LOCO Project

Language **O**f **CO**nspiracy

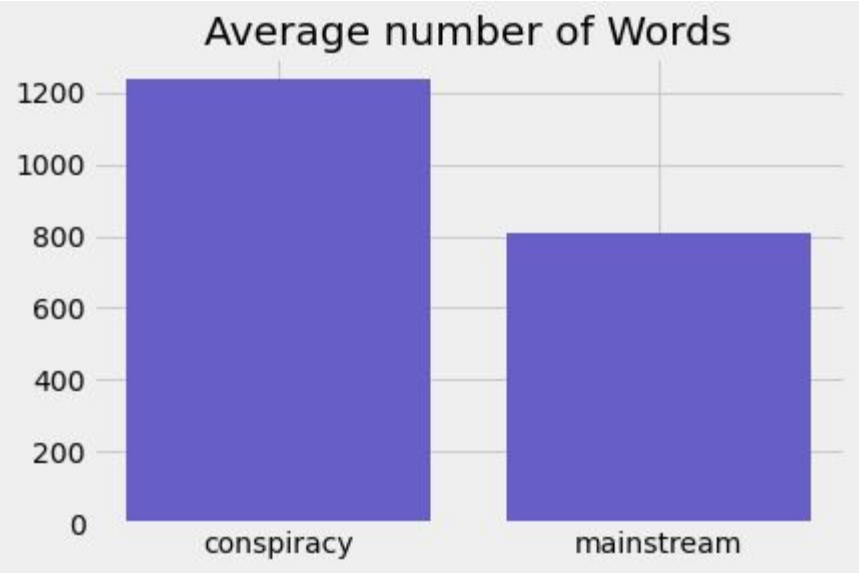
By Marcel Colvin

Intro

- Can we create a model that is able to distinguish whether an article is spreading a conspiracy theory or is taking the mainstream viewpoint
- Researchers may find our model interesting for discovering what conspiratorial topics are being written about
- Large social platforms may be interested in the model to more effectively flag false information on their websites.

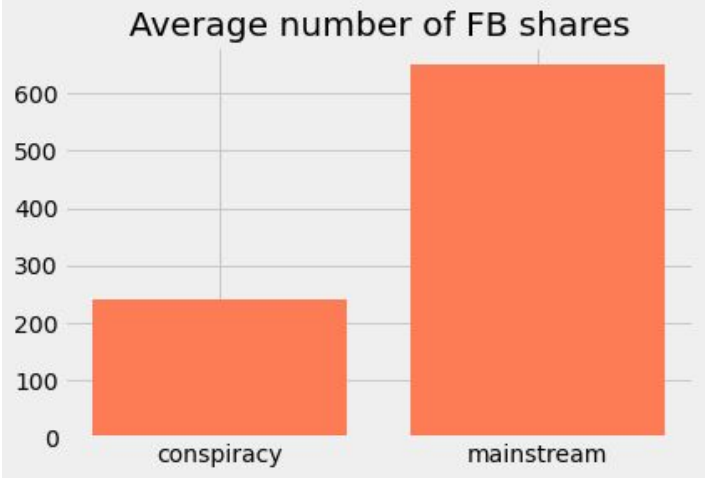
Methodology

- Data: The dataset is 88 million articles, about 20 million conspiracy, 70 million mainstream. Contains text and some social data.
- Tools:
  - Pandas, numpy for feature engineering
  - Sci-Kit Learn for topic modeling (NMF) and classification
  - RE and NLTK for text pre-processing and sentiment analysis
  - Matplotlib for plotting



Methodology

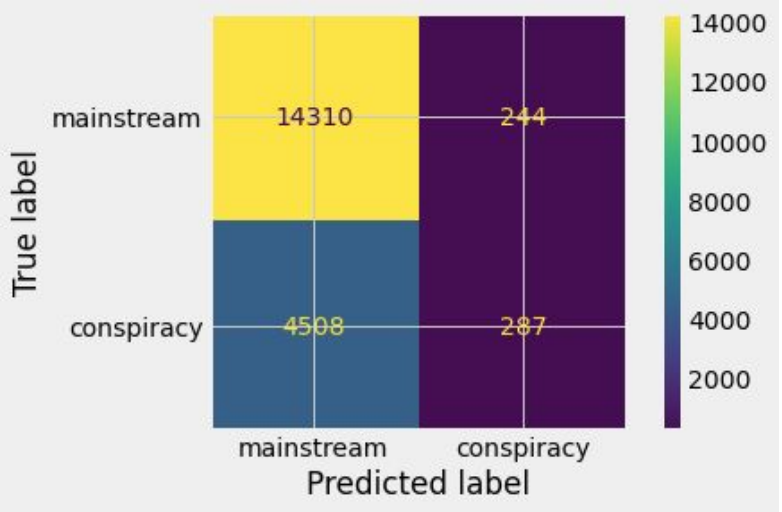
- Metrics
  - Recall is the most important metric for the model, because we want to reduce the number of false negatives (mainstream article, but identified as conspiracy)
  - Used Receiver Operating Characteristic Area Under Curve (ROC AUC) to optimize models.
- Models:
  - Logistic Regression
  - Random Forest Classifier (RFC)



