Detecting Fake News Using a Machine Learning Model Based on Lexical Characteristics of Text

Anne Wu
Advisor: Mr. Nicholas Medeiros

Introduction

- For many, our main source of news and current events is the internet, but it isn't always accurate (Zhang et al., 2019).
- Due to our increased reliance on social media as a source of news as well as its ability to spread news quickly, the proliferation of fake news can be dangerous and difficult to control.
- If we can quickly identify that an article contains fake news, we can prevent or mitigate its spread before it reaches a wider audience (Amoruso et al., 2020).

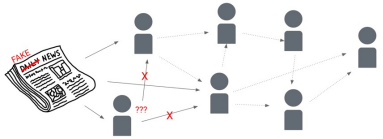


Figure 1: figure showing how the spread of fake news can be mitigated if an article is quickly identified as fake.

	Manual Fact Checking	Automated Fake News Detection
Pros	- Dependable - Accurate Assessments - Can provide reasons	- Quick - Many methods - Inexpensive
Cons	- Time-consuming - Expensive	- Accuracy varies - Currently difficult to provide reasons for assessment

Research Question

Can we effectively classify fake news purely through analyzing the lexical qualities of text?

Hypothesis

Using Linguistic Inquiry and Word Count (LIWC), the vocabulary of various trustworthy and untrustworthy articles can be analyzed by machine learning models in order to effectively distinguish between the two. Certain categories of words, specifically those of the dictionaries in LIWC, are more prevalent in fake news compared to real news and vice versa.

Dataset used

Horne 2017 Fake News Data (Horne et al.)

- Dataset containing 251 news articles, with 123 identified as fake and 128 identified as real
- All news in dataset are from the year 2016 and majority pertain to politics
- All body text from the 251 articles was used for this project

Methodology

All news articles are processed through LIWC, which outputs the percentage of words that belong to certain word categorizes and other lexical information about each text.

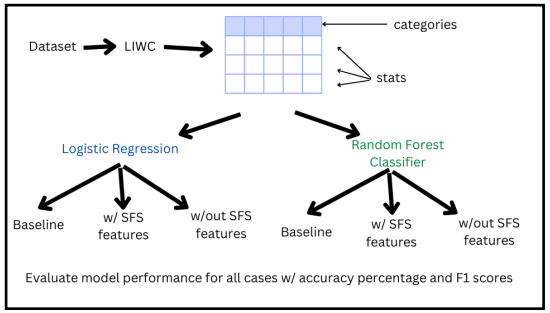


Figure 2: infographic showing methodology process for project.

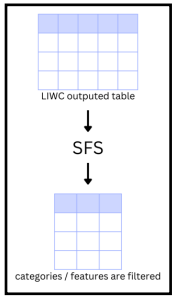


Figure 3: figure describing sequential feature selector, a method in machine learning for feature selection.

Using both machine learning algorithms, 3 tests for model performance were done:

Baseline	Features chosen by SFS	Features not chosen by SFS
Uses entire set of features with no filtration or alterations.	Uses set of features filtered down to features deemed important by SFS.	Uses set of features filtered down to features not chosen by SFS.

All utilized methods use 5-fold cross validation

Results

Model Performances

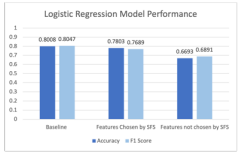


Figure 4: a A bar graph showing the model performance of logistic regression for baseline, using features chosen by SFS for data, and using features not chosen by SFS for data.

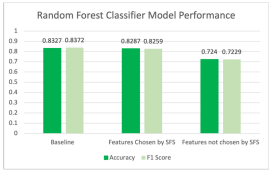


Figure 5: a A bar graph showing the model performance of random forest classifier for the baseline, using features chosen by SFS for data, and using features not chosen by SFS for data.

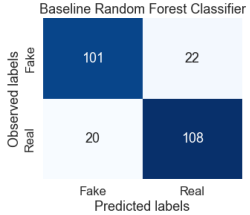


Figure 6: A confusion matrix showing the results of the baseline random forest classifier model, the model that had the best performance

SFS Features Strengths

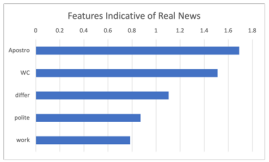


Figure 7: A horizontal bar graph showing the feature strengths of features chosen by SFS that indicate a strong correlation to real news. The higher the value, the more correlation there is.

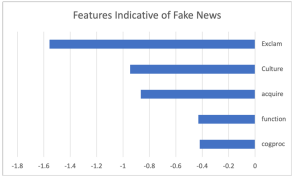
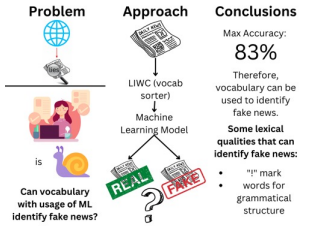


Figure 8: A horizontal bar graph showing the feature strengths of features chosen by SFS that indicate a strong correlation to fake news. The higher the value, the more correlation there is.

Graphical Abstract



Analysis

- The Random Forest Classifier performs better than Logistic Regression in all cases
- All models had slight decreases in accuracy when using data filtered from features chosen by SFS
- All models had significant decreases in accuracy (more than 10%) when using data which was filtered from features NOT chosen by SFS
- Certain features in news, such as word count, apostrophes, exclamation marks, and words used for grammatical structure (function), are more influential in determining if an article is real or fake

Future Work

Use dataset with larger scope

- Current dataset is small and only contains news made in 2016
- Future work can use / make dataset with larger time frame of news

Use a non-binary classification method like in Rashkin et al.

- Current work only uses two categories: "real" and "fake"
- This does not account for news that may lie in between these categories
- Could use categorization method similar to that of PolitiFact.

Conclusion

Though the accuracy of the models aren't perfect, 83% accuracy is sufficient to show that lexical characteristics can indicate if a news article is real or fake.

Using SFS, certain features have also been shown to have a larger influence on whether a news article is fake or not