PEACE SPEECH

Hongou Liu, Pinyi Yang, Xinfu Su, Yibo Chen, Yuwen Zhang, Ziheng Ru

Larry.Liebovitch, Melissa Mannis, Peter Coleman

COLUMBIA UNIVERSITY DATA SCIENCE INSTITUTE

INTRODUCTION

We are interested in finding the differences in speech between high peace and low peace countries. We have expanded the work from the previous teams on two more robust datasets, **TWITTER** and **NEWS**, and several different metrics with more diverse approaches.
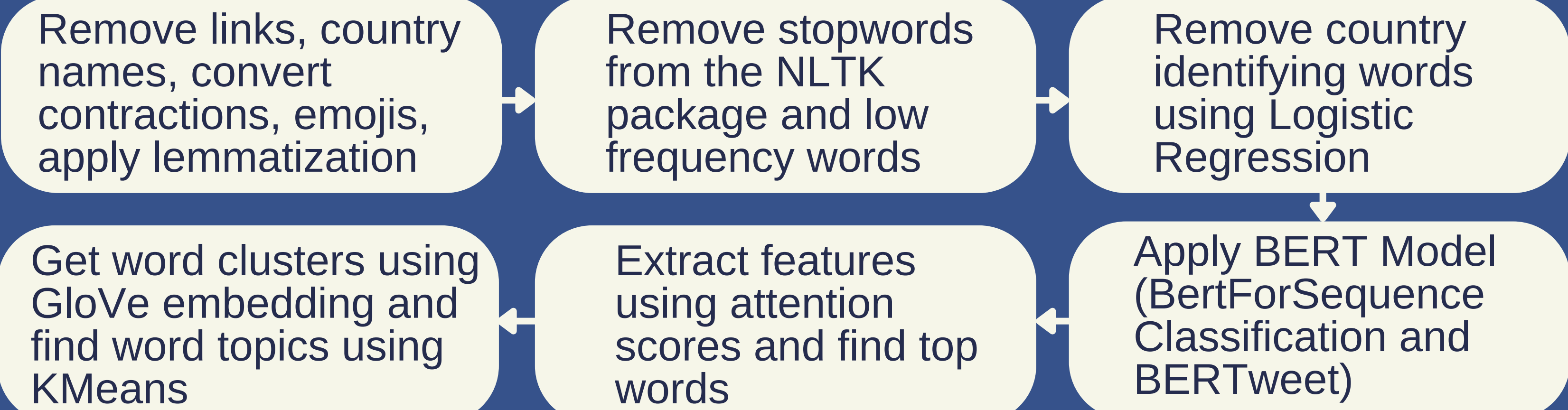
COUNTRY SELECTION

To define High-Peace and Low-Peace Countries, we referred to 5 indexes: GPI, PPI, HDI, WHI, and FSI for 137 countries. We define our peace index using PCA and K-means, and chose 16 countries for the analysis
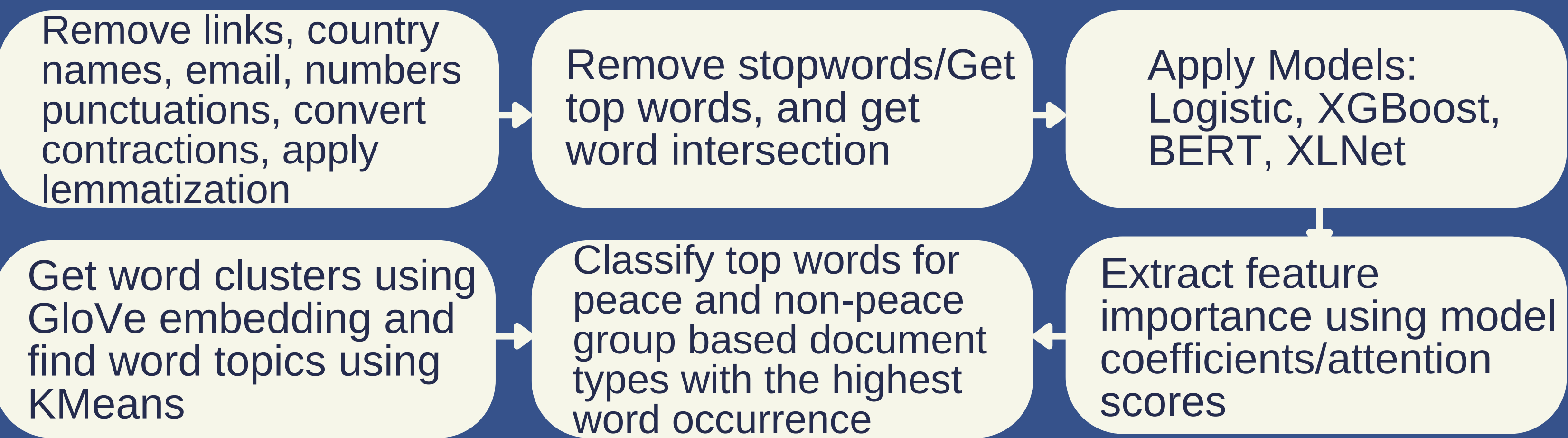
HIGH PEACE	LOW PEACE
Canada, UK, Finland, Norway, Ireland, France, Australia, Singapore	India, Iran, Nigeria, Uganda, Gambia, Libya, Pakistan, Zimbabwe

