A PROJECT REPORT

on

"FAKE NEWS CLASSIFICATION USING NATURAL LANGUAGE PROCESSING"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of
BACHELOR'S DEGREE IN
COMPUTER SCIENCE & COMM ENG

BY

AVINEET JENA	**2029011**	
SUJAL KUMAR	**SANDESH TRIPATHI**	**2029213**

UNDER THE GUIDANCE OF

LIPIKA MOHANTY

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

May 2023

A PROJECT REPORT

on

"FAKE NEWS CLASSIFICATION USING NATURAL LANGUAGE PROCESSING"

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR'S DEGREE IN INFORMATION TECHNOLOGY

BY

AVINEET JENA	**2029011**
SANDESH TRIPATHI	**2029213**
SUJAL KUMAR	**2029036**

UNDER THE GUIDANCE OF

LIPIKA MOHANTY

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAE, ODISHA -751024

May 2022

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

"FAKE NEWS CLASSIFICATION USING NATURAL LANGUAGE PROCESSING "

submitted by

AVINEET JENA	2029011
SANDESH TRIPATHI	2029213
SUJAL KUMAR	2029036

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & System Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2022-2023, under our guidance.

Date: 05/05/2022

(Guide Name)

Acknowledgements

We are profoundly grateful to LIPIKA MOHANTY of **Affiliation** for her expert

guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We are also thankful to School of Computer Science and Communication Engineering

For helping us in acquiring knowledge which was essential for this project and providing us a opportunity to work on this project.

AVINEET JENA
SANDESH TRIPATHI
SUJAL KUMAR

ABSTRACT

News that has misleading and wrong information is reportedly known as fake news. The aim of fake news is to harm a particular person's image or reputation in public. The term "fake news" was first used in the 1890s, a time when dramatic newspaper stories were prevalent, despite the fact that incorrect information has always been disseminated throughout history. In the 2016 U.S. presidential election, for instance, a BuzzFeed News research discovered that the top fake news items generated more Facebook engagement than the top stories from established media outlets. Additionally, it has the potential to erode public confidence in objective media coverage.

It takes a lot of effort and time to combat bogus news. Fake news is having a previously unheard of effect on people's lives as well as on the political and social climate. This situation has increased the popularity of projects dealing with automatic fake news identification, which has peaked academic interest. When trying to develop such solutions, most strategies geared towards English and languages with limited resources encounter obstacles. This study concentrates on the advancement of these investigations, outlining current solutions, difficulties, and findings that were observed by various research groups. We also assess the applicability of current methodologies in the setting while recommending future research areas considering the dearth of automated analysis on false news.

Keywords-: Natural language processing, fake news identification, Natural Language Processing Techniques, classification techniques for fake news identification, news classifier.

Contents

List of Figures

Chapter 1

Introduction

With a growing amount of time spent online on social media, people are increasingly relying on these platforms as a source of news rather than traditional media outlets. This shift can be attributed to the convenience and speed of accessing news through social media, as well as the ability to easily share and discuss news with others on these platforms. The quality of news on social media is nevertheless worse than that of traditional news outlets because fake news, which is material that is purposefully incorrect, is easily disseminated online for a variety of reasons, such as political and financial gain. The widespread impact of fake news can include undermining the authenticity of news, promoting biased or false beliefs, and altering how people interpret and respond to real news. Therefore, it is important to develop methods to automatically detect fake news on social media to help address its negative effects.

Researchers are currently focused on understanding and detecting fake news. While the idea of fully automating the process is still a topic of discussion, most current efforts involve identifying patterns in already disseminated fake news or proposing new features for training classifiers. It is uncertain how effective these proposed features are in detecting fake news. Detecting the fake news on social media platforms is very difficult work. fake news is misleading, making it difficult to detect based solely on its content and social media data is massive, is a no-sql data without proper orientation, generated by users, and often anonymous and noisy. To tackle these challenges, recent research has used social engagement data from users to determine if a news article is fake, yielding promising results. However, current methods for detecting fake news are effective, but do not provide an explanation for why a piece of news was deemed fake. This is significant because an explanation can reveal hidden insights and improve fake news detection performance. Furthermore, it would help to extract explainable features from noisy auxiliary information.

In this Report we have used many different machine learning algorithms such as the naïve Bayes classifier, supervised machine learning algorithms, principal component analysis (PCA), Criterion boost classification algorithm, logistic regression, Random forest, Support Vector Classification algorithm (SVM).
And we have found out the accuracy and compared the accuracy using machine learning methods such as Precision, recall, f1 score and have formed a confusion matrix for them

Chapter 2

Basic Concepts/ Literature Review

Literature Review:

The Literature Review aims to gain an understanding of the most commonly used methods for compiling fake news corpora. fake news classifier would typically include an overview of previous research on the topic of fake news detection and classification, including relevant studies, theories, and methods that have been used in this field.The review also examines the various approaches used in news which is false , like deep learning, macine learning also natural language processing.

Beyond News Contents: The Importance of Social Context for Fake News Identification Author : Kai Shu (2019) in this study it used an approach for detection of false news by utilising the social context of social media's use for news transmission. When compared to conventional content-based methods, the TriFN method improves detection because it models publisher-news and user-news interactions. Experiments in the real world demonstrate its efficacy. The tri-relationship embedding framework, TriFN, suggested for better false news identification, models publisher-news relations and user-news interactions simultaneously. It was discovered to perform better than earlier baseline methods. a model that, in 48 hours on two datasets, achieved an F1 score of above 80%, showing excellent accuracy potential for early false news identification.

According to K Shu, A Sliva, S. Wang,(2017), who wrote False News Identification on Social Media: In Perspective Of Data Mining, identifying false news on social media is a difficult area of research because it is deliberately misleading and difficult to use user engagement data. This study emphasises relevant research topics and potential approaches while providing a thorough analysis of false news detection techniques, including characterizations, algorithms, metrics, and datasets.

