CS 6830
Andrew Curtis
Saichand Thota

Project 5 Report:
Generalizability of a Naive Bayes Fake News Detector

Github Repository: https://github.com/AndrewCurtis27/CS6830_Proj5

Presentation Slides:
https://docs.google.com/presentation/d/1Lj8_RubGZ7IdVoOVxwRxmuy-dwU2BNBW30MyYDStWp8/edit?usp=sharing

Introduction

The goal of this project was to train and test a fake news detector using a Naive Bayes classifier. We then use a trained model to test on another fake news dataset to see if the classifier model will have generalizability. This is an interesting question because the news cycle can change, techniques of writing fake news can change, and the time period and methods of collection can also vary. Having a stable and generalizable fake news detection model would be beneficial as you would not need to collect data as often or spend compute time retraining new models. So how well does the Naive Bayes algorithm perform for this task?

Dataset

There were two news article datasets used for this project, taken from kaggle with collection documentation described below.

The first is called the ISOT Fake News Dataset, documentation for which can be found at: https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/
Citation for this dataset can be found in references. This dataset contains 45,000 articles, true news articles were collected from Reuters and the fake news articles from various sources. The articles are mostly taken from 2016-2017 with the total dataset ranging from 2015-2018.

The second dataset is called the WEL Fake News Dataset, documentation for which can be found at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9395133
Citation for this dataset can be found in references. This dataset contains 72,000 articles, articles were collected from a range of sources including: Kaggle, McIntire, Reuters, and BuzzFeed Political.

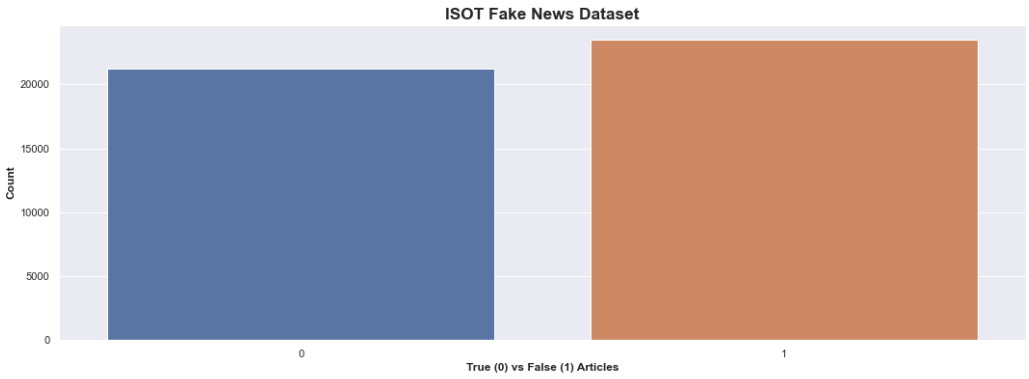
Analysis Technique

Several techniques were used to complete exploratory data analysis as well as for analysis of our naive bayes model. For EDA, we chose to use histogram plots to visualize counts of real

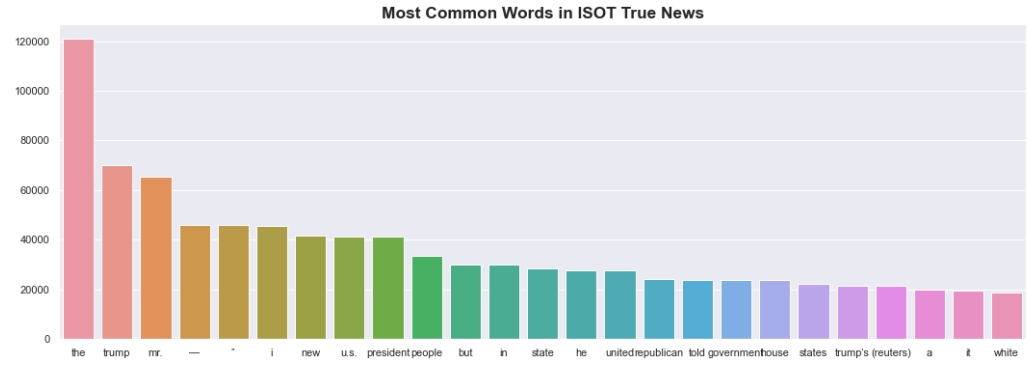
and fake news articles. Frequent words were calculated and count plots were used to visualize those frequent words. This helps show what is present within real vs. fake news articles. Sklearn multinomial naive bayes was used to perform classification of labeled real vs fake news articles. Accuracy, Precision, recall, f1 score and support is calculated.

