ENHANCING VERACITY: LEVERAGING MACHINE LEARNING ENSEMBLE METHODS FOR FAKE NEWS DETECTION

A MAJOR PROJECT-II REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF THE DEGREE
OF
MASTER OF TECHNOLOGY
IN
INFORMATION SYSTEMS

Submitted by:
ANUP KUMAR SRIVASTAV
2K22/ISY/04

Under the supervision of
PROF. DINESH KUMAR VISHWAKARMA AND DR.VIRENDER RANGA

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

May, 2024

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Anup Kumar Srivastav, 2K22/ISY/04 student of M.Tech in Information Systems, hereby declare that the Major Project-II dissertation titled **"ENHANCING VERACITY:LEVERAGING MACHINE LEARNING ENSEMBLE METHODS FOR FAKE NEWS DETECTION"** which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Anup

Place: Delhi
Date: 08/07/2024

ANUP KUMAR SRIVASTAV
2K22/ISY/04

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Major Project-II dissertation titled "**ENHANCING VERACITY: LEVERAGING MACHINE LEARNING ENSEMBLE METHODS FOR FAKE NEWS DETECTION**" which is submitted by Anup Kumar Srivastav, 2K22/ISY/04, Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted inpart or full for any degree or diploma to this University or elsewhere.

08/7/24	
Signature	Signature
Dr.Virender Ranga	Prof. Dinesh Kumar Vishwakarma
SUPERVISOR	**CO-SUPERVISOR**
Associate Professor	Head of Department
Department Of Information Technology	Department Of Information Technology
Delhi Technological University	Delhi Technological University
Place: New Delhi	
Date: 08/07/2024	

ABSTRACT

The emergence of the World Wide Web and the use of social media platforms such as Facebook, Twitter, and Instagram have led to the development of a technique of disseminating information that was not possible before the digital age. Many of the information available on these social media platforms could be false. As such, it is imperative to keep an eye on this data. It is possible to employ a method known as machine learning-based fake news identification to assess the authenticity of fresh articles or facts by feeding them into the model. Before training the dataset, we will preprocess the data (text in this case). The majority of preprocessing involves removing unnecessary data. After that, the dataset is split into two parts: training and testing. Next, the TF-IDF vectorization approach will be used to vectorize the data. The vectorized data is then used to train the different classifiers (like random forest, svm, xgboost, etc.). These findings are then integrated into ensemble models to improve the precision of state-of-the-art false news detection. The timeliness of a dataset affects the model's accuracy since it prevents the model from accurately predicting the authenticity of more recent information because it excludes information that is too old from previous datasets. The model can be promptly tested using the testing dataset after training, at which point it can be put to use.

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to express my sincere gratitude towards my supervisors Prof. Dinesh Kumar Vishwakarma and Dr. Virender Ranga for providing invaluable guidance, comments, and suggestions throughout the course of the project. The results of this thesis would not have been possible without support from all who directly or indirectly, have lent their hand throughout the course of the project. I would like to thank my parents and faculties of the Department of Information Technology, Delhi Technological University, for their kind cooperation and encouragement which helped me complete this thesis. I hope that this project will serve its purpose to the fullest extent possible.

ANUP KUMAR SRIVASTAV
2K22/ISY/04

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LR	Logistic Regression
DT	Decision Tree
RF	Random Forest
NBC	Naive Bayes Classifier
AUC	Area Under Curve
KNN	K- Nearest Neighbors
XGB	XG Boost
CNN	Convolutional Neural Network
BERT	Bidirectional Encoder Representations from Transfers
LSTM	Long short-term memory
GBC	Gradient Boosting Classifier
VGG	Visual Geometry Group
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

Technological progress has brought knowledge to the fingertips of people worldwide. The internet serves as a vast repository of information, yet its dependability hinges on various factors. Each day, a plethora of information inundates print and online platforms, posing a challenge in discerning its authenticity. A meticulous scrutiny and analysis of narratives become imperative, involving the verification of information accuracy through an evaluation of supporting sources, tracing the material back to its origin, and assessing the credibility of authors. “Fake news” can be defined as the manufactured information that deceives people. Such fake messages, news stories, and altered images are frequently seen on social media during Indian election campaigns [11].

In recent years, substantial research has yielded favorable results in this domain. The advancement and triumph of artificial intelligence and machine learning have liberated humans from unnecessary toil. These technologies have the potential to safeguard society against unnecessary turmoil and social unrest by identifying and preventing the spread of fake news.

The goal of this project is to create a classifier that can determine whether a user's claim is true or false.” The "Fake News Detection System" project makes use of natural language processing methods and machine learning algorithms. In the discipline of computer science, machine learning is a subset of artificial intelligence that frequently makes use of statistical methods to give computers the capacity to learn from data without being explicitly programmed [2].

A branch of computer science and intelligence called natural language processing is focused on how computers interact with human (natural) languages, particularly how to programmer computers to process and analyze massive volumes of natural language data[5]. In the current era, misinformation spreads rapidly through social media. Winston Churchill's renowned quote, "A lie gets halfway around the world before the truth has a chance to get its trousers on," underscores the swiftness of falsehood dissemination. The extensive user base on social media accelerates the spread of rumors and inaccurate information. How people react to such news often becomes the determining factor in labeling it as either "fake" or "real." To support or refute the claim, the user offers evidence in the form of video or online links. Adopting a classification system based on this approach would mark significant progress. I conducted an experiment to gauge the frequency of words associated with "fake" in responses, aiming to provide evidence supporting this assertion.