METHOD

TWITTER

Remove links, country names, convert contractions, emojis, apply lemmatization → Remove stopwords from the NLTK package and low frequency words → Remove country identifying words using Logistic Regression

Get word clusters using GloVe embedding and find word topics using KMeans ← Extract features using attention scores and find top words ← Apply BERT Model (BertForSequence Classification and BERTweet)

NEWS

Remove links, country names, email, numbers punctuations, convert contractions, apply lemmatization → Remove stopwords/Get top words, and get word intersection → Apply Models: Logistic, XGBoost, BERT, XLNet

Get word clusters using GloVe embedding and find word topics using KMeans ← Classify top words for peace and non-peace group based document types with the highest word occurrence ← Extract feature importance using model coefficients/attention scores
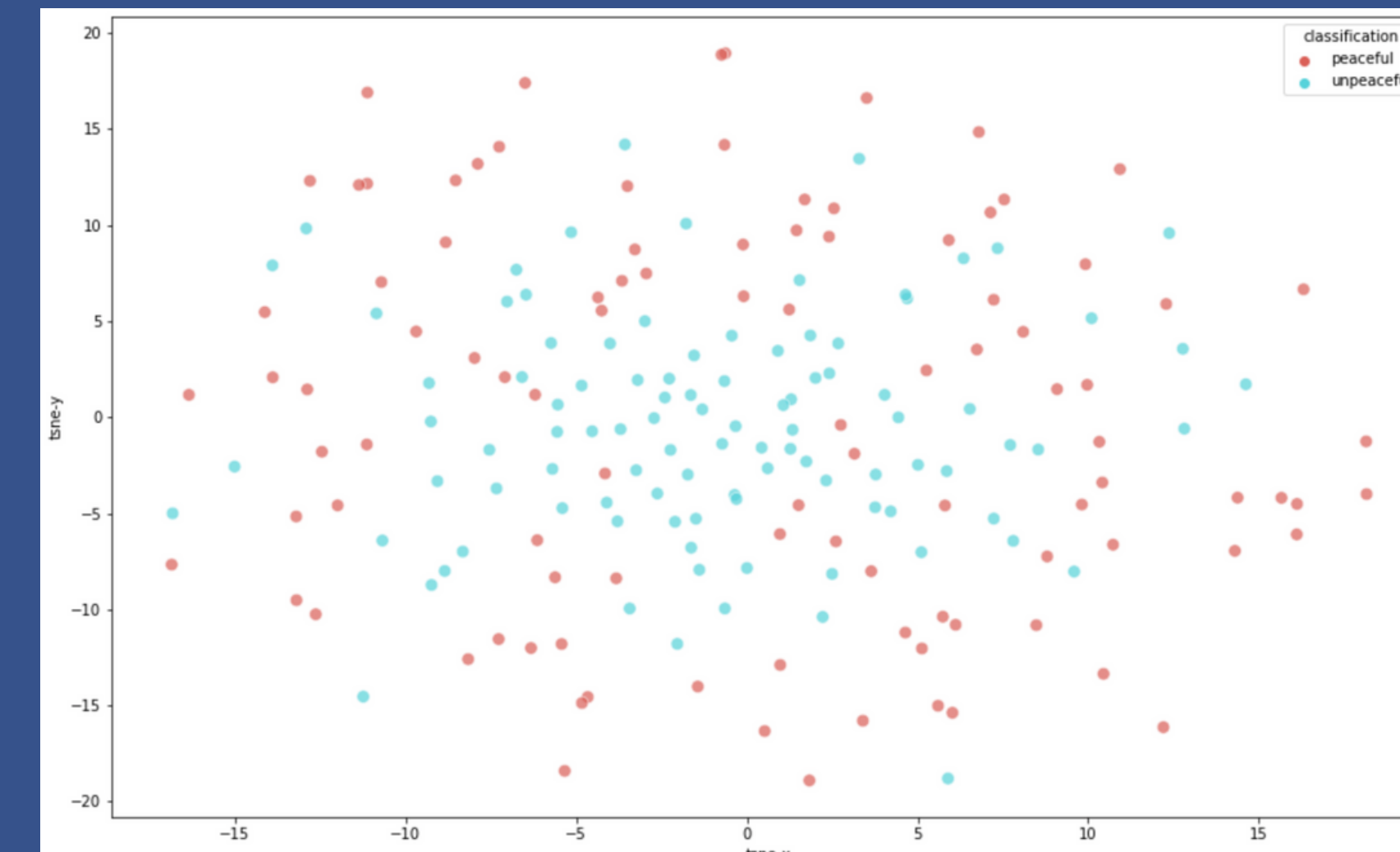

