DSC 680 –   Applied Data Science

News

Krystina Moses

Bellevue University

Abstract

Fake news versus real news. How can we tell the difference? How are they separated and are there key factors that play a role in making the decision? This project aims to answer those questions and more. Additionally, this project will provide knowledge around ways to evaluate the news that we read/hear and determine if it is real or fake. The result of this project will determine which words within the news triggers a fake output and a real output.

The dataset I used had both real and fake news titles and stories within it. The questions I asked were mainly for knowledge around what was fake and what was real. There is also a focus around the importance of being able to identify the difference and the key factors that play a part in the entirety.

Two datasets were combined and analyzed. Models were created to determine different outcomes of several news titles and text. In review of the data, it was determined that politics news was the biggest subject matter within the datasets, followed by world news. In additional review of the data, it was determined that there was a higher amount of words associated with politics. Following through this project has given me more of a push to ensure the knowledge of fake versus real news.

News

Background

There are many ways to look at a news story. Some may just read the title and jump to a conclusion and others may read the title and the story that goes with it, all while drawing their own conclusions. There are different kinds of information out there available to us, all of it coming down to either being real or fake, with a few opinions in there as well. It has been stated that the term 'fake news' is often misunderstood with a variety of definitions. CBC states that they use the terms misinformation and disinformation for the sake of their article about online information. "Disinformation is the deliberate creation and/or sharing of false information in order to mislead and misinformation is the act of sharing information without realizing it's wrong" (Bellemare, 2019). There are many different types of disinformation such as fabricate and manipulated content. With all the variety out there, we need to be careful with what we read and hear from the news.

If we look at the publishing information, claim information, and the author of most articles, it can be determined if it is considered fake news. Additionally, our emotions. Emotions can play a big part in the news and it is important to read them for ourselves. Being able to crack the basics of determining what kind of news we are listening to and reading everyday is just the beginning!

Methods

Data cleaning, exploratory data analysis, and text classification were used within Python for this project. The datasets are from Kaggle, https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset. The target variable was the result of false (fake) or true (real). Additional variables within the datasets include title, text, subject, and date. A chart including information regarding the dataset can be found in Appendix A.

During the data cleaning phase, the two datasets were combined into one and a new column of 'result' was created for the true/false assignment based on the dataset in which the information came. Looking at the false and true values separately, it showed that there were 23,481 fake/false and 21,417 real/true claims. This was interesting to me that the amount of information in the false dataset was so much more. I chose two show two bar charts to compare the number of true/false claims, shown in appendix B, as well as a bar chart for the subject column, shown in appendix C. Within the data itself, stop words were removed for the text and title columns and the text was also made lowercase to be consistent.

The columns that were primarily focused on were text, title, and response. The amount of times a word occurred within the columns was determined with trump, trumpsters, and unprecedented being in the top tier for title and trump, wendywhistles, and wrong being among the top in the text column.

The data was split into a training and testing set for the title column as well as the same for the text column. This was with the target variable of the response. A confusion matrix, logistic regression, and Multinomial Naive Bayes were all used to see what would be the most accurate to determine a fake or real article. Each of the methods came back with great results.

Results

The results show the number of false articles outweigh the number of real articles. It is also shown that the subjects of politic and world news are the big headlines, next to general news.

The confusion matrix results for the title column can be seen in appendix D and the text column in appendix E. Starting with the text column, the Multinomial Naive Bayes model came out with a 94% accuracy. Logistic regression came out to a 98.7% accuracy rating. With the title column, we have the results of the Multinomial Naive Bayes model as 94.3% accuracy and logistic regression as 95.5%. Both show the confusion matrix with a low amount within the true and false negatives.

Discussion/Conclusion

I question a lot of news articles that I open to read daily. Each person has their own views and can take what they read with a grain of salt, although, we want to have the true facts. With all that is happening in the world today and in recent years, there seems to be always something to question. It is important to be able to detect the different types of news available and get our facts correct. Going through the steps of this project and doing the additional research, a lot was gained for me. While there is still a lot of room for opinions as well as teachable moments, we can all learn to better appreciate the truth as well as be better at identifying it.

Acknowledgements

I would like to thank all the news outlets for providing the stories for this project and my boyfriend for always trying to explain and point out all the information in the news to watch out for.

