Sam Loyd

Detecting Fake News

July 2021

Executive Summary

The purpose of this project was to analyze and prepare a dataset using natural language processing to determine if a model could be used to adequately classify fake news in the future. It also sought to determine the relative frequency that the model would require retraining. Fake news is fabricated for many reasons, but typically it is to apply influence. Many in the on-line community have been affected by it. It is often cited as a reason for the eroding trust in our societal institutions and can quickly spread as links are forwarded through various types of social media. The dataset used in this project was acquired from a common repository used in education and open sourced. There were very few dimensions, but title and text were the primary focus. The data was processed to allow for machine learning model selection. Early in the first pass at model selection, the data set yielded several different models with unusually high accuracy scores that did not require retraining for extended periods of time. Further data analysis followed by provenance research determined that this data was biased in its collection. The true sources of news all came from Reuters. The models were quite effective at finding the way Reuters formats its news. While a trusted institution, Reuters is not the only source of accurate news information. At that point, features were removed, to minimize that bias to the extent possible. Of course, this reduced accuracy, but would provide for a better overall model on other unseen data that did not come from Reuters. Ultimately, the final model proved to have a high accuracy level at validating this data set. There are still concerns of remaining bias given the methods used for data collection so I would not recommend using this model in a production environment without further testing and possible retraining using a dataset without these concerns. If used, the model will require at least monthly monitoring for staleness and likely retraining.

Abstract

In 2019, a study performed by the Centre for International Governance Innovation (CIGI) found that 86 percent of people online had been misled by misinformation claiming to be news. (Phys.org, 2019). Given the scale of this impact, being able to classify news in real time could help minimize the dissemination of fake news articles. Fake news can erode trust in our institutions and create real societal problems. This work seeks to restore that faith by providing a simple classification model to label fake news appropriately. The research also seeks to ascertain the likely rate of retraining required.

Introduction

Publishing misleading information as news to sway opinion is nothing new. There have been propaganda campaigns throughout history. However, social media has lowered the bar for entry into this endeavor at scale (Desjardins, 2017). The same phenomena, albeit at a different scale, happened after the printing press was invented in the 1400s (Soll, 2016). Not only are many exposed to false and misleading news articles, but most people are overconfident in their ability to discriminate. In a recent study presented in Proceedings of the National Academy of Sciences of the United States of America, it was determined that 75 percent of Americans rated themselves an average 22 percent better at detecting fake news than their ability warranted (Lyons et al., 2021). This can lead to institutional mistrust. An example of this is highlighted in COVID vaccine hesitancy. A report focused on the residence of Bradford, UK, concludes that vaccine misinformation has led many to avoid the vaccine (Benjamin, 2021).

With all this in mind, a Kaggle dataset was acquired with dimensions suitable for natural language processing to create a machine learning model to classify fake news. The dataset on fake and real news was provided by Clément Bisaillon. It held two files with the four dimensions. Date, subject,

title, and text were included. One file contained fake news articles and the other one contained real news articles. These were merged and a new column created noting classification which became the target variable for model training and validating its accuracy. Date was split into month and year to provide for a reasonable training fold given the over two-year time frame covered in the data.

Methods

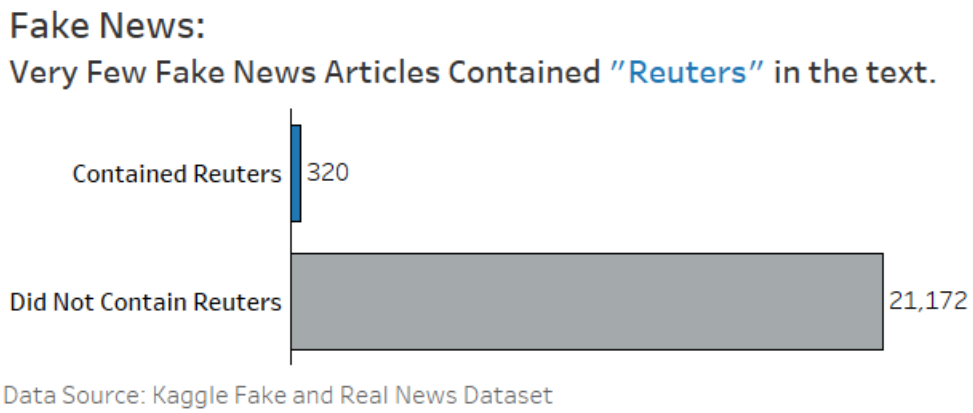
The CRISP-DM methodology was implemented for this project (Siegel, 2016). As such, six iterative stages were followed.