1.1 Classification of Fake News

News can be classified as either true or fake, but fake news is further divided into further categories, which are as follows:

• Propaganda: These articles may be false or misleading, but their main purpose is to further the author's cause. These kinds of articles are nearly often published for political reasons, usually to forward the agenda of the party the author supports or belongs to.[6]

• Clickbait: These are articles that have the potential to be entirely fake or overly dramatic. The goal of these tales is to boost ad revenue and generate revenue.

• Opinion/Commentary: These are some well-known pieces in which the writer essentially seeks to sway the reader's perception of current affairs.

• Satire/humor: These tales may cause readers to rethink certain ideas because they are satirical or humorous, but they also contain some exaggerated material

Based on the purpose behind its spread, fake news can be divided into two categories:

- Misinformation: This happens when someone spreads news that they genuinely think is true even though it is untrue.
- Disinformation is when someone spreads misleading information on purpose in an attempt to deceive others, knowing full well that it is untrue.

1.2 Fake News Detection

Fake news detection is the process of identifying and classifying false or misleading content in order to stop it from spreading. False news propagation is a serious topic that has been made worse by the growth of social media and internet platforms. Encouraging a society that is informed and ensuring the reliability of information sources depend on the detection and suppression of fake news.[9]

Purpose of Detecting Fake News:

Identifying genuine news articles, images, and videos from bogus ones is the main goal of fake news detection. To ascertain authenticity, this technique entails examining a number of content, context, and source factors. The following are some thorough strategies and techniques for spotting false news:

Natural Language Processing (NLP) :

The linguistic elements of news articles are analyzed using Natural Language Processing (NLP) techniques. This calls for a number of advanced techniques:

Sentiment analysis:It looks at the text's overall tone and emotional content to identify any possible cases.

Named Entity Recognition: Resolves conflicts by accurately and pertinently classifying proper names, locations, dates, and other entities mentioned in news articles.

- **Topic Modeling:** Discerns the underlying themes and topics within the text, helping to identify inconsistencies or unusual patterns that might suggest falsehood.

Verification of Source :

Finding misleading information requires confirming the credibility and integrity of the news source. This includes:

Credibility assessment: Assessing the source's past publications, journalistic standards compliance, and overall history and dependability. Assessing the authority of a source by taking into account their qualifications and level of experience related to the topic at hand.

Verifying facts Methods: determining whether the source consistently conducts in-depth fact-checking prior to news publication.

Verifying facts : Cross-referencing the claims and statements stated in the news article with credible, dependable sources is the process of fact-checking. This procedure can be either automated or manual:

Manual fact-checking: Done by groups and individuals who carefully examine the information and compare it to known facts.

Automated Fact-Checking: This method swiftly compares news material with verified information by using databases and algorithms.

Social Media Evaluation:

On social media, false information frequently spreads quickly. Social media activity analysis can be used to find potentially false information:

User Interaction Analysis: Looks for abnormalities in the way people interact with material by looking at likes, shares, and comments.

Social Network Dynamics: Examines how information spreads through a network, searching for odd propagation that could point to fake news.

Source Reputation: Assesses the authority of the source in the social media community by taking into account variables such as the number of followers, engagement metrics, and historical accuracy.

These methods are used in fake news detection in an effort to weed out false information and misinformation while preserving the accuracy of the public's access to information. In the digital age, maintaining the quality of information requires a multifaceted strategy.

Machine Learning Algorithms:

Techniques for machine learning (ML) are widely used to detect misleading information. These methods can classify information and articles as true or fake according to their characteristics. Here are some more machine learning techniques for spotting bogus news:

Naive Bayes Classifier: This probabilistic classifier makes strong (naive) independent assumptions about the features in order to apply the Bayes theorem. For text categorization tasks, such as spam and fake news identification, it is quick and accurate.
Text Classification**:** Because of its ease of use and strong performance on big datasets, Naive Bayes is a particularly useful technique for text-based false news identification.

Random Forest: A technique for ensemble learning that builds several decision trees during training and produces the class mode for classification tasks.
Feature Importance: By revealing which elements of the content are most suggestive of falsity, Random Forests offer insights into feature importance.

Gradient Boosting Machines (GBM): Using this technique, models are constructed one after the other, with each model fixing the flaws of the one before it. High performance in a variety of classification problems is a well-known attribute of GBM, including well-known implementations like XGBoost.

Boosting Trees: GBMs are especially good at managing intricate patterns in false news data since they concentrate on incorrectly categorized instances from earlier iterations.

K-Nearest Neighbours (K-NN):This algorithm is used to categorise instances in the feature space according to how close they are to one another. For smaller datasets or as a baseline comparison, it's easy to use and efficient.
Similarity Measures: k-NN is excellent for identifying patterns that resemble known fake or true news occurrences since it uses distance metrics to classify new examples.