References

Bellemare, A. (2019, July 05). The real 'fake news': How to spot misinformation and disinformation online | CBC News. Retrieved from https://www.cbc.ca/news/technology/fake-news-misinformation-online-1.5196865

Davis, W. (2016, December 05). Fake Or Real? How To Self-Check The News And Get The Facts. Retrieved from https://www.npr.org/sections/alltechconsidered/2016/12/05/503581220/fake-or-real-how-to-self-check-the-news-and-get-the-facts

Elliott, C. (2019, February 21). Here Are The Real Fake News Sites. Retrieved from https://www.forbes.com/sites/christopherelliott/2019/02/21/these-are-the-real-fake-news-sites/

Fake News, Propaganda, and Misinformation: Learning to Critically Evaluate Media Sources.: Infographic: Spot Fake News. (n.d.). Retrieved from https://guides.library.cornell.edu/evaluate_news/infographic

Graham, D. (2019, June 12). Some Real News About Fake News. Retrieved from https://www.theatlantic.com/ideas/archive/2019/06/fake-news-republicans-democrats/591211/

McCarthy, N., & Richter, F. (2016, December 09). Infographic: Most Americans Believe Fake News Headlines. Retrieved from https://www.statista.com/chart/7153/most-americans-believe-fake-news-headlines/

McClure, L. (2017, January 23). How to tell fake news from real news. Retrieved from https://blog.ed.ted.com/2017/01/12/how-to-tell-fake-news-from-real-news/

Northern Essex Community College. (n.d.). Fake News vs. Real News: Tips for Evaluating Information. Retrieved from https://library.nwacc.edu/fakenews/evaluating

Pierce College. (2020). FAKE NEWS vs. REAL NEWS: How to Determine the Reliability of Sources: Fake News. Retrieved from https://library.piercecollege.edu/c.php?g=598055

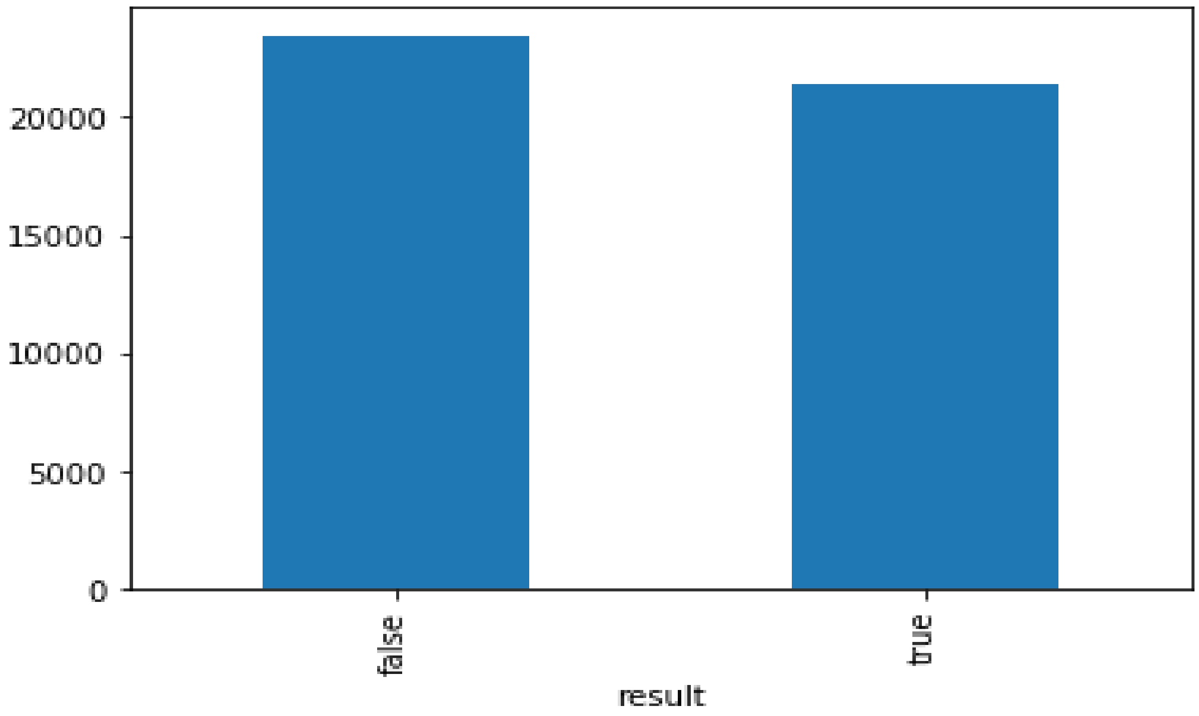
University of New Hampshire. (n.d.). Fake News, Misleading Information, and Evaluating Sources: Problem. Retrieved from https://libraryguides.unh.edu/fakenews

