Running head: FAKE OR REAL: AN ALGORITHM TO FIND FAKE NEWS

Fake or Real: An Algorithm to Find Fake News

Jessica Everett

Bellevue University

October 18, 2020

Abstract

Confirmation bias happens when our beliefs influence our interpretation of the world around us (Casad, n.d.). For example, if we believe someone is an evil person, we might believe a fake news story about evil things they have done. This is exactly what happened with a fake news story known as Pizzagate. One man, convinced that the story about Hillary Clinton heading up a secret society of child sex traffickers was true, shot up a pizza place thought to be the headquarters of her evil society. Thankfully, no one was injured, but this story illustrates just how dangerous fake news stories can be. Is there a way to stop fake news stories before they cause damage? One way to do this is to use machine learning algorithms to stop fake news from being spread on social media sites.

Fake or Real: An Algorithm to Find Fake News

The damaging effects of fake news are personified by the pizzagate shooting. No one was injured, but it shows how believable and dangerous fake news articles can be. They can spread false conspiracy theories, libel, and defamation. Fake news may have affected the 2016 election. Is there a way for social media companies to help prevent the spread of fake news?

Data Source

I used two datasets from Kaggle: one containing fake news stories and another containing real news stories (Bisaillon, 2020). Both datasets include four columns: the title, text, subject, and date of the news story. There were 23,502 fake news stories and 21,417 real news stories. The date range for the news articles is January 2016 through December 2017. The true news stories were collected from official Reuter's sources. The fake news stories were collected from known fake news sources as determined by PolitiFact and Wikipedia.

To work with the data, I removed the subject column because the fake and real news story datasets used different naming conventions. I also removed the date column as irrelevant. I created a new column that showed whether the news story was fake or real so I could combine the datasets for analysis. The title and text of each news story was combined for analysis.

Methodology

I completed some basic exploratory analysis on the data sets. Here are word clouds I created for the fake and real news stories:

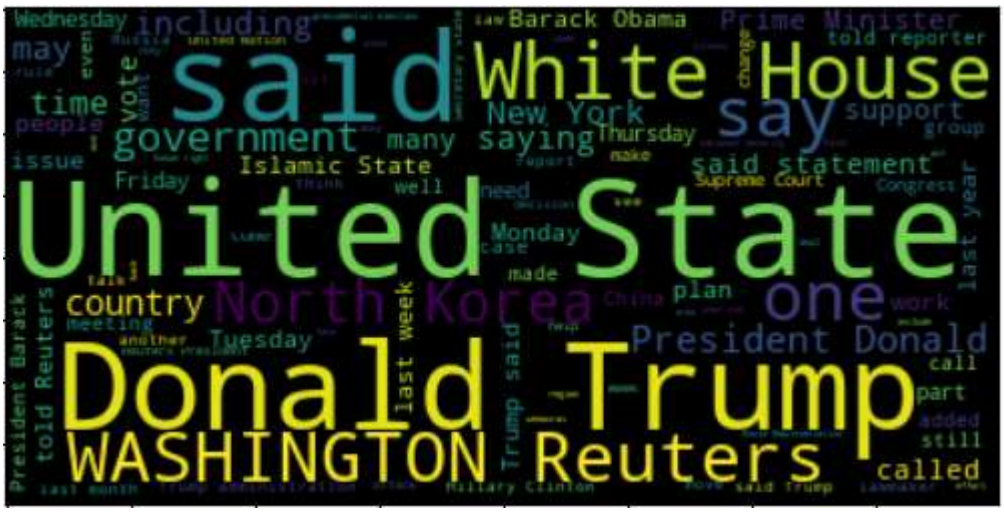

Figure 1: Real news word cloud

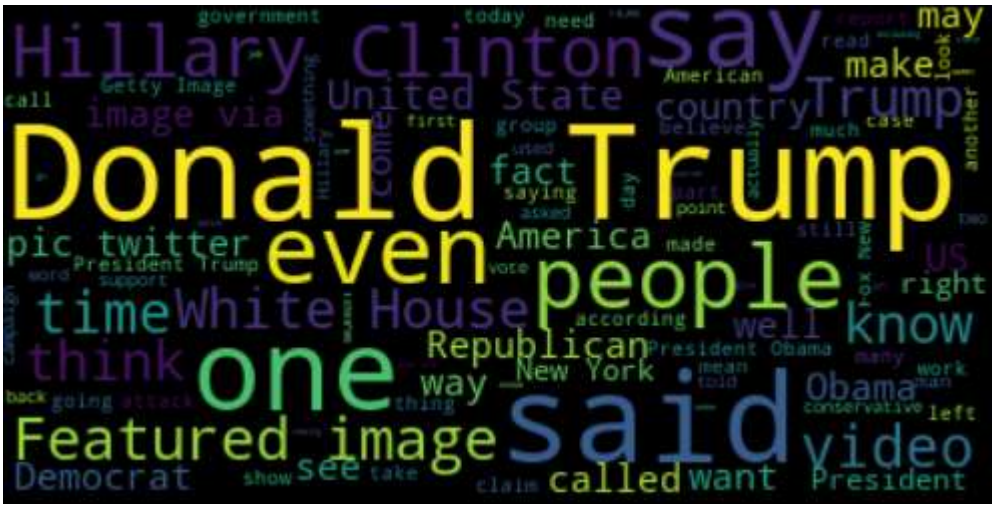

Figure 2: Fake news word cloud

After further analysis, I noticed that the real news articles tended to be shorter than the fake news articles:

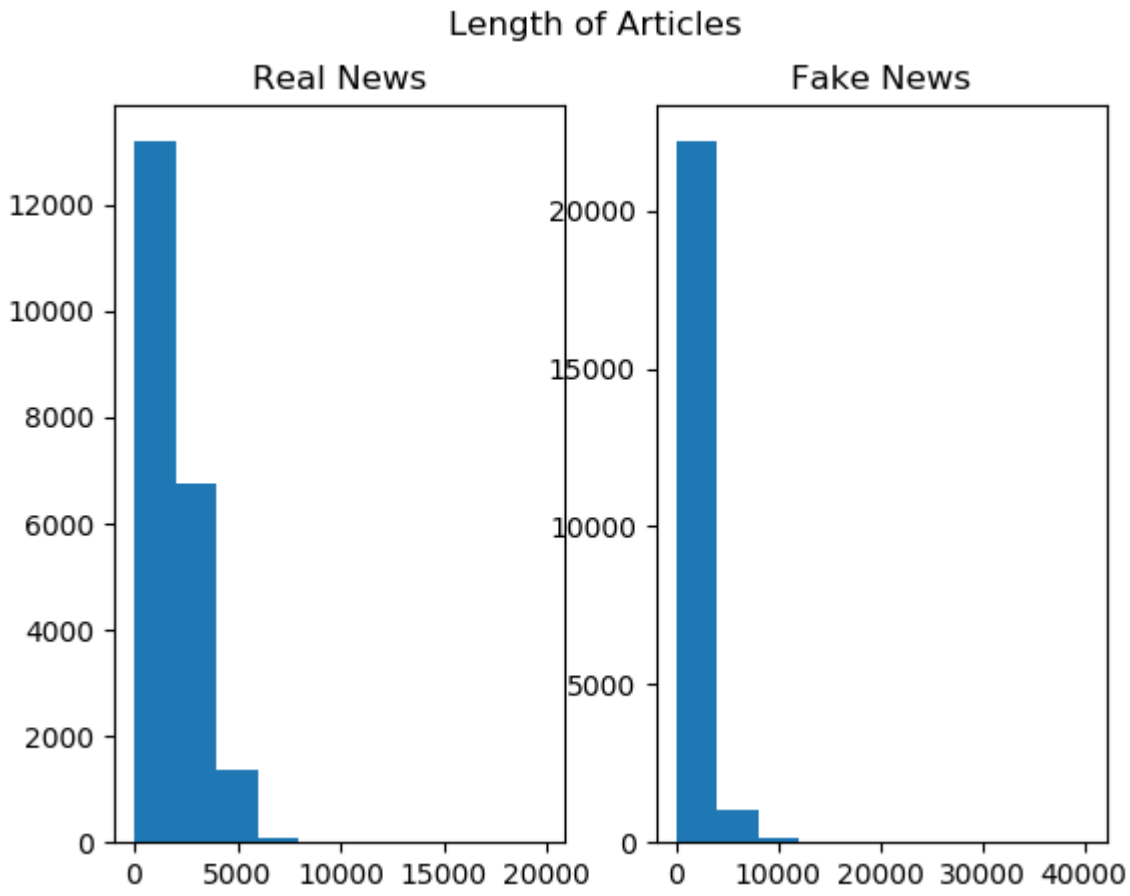

Figure 3: Real news article length versus fake news article length

In addition, the word length tended to be longer for real news than fake news:

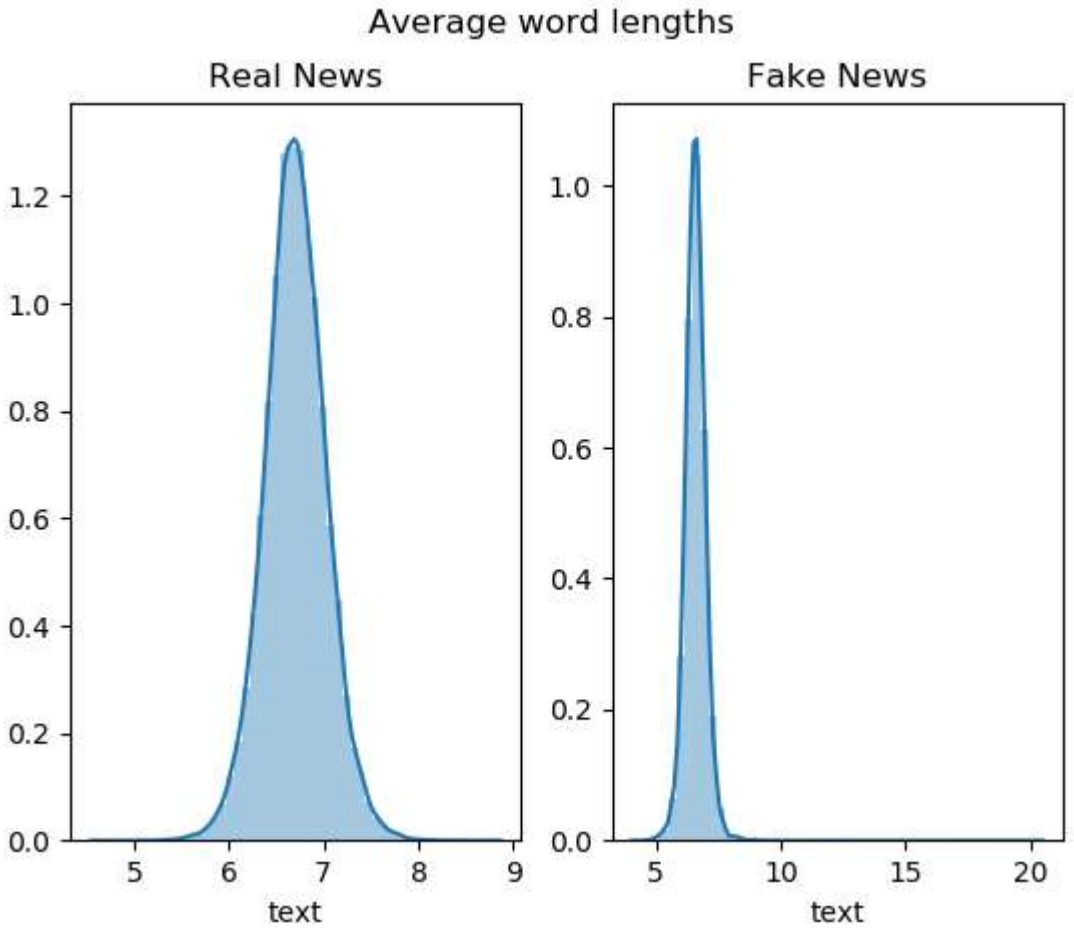

Figure 4: Average word length in real and fake news articles

I also completed an N-gram analysis on the words in the combined fake and real news stories dataset:

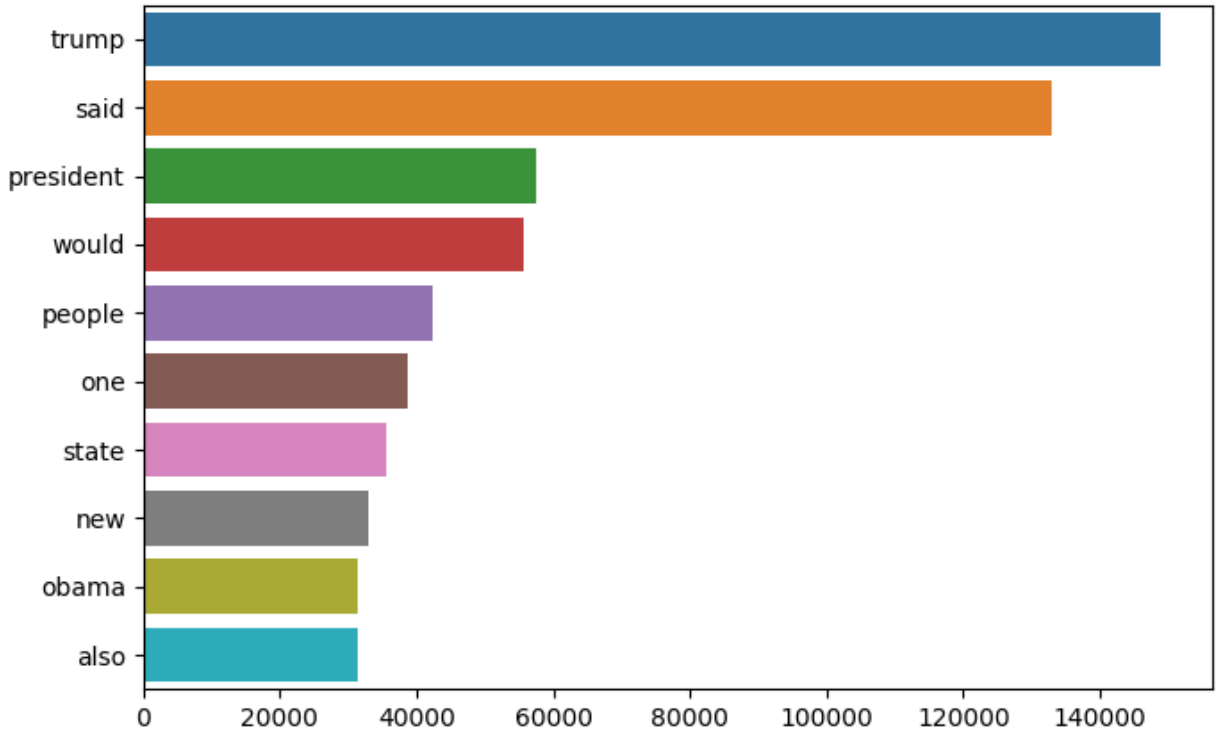

Figure 5: Unigram analysis

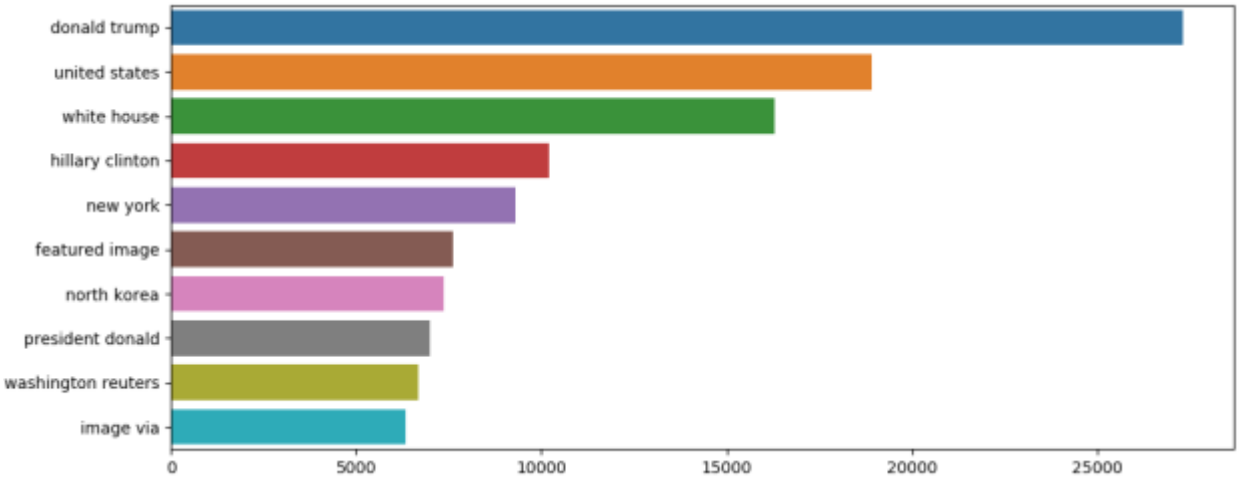

Figure 6: Bigram analysis

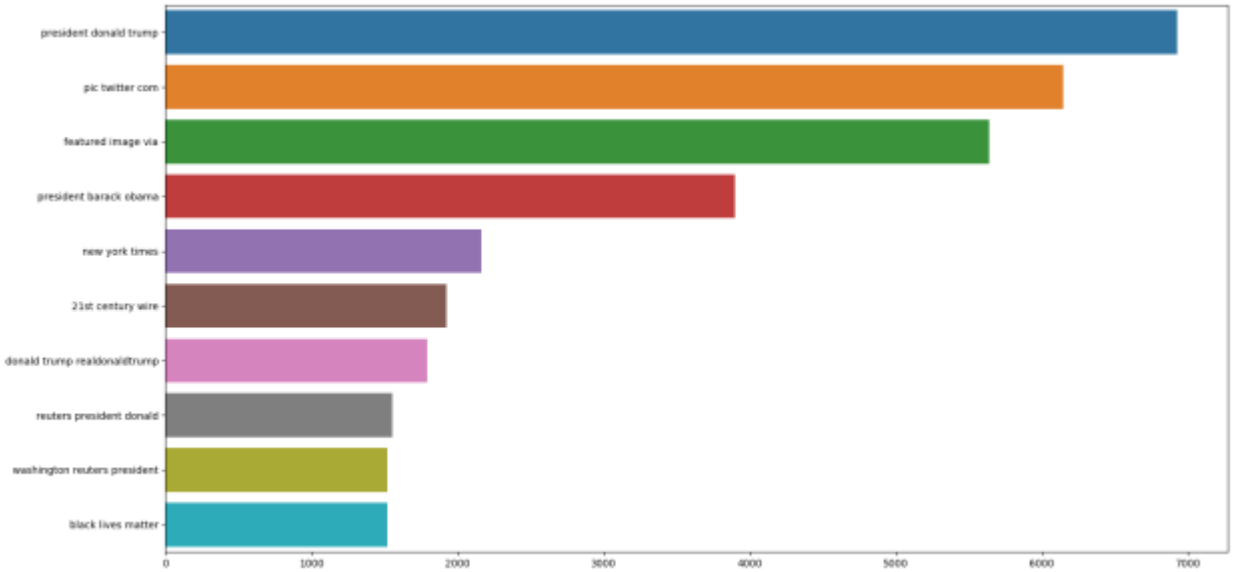

Figure 7: Trigram analysis

Finally, I fitted a logistic regression model that returned 98.85% accuracy with the following confusion matrix:

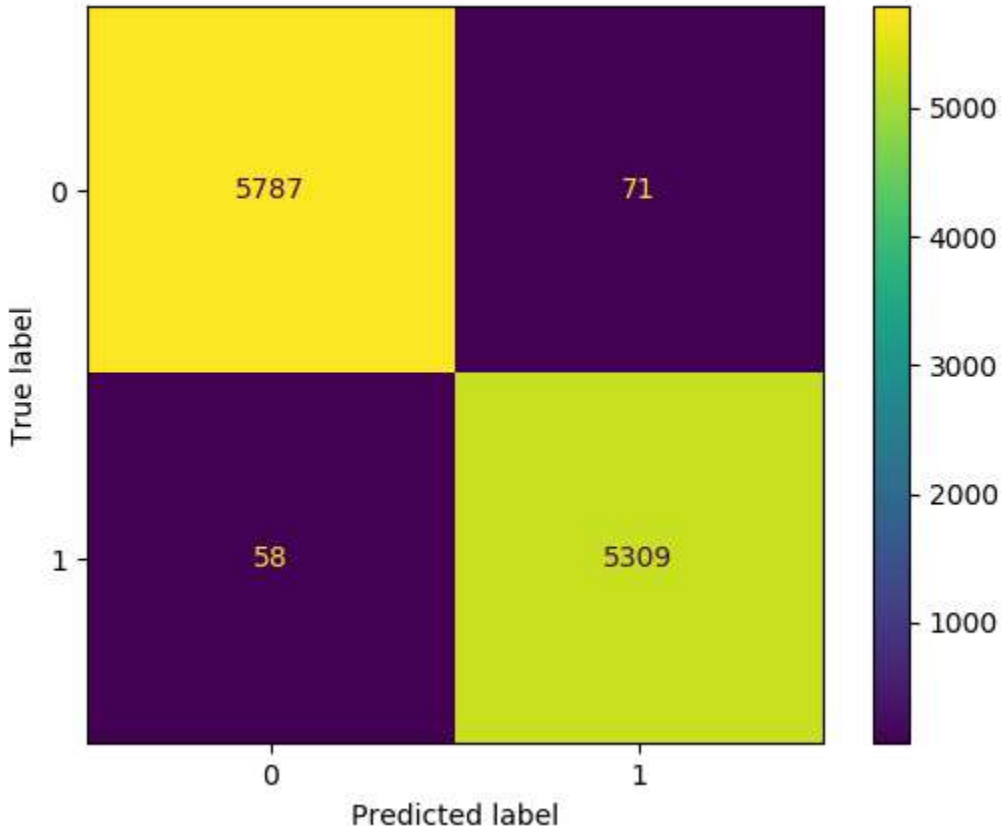

Figure 8: Confusion matrix

Conclusion

Using this logistic regression model, social media companies can catch almost all fake news stories and prevent them from being shared on their websites. This will not stop people from creating their own websites with fake news, but at least the fake news won't be generally available and accepted. I hope that algorithms like this one can prevent further damage to our election process and prevent people from being harmed by fake news.

References

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).

Ahmed H, Traore I, Saad S. (2018) “Detecting opinion spams and fake news using text classification”, Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018. 2.

Akpan, N. (2016). The very real consequences of fake news stories and why your brain can’t ignore them. Retrieved October 3, 2020, from https://www.pbs.org/newshour/science/real-consequences-fakenews-stories-brain-cant-ignore

Bisaillon, C. (2020). Fake and Real News Dataset. Retrieved October 3, 2020, from https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset

Casad, B. J. (n.d.) Confirmation bias. Retrieved October 18, 2020, from https://www.britannica.com/science/confirmation-bias

"Fake News," Lies and Propaganda: How to Sort Fact from Fiction. (2020). Retrieved October 3, 2020, from https://guides.lib.umich.edu/fakenews

Hand, J. (2020). ‘Fake news’ laws, privacy & free speech on trial: Government overreach in the infodemic?. Retrieved October 3, 2020, from https://firstdraftnews.org/latest/fake-news-lawsprivacy-free-speech-on-trial-government-overreach-in-the-infodemic/

How to Identify Fake News in 10 Steps. (n.d.). Retrieved October 3, 2020, from http://library.pfeiffer.edu/Fake-News-Worksheet.pdf

ISOT Fake News Dataset. (n.d.). Retrieved October 3, 2020, from https://www.uvic.ca/engineering/ece/isot/assets/docs/ISOT_Fake_News_Dataset_ReadMe.pdf

Kumar, V.P. (2020). Basic Text cleaning, WordCloud and N-gram analysis. Retrieved October 10, 2020 from https://www.kaggle.com/madz2000/nlp-using-glove-embeddings-99-87-accuracy

Madz2000. (2020). NLP using GloVe Embeddings(99.87% Accuracy). Retrieved October 10, 2020 from https://www.kaggle.com/madz2000/nlp-using-glove-embeddings-99-87-accuracy

Shane, T. & Noel, P. (n.d.). Data deficits: why we need to monitor the demand and supply of information in real time. Retrieved October 3, 2020, from https://firstdraftnews.org/long-form-article/datadeficits/