ReThink Media

Classifying Speaker Types in News Articles

Researchers:
Carla Zhao, Yash Dave, Sunya Abbasi

Introduction

Our project aimed to build machine learning models to classify quote sources in news articles. This meant building on the work done in previous semesters, engineering specific features that could help discern the speaker group given the quote text, and evaluating the newer models' performance in classifying quote sources.

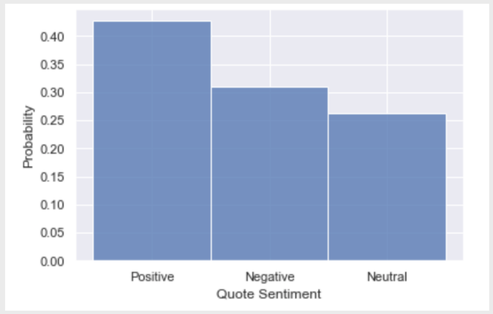
The Dataset

The dataset includes over 400,000 quotes extracted from around 77,000 manually-annotated articles about nuclear weapons issues published in U.S. news outlets from 2011-2022. There are over 30 data attributes with some of the most useful being: the article's full text, the text of quotes within the articles, the name of the quote's speaker, and the type of speaker, e.g. Academic, Blogger, Military, Federal Official, etc.

Feature Creation

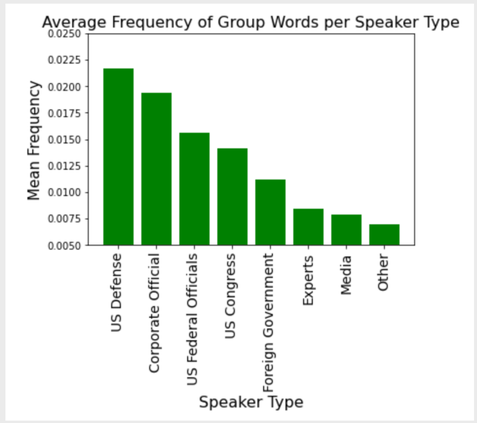
This semester we created 10 new features, some of which include:

Quote Sentiment Analysis



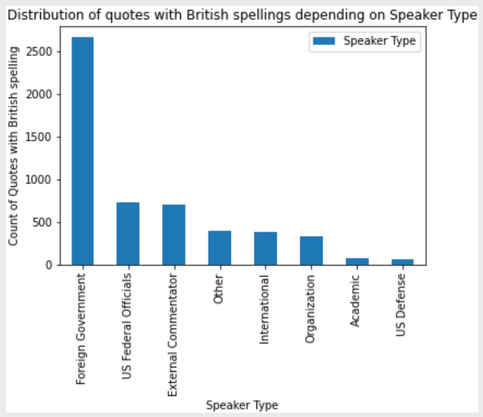
Our quote sentiment feature uses VADER to measure the sentiment of a given quote (negative, positive, or neutral). In order to normalize the sentiment scores, we controlled for quote length by splitting quotes longer than 15 words into multiple spans and averaging the sentiment scores. The chart above shows the sentiments of quotes from "Corporate Official" speakers.

Frequency of "Group" Pronouns



Another feature we looked at was the frequency of the words "our", "ours", "we", and "us" within the quote text of an article. As shown by the visual above, instances where the speaker is representing a group have higher frequencies of the listed words.

Occurrences of words with British spelling



Using the insight that the British and American spelling of words differ (eg. color vs colour) and might give a clue as to the speaker group, the 'count British words' feature was developed. The plot above displays how the usage was distributed across the various speaker groups.

The Model

We added on to models created by last semester's team by creating new features to improve performance. Last semester's team tried several different classifier models including Logistic Regression, Random Forest and Support Vector Machine. The Support Vector Machine model achieved the highest accuracy.

Classification Model	precision	recall	F1
Logistic Regression	0.62	0.62	0.61
Random Forest	0.58	0.60	0.59
SVM	0.63	0.63	0.62

We found that our two most important features were features measuring quote sentiment and quote subjectivity. After hyperparameter tuning, the accuracy of the Support Vector Machine model increased by around 1%.

Takeaways

1) The preprocessing stage of classification significantly influenced the performance metrics obtained in non-trivial ways – dropping certain rows could improve or hurt the accuracy of the classifier. The rows dropped were a hyperparameter to tune when evaluating model performance

2) Creativity drives feature engineering, especially when dealing with text as data. The patterns in text and grammar are more difficult to see, but can be very useful in picking up mannerisms broadly observed in particular speaker types

Next Steps

1) Trying even more classification models other than Logistic Regression, Support Vector Machine and Random Forest Classifiers

2) Experimenting with more state-of-the-art techniques of natural language inference (eg. using BERT to create word embeddings) and discerning whether they could work at scale