Top 10 words of High-Peace and Low-Peace words based on SVM

RESULT

Words used in high peace countries are more diverse compared with those in low peace countries

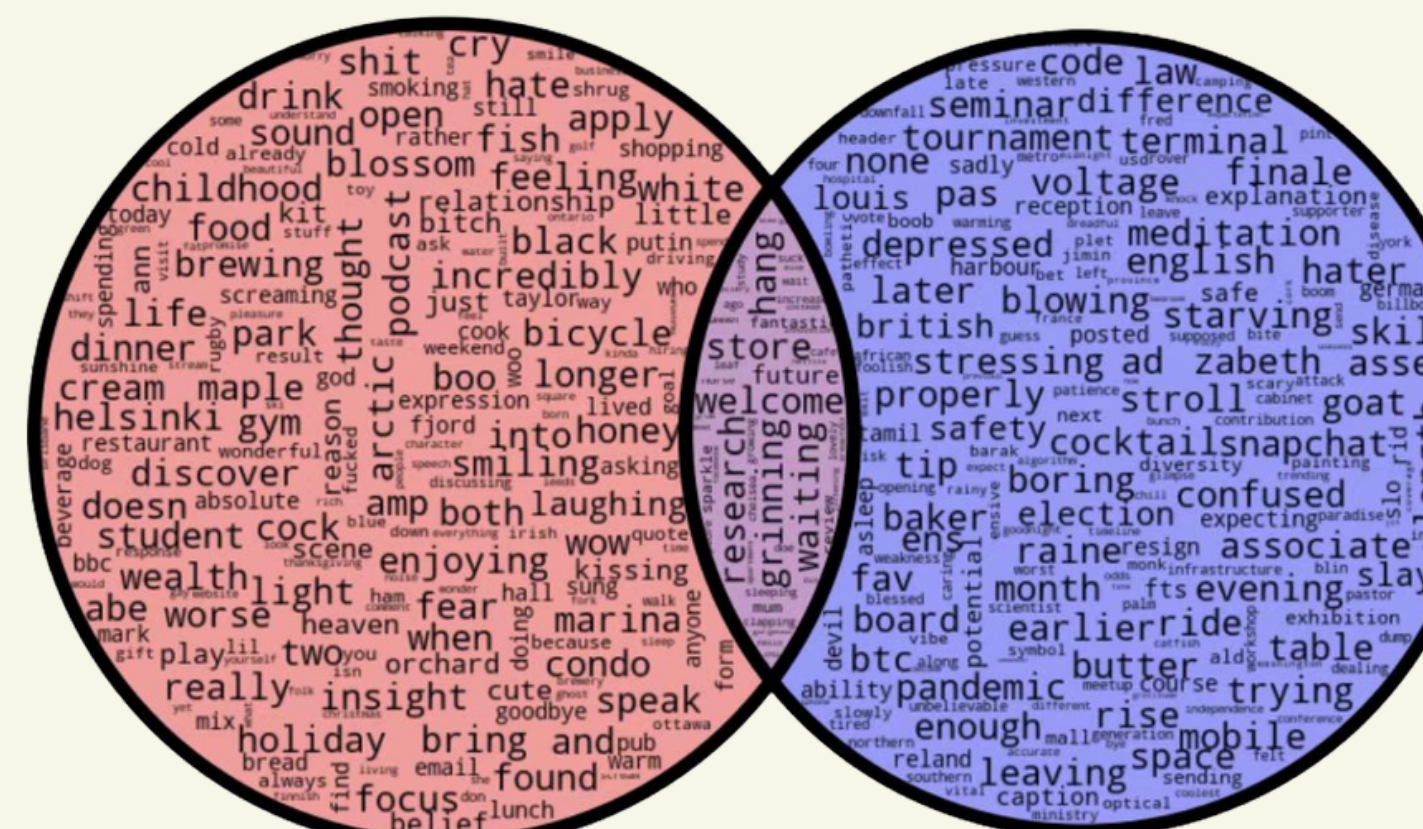t-SNE Visualization of Dimension-Reduced Word Embeddings (News)