Neural Networks: A variety of text and picture classification tasks, including the detection of fake news, are performed using these models, particularly feedforward neural networks. Multi-layer perceptrons, or MLPs, offer a reliable solution for classification issues since they can identify non-linear relationships in the data.

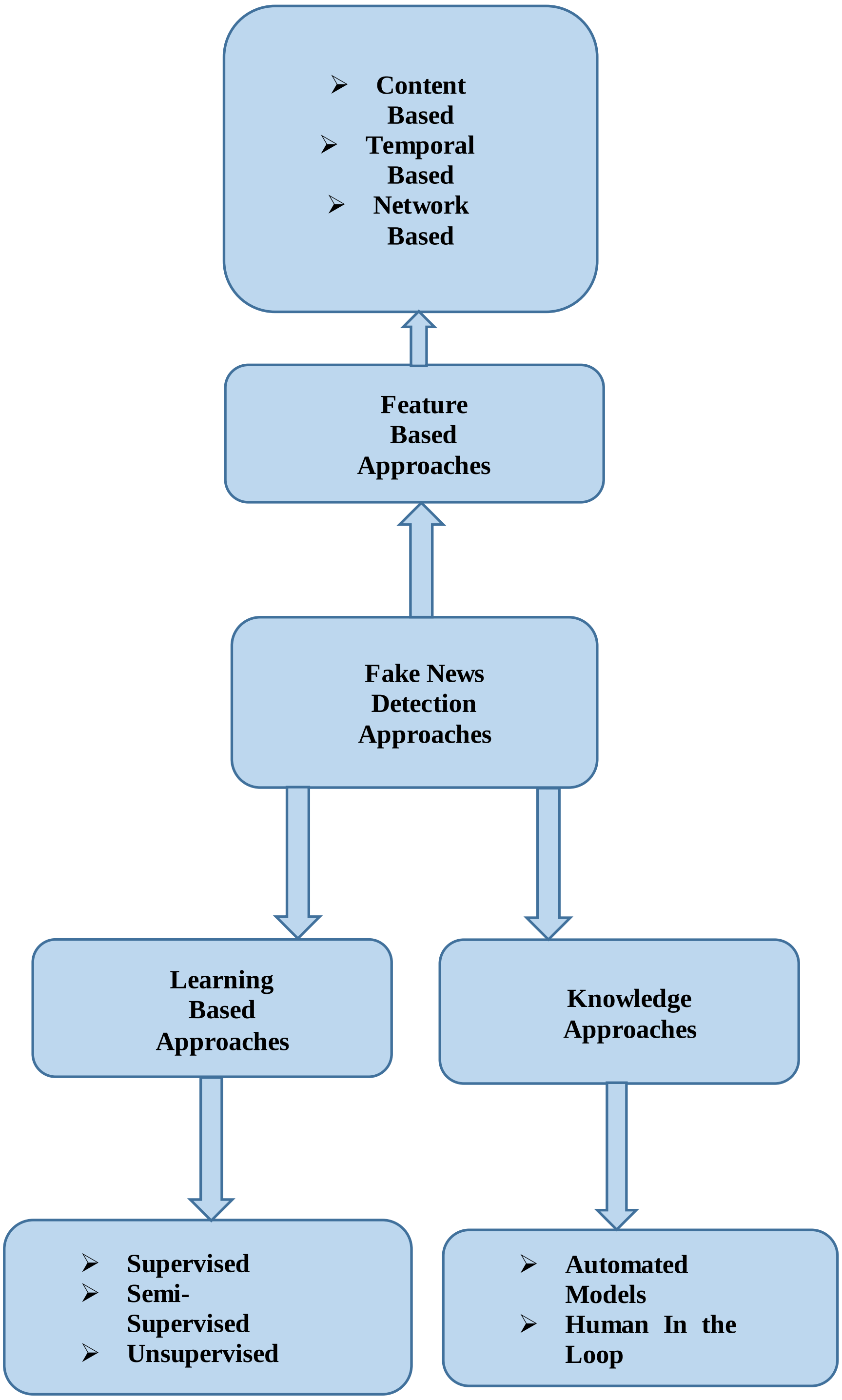

Figure 1.1 Various Approaches for Fake News Detection

Further tactics such as fact-checking, crowdsourcing, and textual modality hold considerable potential for enhancing the accuracy and reliability of false information detection systems. Textual modality analysis uses state-of-the-art natural language processing techniques to detect deceptive content by detecting linguistic cues and inconsistent text. Crowdsourcing makes textual and visual material verification scalable and fast by harnessing the collective intelligence of volunteers. To improve detection, ensemble models use the opinions and assessments from the population. Ensemble models can also benefit from the knowledge and data that fact-checking procedures—which entail extensive investigation and cross-referencing with credible sources provide.[7]

CHAPTER 2

LITERATURE REVIEW

After reviewing different research papers, we observed different types of approaches and we can classify them into the following categories based on the characteristics that they have used for detection:

Sakib Hakak et al.[3]emphasizes that use of digital media has greatly aided in the dissemination of false information, making it more difficult to identify because of problems with feature selection, parameter tuning, and dataset imbalance. Accuracy is still subpar despite a large number of experiments using supervised and unsupervised learning techniques. To overcome these difficulties, an ensemble classification model that combines Random Forest, Decision Tree, and Extra Tree Classifier has been presented. It achieves 100% training accuracy on the ISOT dataset and 99.8% training accuracy on the Liar dataset. For the Liar dataset, however, the testing accuracy was significantly lower at 44.15%, showing overfitting and inadequate generalisation to previously unknown data. This discrepancy highlights the necessity of improving feature selection, parameter tuning, and imbalanced dataset handling in order to increase model efficacy and dependability in identifying fake news.

