### SAVITRIBAI PHULE PUNEUNIVERSITY



**A MINI PROJECT REPORT ON**

# FAKE NEWS DETECTION

Submitted by

## Name : Manasi.Renuse Roll no :25

**CLASS: TE DIV:B**

## Under the Guidance of

Ms. Pradnya Kasture



**DEPARTMENT OF COMPUTER ENGINEERING RMD SINHGAD SCHOOL OF ENGINEERING** WARJE, PUNE 411058

### 2023 - 24



**DEPARTMENT OF COMPUTER ENGINEERING RMD SINHGAD SCHOOL OF ENGINEERING** WARJE, PUNE 411058

# CERTIFICATE

This is to certify that the project report entitles

### Fake News Detection

*Submitted by*

Name:Manasi.Renuse PRN No :72217955D

is a bonafide work carried out by them under the supervision of Ms. Pradnya Kasture and it is submitted towards the partial fulfillment of the requirement of University of Pune for Third Year.

### (Ms. Pradnya Kasture ) (Dr. Vina M. Lomte)

Guide Head,

Department of Computer Engineering Department of Computer Engineering

Pune

Date: 16/4/24

**(Dr. V. V. Dixit)** Principal,

RMD Sinhgad School of Engineering Pune – 58 Place:

**Certificate by Guide**

This is to certify that Ms.Manasi.Renuse has completed the MINI Project work under my guidance and supervision and that, I have verified the work for its originality in documentation, problem statement, implementation and results presented in the Project. Any reproduction of other necessary work is with the prior permission and has given due ownership and included in the references.

Place: Pune Date: 14/4/2024

Signature of Guide

**(Ms. Pradnya Kasture )**

## ACKNOWLEDGEMENT

It is our pleasure to acknowledge sense of gratitude to all those who helped us in making this seminar. We thank our Mini Project Guide **Ms. Pradnya Kasture** for helping us and providing all necessary information regarding our project. We are also thankful to **Dr. Vina M. Lomte (Head - Department of Computer Engineering)** for providing us the required facilities and helping us while carrying out this seminar work.

Finally we wish to thank all our teachers and friends for their constructive comments, suggestions and criticism and all those directly or indirectly helped us in completing this seminar.

**NAME OF THE STUDENTS**

## Manasi.Renuse

**ABSTRACT**

The purpose of this thesis is to assist in automating the detection of Fake News by identifying which features are more useful for different classifiers. The effectiveness of different extracted features for Fake News detection are going to be examined. When classifying text with machine learning algorithms features have to be extracted from the articles for the classifiers to be trained on. In this thesis, several different features are extracted: word counts, ngram counts, term frequency-inverse document frequency, sentiment analysis, lemmatization, and named entity recognition to train the classifiers. Two classifiers are used, a Random Forest classifier and a Naïve Bayes classifier. Training on different features combined with different machine learning algorithms yields different accuracies. By testing the different features on different classifiers, it can be determined which features are the best for Fake News detection. Classifying news articles as either Fake News or as not Fake News is explored using three datasets, which in total contains over 40,000 articles. One of the datasets is used to partly to train the classifiers and partly to test the classifiers. The remaining two datasets are used purely for testing the classifiers. All the code used in conjunction with thesis can be found in Appendix B

## CONTENTS

### I Certificate

**III Certificate by Guide IV Acknowledgement V Abstract**

### Introduction page no 1

1. **Technology Used page no 2**

### Filter page no 3

1. **Algorithm page no 4**

### 5. Implementation…………………………………………………………..page no 7

1. **Output… page no 8**
2. **Conclusion page no 9**

# INTRODUCTION

The purpose of this thesis is to assist in automating the detection of Fake News by identifying which features are more useful for different classifiers. The effectiveness of different extracted features for Fake News detection are going to be examined. When classifying text with machine learning algorithms features have to be extracted from the articles for the classifiers to be trained on. In this thesis, several different features are extracted: word counts, ngram counts, term frequency-inverse document frequency, sentiment analysis, lemmatization, and named entity recognition. Two classifiers are used, a Random Forest classifier and a Naïve Bayes classifier. Training on different features combined with different machine learning algorithms yields different accuracies.

By testing the different features on different classifiers, it can be determined which features are the best for Fake News detection. Classifying news articles as either Fake News or as not Fake News is explored using three datasets, which in total contains over 40,000 articles. One of the datasets is used to partly to train the classifiers and partly to test the classifiers. The remaining two datasets are used purely for testing the classifiers. All the code used in conjunction with thesis can be found in Appendix B. The

term Fake News has many definitions, for this paper we will be using Axel Galfert’s .

“Fake news is the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design.” Although some form of Fake News has been around for many years, it is now mainstream and is widely considered to be a major issue .

The 2016 presidential election and Brexit are clear 6 examples of the relevance of Fake News in modern society . With the nature of the Internet as it is, anybody can spread untrue and biased information. It is virtually impossible to prevent Fake News from being created. Therefore, the next best thing is to find a way to identify and differentiate Fake News from real news.

