LIVE

DS3000: FAKE NEWS

By: Cici Ling, Miles Vollmer, Julian Savini & JingXuan Liu

TABLE OF CONTENTS

LIVE

01

ABSTRACT

As social media is continuing to become more immersed in people's lives, so is misinformation and fake news. It's a challenge for journalists to relay information that does not have their own bias in it, and the digital era allows for misinformation to be shared with millions of people within a single day. It is challenging to identify what is not and is fake news, so if a machine learning model could accurately predict whether a news article is reliable or unreliable, that would be very beneficial for what media one chooses to engage with.

ABSTRACT


02
LIVE
INTRODUCTION

INTRODUCTION

This is an issue you probably run into every day if you use social media or browse news on the internet.

Recently, there are some high-profile examples of the damage that fake news can cause from the phony news smear campaigns aimed to influence the 2016 & 2020 elections to the swath of misinformation during the COVID-19 epidemic

- People are using Ivermectin (a drug used on livestock), to try and treat covid or even prevent it.
- Proven to be completely ineffective in treating covid, especially when compared with acetaminophen or ibuprofen, but people still took it because of fake news.

We believe using fake news data to create a phony news classifier would be a way to combat this issue.

03

LIVE

METHODOLOGY

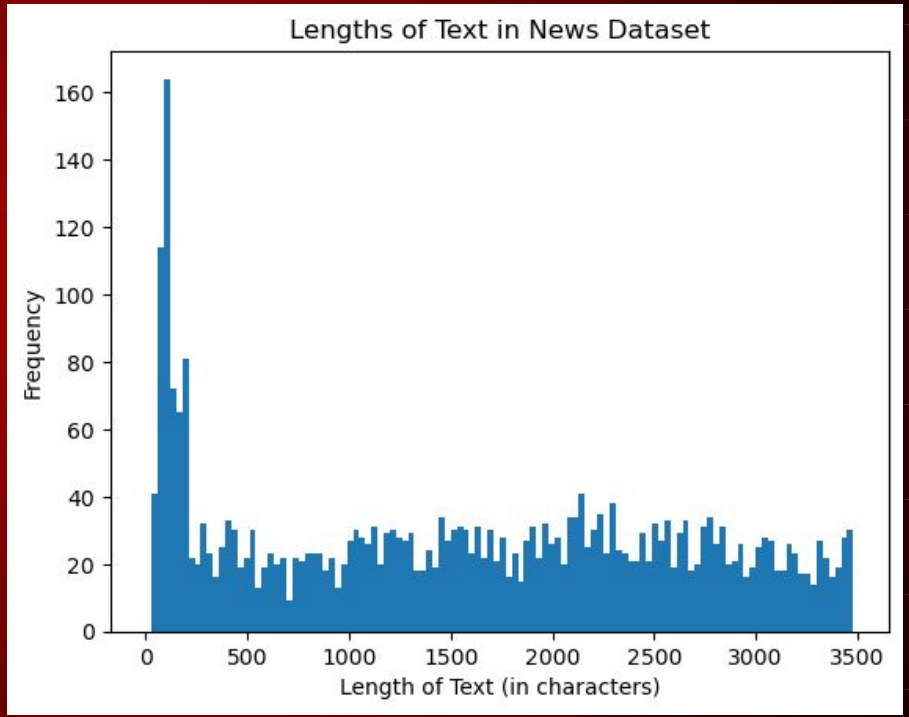
Methodology

- Obtained our dataset from an article called "Detecting Fake News with Python and Machine Learning" by DataFlair. The dataset contains the title of news articles, the text of news articles, and whether each news article is real or fake.
- A News ID column which has corresponding numbers relating to each news source to identifies which articles were real or fake.
- For data preparation, null values were removed from the dataset, and any articles that contained less than fifty words were also removed because having more text data in an article improves a model's accuracy.
- In our EDA phase, one constraint is that the news source is represented by an identification number, so we weren't able to analyze the degree of misinformation in particular well-known news sources.
- In order to conduct text classification, we needed to convert the text columns into numerical feature vectors.
- We used Term Frequency–Inverse Document Frequency to reduce the weightage of common words.
- We used hyperparameter testing using n_samples, bootstrap, max_depth, and min_samples_split, and with the CV grid, we found the most optimal parameters.