HIGH PEACE	LOW PEACE
20.53	10.80

F-Norm of Top Words (News)

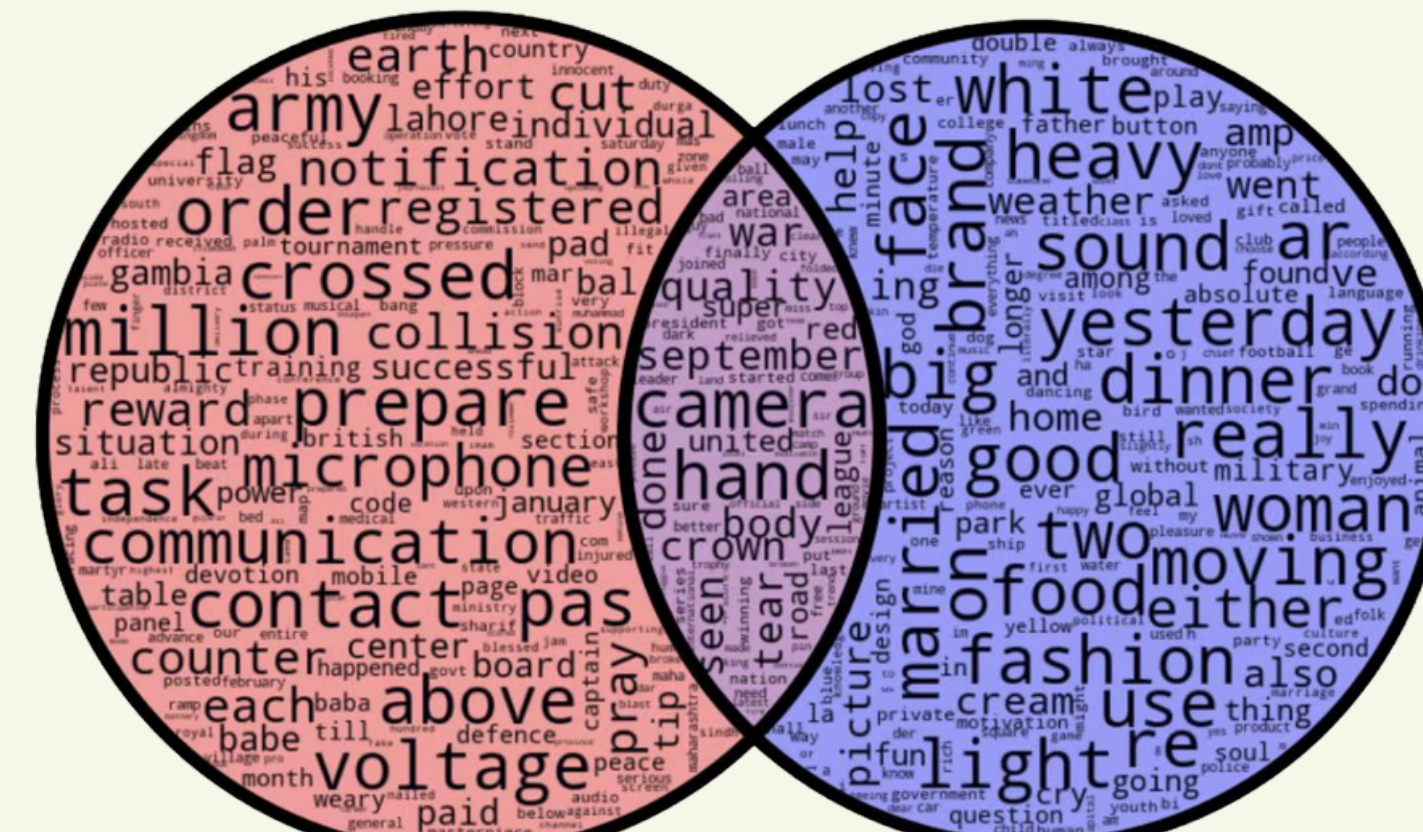|	HIGH	LOW
HIGH	26169	7476
LOW	5852	24490