One of the ways to determine validity is to fact check, but this is time consuming and requires skills that are not shared by everyone. The next best thing is to automate the detection of Fake News by using the methods and techniques of Data Science

# TECHNOLOGY USED

### Machine Learning Library

The engine of the recommendation system filters the data via different machine learning al- gorithms, and based on that filtering, it can predicts the most relevant entities to be recom- mended. After studying the previous behaviours of the users, it recommends

* + - Machine Learning Library:

1. pandas
2. numpy
3. Difflib
4. AST
5. scikit-learn

pandas numpy difflib AST scikit-learn

### 2.Requirement

Python 3.6

# FILTER

### Text Analysis Filters:

TF-IDF (Term Frequency-Inverse Document Frequency): This filter evaluates the importance of a word in a document relative to a collection of documents. Words that are common in a specific document but rare in the overall corpus may indicate uniqueness or relevance, which can be used to detect fake news.

N-grams: Analyzing sequences of words (n-grams) can help identify patterns and linguistic characteristics associated with fake news.

Semantic Analysis: Leveraging natural language processing (NLP) techniques to understand the meaning and context of text can aid in identifying deceptive language or misleading content.

### Social Context Filters:

Source Credibility: Assessing the reputation and reliability of the news source can be crucial. This filter may consider factors such as domain authority, history of misinformation, and journalistic standards.

User Engagement: Analyzing user interactions such as likes, shares, and comments can provide insights into the virality and credibility of the news content.

Sentiment Analysis: Understanding the sentiment expressed in the news article or social media post can help discern biased or emotionally manipulative content.

### Network Analysis Filters:

Propagation Patterns: Analyzing the spread of news across social networks can reveal suspicious dissemination patterns associated with fake news campaigns.

Bot Detection: Identifying automated accounts (bots) involved in spreading misinformation can be crucial for filtering out fake news sources.

Community Detection: Identifying clusters or communities of users who frequently interact with each other and share similar content can help identify potential echo chambers or sources of misinformation.

### Multimodal Filters:

Image and Video Analysis: Analyzing visual content for signs of manipulation, such as image tampering or deepfake videos, can complement text-based analysis in detecting fake news.

Audio Analysis: Examining audio content for anomalies or discrepancies can help detect fake news in multimedia formats.

### Machine Learning Filters:

Supervised Learning: Training machine learning models on labeled datasets to classify news articles or social media posts as real or fake based on extracted features.

Unsupervised Learning: Clustering techniques can be used to identify patterns or anomalies in large datasets, potentially uncovering fake news campaigns or coordinated disinformation efforts.

### Temporal Filters:

Temporal Analysis: Examining the temporal characteristics of news propagation, such as the timing of publication or the rate of dissemination, can provide insights into the credibility of the information.

**ALGORITHM**

### Naive Bayes Classifier:

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. It's often used for text classification tasks, including fake news detection.

### Logistic Regression:

Logistic regression is a linear model used for binary classification. It's well-suited for fake news detection when the relationship between input features and the target variable is linear.

### Random Forest:

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's known for its robustness and effectiveness in handling high-dimensional data.

### Support Vector Machines (SVM):

- SVM is a supervised learning algorithm that can be used for classification tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVMs are effective in highdimensional spaces and when the number of dimensions exceeds the number of samples.

### Deep Learning Models:

-Convolutional Neural Networks (CNN): CNNs are commonly used for image analysis, but they can also be applied to text classification tasks by treating words or characters as spatial features.

* Recurrent Neural Networks (RNN): RNNs, particularly Long Short-Term Memory (LSTM) networks, are well-suited for sequential data like text. They can capture contextual information and dependencies between words in sentences.
* Transformer Models: Transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) have achieved state-of-the-art performance in natural language processing tasks, including text classification and sentiment analysis. They can effectively capture longrange dependencies in text data.

### Ensemble Methods:

- Ensemble methods combine multiple models to improve predictive performance. Techniques like bagging (e.g., Bootstrap Aggregating) and boosting (e.g., AdaBoost) can be used to combine the predictions of multiple weak classifiers into a strong classifier, thereby enhancing the overall fake news detection accuracy.

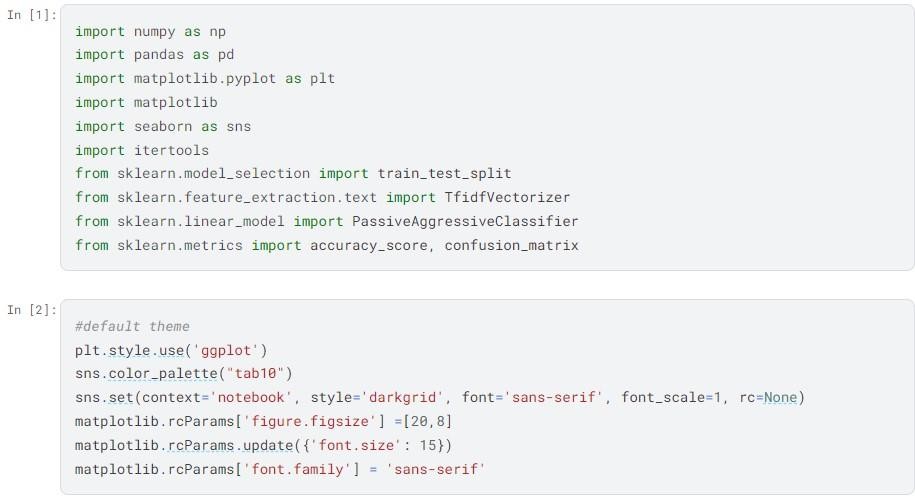
### Clustering Algorithms:

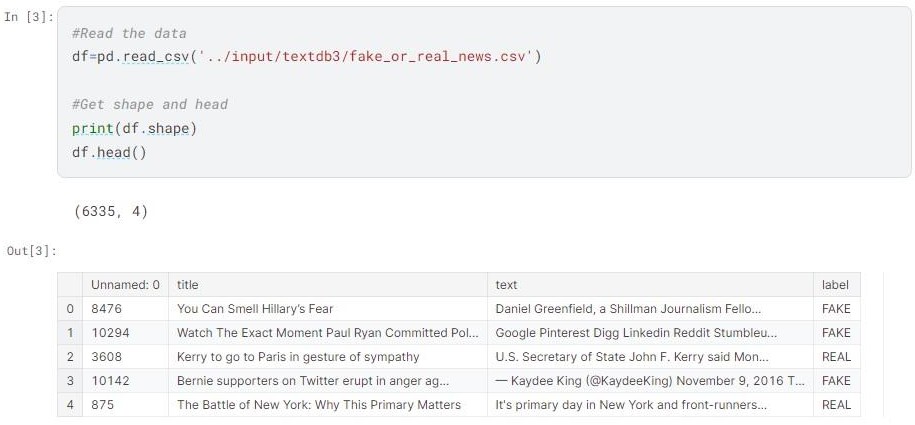
- Unsupervised clustering algorithms like k-means or hierarchical clustering can be used to identify clusters of similar news articles or social media posts. Anomalies or outliers within these clusters may indicate potential instances of fake news.

### Graph-Based Algorithms:

- Graph-based algorithms such as PageRank or HITS (Hyperlink-Induced Topic Search) can be used to analyze the network structure of news sources or social media platforms. Suspicious nodes or communities with low credibility may be indicative of fake news propagation.

# IMPLEMENTATION



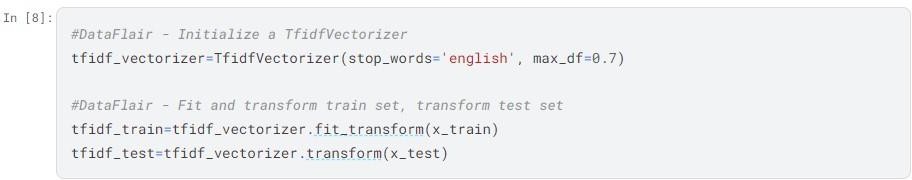


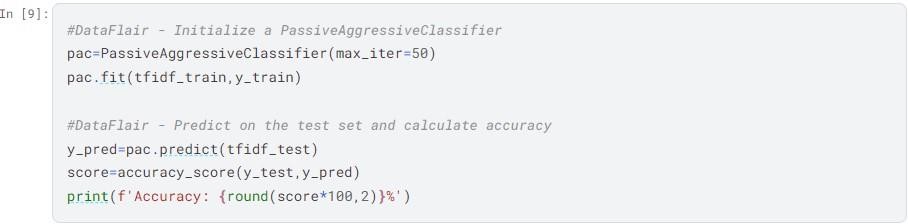


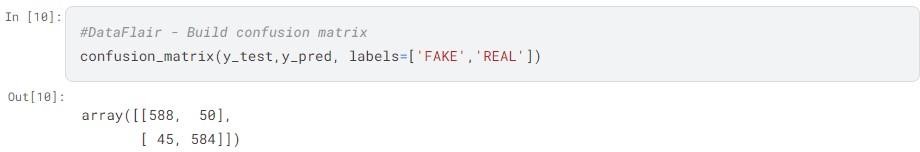


\









### CONCLUSION

With the nature of the Internet as it is, Fake News is easily created and distributed. Fact checking is tedious and time consuming, so automating Fake News detection is critical. Thus Fake News classifiers should be created. However, a classifier does not come out of thin air, it must be trained on already existing data. The quality and quantity of the data is important. Three datasets were used for the research in this thesis. ISOT, a huge dataset of over 40,000 articles. FakeNewsNet is another, much smaller dataset containing 422 articles. Lastly, the Original dataset, containing 180 articles, that was gathered specifically for this research. However, a classifier cannot read, so it must have features extracted for the articles. A feature is a numeric value extracted from the article. Such as a word count, or a count of parts of speech, or more complicated features. Such as a count of the named entities, like businesses or organizations.