- Business (Domain) Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

In the context of this paper, the first four will be covered in the current methods section. The evaluation and deployment phases of CRISP-DM can be found in the results and conclusion sections. Domain knowledge was referenced in the introduction. The initial analysis indicated that the date dimension was skewed with most of the data found in 2016 and 2017 (see Appendix A). In natural language processing, text data is converted into a numeric format usable by machine learning models. This transformation involves tokenizing the text which breaks the words apart and vectorizing those words. The latter converts them to a numeric value based on the word's representation in each entry. This was performed against both the text and title columns. The data was then split based off time into folds using year and month. When a fold was used to train the model, it was tested against data a month newer. This avoids overfitting the data to a given time frame and providing a false sense of

accuracy. During this stage, a problem was discovered. Several models had extremely high accuracy after small groups of training data. This persisted over time. That does not seem likely in a real-world scenario. The term "Reuters" emerged as a common feature found in most of the models as being the most important (see Appendix B). This forced the project back to the analysis phase. "Reuters" was found in almost all the real news and in very little of the fake news. Figure 1 shows the prevalence of the word Reuters in the fake news articles.

Figure 1

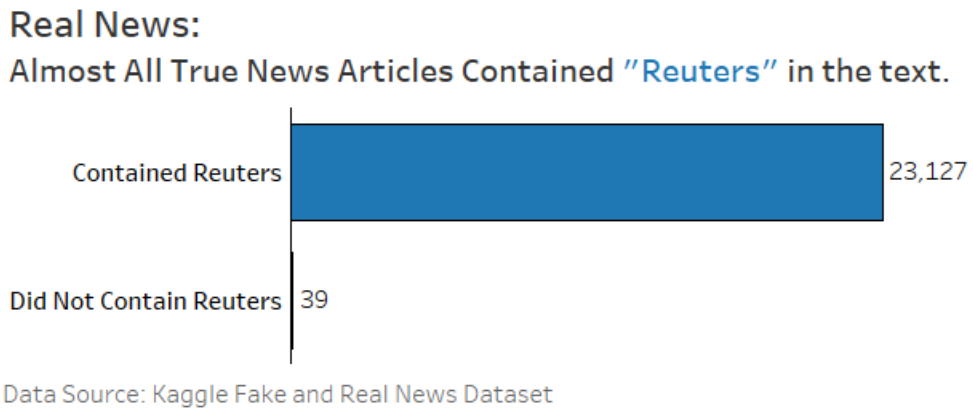


Here is an example text from one of the real news articles.

> "WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts …"

This formatting was common in most of the real news articles. Figure 2 shows the prevalence of real news stories containing Reuters as a feature.