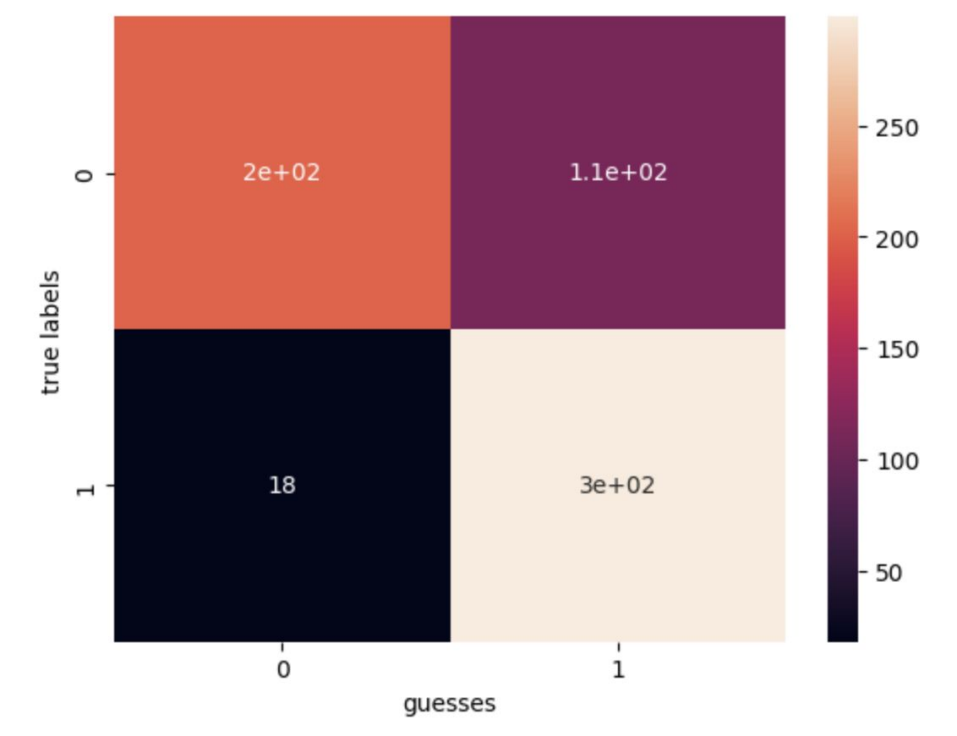
GRAPHS

Lengths of Text

The graph to the right shows the lengths of the text after condensing the dataset. The limit of the length of the text is 50.



GRAPHS



Confusion Matrix

The graph to the left shows the it is rarely classifying real news as fake but some fake news is getting misclassified as real.

04

LIVE

RESULTS

Results

In conclusion, we were able to generate models for fake news classification. For each model, we have different accuracy scores.

For Naive Bayes, our accuracy was 84.3%.

For KNN, our highest accuracy was 82.4%.

Lastly, after hyperparameter testing, the Random Forest Classifier Model had an accuracy of 89.2%.

05

THE IMPACTS

Impacts people in the way that it identifies fake and true news with comparatively high accuracy

People who want to check the validity of certain news

Impacts the way people trying to send out phony news across different social media platforms

LIVE

06

REFERENCES

Related Works

Manzoor, Singla, and Nikita's article, titled "Fake News Detection Using Machine Learning approaches: A Systematic Review"

- a paper review on the various machine learning approach models for fake news detection
- analyzes previous research and identifies types of data in social media posts as well as types of fake news
- address two main components: representation feature learning and credibility label inference

Ahmad, Muhammad Yousaf, Suhail Yousaf, and Ovais Ah Ahmad's article, "Fake News Detection Using Machine Learning Ensemble Methods"

- discusses fake news classification using machine learning models and ensemble techniques.
- it opens up issues that are interesting to explore: identifying key elements involved in the spread of news as well as real-time fake news identification in videos

DataFlair is a guide to help build a project to test if a piece of the news article is real or fake.

- DataFlair explained code terminology like TfidfVectorizer, which can convert a collection of raw data into TF-IDF features
- created a machine-learning model with a 92.82 percent accuracy in magnitude.

Related Works

Piero Paialunga's "Fake News Detection with Machine Learning"

- discusses how to code Fake News detection using BERT, TensorFlow, and PyCaret. Paialunga goes over easier and faster Machine Learning to a pre-trained neural network, fine-tunes it
- obtains state of art results on a dataset
- able to obtain 88% accuracy, 88% recall, and 89% precision.

PythonCode : Rockikz and Payong's "Fake News Detection in Python"

- a website that provides a collection of resources and tutorials on python topics to beginner and intermediate programmers
- allows readers to explore fake news datasets and build a fake news detector using a transformer library.
- able to create a model with 99.78 percent accuracy on private and public news.

Helmstetter and Paulheim, "Weakly supervised learning for fake news detection on Twitter".

- developed a machine learning model on fake news detection on Twitter
- manually collected datasets using Twitter API, and DMOZ and worked with large datasets
- detect fake news with an F1 score of up to 0.9.

CONCLUSION

We reached our goal of building a fake news classifier. With the three models we built, we were able to get the highest accuracy of around 90% for our results from the random forest classifier. For future research, we would like to work alongside information analysts, who could create our data set, so we have an understanding of what goes into the process of determining whether an article is real or fake. Thus, we would have full confidence if our articles were properly labeled. Also, if possible, we would want to utilize and experiment with another text classification other than the bags of words(CountVectorizer) model. We would also want to employ keywords search function if given more time.

References

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousad, and Muhammad Ovais Ahmad. 2020. Fake News Detection Using Machine Learning Ensemble Methods. 2020,, 11 pages. https://www.hindawi.com/journals/complexity/2020/8885861/

DataFlair 2021. Detecting Fake News with Python and Machine Learning. Retrieved Oct 15, 2022 from https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/

Paulheim Stefan Helmstetter and Heiko Paulheim year = 2018 title =. [n. d.].

S.I. Manzoor, J. Singla, and Nikita. 2019. Fake News Detection Using Machine Learning approaches: A systematic Review. 1 (2019), 230–234. https://doi.org/10.1109/ICOEI.2019.8862770

Piero Paialunga. 2022. Fake News Detection with Machine Learning, Using Python. Retrieved Oct 14, 2022 https://towardsdatascience.com/fake-news-detection-with-machine-learning-using-python-3347d999ad1

Abdou Rockikz and Adrien Payong. 2022. Fake News Detection in Python. Retrieved Oct 15, 2022 from https://www.thepythoncode.com/article/fake-news-classification-in-python

LIVE

THANKS!

Do you have any questions?