Confusion Matrix for Bert model (Twitter)

INTERSECTION OF HIGH PEACE WORDS IN TWITTER

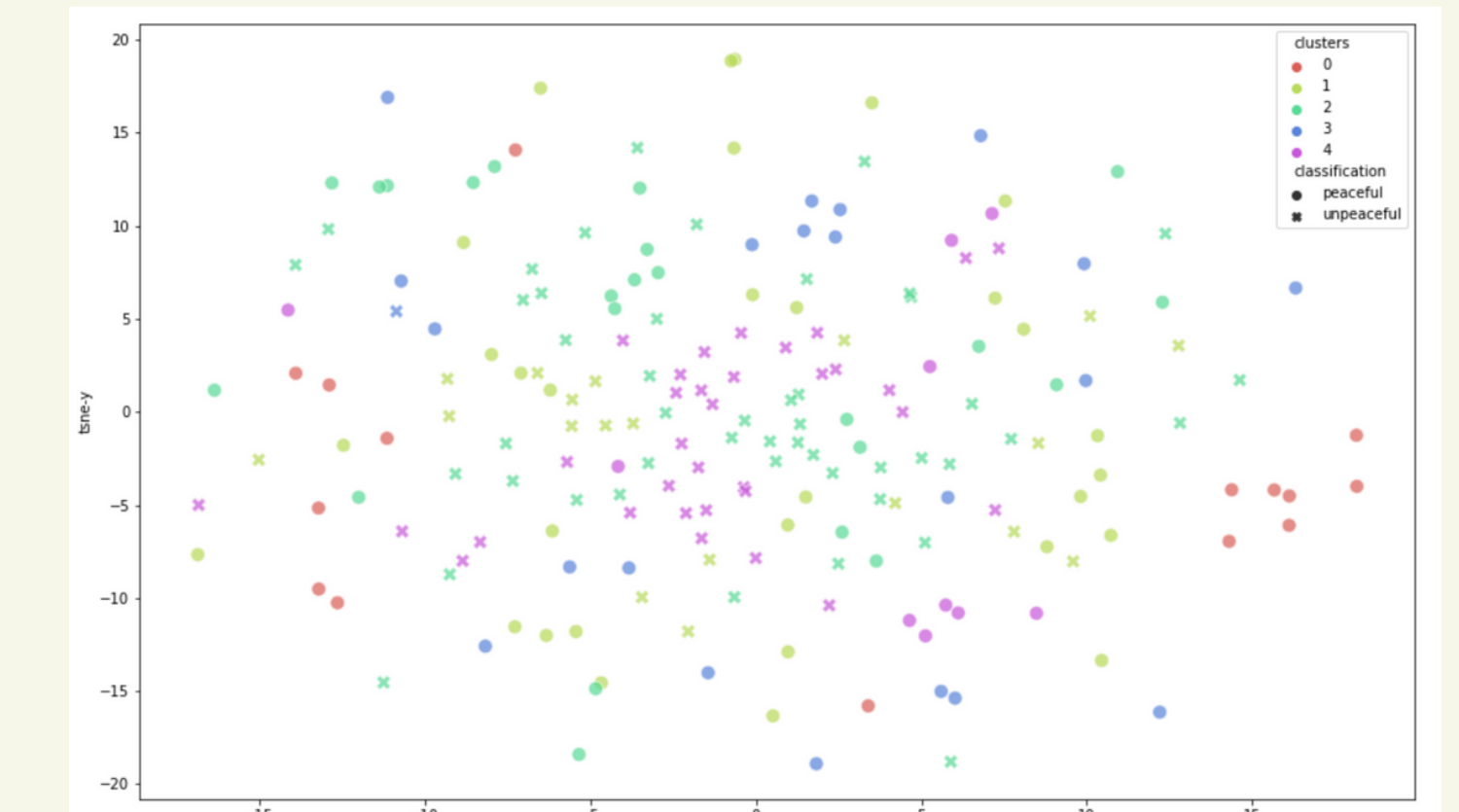BertForSequenceClassification   BERTweet