Figure 2

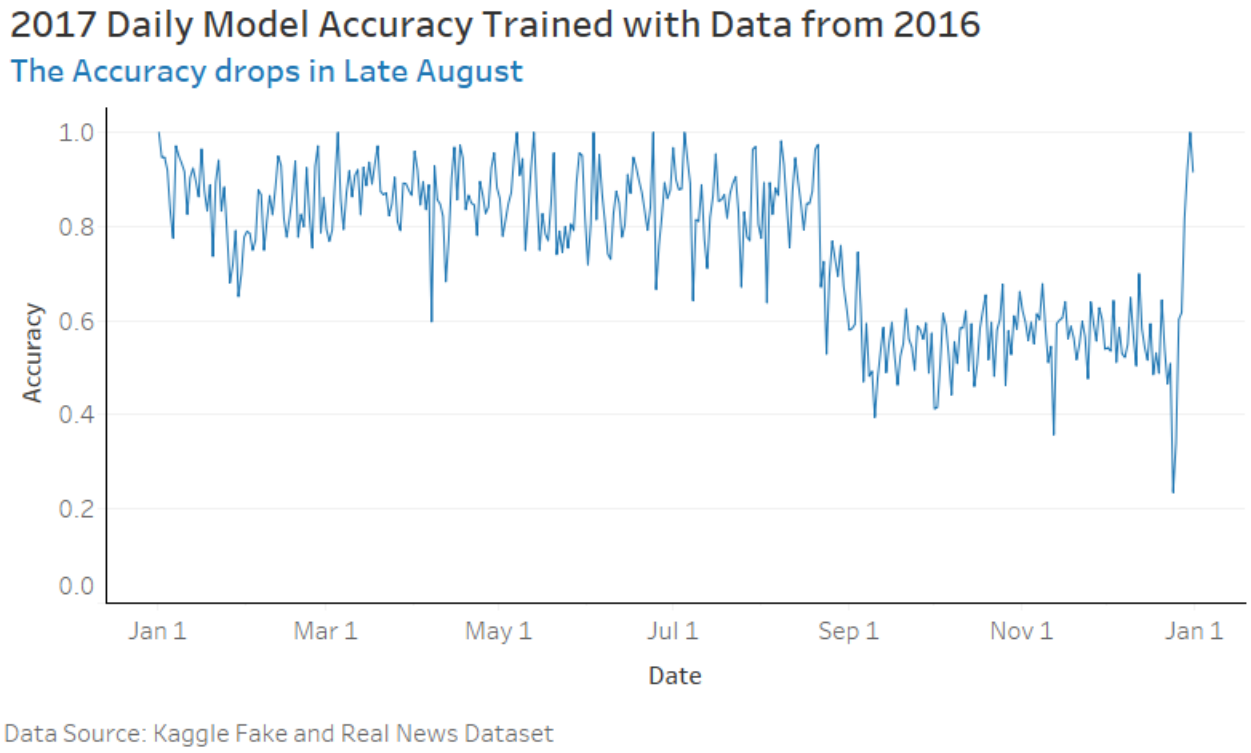


More provenance research was then performed on the dataset, and it was discovered that the original authors only pulled real news stories from Reuters (Ahmed et al., 2018). This caused a major problem for this research as it biased the data. Classifying data from Reuters was not the desired outcome. A mitigating decision was made to remove the feature. This process was repeated three more times as the models found other features common or uncommon in the way Reuters formatted their news. These were all removed as well. Unfortunately, there may still be bias tugging at the models with less important features but exploring each one is very time consuming so at that point a practical decision was made to move forward documenting that concern. The model was tuned and validated.

Results

The final model proved 90 percent accurate at classifying news labelled as real. It proved 98 percent accurate at predicting news labelled as fake. This provided a more realistic 91 percent overall accuracy. The validation set accuracy remained consistent for the four months that were originally withheld. At that point, the model was retrained on a year of test data and the validation was run daily over the following year. At around eight months there was a significant drop in accuracy. Figure 3 shows this drop.

Figure 3

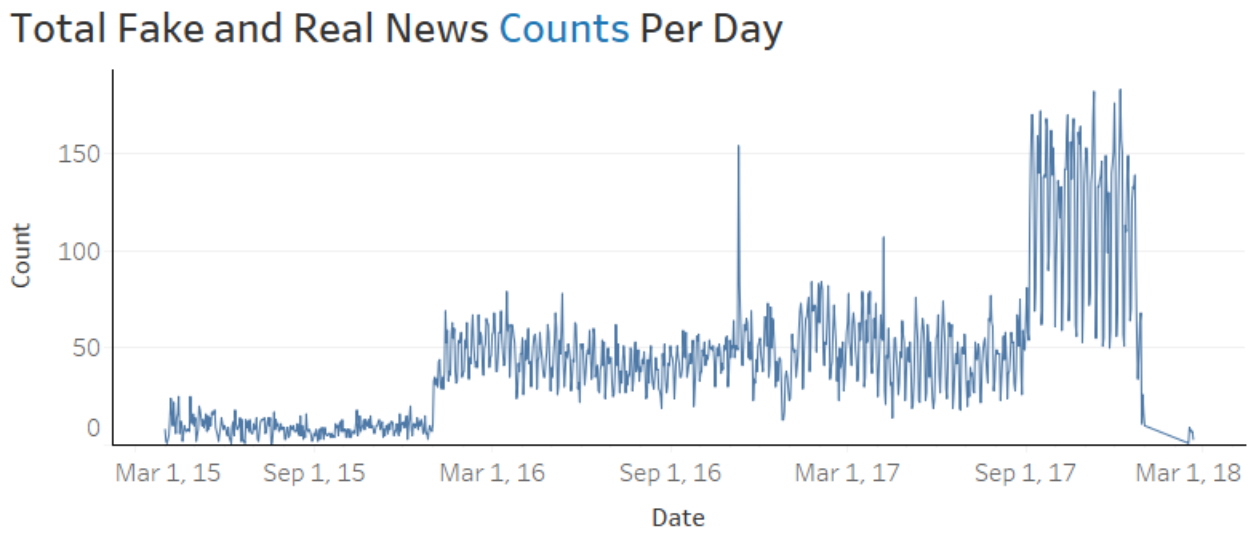


However, there was also an increase in the overall count of news articles relatively close to the same time (see Appendix A). Upon further test runs that drop in accuracy persisted unless retrained with data closer to the timeframe impacted. With that in mind, it is recommended that the training data be evaluated at least monthly for any drops in accuracy that might require retraining. There does not appear to be an amount of time for when the model will become stale, but testing confirmed that it likely will become stale without some amount of retraining. There might be an inverse relationship with overall news volume or some seasonal aspect that changed during that time worth further research, but that is beyond the scope of this work and in the time frame allotted.