Appendix A

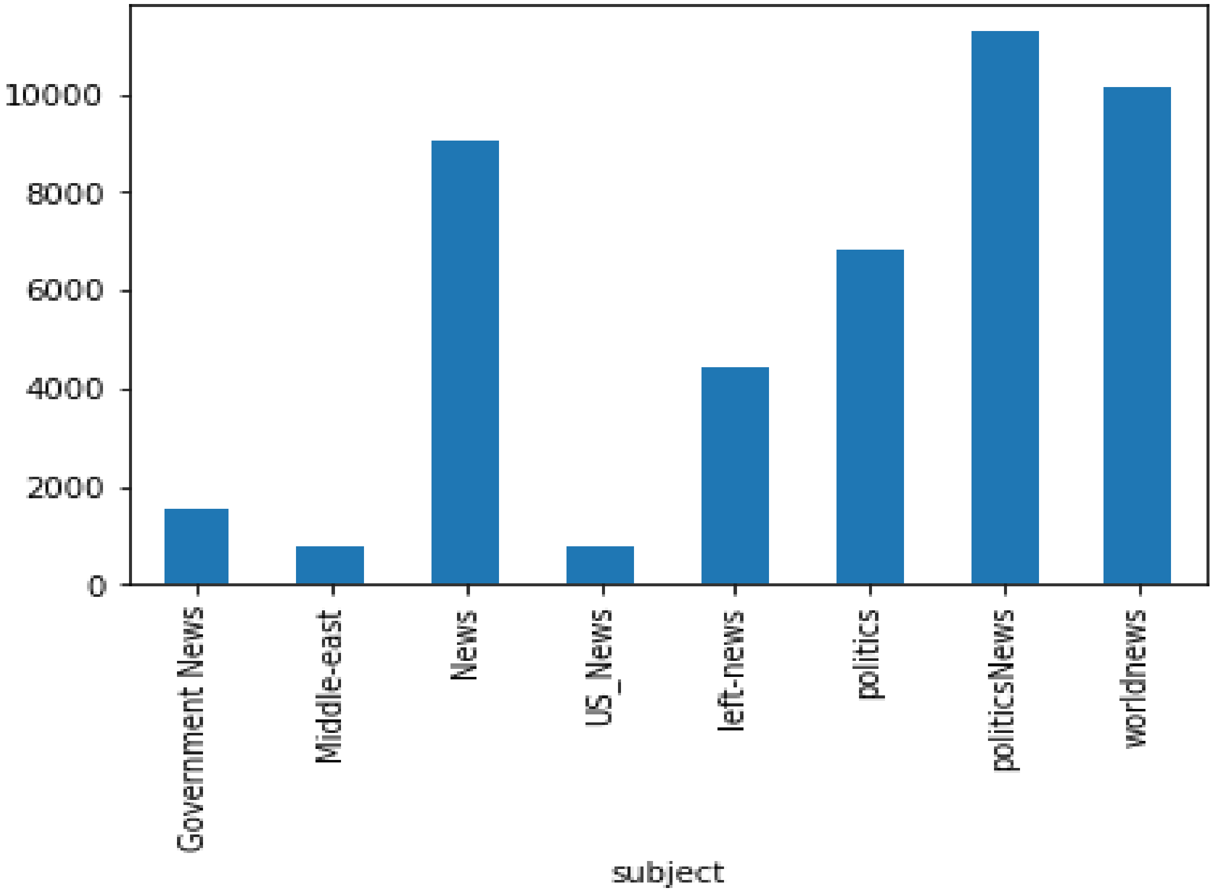
News dataset columns and descriptions

Column Name	Data Type	Description
title	object	Title of the article
text	object	Text of the article
subject	object	Subject of the article
date	object	Date when the article was posted
result	object	News type True (real) / False (fake)