Nitish Kumar et al. [4] have observed that the proliferation of social media platforms has contributed greatly to the emergence of fake news by offering widely accessible means of disseminating information, whether it be real or false. For specialists in machine learning (ML) and natural language processing (NLP), the spread of fake news is a significant challenge. Nowadays, spotting bogus news requires careful consideration of the facts. This study evaluates different approaches used in the field of false news detection by doing a thorough analysis of recent literature. The survey provides information about the detection process, emphasizing the use of ML, DL, and NLP algorithms to detect false information. The results highlight the difficulty in identifying fake news and the need for cutting-edge techniques to increase precision and dependability.

Dinesh Kumar Vishwakarma et al.[8] have highlighted that due to its quick and affordable dissemination, social media is becoming a more important source of news, which has a negative influence on both society and people. As a result, real-time methods for the detection of fake news have emerged as a major area of research attention. Conventional supervised techniques require large amounts of labelled datasets and are resource-intensive. In this work, a text-based framework for detecting fake news with restricted labels is presented, utilising Graph Convolutional Networks (GCN) in a semi-supervised learning method. The framework combines three essential elements: using GloVe to extract word embeddings, Word Mover's Distance (WMD) to build a similarity graph, and GCN for binary classification. This approach outperformed previous recent methods and was evaluated on three datasets, with the Real or Fake dataset yielding the greatest accuracy of 95.27%.

Arush Agarwal et al. [10] presents a strategy to identify fake news by assessing the veracity of reports and estimating their authenticity using textual data feature extraction and credibility score. An ensemble network using different machine learning techniques (SVM, CNN, LSTM, KNN, and Naive Bayes) was created to analyse news stories, authors, and titles at the same time. According to the model, LSTM had the best accuracy, coming in at 97%. These classifiers' performance was assessed using measures including precision, recall, and F1-score, which showed how well the various algorithms performed on the dataset and highlighted the greater accuracy of LSTM in detecting bogus news.

Rohit Kumar Kaliyar et al.[12] explore the use of content and context-level characteristics in a tree-based ensemble machine learning framework—specifically, Gradient Boosting with optimized parameters—to identify false news. Recently developed as gradient descent algorithms, adaptive boosting techniques for classification maximize a single objective function by adjusting critical components and parameters. Experiments with different machine learning models on a multi-class dataset (FNC) showed the efficacy of the ensemble architecture. The Gradient Boosting approach outperformed previous benchmark results and demonstrated the promise of ensemble methods in addressing the complexity of false news detection beyond binary classification, achieving an accuracy of 86% for multi-class classification of fake news with four classes.

Harita Reddy et al.[13] highlight the need for effective detection mechanisms on social media platforms. Given the high volume of news generated, distinguishing genuine news from hoaxes is challenging without considering the news source. This study [13] focuses on detecting fake news by analyzing only the textual features, excluding other metadata. The researchers found that combining stylometric features with text-based word vector representations using ensemble methods can effectively predict fake news. Their approach achieved an accuracy of up to 95.49%, demonstrating the potential of text-based analysis in identifying fake news without relying on additional metadata.

A thorough summary of the most recent developments in fake news identification can be found in the paper "Recent State-of-the-art of Fake News Detection: A Review" by the author Dinesh Kumar Vishwakarma et al.[14] It looks at many approaches used to detect fake news, including both established and novel methods. The paper compares the efficacy and drawbacks of various detection techniques, highlighting how they have evolved. It talks about methods that make use of picture verification, hybrid models that combine several data kinds, and textual analysis. The study emphasizes how crucial sophisticated machine learning techniques are for improving detection accuracy, including ensemble approaches and deep learning. The review provides useful insights into the advantages and disadvantages of cutting-edge fake news detection techniques by summarizing current trends and technology, highlighting the necessity for ongoing innovation.

Akshay Jain et al. [15] uses the Naive Bayes classifier to attempt to categorise bogus news. The dataset utilised in this method was gathered from github and comprises about 11,000 news stories organised into rows with four columns: text, label (fake or true), title, and index. Word embeddings are produced using the vectorization techniques bag of words and n-grams. The author completed the classification by initially simply considering the title column, and then further considering the text column. After comparing the AUC scores for the two columns using the two classifiers, it is discovered that the AUC score is enhanced by the quantity of words in the text that is fed to the classifier. This paper also talks about the use of web scrapping to keep our datasets updated.