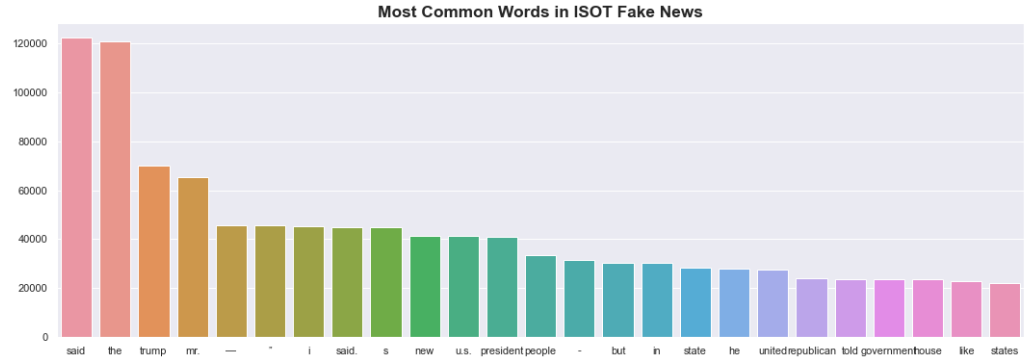
Results

The ISOT dataset contains a fairly close balance of true and false news articles with a total of about 45,000 non-duplicate entries.



Top common words between the true and fake news articles in the ISOT dataset have many similar entries. This makes it seem as though it might be difficult for the naive bayes algorithm to take the words from the text and find features that predict well for classification of real vs fake. So let's see how the modeling performs.



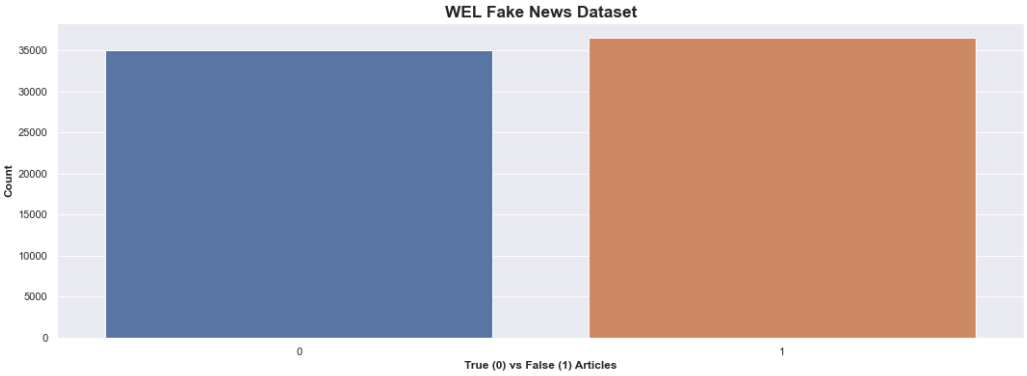


After performing the model training on 80% of the data and testing on 20%, the naive bayes modeling achieved an accuracy of 93.4% on the dataset. For our other statistics

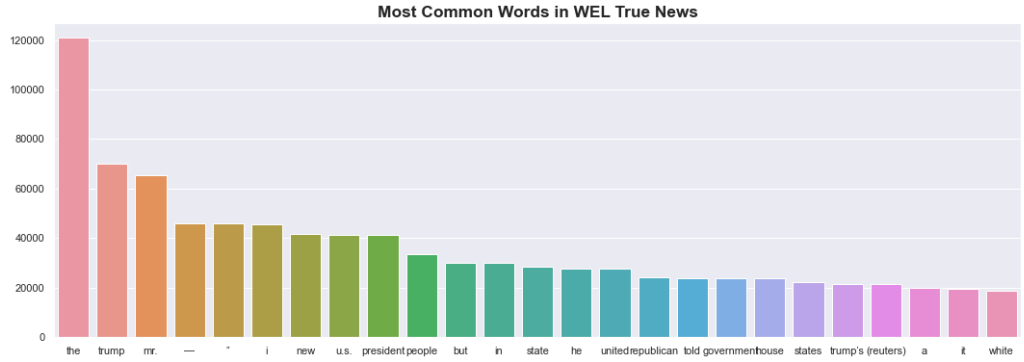
	precision	recall	f1-score	support
True	0.94	0.93	0.93	4200
Fake	0.94	0.94	0.94	4738