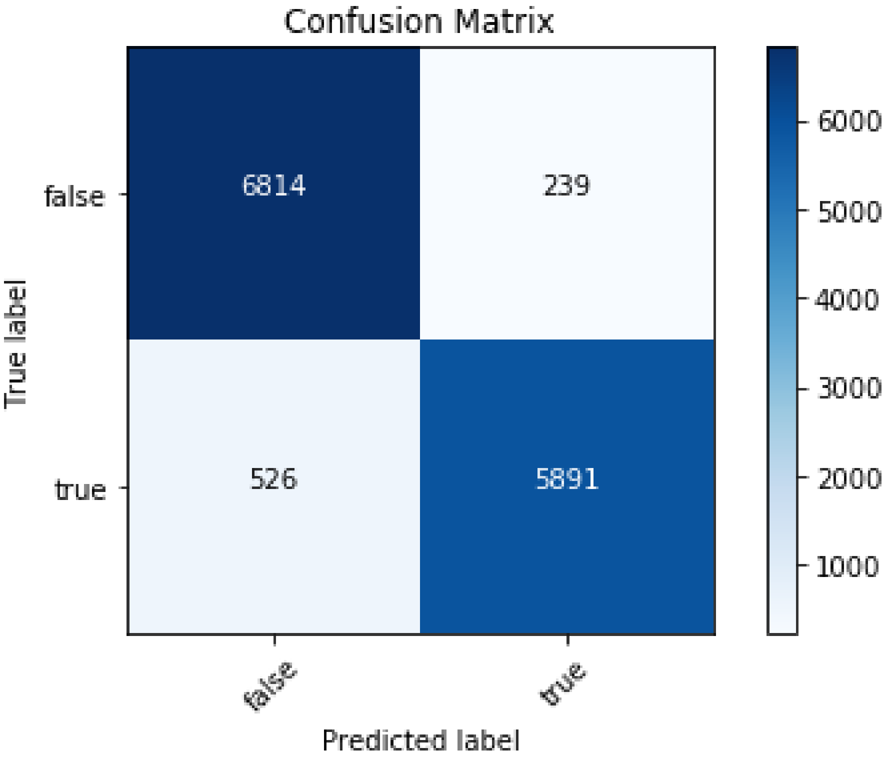
Appendix B



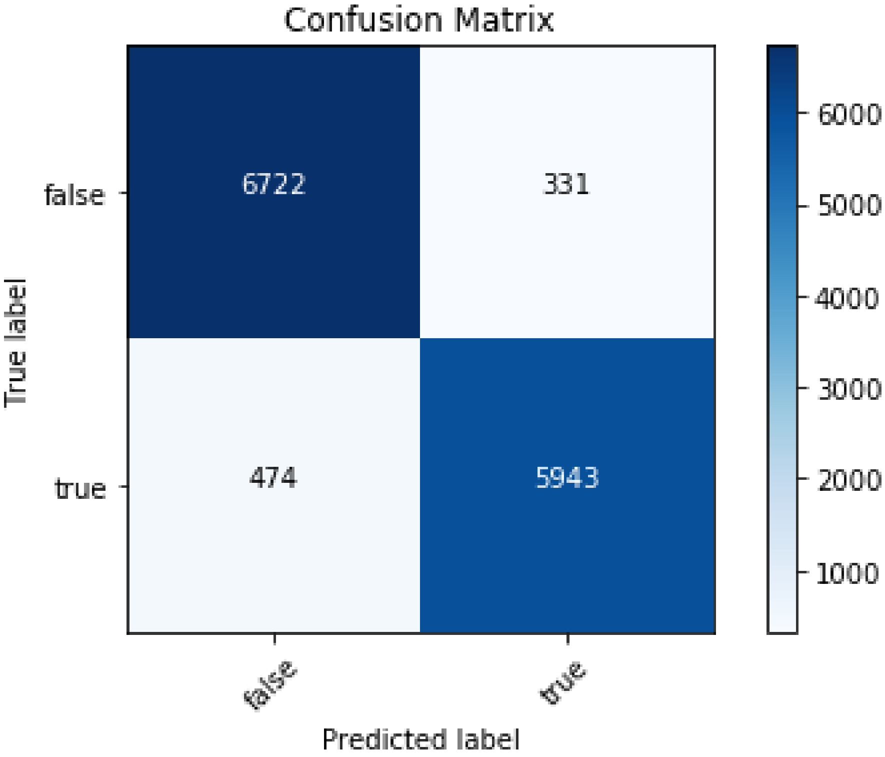
Appendix C


10000
8000
6000
4000
2000
0
Government News
Middle-east
News
US_News
left-news
politics
politicsNews
worldnews
subject

Appendix D



Appendix E


Confusion Matrix
True label
false
true
Predicted label
false
true