Results

- The models took in 20 topics from the Non-negative Matrix Factorization, plus facebook shares, likes, and word count.
- Logistic Regression performed okay, but when the hyperparameters were optimized, it did not perform as well as the RFC

Un-tuned RFC model Confusion Matrix



Results

- After tuning the optimized RFC model, the accuracy was high (~80%), but the recall was still only at 50%. Moved threshold to 40% instead of 50%

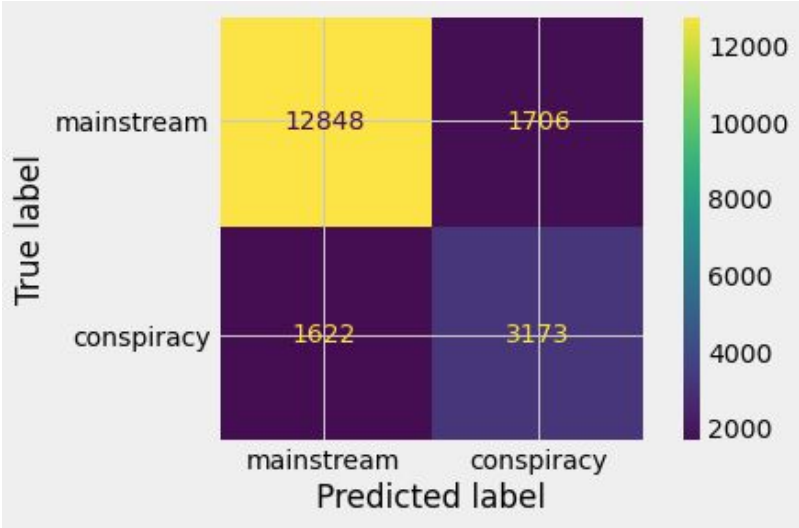
Accuracy: 0.83

Precision: 0.65

Recall: 0.66

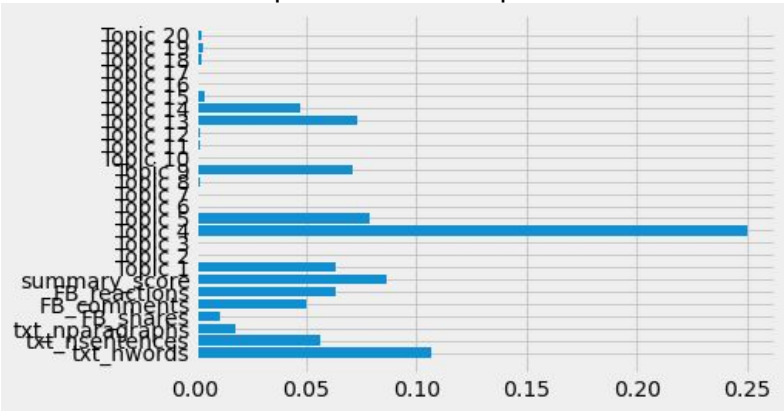
F1: 0.66

Final RFC Confusion Matrix

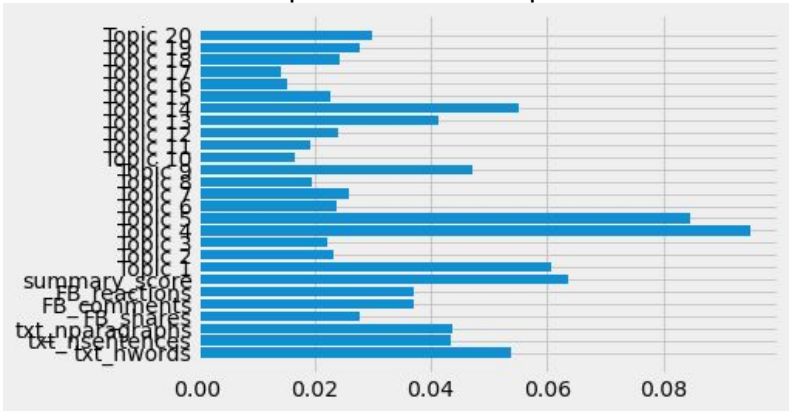


Results

RFC Feature Importance Pre Optimization



RFC Feature Importance Post Optimization



Topic 1
people, like, time, know, think, god, going, way, world, life

Topic 4
world, states, war, united, new, government, united states, american, order, international

Topic 5
said, told, year, people, according, coronavirus, country, officials, state, jackson

Topic 9
11, bin, laden, bin laden, cia, attacks, al, intelligence, september, 2001

Topic 14
trump, president, obama, house, clinton, campaign, white, state, american, news

Conclusions

- Machine learning models are useful in identifying conspiratorial articles, but may need more tuning with more topics and lexical features outside of the data set.
- The models can perform well in some capacity, but may require more updated data to continue to operate at a high level

Future Work

- Start creating new lexical features or using some from the LOCO_LF data set to test new models
- Try different topic modeling techniques
- Use a Clustering model to see if it can work well on the corpus

Thank You