WEB APIS & NLP: TRUTH OR TRUTH

Brian Pinto

Table of Contents

- Background Information
- Problem Statement
- Exploratory Data Analysis
- Modeling
- Conclusion

Background Info

r/conspiracy

- Over 21 million members
- 9th Largest Subreddit
- Created Jan 25, 2008

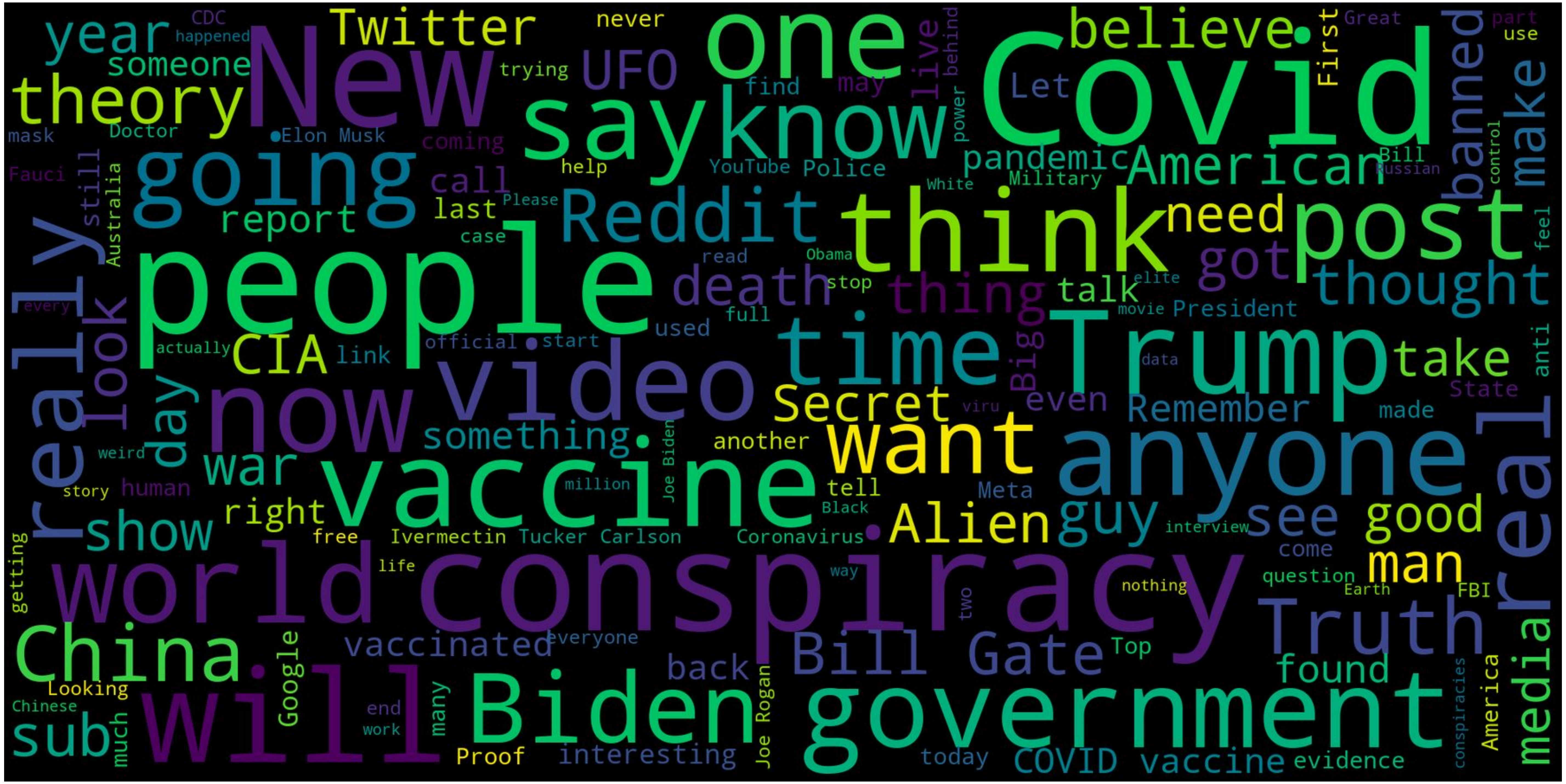
r/news

- Nearly 2 million members