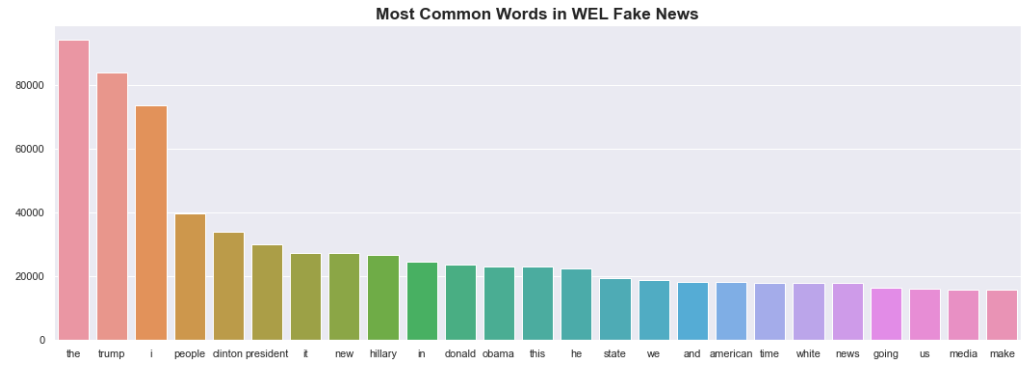
Perhaps surprisingly, our model did fairly well on predicting for real vs. fake news. But the question is, how well will it generalize for our second dataset?

Again the distribution of real vs. fake news articles is balanced for the WEL news dataset, containing around 72,000 articles in total.



This time for our common word counts there appear to be a few more differences between the true and fake news as compared to the ISOT dataset.





For validation, the WEL fake news dataset was analyzed by both performing a train test split and modeling on itself for prediction and taking the pretrained model from the ISOT dataset and only doing predictions with that.

When naive bayes was modeled using a train test split and the WEL fake dataset itself, it achieved an accuracy of 86.8% The remaining statistics were as follows:

	precision	recall	f1-score	support
True	0.86	0.87	0.87	7081
Fake	0.87	0.86	0.87	7227

When tested with the ISOT model the accuracy dropped to 80.1%. With the remaining statistics as follows:

	precision	recall	f1-score	support
True	0.86	0.71	0.78	35028
Fake	0.76	0.89	0.82	36509

The results show that the overall accuracy only decreased about 7% when using the fake news detection model from the first dataset on the second. The generalizability overall seems good and still is able to capture most of the fake news articles for classification. Interestingly when

comparing the model trained on itself vs. the ISOT model, precision on true articles did not drop at all on true news articles. The conclusion is that fake news detection using the naive bayes classifier can generalize and be useful between datasets but will likely lose some performance.

Technical

Both datasets were preprocessed first by using Sklearn count vectorizer and removal of stop words. We also chose to use a tf-idf encoding for the text before passing to the model. The ISOT model was trained first with a 80/20 train test split on the data. The Sklearn multinomial naive bayes classifier was then used and is appropriate for our 0/1 encoding of true and fake news articles. The analysis was then conducted from predictions using this model. First on the ISOT data set. Then using the same pre-trained model we tested for predictions on the full WEL fake dataset. Finally for validation, we did the same train/test approach training a model for the WEL fake using its own data. Analysis was conducted for both sets of predictions on the WEL fake dataset.

References

1. Ahmed H, Traore I, Saad S. “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

2. Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127- 138).

3. Pawan Kumar Verma, Prateek Agrawal , Ivone Amorim, Radu Prodan. “WELFake: Word Embedding Over Linguistic Features for Fake News Detection”. IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, VOL. 8, NO. 4, AUGUST 2021.