In 2019 a research and analysis by Julio C. S. Reis, Andre Correia, Fabrcio Murai employs a number of elements for spotting bogus news on social media and offers fresh features for assessment. The findings offered new perspectives on the value of various variables and highlighted difficulties in the actual deployment of false news detecting techniques. In addition to the k-Nearest Neighbours (KNN), Naive Bayes (NB), Random Forests (RF), Support Vector Machine with RBF kernel (SVM), and XGBoost (XGB) classifiers, they also employed the supervised learning technique.While misclassifying 40% of the true news with a false positive rate of 0.4, RF and XGB classifiers produced the best results, statistically tied with 0.85 (+/- 0.007) and 0.86 (+/- 0.006) for AUC.

Using two new datasets, machine learning approaches, and a comparison of manual and automatic identification**, Verónica Pérez-Rosas, Bennett Kleinberg, (2017)** developed a novel method for recognising false information in online news sources. It offered a thorough answer to the false news issue. The report was sourced from ABCNews, CNN, USAToday, NewYorkTimes, FoxNews, Bloomberg, and CNET, among other major news outlets. crowdsourcing using FAKENEWS AMT and Amazon Mechanical Turk. With accuracy, precision, recall, and F1 measures averaged over the five iterations and machine learning classification, the model used an SVM classifier and five-fold cross-validation

It was suggested by **Robiert Seplveda Torres, Marta Vincente, Estela Saquette,(2021)** to utilise automated news summaries rather than the complete text to establish the position of a headline in relation to the body of text. The strategy seeks to save crucial information while reducing information processing. The suggested method may identify several stance categories with competitive results by using just pertinent information from automatic summaries, which suggests a good influence for figuring out the stance of brief material (such as a headline or phrase) in relation to its entire content. The two levels of the architecture were Relatedness Stage and Stance Stage. The TF-IDF vector, cosine similarity, and Python tools like NLTK are employed.

Mykhailo Granik and Volodymyr Mesyura (2017) published a research article with a straightforward methodology, implementing the Naive Bayes classifier as a software system for false news identification, and stating that artificial intelligence may be utilised to solve the fake news detection problem. The purpose of the study is to support the premise of utilising machine learning to detect false news by examining how well this particular approach performs for the specific problem with a manually tagged (fake or real) dataset. This article differs from others like it in that it largely relies on a Naive Bayes classifier, which is used to distinguish between false news and actual news.

In **2020, Jie Cao, Peng Qi, Qiming Sheng** looked at the use of visual content in the identification of bogus news. To identify bogus news, the authors suggest using a multimodal technique that integrates textual and visual data. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), used in conjunction, are used in the model's architecture to extract features from visual and textual data, respectively. It featured a dataset of news stories with labels indicating whether they are true or false, as well as the photographs that go with them. They evaluate how well their method performs in comparison to a number of cutting-edge false news detection techniques that rely solely on textual data.

A thorough overview of the current state of research on the use of machine learning algorithms for false news identification was published in **2019 by Syed Ishfaq Manzoor, Jimmy Singhla, and Nikita.** The authors conduct a review of the literature on the identification of false news and categorise the various methods into groups including feature-based, rule-based, and hybrid approaches. The accuracy of the models tested by the authors was only between 63 and 70 percent. The paper offers a thorough overview of current methods for machine learning-based false news identification.

A multi-modal system for identifying false news called **Spotfake** was created by **Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, in 2019.** According to the authors, this framework is distinctive since it can identify a greater variety of fake news than text-based techniques because it employs various modalities (text, picture, and video) for the purpose of detecting false news. The framework used by the authors has a three-stage design. Pre-processing is the initial stage, when text, picture, and video data are taken from the news story. The authors extract a collection of features from the text, picture, and video data during the feature extraction step, which is the second stage. The third stage is the classification stage, when the authors categorise the news as false or authentic using a machine learning classifier.

In their paper, Mhatre and Masurkar (2021) proposed a new hybrid technique for identifying false news. To validate the accuracy of the method, they utilized web-scraped data to assess the truthfulness of two news articles. The first step involved preparing the data with NLP techniques such as removing special characters, white spaces, stop words, and extracting text. Then, the words were organized based on their semantic similarities through lemmatization. This resulted in the creation of a corpus using TF-IDF vectorization, which was utilized to train the models. The authors suggested using cosine similarity calculated from topic modeling and the corpus to enhance the classification accuracy. Several classifiers, including logistic regression, passive-aggressive classifier, KNN, decision tree, naive Bayes, and SVM, were used to assess the news's veracity. The passive-aggressive classifier, which is used often to identify bogus news, received additional attention to boost its accuracy among all the classifiers.

Chapter 3

Problem Statement

Fake news is often misleading and is tried to be pushed as a legitimate news. It has a negative impact on the societal values. It can harm a person image or target a particular community. As a result, the detection of fake news on social media has recently emerged as an area of study that is receiving a lot of attention. The problem of fake news detection on social media is characterized by the intentional spread of false information, making it difficult to detect based solely on the news content.

Identifying fake news on social media is a complex task due to its unique features and challenges. The existing news media detection methods fail to detect it accurately. To identify fake news, additional information like user's social engagement on social media is required as the fake news is designed to deceive the reader with false information. However, even the use of additional information is challenging as the data generated by users' interaction with fake news is vast, inconsistent, unorganized, and noisy.

3.1 Project Planning

We are undertaking a project that involves analyzing two datasets: genuine news articles and fake news articles. The datasets consist of around 35,000 news stories, with an equal number of stories classified as true or false. The real news articles are sourced exclusively from Reuters and The Guardian, while the origin of the fake news articles is unknown. Our first step is to import necessary libraries and the Natural Language Tool Kit. Then, we will load the datasets. Exploratory Data Analysis (EDA) is an important part of data analysis in such projects. EDA involves examining the data to uncover patterns, verify assumptions, and test hypotheses using summary statistics and visualizations.

