A Project Report for Natural Language Processing

**FAKE NEWS DETECTION**

Submitted to Manipal University, Jaipur

Towards the partial fulfilment for the Award of the Degree of

**BACHELOR OF TECHNOLOGY**

In Computer Science

2022-2023

By

Yash Kishore 209301512



Under the Guidance of

Ms. Shikha Mundra

**Department of Computer Science**

**School of Computer Science**

**Manipal University Jaipur**

**Jaipur, Rajasthan**

(Jan-May 2023)

**CERTIFICATE**

Date: 15/04/2023

This is to certify that the project work entitled **“Fake News Prediction”** submitted by in fulfilment for the requirements of the award of Bachelor of Technology Degree in Computer Science at Manipal University Jaipur is an authentic work carried out by them under my supervision and guidance. To the best of my knowledge, the matter embodied in the project has not been submitted to any other University / Institute for the award of any Degree.

Ms. Shikha Mundra

Professor & Project Guide

Department of CSE

Manipal University Jaipur

**ABSTRACT**

With the recent social media boom, the spread of fake news has become a great concern for everybody. It has been used to manipulate public opinions, influence the election - most notably the US Presidential Election of 2016, incite hatred and riots like the genocide of the Rohingya population. A 2018 MIT study found that fake news spreads six times faster on Twitter than real news. The credibility and trust in the news media are at an all-time low. It is becoming increasingly difficult to determine which news is real and which is fake. Various machine learning methods have been used to separate real news from fake ones. In this study, we tried to accomplish that using Passive Aggressive Classifier, LSTM and natural language processing. There are lots of machine learning models but these two have shown better progress. Now there is some confusion present in the authenticity of the correctness. But it definitely opens the window for further research. There are some of the aspects that has to be kept in mind considering the fact that fake news detection is not only a simple web interface but also a quite complex thing that includes a lot of backend work.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **CONTENT** | **PAGE.NO** |
| 1. | CERTIFICATE | 2 |
| 2. | ABSTRACT | 3 |
| 3. | INTRODUCTION | 5 |
| 4. | MOTIVATION | 6 |
| 5. | OBJECTIVE | 7 |
| 6. | METHODOLOGY | 8-10 |
| 7. | RESULT | 11-15 |
| 8. | CONCLUSION | 16 |
| 9. | REFERENCES | 17 |

**INTRODUCTION**

Everyone makes decisions based on their understanding or based on the data, information provided to them. Machines and artificially intelligent agents work in a similar way. They make their decision in the basis of the information given to them and the rule base provided to process and access the data.

The focus of this project will be fake news dataset.A double-edged sword is social media for news consumption. On the one hand, its low cost, convenient access, and rapid information sharing lead people to search for and consume social media content. On the other hand, it allows "fake news" to spread widely, i.e. low-quality news with purposely misleading facts. The widespread distribution of fake news has the potential to have highly negative effects on people and culture. Fake news identification on social media, therefore, has recently become an evolving study that attracts immense interest. Fake social media identification of news poses specific features and obstacles that make conventional news media detection algorithms ineffective**.**

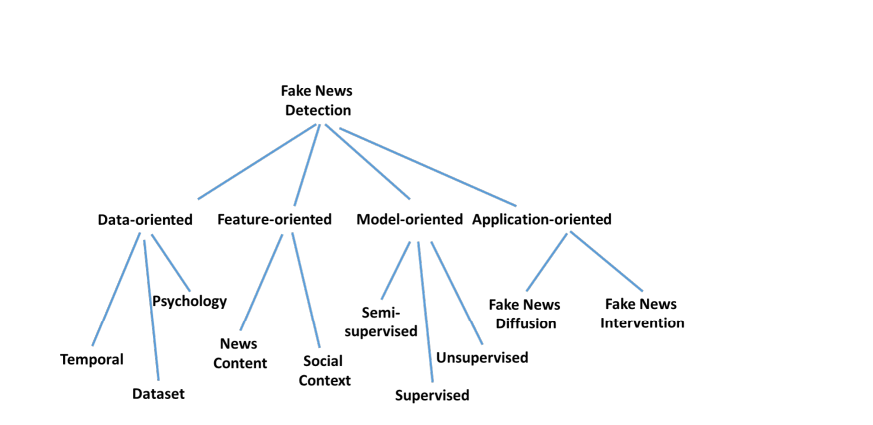
.

**MOTIVATION**

The motivation for research on this topic was that this is a relatively new area of research with many opinions but not many concrete solutions. Many implementations focus primarily on the host of the article, but even articles hosted on otherwise trustworthy websites can be classified as fake news. The primary motivation of this project was to bring awareness, propose a solution, and work towards minimizing the effects of fake news.