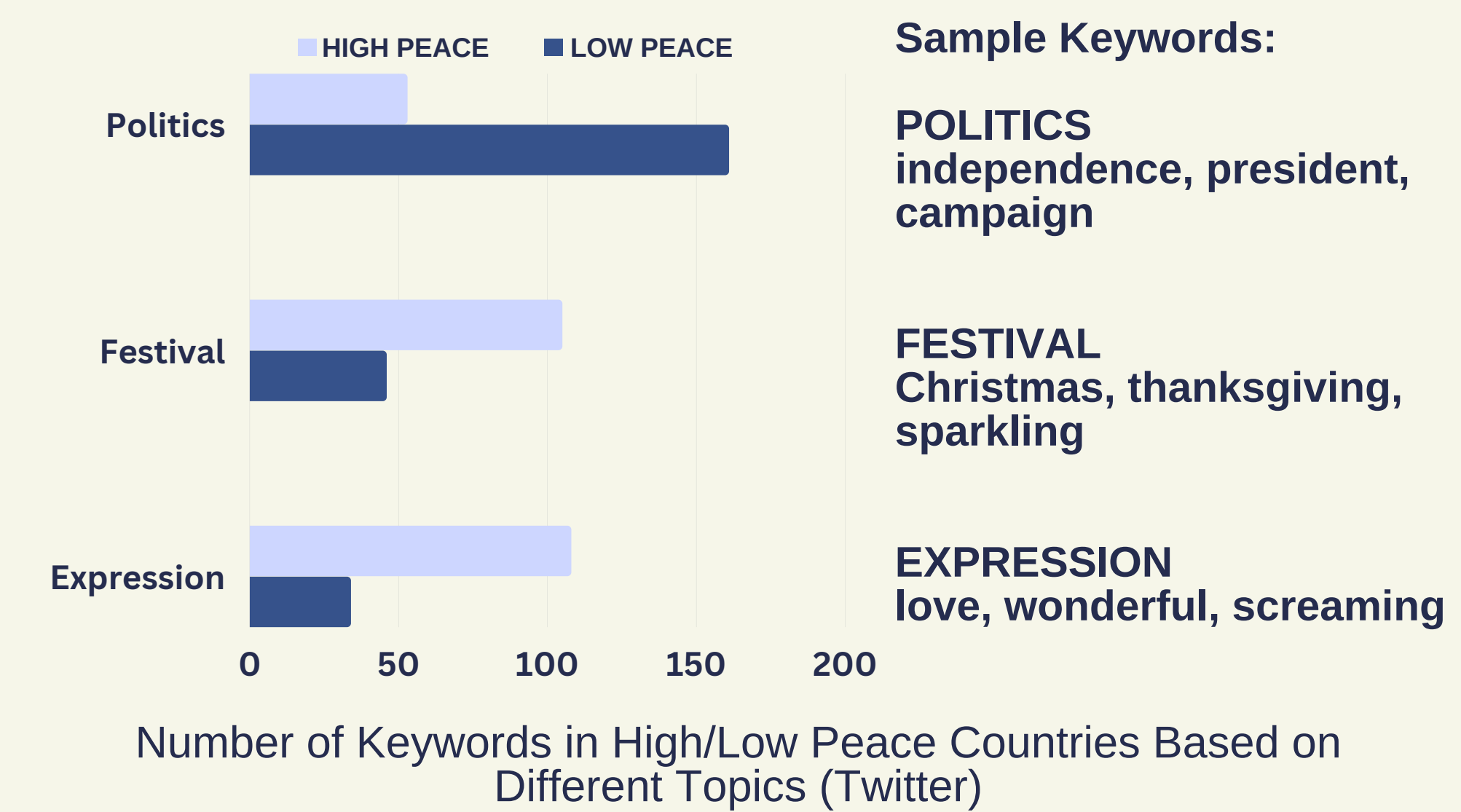### INTERSECTION OF LOW PEACE WORDS IN TWITTER

BertForSequenceClassification   BERTweet

Words used in high peace and low peace countries have different interests in topics