In the next phase of the project, we will perform pre-processing on the text data by dividing it into tokens, reducing words to their base form i.e. stemming, and removing commonly used words. After that, we will use TF-IDF vectorization to convert the text into numerical representations. Then, we will classify the news articles using machine learning algorithms, specifically Logistic Regression and

Passive-Aggressive classifier. To evaluate the performance of the models, we will generate a classification report and a confusion matrix, which will help us determine which model is more effective.

3.2 System Architecture **OR** Block Diagram

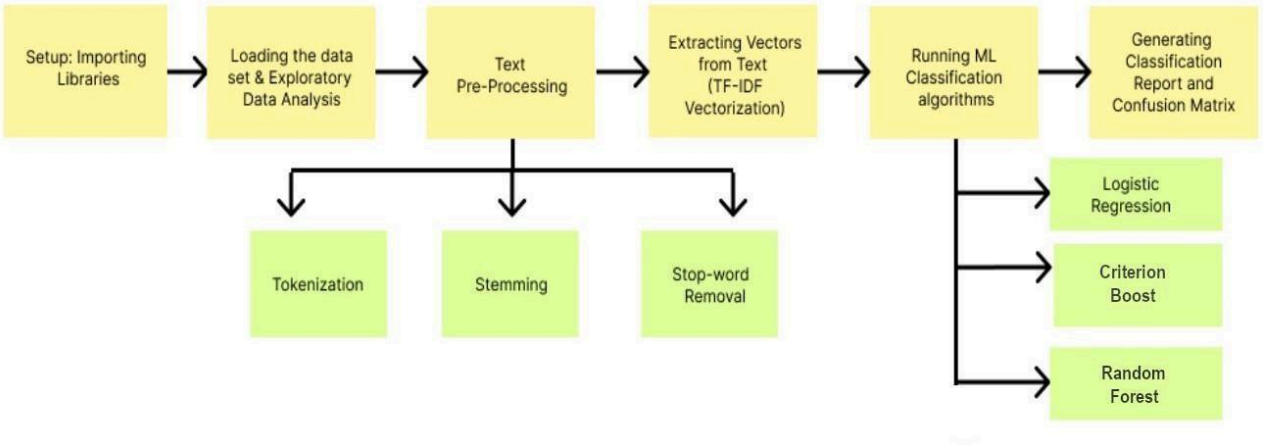

Fig:3.1 Architecture Diagram

- **Data Collection** : In this step we gathered a large and diverse dataset of news articles, labelled as either fake or real from various sources.
- **Text Preprocessing** : The data is cleaned and preprocessed to remove noise and stop-words and convert the text suitable for machine learning models.
- **TF-IDF Vectorization** : After text preprocessing we represent the text as a numerical vector. The TF-IDF value measures the importance of a word in a document relative to its importance in the corpus.
- **Training and Validation** : Now various datasets are split into training and validation sets. They undergo training under various machine learning algorithms like Random Forest, Logistic Regression etc.
- **Classification Report and Confusion Matrix** : Here we summarize the performance of a model on a set of test data with the help of various metrics like F1-score etc. and then a classification report is prepared.

Chapter 4

Implementation

In this section, present the implementation done by you during the project development.

4.1 Methodology OR Proposal

We have reviewed some selected studies to give a summary of how ML models are being applied. Some key terminologies that we identified were: Fake news, supervised learning, NLP and its libraries, text pre-processing, vectorization and classification algorithms.

4.1.1 Terminologies

1. **PCA- Principal Components Analysis (PCA)** : is frequently used to decrease the amount of associated variables (features) within the set of data. Weighted linear combinations of the original variables, which roughly maintain the statistical content of the original dataset, make up the decreased number of variables.
The original variables' physical significance, if any, is lost in the linear combinations of the initial variables.
PCA tries to get the best data transformation.Redundancy and noise in the data are possible issues

2. Dealing with noise:
Signal-to-Noise Ratio, or SNR, is a metric used to assess how strong a signal is in comparison to noise in the background in a medium of communication or system

$$\mathrm{SNR} = \frac{\sigma^2_{\mathrm{signal}}}{\sigma^2_{\mathrm{noise}}}$$

3. Supervised Learning-**Directed/Supervised learning**: Classification and regression in which output is known.Learning from a teacher is called supervised learning, it involves training a model on labelled dataset, where the correct output is provided by each input.

- **Classification**: In classification, it is expected that the machine will learn technique to classify the observation data. Output is discrete in nature.
- **Regression**: In regression , the output is continuous in nature.
- Decision trees classifier
- Naïve bayes classifier
- Support Vector Machine

4. **KNN- K-Nearest Neighbour**: founded on the Supervised Learning approach, it is one of the basic Machine Learning algorithms.

- The K-NN algorithm places the new case in the category that is most comparable to the available categories, supposing that the new instance/data and the existing cases are identical
- It is referred to as a lazy learner algorithm since it saves the training dataset rather than learning from it instantaneously. Instead, it uses the dataset to execute a task when classifying data.
- Setting k to a little value will result in a significant variance in predictions for any given problem; conversely, putting k to a big value could result in a model with a lot of bias.

5. Bayes Theorem:

- The Bayes theorem offers a method for calculating a hypothesis' probability based on the probabilities of its prior likelihood, the probability of discovering various sorts of information that the hypothesis predicts, and the real data that has been collected.
- Discrete random variable y's distribution of all potential values is represented by a probability distribution.

$$P(y_k|\mathbf{x}) = \frac{P(y_k)\,P(\mathbf{x}|y_k)}{P(\mathbf{x})}$$

$$P(\mathbf{x}) = \sum_{q=1}^{M} P(\mathbf{x}|y_q)\,P(y_q)$$

6. Naïve bayes classifier

• Many situations require us to calculate class-conditional likelihood P(x|yq).

$$P(\mathbf{x}|y_q) = \frac{\text{Number of times pattern } \mathbf{x} \text{ appears with } y_q \text{ class}}{\text{Number of times } y_q \text{ appears in the data}}$$

