**Mini Project Report on**



**Fake News Detection Using ML**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Name: University Roll no.**

**Himanshu Vashishtha 220211303**

***Under the Mentorship of***

**Dr. Pawan Kumar**

**Associate Professor**

**Graphic Era (Deemed to be University)**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2024**



**CANDIDATE’S DECLARATION**

I hereby certify that the work being presented in the project report entitled **“Fake news detection using machine learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Pawan Kumar,** Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Himanshu Vashishtha 220211303

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1** |
| Chapter 2 | Literature Survey | **2-3** |
| Chapter 3 | Methodology | **3-6** |
| Chapter 4 | Result and Discussion | **7-10** |
| Chapter 5 | Conclusion and Future Work | **11-12** |
|  | References | **13** |

**Chapter 1**

**Introduction**

* 1. **Problem Statement**

As our world becomes more connected to the internet, we are also becoming more diverse. This leads to an increasingly interconnected and diverse global society. In this society, accurately predicting whether news is fake or not is more crucial. This project addresses the challenge of Fake News Detection using ml, this project aims to develop a solution that enhances the accuracy of predicting fake news.

* 1. **Introduction**

The project demonstrates Fake news detection implemented in Python using machine learning techniques. The project involves preprocessing textual data, building a classification model, and saving it for future sentiment predictions on new, unseen news. The project utilizes a logistic regression model with the 'count-vectors' method for classifying news into fake and real. Key steps include data preprocessing, column merge, and feature extraction through the Count Vectorizer. The model is trained on a labeled dataset and evaluated using cross-validation to measure its accuracy.

* 1. **Objectives**
* Data acquisition: Retrieve and load news datasets (train and test) to analyze credibility. Understand the structure and characteristics of the data.
* Data preprocessing: Cleanse and preprocess news text, including lowercasing, and removal of special characters and digits. Remove stopwords to focus on essential words for sentiment analysis. Apply stemming to standardize word forms.
* Exploring Data Analysis (EDA): Explore the distribution of sentiment labels in the training set. Visualizing the key features and patterns within the data.
* Feature Engineering: utilize CountVectorizer.

**Chapter 2**

**Literature Survey**

A fake news detector is an informally titled data science implementing a model that is capable of detecting and classifying fake and true news from provided data.

The project focuses on creating a method for identifying fake news that is based on machine learning. The distribution of false information and fake news has grown to be a significant concern as a result of the development of online news outlets and social media. The goal of this research is to develop an automated system that can tell the difference between authentic and fake news reports.

With the popularization of digital technology, the problem of information pollution caused by fake news has become more common. Malicious dissemination of harmful, offensive, or illegal content may lead to misleading, misunderstanding, and social unrest, affecting social stability and sustainable economic development. With the continuous iteration of artificial intelligence technology, researchers have carried out automatic and intelligent news data mining and analysis based on aspects of information characteristics and realized the effective identification of fake news information. However, the current research lacks the application of multidisciplinary knowledge and research on the interpretability of related methods. This paper focuses on the existing fake news detection technology. The survey includes fake news datasets, research methods for fake news detection, general technical models, and multimodal-related technical methods.

To achieve precise fake news identification, the research combines data pretreatment, feature extraction, and machine learning methods. The dataset consists of tagged news items, both real and false samples being included. The data pretreatment step entails text cleaning, the elimination of stop words and special characters, and formatting the material in a way that makes it appropriate for analysis.

Several machine learning techniques, such as Gradients, Decision Trees, and Logistic Regression.

**2.1 Related Works**

**2.1.1"A Survey on Fake News Detection Using Deep Learning"**

**Information Processing & Management, 2021** (Karimi, F., Gupta, D., & Alhameed, R. A.)  
This survey thoroughly summarizes recent work on deep learning-based fake news identification. The authors evaluate the benefits and drawbacks of neural network models for detecting fake news based on textual and visual material, such as LSTM, CNN, and BERT.

**2.1.2"Detecting and Tracking Political Fake News**"

(Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. - arXiv preprint arXiv:1708.06233, 2017) is a research paper that was published in 2017.