- 150th Largest Subreddit
- Created Jan 25, 2008

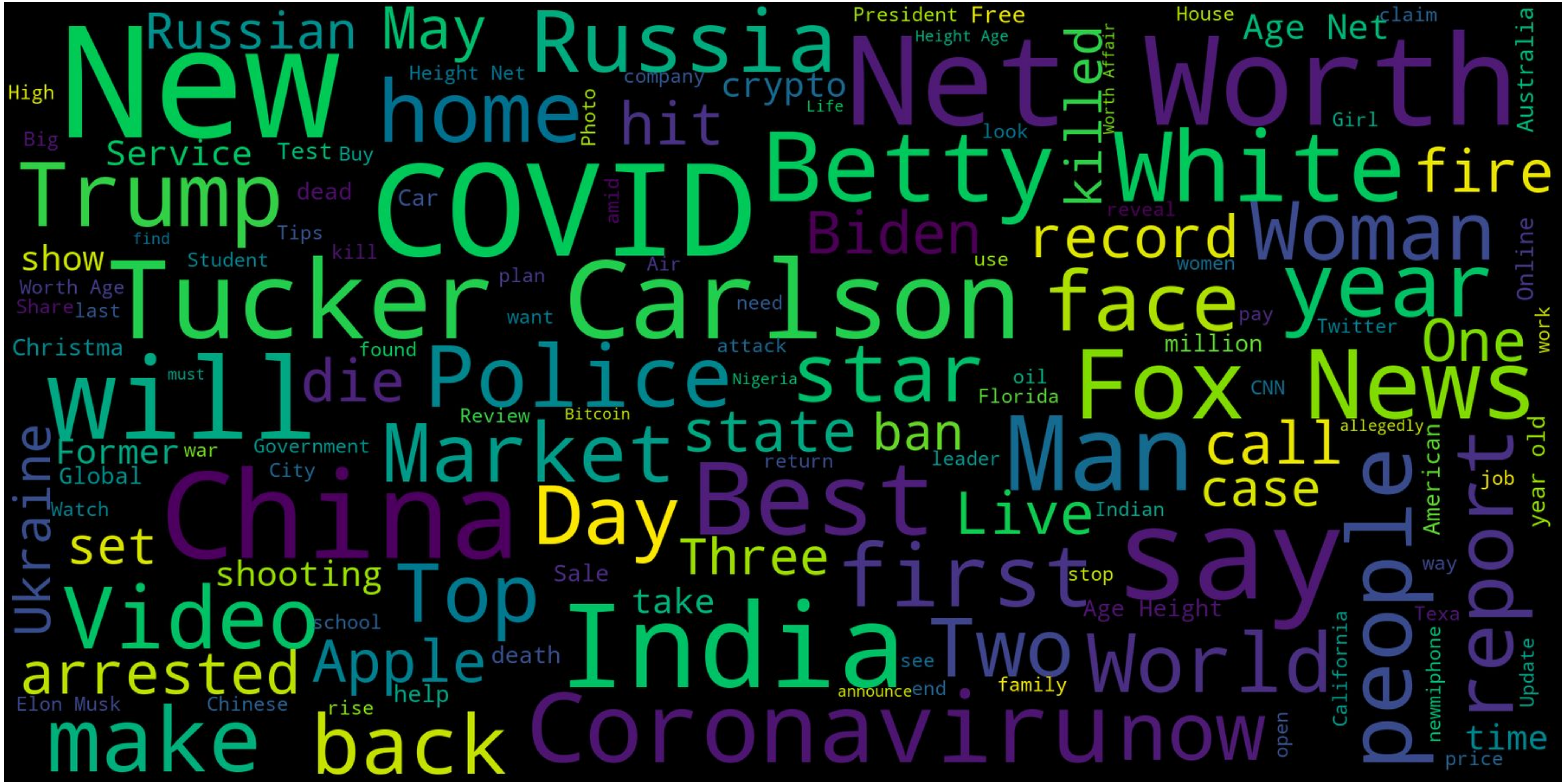
Problem Statement

The aim of this project is to build a model that can take text content from two different subreddits and accurately classify the origin of each piece of content. The main metrics to maximize are both the accuracy and F1 scores of the models.