Conclusions

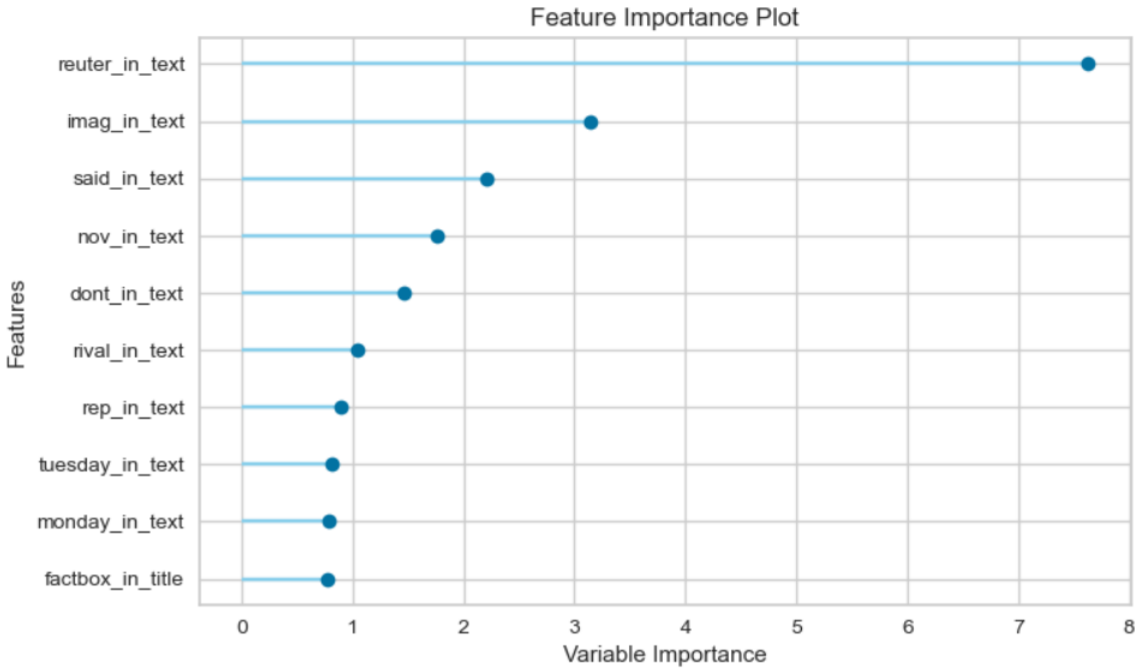
While yielding a high accuracy against testing and validation data from this dataset, it is still unknown how the model might perform on completely unrelated data given the bias used to create the sample. It is recommended that the model be measured against random data from other valid news sources as well. As the research stands, it is not recommended to use this model in a production environment without further testing. The model has been pickled for further investigation. If the model were used, it would occasionally require retraining, but there is not a time-based frequency that could be determined with the data provided. Bias is an ongoing concern for this dataset. It is strongly encouraged that other data sets for real news be added to the training data for this model to reduce the bias currently documented. Given the problems with this dataset, commissioning or finding a new sample with more recent data is desirable as well. Understanding what precipitated the drop in accuracy shown in Figure 3 and discussed in the results section is also a worthy endeavor for future research.

Appendix A



Data Source: Kaggle Fake and Real News Dataset

Appendix B



References

Ahmed, H, Traore, I, Saad, S. Detecting opinion spams and fake news using text classification, Security and Privacy, 2018; 1: e9. https://doi.org/10.1001/spy2.9

Desjardins, J. (2017, February 10). The Fake News Problem in One Chart. Retrieved from https://www.visualcapitalist.com/fake-news-problem-one-chart/

Lockyer, B., Islam, S., Rahman, A., Dickerson, J., Pickett, K., Sheldon, T., Wright, J., McEachan, R., & Sheard, L. & Bradford Institute for Health Research Covid-19 Scientific Advisory Group. (2021). Understanding COVID-19 misinformation and vaccine hesitancy in context: Findings from a qualitative study involving citizens in Bradford, UK. Health Expect. 2021 May 4. Epub ahead of print. https://doi.org/10.1111/hex.13240

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B. & Reifler, J. Overconfidence in news judgments is associated with false news susceptibility. Proceedings of the National Academy of Sciences Jun 2021, 118 (23) e2019527118; DOI: https://doi.org/10.1073/pnas.2019527118

Phys.org. (2019, June 12). 86 percent of internet users admit being duped by fake news survey. Retrieved from https://phys.org/news/2019-06-percent-internet-users-dupedfake.html

Siegel, E. (2016). Predictive Analytics. Hoboken, NJ: Wiley.

Soll, J. (2016, December 18). The Long and Brutal History of Fake News. Retrieved from https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/