The goal of this study is to identify and monitor political fake news on social media. It proposes a diffusion-based approach for locating the origins of false information and its channels of dissemination inside online networks.

**2.1.3 "Automatic Detection of Fake News"**

(Rubin, V. L., Chen, Y., & Conroy, N. J. - Proceedings of the 80th Annual Meeting of the Association for Information Science & Technology, 2017) was published in this journal.

The study offers a machine learning method for automatically identifying whether news articles are authentic or not. It uses linguistic and structural factors to train classifiers like SVM and Random Forests to discern between trustworthy and deceptive news material.

**2.1.4 "Fake News Detection: A Deep Learning Approach"**

To identify bogus news, this study investigates a deep learning architecture using convolutional and recurrent neural networks. The authors show how word embeddings and character-level representations work well for identifying authentic news articles from false ones.

**2.1.5"Detect Rumors using Time Series of Social Context Information on Microblogging Websites"**

(Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. - Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011) is an article that was published in the conference proceedings.

This work examines rumor detection on microblogging sites, comparable to misinformation detection even though it is not specifically focused on fake news. It emphasizes the significance of temporal characteristics and context details for spotting false information.

**2.1.6 Fake News Early Detection: A Graph-Based Approach"**

(Wu, S., Yang, D., Chen, Z., & Zhou, X. - World Wide Web Conference 2018 Proceedings, 2018)In this study, a graph-based method is investigated for the early detection of bogus news. The authors provide a strategy to pinpoint disinformation sources and their influence by looking at how news pieces spread within the social network.

**Chapter 3**

**Methodology**

**3.1 Data Collection**

* Gathered a labeled dataset from Kaggle containing news with their corresponding labels (true or false).
* The dataset should be diverse and representative of the target domain.

**3.2 Data Preprocessing**

* Load the dataset using pandas and explore its structure.
* Handle any missing values in the dataset.
* Clean the text data by converting it to lowercase, removing special characters, and eliminating numbers
* Remove stopwords and apply stemming to reduce words to their base or root form.

**3.3 Data Exploration**

* Explore the distribution of sentiment labels in the training dataset.

**3.4 Text Vectorization**

* Utilize techniques like Count Vectorization or TF-IDF to convert the processed text data into numerical vectors.
* This step is crucial for machine learning models to process and understand textual information.

**3.5 Model Development**

* Choose a suitable machine learning model for sentiment analysis. In this case, a Logistic Regression model is used.
* Split the dataset into training and testing sets.
* Train the model using the training set and evaluate its performance on the testing set.

Data Collection

Text Vectorization

Data Preprocessing

Data Exploration

**Credibility Prediction**

**Graphical representation and confusion matrix**

Model Evaluation

Model Development

**3.6 Model Evaluation**

* Employ cross-validation techniques to assess the model’s performance robustly.
* Evaluate metrics such as accuracy and make a predictive system to gauge the model’s effectiveness.
  1. **Graphical representation and Confusion matrix**
* Plot a graph between total fake and true values.
* Built a confusion matrix among data.

**3.8 Model Saving**

* Save the trained model using.

**3.9 Credibility Prediction**

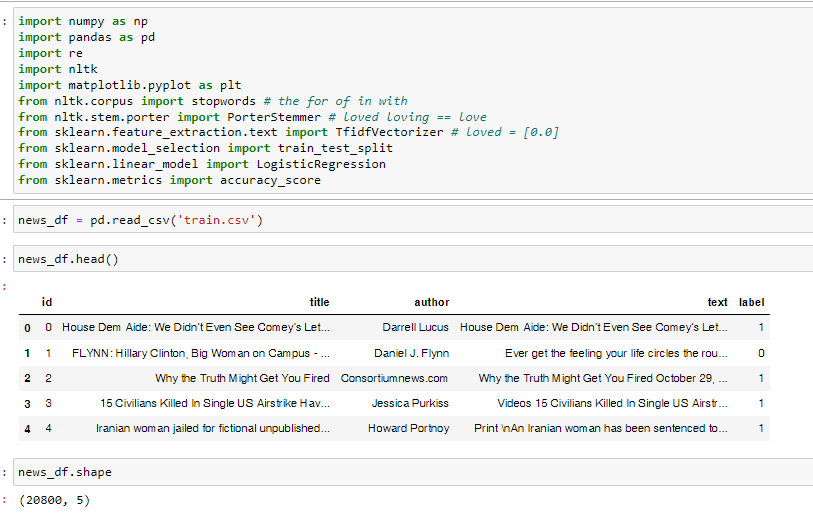
* Load the saved model when needed.
* Process new news data using the same preprocessing steps applied during training.
* Use the loaded model to predict the credibility of the data.