The author Julio C. S. Reisin et al. [16] concentrated on three distinct feature sets for classification: features taken from news articles, features taken from news sources, and features taken from news environments. Language features (obtained using POS tagging), Lexical features (number of unique words and their frequency in the text), Psycholinguistic features (obtained through Linguistic enquiry and word count (LIWC)), and Subjectivity (obtained using Textblob's API) are the features that were extracted from news content. The following features were taken from the news source: domain location (obtained using ipstack API), credibility and trustworthiness (obtained by collecting rankings of various newspapers and websites using Facebook and Alexa's APIs), and bias. Features of the ecosystem include: Engagements (likes and comments),temporal patterns(obtained by calculating the frequency of comment posting). The dataset utilised in this work includes 2282 Buzzfeed news stories about the US elections of 2016. Stories with "non-factual content" are eliminated from the dataset, and the remaining articles are all classified as real news. All articles falling into the categories of "mostly false" and "the mixture of true and false" are combined into a single class that is called the fake news class. Five classifiers are used in the classification process: Random Forest (RF), k-nearest neighbours (KNN), Naive Bayes (NB), Support Vector Machine With RBF kernel (SVM), and XGBoost (XGB). Additionally, AUC is used to calculate performance.

Yang Yang et al. [17] uses convolution neural networks and attempted to categorise the news while taking into account both the text and picture components of the news story. This method gathers the 20,015 news articles from roughly 240 websites that have been scraped from Kaggle. There are about 8,000 accurate news stories and 12,000 fake ones in it. For the textual data this paper considers some linguistic features like number of words and sentences in a news article (generally less in case of fake news), punctuation marks which tells us how confident the writer is while writing that article (usually found to be more in case of fake news), cognitive perspective which includes the use of negative words in the article (used less by the fake news creators to avoid contradictions), lexical diversity (more diverse the use of words, more likely it is a real news) and sentiments analysis (usually negative in case of fake news due to the mindset of the creator). Additionally, an image analysis was conducted, and it was discovered that genuine news images had a greater number of faces than fake news images. Furthermore, more unrelated pictures of scenery and animals that have nothing to do with the content are included in the false articles. Two CNNs are used in parallel for classification; one for textual input and the other for picture analysis. Textual explicit features and textual latent features are the two features used by the text branch. The linguistic features that were previously mentioned make up the textual explicit features. CNN creates the textual latent features by word embeddings, which may then be concatenated to produce feature vectors. The image branch is also utilizing two features which include visual latent and visual explicit features. The visual explicit feature is used to extract the resolution and the number of faces in the image and the visual latent features are used to learn from raw images and derive some more powerful features.

Sonia Castelo et al.[18] classifies news articles by taking into account both their linguistic and web mark characteristics. The methodology is predicated on a baseline study named FNDetector, which also takes into account these important news article components. The morphological features are obtained by part-of-speech tagging, which groups words according to their context. The psychological features are obtained by Linguistic Enquiry and Word Count (LIWC). The readability features are obtained by Textstat, an integrated Python library that provides an ease score for an article's readability. Beautiful Soup and The Newspaper, two Python libraries, are used to extract the web markup features. This method makes use of the Celebrity, US-Elections2016, and Political News datasets. Three classifiers are used for the classification process: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF). Each feature combination is utilized independently to see how the model's accuracy varies. Additionally, the news article's headline and text are provided to the model both independently and together so that any differences in accuracy can be observed. This paper greatly improves upon its basis paper, FNDetector, in terms of accuracy.

Sajjad Ahmed et al. [19] suggested integrated strategy that consists of three primary steps: knowledge engineering-based fact checking, classification, and user stance detection. The dataset, which comprises 17,946 news articles—12,460 biased, 572 fake, 870 conspiracy, and 2,059 non-fake—is gathered using this approach from Kaggle. For the classification, the widely utilised Support Vector Machine (SVM), a popular classifier in machine learning, is employed. The limits of various classification techniques, such as neural networks and Bayesian classifiers, were also covered by the author. One way to identify stance is to just look at user opinions on the page. They fall into two categories: implicit (which can be taken from social media) and explicit (when the user provides a direct impression). Three approaches can be used for the last phase, fact-checking: Computational Oriented (knowledge engineering is used here where several rules are given to a machine so that it can imitate the thought process of a human expert): expert-based (human expertise is required to check facts in the article), crowd-sourcing-based (reader can read the article and after understanding he/she can flag the article as real or fake).

Shivangi Singhal et al.[20] uses textual and graphic components of the news material as the main focus of this method. This method makes use of the Weibo and Twitter datasets. There are 17,000 in the Twitter dataset distinct tweets about a range of occasions. Every tweet includes both the text content and any related images. There are roughly 10,000 bogus tweets and 7,000 genuine tweets in this collection. Conversely, Weibo is a dataset compiled from reputable Chinese news sources. This dataset contains bogus news that was gathered between 2012 and 2016. In this method, Bidirectional Encoder Representations from Transformers (BERT) are used to extract textual information. BERT has twelve encoding layers in order to encode the contextual features as vectors. e model. The visual features are obtained using the pre-trained VGG-19 convolutional network on the ImageNet dataset. At last, the feature vector is reduced to 32 dimensions. An integrated vector representation of the article's text and image is then created by fusing the two feature vectors that were acquired from the two feature extractors using the concatenation technique. The text data length is fixed during the pre-processing stage of the data by padding zeros to anything below the fixed length and cutting anything above it. Every image in the data is downsized to 224x224x3 for the image components. Additionally, hyperparameter adjustment is carried out to raise the model's accuracy.