- The number of different P(xj | yq) words that must be calculated from the training data is exactly the number of unique characteristics (n) times the number of distinct groups (M), where yNB signifies the class output by the naive Bayes classifier.

$$y_{\text{NB}} = \arg\max_q P(y_q) \prod_j P(x_j|y_q)$$

7. NLP and it's Libraries

❖ NLP refers to the automatic or semi-automatic processing of human language, which can encompass tasks such as information retrieval and machine translation. It is sometimes referred to as more applied compared to computational linguistics. The distinction between NLP and speech processing is often made, with NLP broadly encompassing both. NLP is a subset of Artificial Intelligence.
Natural Language Toolkit features include:

- Text classification
- Part-of-speech tagging
- Entity extraction
- Tokenization
- Parsing
- Stemming
- Semantic analysis
- Information Retrieval
- Lemmatisation

❖ Scikit learn

It is a very helpful library for NLP and machine learning. It includes numerous algorithms like support vector machine, random forests, and k-nearest-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy. It provides various functions for solving text classification problems using the bag-of-words and TF-IDF method of feature selection.

8. **Classification algorithms:** For this research, we employed four ML models: the Criterion boost classifier and logistic regression. In addition to these models, Random Forest, Decision Tree, Naive Bayes .
To distinguish between fake and real news, classifiers such as Stochastic Gradient Descent and Support Vector Machine can be utilised.
A decision tree is a type of tree-structured classifier in which internal nodes represent dataset features, branches represent decision rules, and each leaf node represents the outcome. The Naive Bayes Classifier is a probabilistic classifier based on the Bayes' theorem, meaning it predicts based on an object's probability.

Explaining the Logistic Regression ,Criterion boost and random forest classifier

Logistic regression :- a statistical analysis is used for binary classification, or determining which of two possible groups an observation belongs to. The modelling of the link between a set of input variables and a binary output variable is widely utilised in machine learning and data analysis.

By determining the likelihood that the output variable belongs to one of the classes, logistic regression seeks to determine the input variables and output variable that fit together the best. The logistic function, on which the logistic regression model is built, transforms every real-valued input to a number between 0 and 1, which represents the

likelihood that the output variable belongs to one of the two classes

By determining the likelihood that the output variable belongs to one of the classes, logistic regression seeks to determine the input variables and output variable that fit together the best. The logistic function, on which the logistic regression model is built, transforms every real-valued input to a number between 0 and 1, which represents the likelihood that the output variable belongs to one of the two classes.

Criterion Boost classifier A machine learning method called Criterion Boost is used to enhance the performance of categorization models. It entails altering a classifier's decision boundary by giving various samples varying weights in accordance with how challenging it is to accurately categorise each one of them.
In order to give these samples more weight during training, criterion Boost seeks out misclassified samples. The classifier concentrates more on learning from the challenging examples and enhances its capacity to accurately classify them. Each time the training process is repeated, the weights are adjusted, and the algorithm attempts to minimise a cost function that accounts for the weighted samples.

Random Forest Classification- The machine learning algorithm, Random forest classification uses the supervised learning method. Based on the consensus of several decision trees, it can be used to categorise data into different groups. It is based on the idea of ensemble learning, which is the act of integrating various classifiers to address a difficult issue and enhance the functionality of the model.

9. Text Pre-processing:-

Text pre-processing is Cleaning and preparing textual data for subsequent analysis or machine learning tasks.
This process entails a number of processes, including eliminating punctuation, changing the text to lowercase, getting rid of stop words, stemming or lemmatizing the text, and turning it to a numerical representation. These procedures aid in standardising and streamlining the text data, making it simpler to model and analyse. In natural language processing (NLP) applications including sentiment analysis, text categorization, and information retrieval, text pre-processing is a crucial step.

Tokenization- A text or document is tokenized when it is divided up into tiny pieces called tokens, most often words or phrases. Tokenization aims to transform unprocessed text into a form that is simple for machines to understand and interpret. In natural language processing (NLP) activities including text categorization, sentiment analysis, and machine translation, tokenization is a crucial step.

Stemming- is the process of stripping words of any suffixes or prefixes to leave only the stem, which is the basic or root form of the word. Reduce the amount of unique words in a text while maintaining the words' primary meanings is the aim of stemming. Natural language processing (NLP) tasks like text categorization, sentiment analysis, and information retrieval frequently involve stemming.
Stemming algorithms operate by deleting frequent suffixes like -ing, -ed, -s, and -es from words in a text according to a set of guidelines or heuristics. Even though the final stemmed words may not always be actual words, they can still convey the original words' fundamental meaning.

Stop word elimination: Stop words are words that fail to add any additional meaning to sentences and can be removed with no impact on way content goes through processing for the reason it was included. They are removed from our lexicon in order to lessen noise decreasing size of the set of features.

4.2 Coding

4.2.1 SET UP: IMPORTING LIBRARIESAND LOADING DATASET

Fig 4.1

Fig 4.2

Fig-4.1and 4.2 Importing Libraries and Loading dataset

4.2.2 EXPLORATORY DATA ANALYSIS

Fig 4.3 Displaying first 10 rows of both Datasets

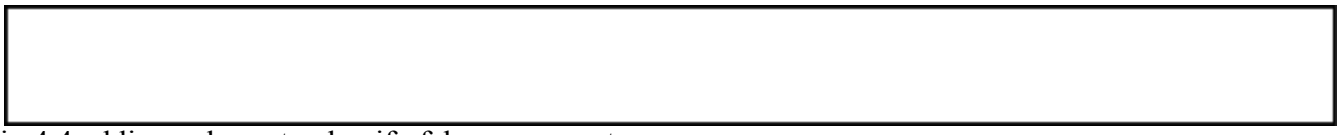

Fig 4.4 adding column to classify fake news ans true news

Fig 4.5 Diplaying shape of dataset

Fig 4.6 creating two smaller datasets for manual testing and to remove specific rows from the original datasets for evaluation and testing purposes.