**Chapter 4**

**Result and Discussion**

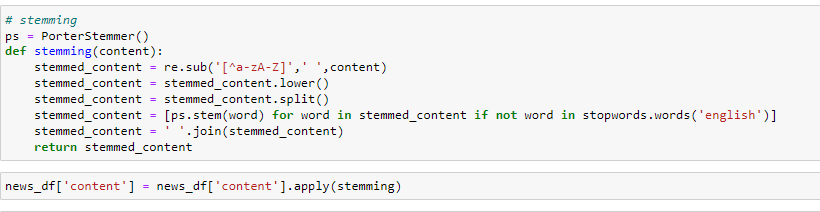
**4.1 Data Preprocessing**

* The training and test sets are loaded and explored

****

**Fig. 4.1 Loading and exploring dataset**

* Text data is cleaned by converting to lowercase, removing special characters, and eliminating spaces.



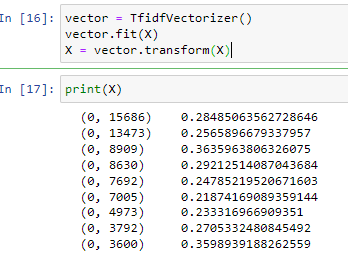
* **Fig. 4.2** Stemming content

**4.2 Data Exploration**

* The distribution of credibility labels in the training dataset is explored.
* Visualization may indicate class imbalances or provide insights into the dataset.

**4.3 Text Vectorization**

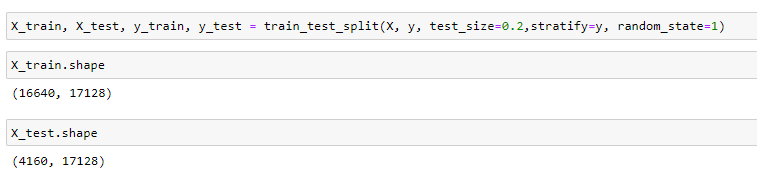
* The news data is transformed into numerical vectors using Count Vectorization.

****

**Fig 4.4 Text Vectorization**

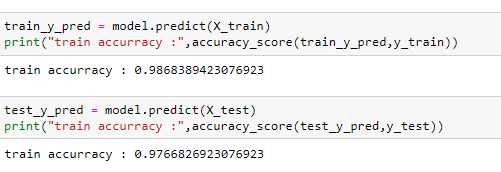
**4.4 Model Development**

* Logistic Regression model is trained on the transformed text data.