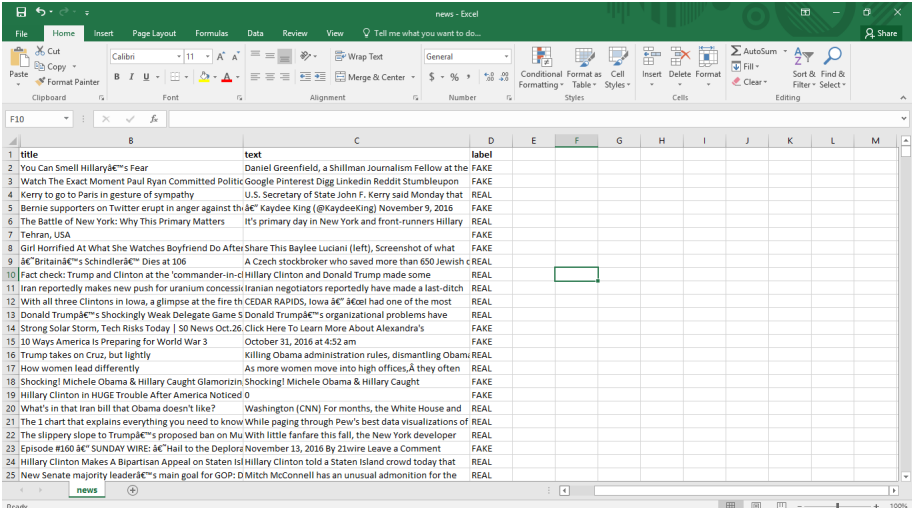
**OBJECTIVE**

Since this problem is a kind of text classification, Implementing a Naive Bayes classifier will be best as this is standard for text-based processing. The actual goal is in developing a model which was the text transformation (count vectorizer vs tfidf vectorizer) and choosing which type of text to use (headlines vs full text). Now the next step is to extract the most optimal features for count vectorizer or tiff-vectorizer, this is done by using a number of the most used words, and/or phrases, lower casing or not, mainly removing the stop words which are common words such as “the”, “when”, and “there” and only using those words that appear at least a given number of times in a given text dataset.



**METHODOLOGY**

The Dataset

****

The dataset is simple. It contains the titles of the news, the body text, and a label field, which, if the news is authentic, shows REAL and if inauthentic, shows FAKE.

There are 3 main segments of the methodology:

* The core Machine Learning model.
* The web interfaces.
* The common platform that brings the model and the interface together.

The Machine Learning Model

There are two parts to the ML Model building. Machine Learning is a part of our life that can help us in predicting. We are using two types of models in this case. For the first part, we used passive-aggressive classifiers.

And the steps include:

1. Data Loading: We are loading a CSV file for the data sorting and training-testing part of the model. The CSV file is turned into an array for easier work purpose.

2. Vectorization: Vectorization is needed for determining the frequency of the words present in a passage. This is needed to determine which words are used often.

3. Classifier: Passive-aggressive algorithms are a family of great learning algorithms. They are similar to Perceptron because it does not require a reading scale. However, unlike Perceptron, they include parameter correction. Passive is used when the prediction is correct and there is no Fake News Detector 16 change in the model. But if there is any kind of change in the model, that is if the prediction is not correct then the aggressive part is called, which changes the model accordingly. The aggressive part of the model changes the model according to its wish on the backend.

Diagram

Description automatically generated

4. Model Building: The model is built through the train and test of the dataset, by ensuring that the training is done for 80% of the dataset and testing is done in the rest of the 20% of the dataset.

Diagram

Description automatically generated

**SYSTEM ARCHITECTURE**

Import Libraries

Extract Data Set

Analyse Data

Validation of the Dataset

Splitting Data

Training the dataset

Calculating the accuracy

Making the prediction

**CODE & OUTPUT**

**Graphical user interface, text

Description automatically generated**

**Text

Description automatically generated**

**A screenshot of a computer

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

**Text

Description automatically generated**

**vText

Description automatically generated** **Text

Description automatically generated**

**Text

Description automatically generated**

**Graphical user interface, text

Description automatically generated**

**Text

Description automatically generated**

**Graphical user interface, text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

**CONCLUSION**

Our project can ring the initial alert for fake news. The model produces worse results if the article is written cleverly, without any sensationalization. This is a very complex problem, but we tried to address it as much as we could. We believe the interface provides an easier way for the average person to check the authenticity of a news. Projects like this one with more advanced features should be integrated on social media to prevent the spread of fake news. A python is open-source software, and the machine computation is also faster than many others and the cost might decrease dependently. Through this project, we want to make the process a bit easy and simple to achieve and use.

**REFERENCES**

* Kaggle : <https://www.kaggle.com/>
* YouTube : <https://www.youtube.com/>
* Wikipedia : <https://www.wikipedia.org/>
* A. Douglas, “News consumption and the new electronic media,” *The International Journal of Press/Politics*, vol. 11, no. 1, pp. 29–52, 2006.

View at: [Publisher Site](https://doi.org/10.1177/1081180X05284317) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=News%20consumption%20and%20the%20new%20electronic%20media&author=A.%20Douglas&publication_year=2006)

* J. Wong, “Almost all the traffic to fake news sites is from facebook, new data show,” 2016.

View at: [Google Scholar](https://scholar.google.com/scholar_lookup?title=Almost%20all%20the%20traffic%20to%20fake%20news%20sites%20is%20from%20facebook%2C%20new%20data%20show&author=J.%20Wong&publication_year=2016)

* D. M. J. Lazer, M. A. Baum, Y. Banker et al., “The science of fake news,” *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

View at: [Publisher Site](https://doi.org/10.1126/science.aao2998) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=The%20science%20of%20fake%20news&author=D.%20M.%20J.%20Lazer&author=M.%20A.%20Baum&author=Y.%20Benkler&publication_year=2018)

* Kaggle, *Fake News*, Kaggle, San Francisco, CA, USA, 2018, <https://www.kaggle.com/c/fake-news>.
* Kaggle, *Fake News Detection*, Kaggle, San Francisco, CA, USA, 2018, <https://www.kaggle.com/jruvika/fake-news-detection>.