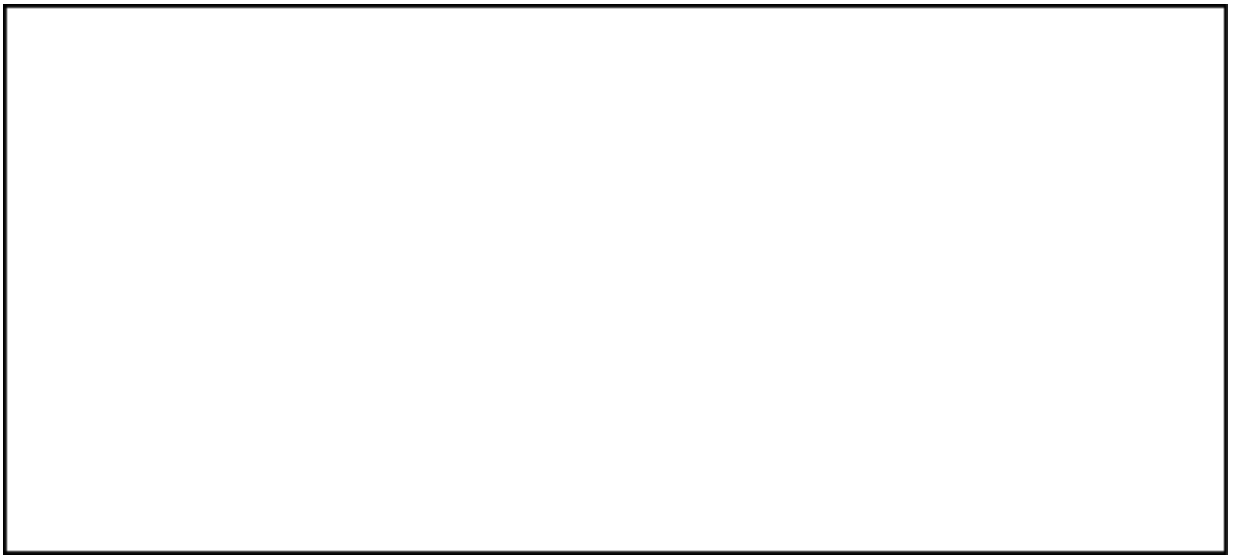

Fig 4.7 Merging the two datasets into a single dataset and displaying the data

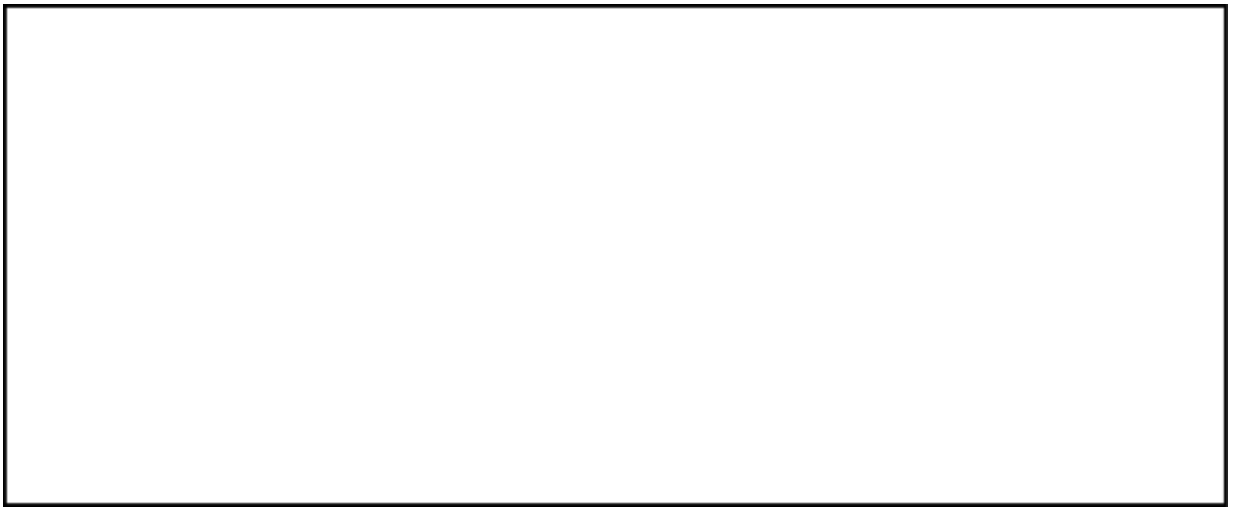

Fig 4.8 Removing the 3 columns ‘Title’, ’Subject’, ’date’ and displaying Data

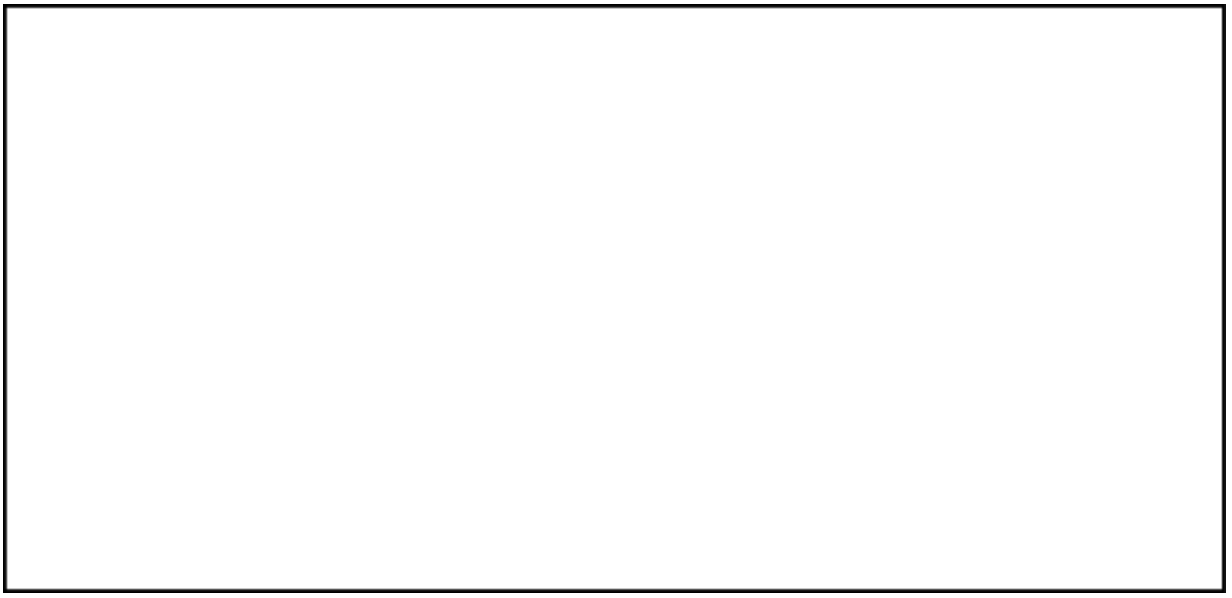

Fig 4.9 shuffling the dataset

Fig 4.10 checking for Null Values

Fig 4.11 Removing Unnecessary characters

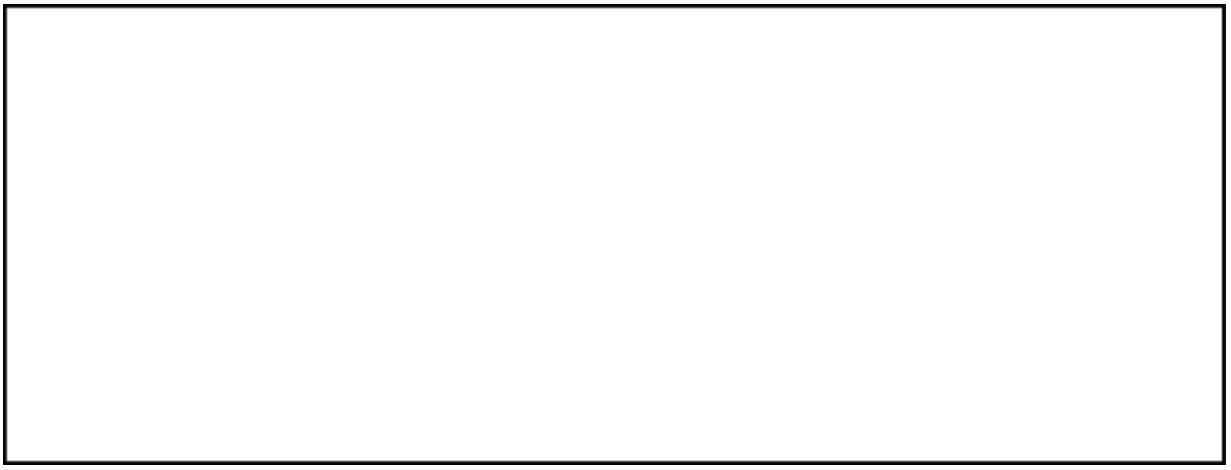

Fig 4.12 Displaying data after removing unnecessary characters and stopwords

Fig 4.13 defining Variables

Fig 4.14 Splitting dataset into train and test

4.2.3 EXTRACTION OF VECTORS FROM DATASET USING TF-IDF VECTORIZATION

Fig 4.15 Doing TF-IDF Vectorization

4.3 ANALYSISING RESULTS AND ACCURACY CHECK

Fig 4.16 Applying Logistic Regression , defining the Object and Fitting the dataset

Fig 4.17 Generating report of classification and confusion matrix for Logistic Regression

Fig 4.18 Applying Decision Tree classification , and finding accuracy using Decision tree Classsification

Fig 4.19 generating confusion matrix and decision tree classification report

Fig 4.20 Applying Gradient Boostong classification and finding accuracy.