Aswini Thota et al.[21]uses the approach which predicated on a textual analysis of the information in the news story. This approach focuses on determining the news story's point of view and only illustrates how closely the article's title and body relate to one another. This method made use of the FNC-1 dataset. It contains the headline, the content of the news piece, and a description of the two parties' relationship (stance). The dataset consists of 49,973 unique combinations of news headlines and content that fit into one of four categories: discuss, disagree, agree, or unrelated.

Federico Monti et al.[22] follows several pre-processing processes which are performed in this technique to prepare the data for modelling, some of which include stemming (removing prefixes and suffixes from a word), stop word removal (removing the most common terms used in a language), and punctuation removal (removing marks like:,?!... This method uses the following vectorization techniques: word2vec, bag-of-words, Tf-idf, and GloVe. Three main types of neural networks—Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN)—are employed with these vectorization approaches. The vectors for the headline, body, and cosine similarity between both vectors—all three in concatenated form—are included in the input sent to the neural networks. Also the activation functions contains ReLU, Tanh and Softmax. When employing Tf-idf on bigrams and unigrams with cosine similarity fed to a dense neural network, the best results are obtained.

Niraj Sitaula et al.[23] focuses on the method limited to the textual components of news articles. This strategy's datasets came from Buzzfeed News and Politifact. These databases include details about the users' social networks in addition to the labels and content of the news. There are 240 news articles in the Politifact dataset; 120 of them are false, and the remaining 120 are true. There are 182 news stories in the dataset that was gathered from Buzzfeed News; 91 of them are false, and the remaining 91 are true. Source Credibility and Content Credibility are the two basic categories into which the features retrieved from the data are separated. .. The following characteristics contribute to the reliability of the source: Authors (pieces with multiple authors are typically authentic); Co-authorship (does the author have a connection to either or both real or fake news items?). Sentiment, readability, argumentation (constructed by supplying evidence and references), character, word, and phrase count, typos, and other factors are included in the content credibility features. Seven machine learning classifiers—SVM (RBF kernel), Logistic Regression, Linear SVM, Adaboost, Random Forest, Gradient Boosting Decision Tree, and Naive Bayes—are given these extracted characteristics. F1-macro, F1-weighted, and F1-micro are the performance metrics that were utilised to determine the performance. The scores for source credibility and content credibility aspects are computed independently for each of the two datasets.

Anastasia Giachanou et al.[24] uses SpotFake as a baseline paper to built the technique. For classification purposes, it takes both textual and picture data into account. The dataset was gathered from the collection of FakeNews Net. GossipCop entries, which contain news about celebrities and entertainment, are specifically used. A total of 5459 news stories with at least one photograph were gathered; of these, 2745 were false and 2714 were true. This technique consists of three parts: text and picture similarity, visual similarity, and language similarity. The textual component uses BERT (Bidirectional Encoder Representation from Transformers) to gather contextual data. There are two different systems: an encoder and a decoder. Once the encoder's input has been read, the decoder generates the task prediction. .. The 768-word vector is sent to the pre-trained BERT along with the padding text. The pre-trained VGG-16 on the visual dataset ImageNet is used for the picture content in the visual component. Moreover, VGG-16 activations are employed by LSTM to ascertain the temporal picture order. Lastly, the LSTM output is subjected to mean pooling in order to produce a single temporal component. The third component, text and image similarity, is computed by extracting the top ten picture tags from the pre-trained VGG-16 model. After that, word2vec is used to construct the word embeddings, and the embeddings are averaged to create a 300 dimension vector. Utilizing the output layer, the Following the computation of a probability representation for each feature using the Softmax function, the concatenated features are multiplied by a Soft Mask with values ranging from 0 to 1. The greatest results are obtained when contextual features are extracted using a 3-image VGG-16 with LSTM and BERT. The similarity is also estimated as previously mentioned, and all of these features are fused using the attention method.

Table 2.1: Comparison Table for Various Approaches

Year	Technique	Performance Metrics	Performance Score
2017[11]	KNN, NB, RF, SVM, XGB	AUC, F1 Score	KNN(0.80,0.75),NB (0.72,0.75),RF(0.85,0.81),SVM (0.79, 0.76),XGB (0.86, 0.81)
2018[12]	CNN for both textual and visual data	Precision, Recall,F1-score	0.9220, 0.9277, 0.9210
2017[20]	Linear SVM	Accuracy	Collected Dataset 0.74
2020[18]	LR, RF, Adaboost, NB,GBC, DT	F1-score	0.80
2020[19]	BERT(For textual Content)	F1-score	0.7955
2018[10]	NBC	AUC Score	Title (0.807 with N-grams)Text (0.912 with countvectorizer)Text(0.931 with N-grams)
2020[25]	LSTM, RNN,GRU	Accuracy	75%-LSTM45%-GRU62%-RNN
2019[26]	BiLSTM-CNN	Accuracy	86.12%
2019[27]	DT,NB	Accuracy	DT-96.65%, NB-91.52%
2019[27]	KNN,NB, RF,SVM, XGBoost	AUC	0.80, 0.72, 0.85, 0.79, 0.86

CHAPTER 3
METHODOLOGY

We will talk about the dataset that was utilised, the workflow, and other theoretical topics like the vectorizers and classifiers that were used in the coding part in the methodology section. Let's start by talking about the dataset that was used.