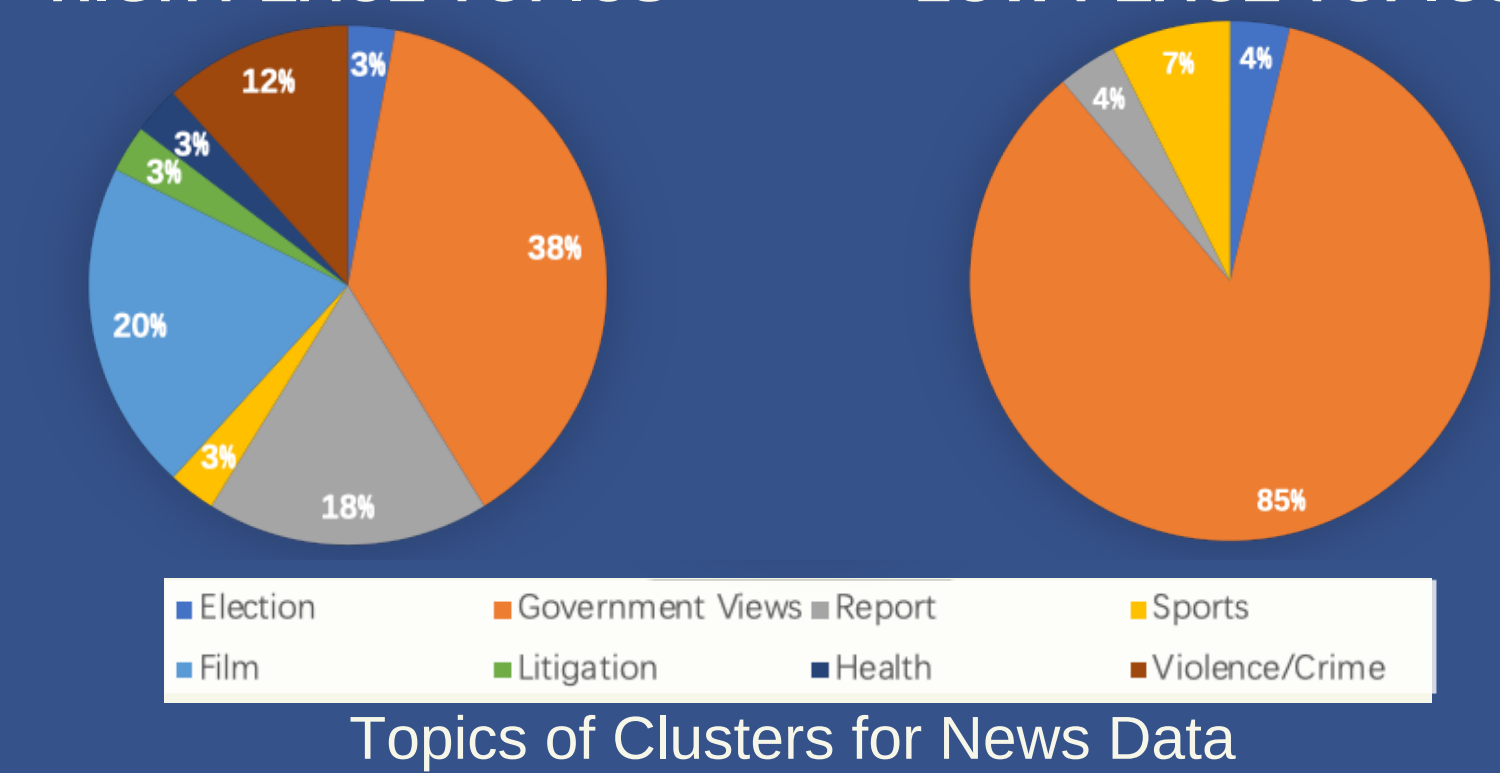- Keywords have distinct distributions in clusters of different topics

t-SNE Visualization with Different Topics: Cluster 3 Represents Business Topic, Cluster 4 Represents Political Topic (News)

Sample Keywords:

POLITICS
independence, president, campaign

FESTIVAL
Christmas, thanksgiving, sparkling

EXPRESSION
love, wonderful, screaming

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

HIGH PEACE TOPICS   LOW PEACE TOPICS

■ Election ■ Government Views ■ Report ■ Sports
■ Film ■ Litigation ■ Health ■ Violence/Crime

Topics of Clusters for News Data

CONCLUSION

- High and low peace countries use different sets of words focusing on different topics

- High peace countries talk about more topics than low peace countries

- High peace countries discuss more about topics in daily lives, such as festivals, films, and feelings, whereas low peace countries focus more on political subjects