country.
- We performed normalization within countries and then averaged across countries in peaceful and non peaceful data to get the required training data with n = 2 and d ~ 1000.
- Inorder to find the optimal classification, we tried Logistic Regression, Random Forests, Decision Trees and SVMs(Linear Kernel). We used Optuna to find the optimal hyperparameters and used a hold out country to perform the cross validation.
- We got 100% accuracy for SVMs and Random Forests, and used their coefficients and feature importances to get the relevant results.

Peaceful





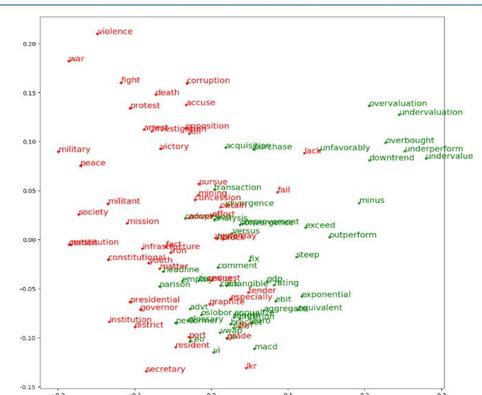
RF Themes Identified by GPT-4

Finance, Personal Social, Health Wellbeing, Tech Development, Markets

SVM Themes Identified by GPT-4

Financial Metrics and Ratios, Corporate and Employment, Market Analysis

Non-Peaceful



RF Themes Identified by GPT-4

Government Public, Legal Conflict, Geopolitical, Media, Analysis Decision

SVM Themes Identified by GPT-4

Political and Social Issues, Legal Processes, Governance, Leadership, Natural Resources

Methodology

SVM Word Cloud

The SVM word clouds are made from the top 75 words contributing to peaceful and non-peaceful countries by their coefficients. The size of the words are dependent on their occurrences in the respective peaceful and non peaceful countries.

Random Forest Word Cloud

The Random Forest word clouds are made from the top 50 words which occur in peaceful and non-peaceful countries respectively. These 50 words have been selected from a set of which which the model considers to have a high feature importance. The size of the words are dependent on their occurrences in the respective peaceful and non peaceful countries. The words in blue are those which occur for both peaceful and non peaceful countries.

PCA + K-Means Clustering

We found the semantic similarity between the words in the SVM and Random Forest word clouds using embeddings (multilingual-e5-large-instruct) and used PCA to transform them into 2-dimensional space. We also performed K-Means clustering with 3 clusters for both the graphs, and manually tried to identify the topics.

SVM Cluster Identification Themes:-

● - Warfare, political strife or news reporting on conflicts
✗ - Stock Market Performance, Company Performance Reviews
▲- Business, Economics, Governance, Policy-making

Random Forest Cluster Identification Themes:-

● - Processes, Analytical methods, Legal and Investigative actions
✗ - Cultural, Political and Global Dynamics
▲- Economic and Financial Activities

GPT-4 Clustering

We also experimented with using GPT-4 to cluster the words from the SVM and Random Forest word clouds. Essentially asking GPT to do topic modeling for us.

Next Steps

- Removing the financial bias in the dataset.
- Performing article level analysis.