3.1 Dataset Used

WELFake Dataset is the name of the dataset we utilised for the coding portion. Of the 72134 news article entries, 35028 are authentic while 37106 are fraudulent. The four most well-known news datasets—Kaggle, McIntire, Reuters, and Buzzfeed Political—were combined to create this dataset. Combining these datasets is primarily done to increase the amount of training data available and avoid overfitting of the classifiers. Four columns make up this dataset: text, label, serial number, and title. Index 0 is the first number in the serial number column. The news headline is found in the title column, and the news substance is found in the text column. There are two types of labels present in the label column i.e, 0 and 1. Label 0 is for the fake news and 1 is for the real news. There are five classifiers that we have used to classify during coding part.

3.2 Classifiers Used

3.2.1 Random Forest

• A supervised learning method called random forest can be applied to the regression problem as well as classification problems.

• The random forest classifier is a kind of ensemble learning technique that combines several classifiers to enhance model performance and address challenging issues.

• In a random forest classifier, we create distinct decision trees based on different dataset subsets. Random forest makes final output predictions based on majority votes from various trees.

•Additionally, it raises the decision tree classifier's predictive accuracy.

- In python we first need to import random tree classifier from sklearn.ensemble li- brary and then we need to feed it with the vectorized data and the output label.

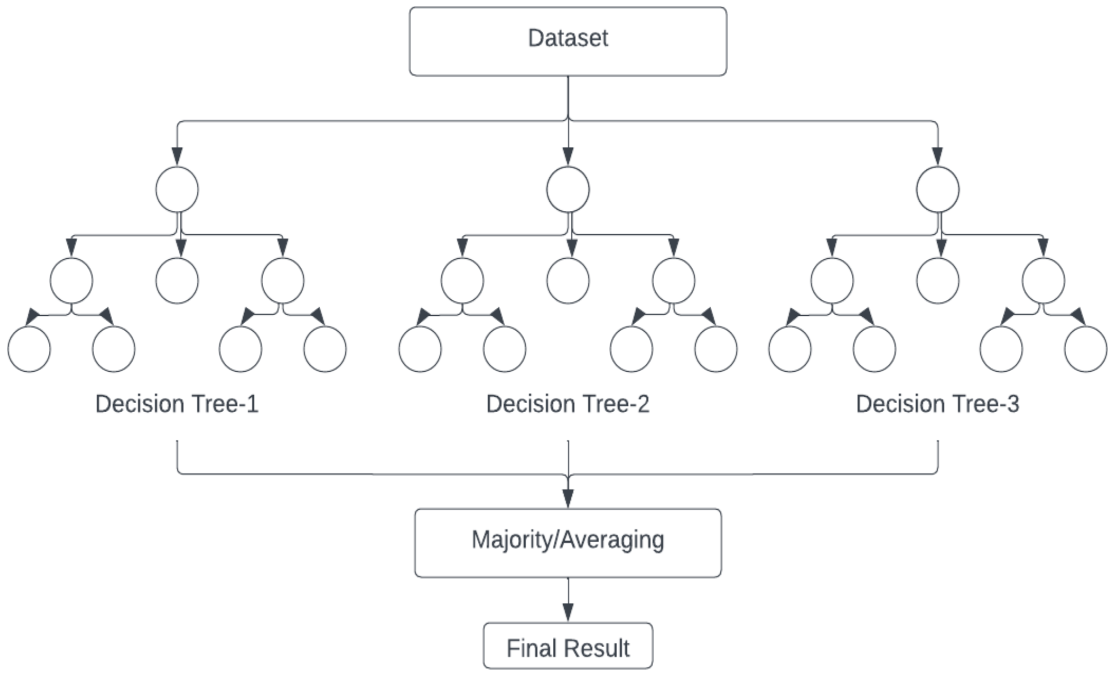

Figure 3.1 Random Forest Approach

3.2.2 Support Vector Machine (SVM)

A generalization of linear classifiers, support vector machines (SVM) are a collection of supervised learning algorithms used to address regression and classification issues. SVM were created in the 1990s, and because of their practical success, minimal number of hyperparameters, theoretical assurances, and capacity to handle massive amounts of data, they were swiftly embraced. In contrast to other learning algorithms, the Support Vector Machine (SVM) algorithm aims to identify the most comparable cases between classes in order to generate a set of support vectors. Subsequently, by determining the best margin of the hyperplane, the SVM algorithm determines the ideal hyperplane for class division.[28]

SVM can be used to predict a variable's numerical value in regression problems or to solve classification problems by determining which class a sample belongs to. The creation of a function f with an input vector (X) that matches an output (Y) is required to solve these two kinds of challenges.

$$Y = f(X)$$

SVM algorithms make use of kernel functions. The linear kernel, which is frequently suggested for text classification issues, was employed in our investigation. Compared to most other kernel functions, such as polynomial and radial functions, the linear kernel function requires fewer parameters and operates more quickly. The linear kernel function found in the formula below defines the decision boundary that the SVM returns.

$$f(X) = w^T X + b$$

where X is the data to be classified, b is the estimated linear coefficient, and w is the weight vector to minimise.The hyperplane is defined by the two parameters, w and b.