Fig 4.21 Generating confusion matrix and gradient boosting classification report

Fig 4.22 Applying and Checking accuracy using Random Forest classification

Fig 4.23 Confusion matrix and accuracy report of Random forest classifier

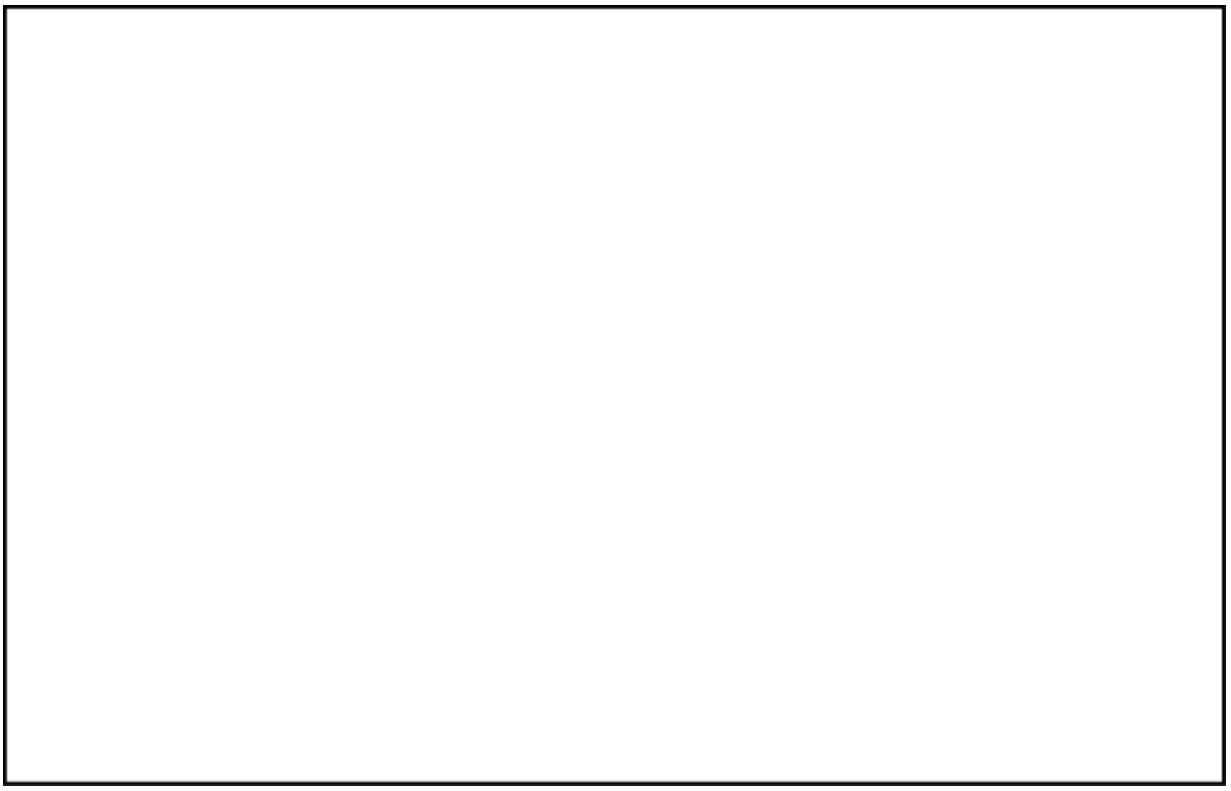

Fig 4.24 Graph for subject distribution ‘Data’

```
In [51]: plt.figure(figsize= (12,9))
         sns.countplot(x='subject',data = df_fake)

Out[51]: <AxesSubplot:xlabel='subject', ylabel='count'>
```

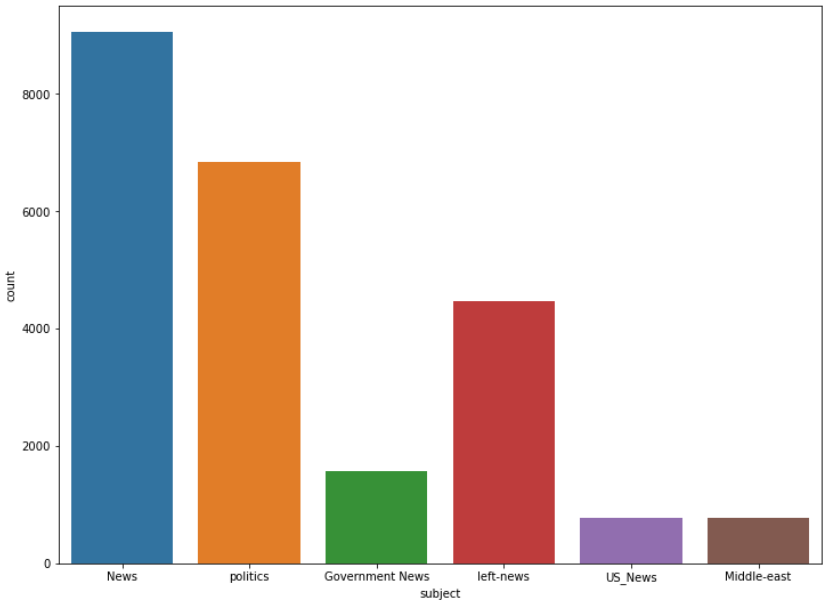

Fig 4.25 distribution in graph for distribution of subject in ‘fake’

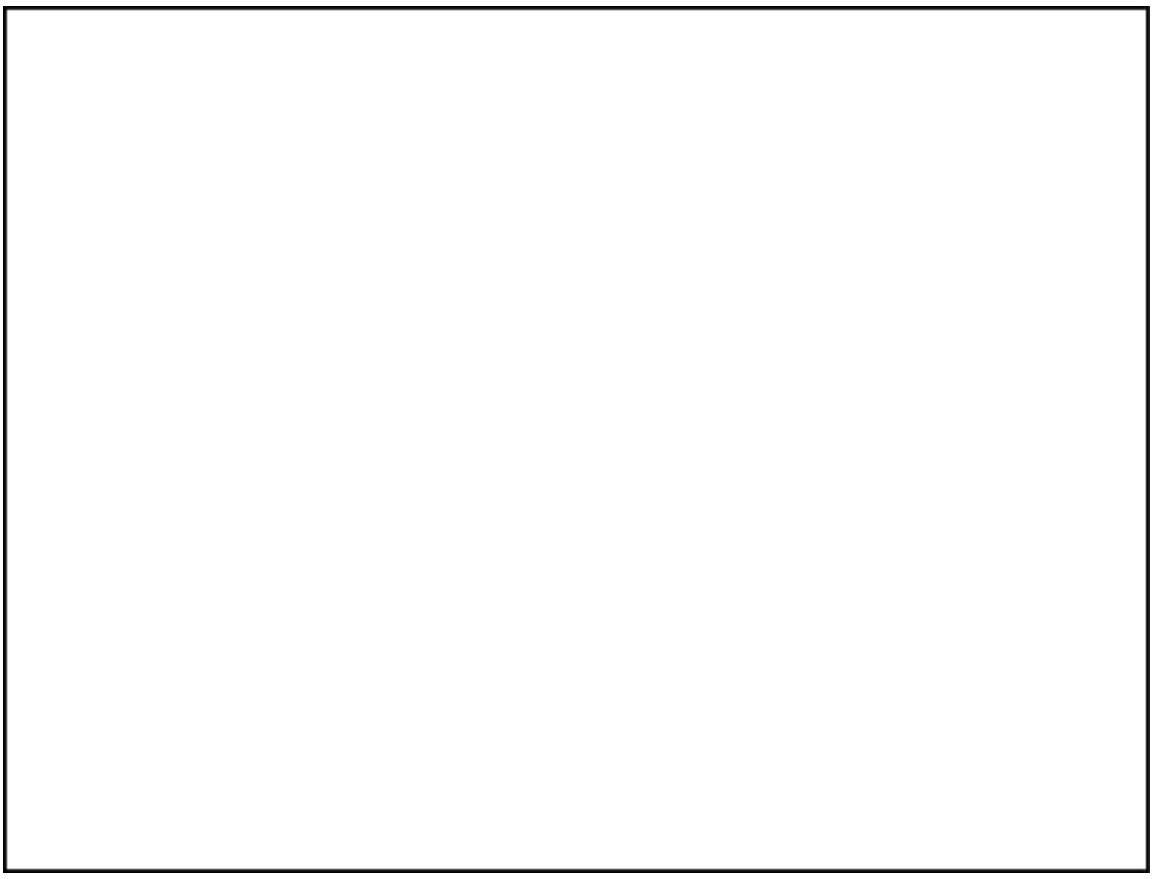

Fig 4.26 Graphic display for Subject Distribution in ‘Real News’