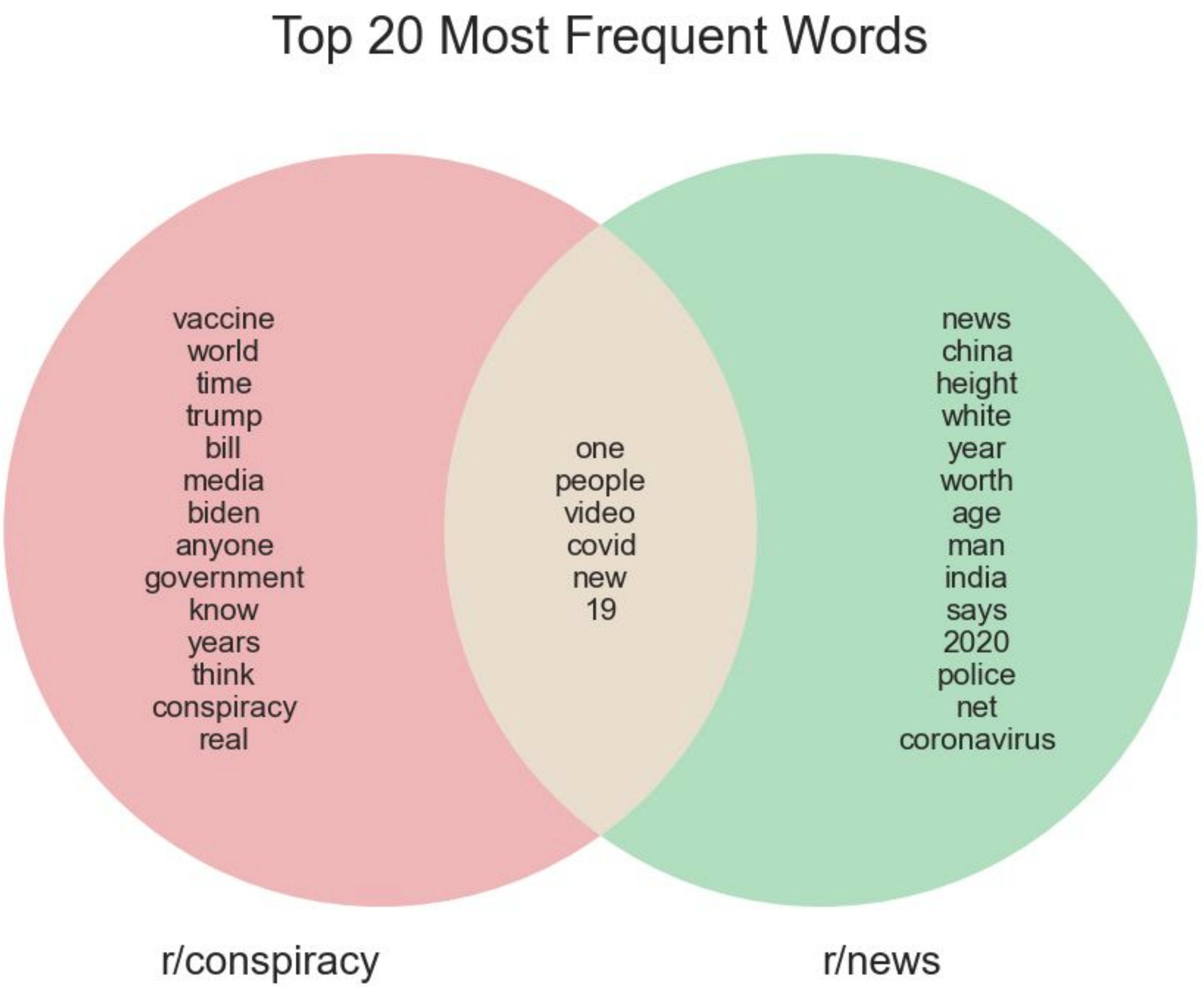
Exploratory Data Analysis



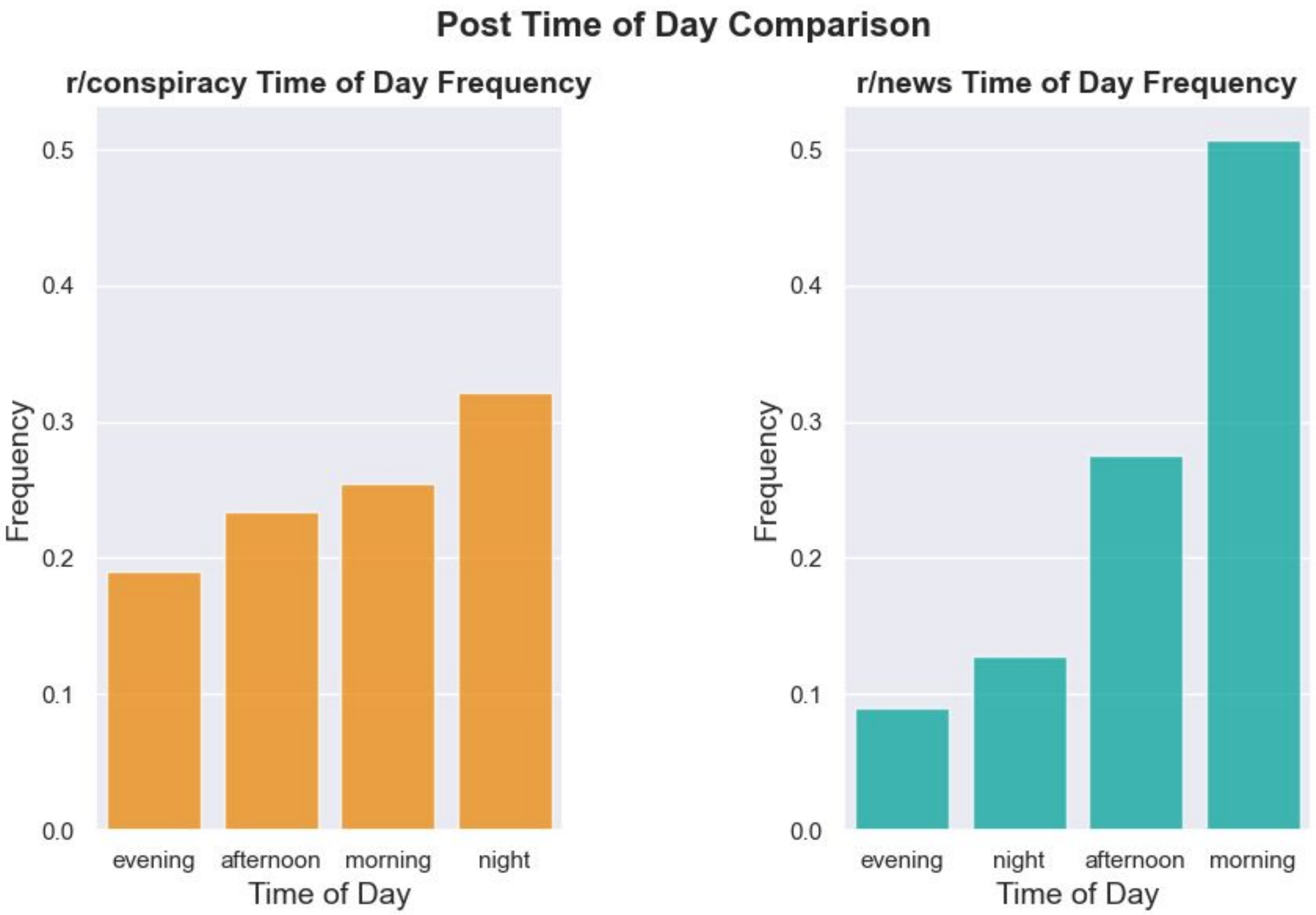
Exploratory Data Analysis



Exploratory Data Analysis



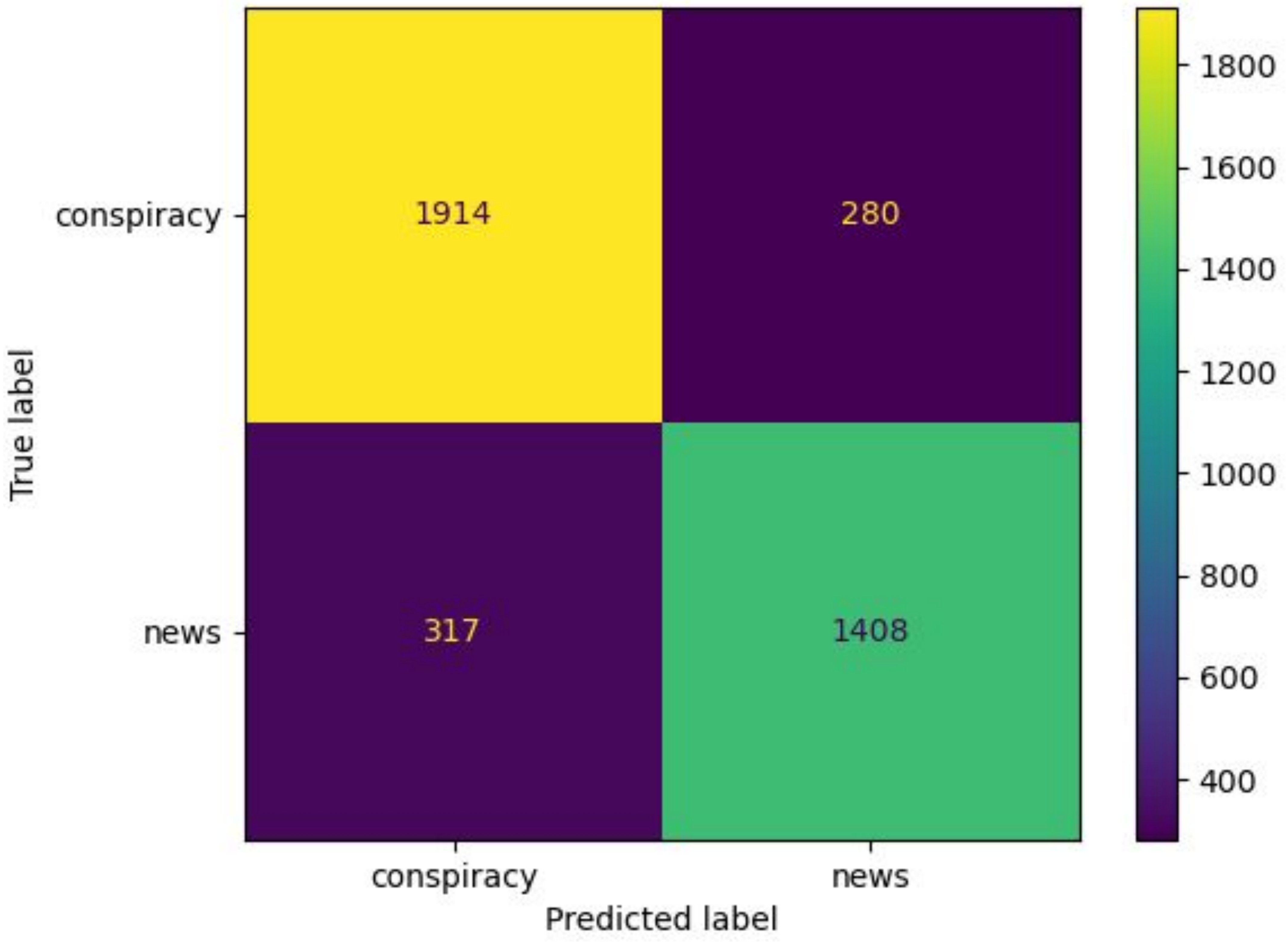
Exploratory Data Analysis



Modeling

Model	Training Accuracy	Testing Accuracy	F1 Score
Logistic	.892	.842	.822
MNB	.889	.848	.825
XGBoost	.9	.810	.786
Voting Classifier	.902	.847	.826

Modeling



conspiracy
news
True label
1914
280
317
1408
conspiracy
news
Predicted label
1800
1600
1400
1200
1000
800
600
400

Conclusion

- MNB chosen to be production model
- Solid foundation but needs improvement before deployment
- Changes:
  - Feature Engineering
  - Hyperparameter Tuning
  - Data Collection and EDA
  - Explore other models