****

**Fig 4.5 Model Development**

**4.5 Model Evaluation**

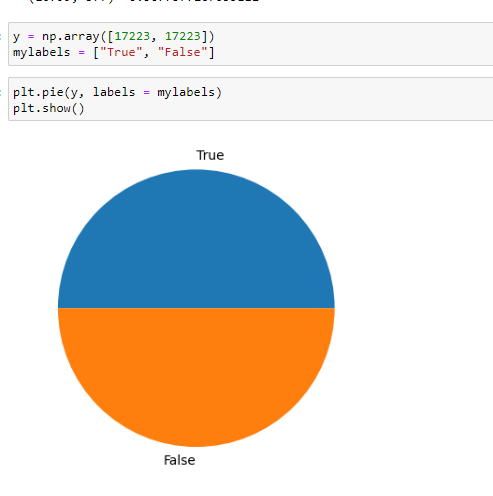
****

**Fig 4.6 Model Evaluation**

**4.6 Grpahical reprsentatrion and confusion matrix**

****

**Fig 4.6 Confusion matrix**

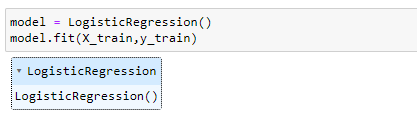
****

**Fig 4.6 Graphical representation**

**It is a graphical representation between 0 and 1 where 0 means fake and 1 for real.**

**4.7 Model Saving**

The trained Logistic Regression model is saved for future use

****

**Fig 4.7 Model Saving**

**4.7 Credibility Prediction**

* A new data is processed using the same preprocessing steps.
* The processed data is vectorized using the previously fitted vectorizer.
* The loaded model predicts the credibility of the news(true or false).
* The predicted credibility is displayed.

****

**Fig 4.8 Credibility Prediction**

**Chapter 5**

**Conclusion and Future Work**

**5.1 Conclusion**

In this project, we discussed the prevalence of fake news and how technology has changed over the last years enabling us to develop tools that can be used in the battle against fake news. We also explored the importance of identifying fake news, the influence that misinformation can have on the public’s decision-making, and which approaches exist to combat fake news.

With the popularization of digital technology, the problem of information pollution caused by fake news has become more common. Malicious dissemination of harmful, offensive, or illegal content may lead to misleading, misunderstanding, and social unrest, affecting social stability and sustainable economic development. With the continuous iteration of artificial intelligence technology, researchers have carried out automatic and intelligent news data mining and analysis based on aspects of information characteristics and realized the effective identification of fake news information. However, the current research lacks the application of multidisciplinary knowledge and research on the interpretability of related methods. This paper focuses on the existing fake news detection technology. The survey includes fake news datasets, research methods for fake news detection, general technical models, and multimodal-related technical methods.

**Future Work**

The project's future work will focus on several areas for development and enhancement to increase the system's precision and efficiency. For upcoming work, consider the following perspective directions**:**

**1. Incorporate deep learning**  It is possible to investigate networks (RNNs) or transformer models (like BERT) to capture more intricate patterns and contextual information in text data. These models demonstrated promising outcomes in challenges involving natural language processing and may improve the effectiveness of fake news detection.

**2. Multimodal Data Analysis:**

Taking into account additional modalities, such as photographs, videos, or user metadata, can help us understand bogus news more thoroughly. The accuracy and robustness of the model can be increased by combining textual data with visual and behavioral inputs.

**3. Adversarial Defense Techniques:** Examining adversarial attacks and creating defense mechanisms might strengthen the model's resistance to complex methods used to produce false information. Examining adversarial attacks and creating defense mechanisms might strengthen the model's resistance to complex methods used to produce false information.

**4. Real-Time Monitoring and Feedback:**

By extending the model to function in real-time or almost real-time circumstances, fake news dissemination can be actively identified and stopped before it spreads. It may be possible to stop the spread of false information by developing a scalable system that continuously scans news outlets and social media websites.

We can progress in the field of fake news identification and create more complex and efficient solutions by tackling these upcoming development topics. To remain ahead of emerging strategies and preserve the integrity of information distribution, continual research and innovation are crucial in the ongoing battle against fake news.

**References:**

**[1]**[**https://www.kaggle.com**](https://www.kaggle.com) **for the dataset(Train.csv).**

[2] <https://www.mygreatlearning.com/blog/open-source-python-libraries/> for basic ideas about Python libraries.

[3] **Information Processing & Management, 2021** (Karimi, F., Gupta, D., & Alhameed, R. ) refers to a literature survey.

[4] (Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. - arxiv preprint arXiv:1708.06233, 2017) is a research paper that was published in 2017.

[5] (Rubin, V. L., Chen, Y., & Conroy, N. J. - Proceedings of the 80th Annual Meeting of the Association for Information Science & Technology, 2017) was published in this journal.

[6] Abirami, A. M., and V. Gayathri. "A survey on sentiment analysis methods and approach." In 2016 Eighth International Conference on Advanced Computing (ICoAC), pp. 72-76. IEEE, 2017.

[8]<https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/7438/1628083317.pdf>.

[9]<https://www.streamlit.com> for graphical user interphase.