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

Decision tree classifier was the model that we discovered to be the most capable of being generalised returning an accuracy of 99.65% and F1 scores of 0.99 for both false and true results.

The Confusion Matrix displays the number of accurate and incorrect classifications that the Classification Report metrics would yield when applied to a dataset, while the Classification Report displays the high accuracy and f1 scores used as performance metrics.

5.2 Future Scope

5.2.1 Obtaining larger set of labelled news dataset to improve model

Although there was a balanced distribution of classes among the 35,000 news stories in this dataset, only two sources—Reuters and the Guardian—were used to create the real news stories, and it is unknown where the fake news stories came from.
The use of social media data will allow for considerations of language variations. We aim to delve deeper and analyse the impact of fake news on readers and develop faster prediction methods. The research can also incorporate successful models from other fields and re-assess the feature engineering and pre-processing techniques used.

5.2.2 Identification of Source of News Stories

The source URLs for the tales that are already in our dataset should, to the maximum extent feasible, be identified. More news data set stories should be collected along with proper information about their source which will eventually help in proper analysis. It is likely to be helpful information that can be added into the classification process to know where the news report originated.

References

[1] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), 22-36.

[2] Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.

[3] Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 53(5), 1-40.

[4] Shu, K., Wang, S., & Liu, H. (2019, January). Beyond news contents: The role of social context for fake news detection. In Proceedings of the twelfth ACM international conference on web search and data mining (pp. 312-320).

[5] Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. IEEE Intelligent Systems, 34(2), 76-81.

[6] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. arXiv preprint arXiv:1708.07104.

INDIVIDUAL CONTRIBUTION REPORT:

NEWS CLASSIFIER USING NLP

AVINEET JENA
2029011

Abstract: fake news can be very harmful and can affect people's daily life and have adverse effects on it. Social media has become a daily part of peoples lives and is a major source for spreading of rumors and news which can be fake. We have built a project which tries to classify news as real/genuine or fake using machine learning algorithms and natural language processing

Individual contribution and findings: I have contributed for the following part:- Machine learning algorithms and supervised learning algorithms including classification methods. I have used NLP and it's libraries for this purpose such as NLTK and NUMP, Scikit learn. I have also performed Text preprocessing which included lemmatization and removal of unnecessary words and performed data cleaning. along with that I contributed in the ideas and the framework and formation of this project "NEWS CLASSIFIER USING NLP".

Individual contribution to project report preparation: I made a contribution to the report about the project by reading a lot of research papers and online resources to find the ideas and information needed to construct the project. I focused on the introduction, abstract, and literature review and in this project report's implementation and analysis sections**.**

Individual contribution for project presentation and demonstration: I made the libraries slide's , explained the architecture and workflow of this project

Full Signature of Supervisor:

…….

Full signature of the student:

AVINEET JENA
2029011

INDIVIDUAL CONTRIBUTION REPORT:

NEWS CLASSIFIER USING NLP

PRIYAVRAT TRIPATHI
2028070

Abstract: fake news can be very harmful and can affect people's daily life and have adverse effects on it. Social media has become a daily part of peoples lives and is a major source for spreading of rumors and news which can be fake. We have built a project which tries to classify news as real/genuine or fake using machine learning algorithms and natural language processing

Individual contribution and findings: My contribution was to study NLP and it's a implementation. I also studied Machine learning algorithms and used the classification methods such as logistic regression, decision tree classification, random forest classifier, gradient boost classifier in our project to find the most optimum and accurate results. Along with that I contributed in the ideas and the framework and formation of this project "NEWS CLASSIFIER USING NLP".

Individual contribution to project report preparation: I made a contribution to the report about the project by reading a lot of research papers and online resources to find the ideas and information needed to construct the project. I focused on the introduction, abstract, and literature review and in this project report's implementation and analysis sections**.**

Individual contribution for project presentation and demonstration: I explained the Machine learning algorithms and classification implied in this project and showed the accuracy results and explained the future scope of this project.

Full Signature of Supervisor: Full signature of the student:

…….

SANDESH TRIPATHI
2029213

INDIVIDUAL CONTRIBUTION REPORT:

NEWS CLASSIFIER USING NLP

HEEMANSHU ACHARYA
2028124

Abstract: fake news can be very harmful and can affect people's daily life and have adverse effects on it. Social media has become a daily part of peoples lives and is a major source for spreading of rumors and news which can be fake. We have built a project which tries to classify news as real/genuine or fake using machine learning algorithms and natural language processing

Individual contribution and findings: I contributed in the project by looking into and assessing numerous vectorization techniques as part of After doing a thorough analysis of several techniques, TF-IDF vectorization seemed to be the most appropriate choice for our project. Along with that I contributed in the ideas and the framework and formation of this project "NEWS CLASSIFIER USING NLP".

Individual contribution to project report preparation: I made a contribution to the report about the project by reading a lot of research papers and online resources to find the ideas and information needed to construct the project. I focused on the introduction, abstract, and literature review and in this project report's implementation and analysis sections**.**

Individual contribution for project presentation and demonstration: I explained the working and the basic architecture of this project the and explained the importance off the source of dataset.

Full Signature of Supervisor: Full signature of the student:

……

SUJAL KUMAR