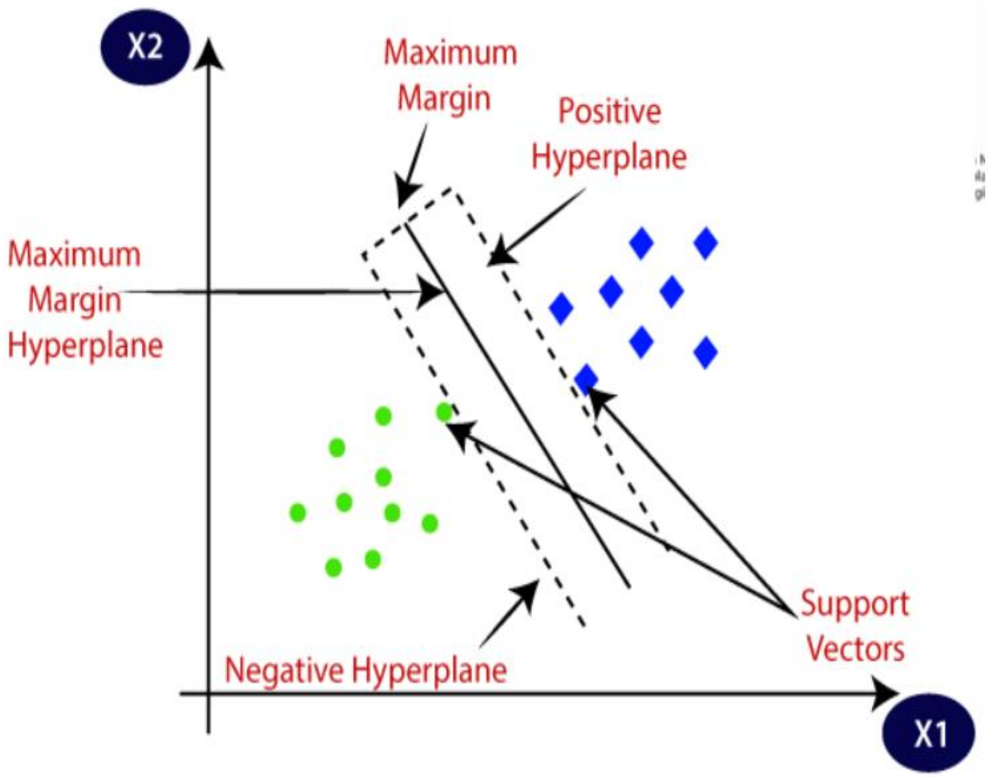

Figure 3.2 Classification of data points using SVM

3.2.3 XGBoost Classifier

- XGBoost classifier stands for Extreme Gradient Boosting classifier.
- XGBoost classifier is a decision tree based ensemble machine learning algorithm, it uses a gradient boosting framework.

- XGBoost uses parallel processing, tree pruning and handles the missing values too.
- In python we first need to import XGBoost Classifier from XGBoost library and then we need to feed it with the vectorized data and the output label.

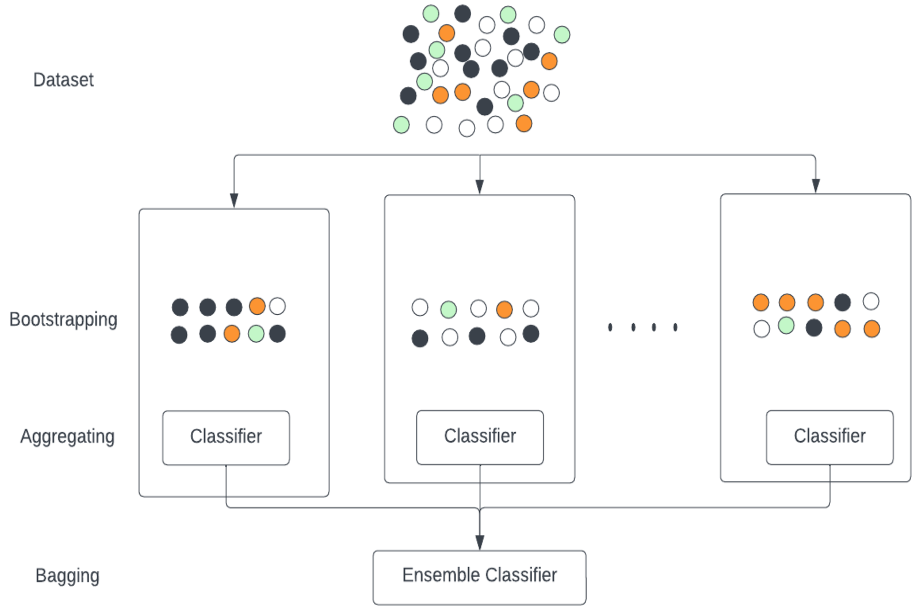

Figure 3.3 XGBoost Approach

3.2.4 Ada Boost Classifier

- Adaptive Boosting is referred to as AdaBoost.
- The outputs of several weak classifiers are combined in this ensemble learning technique to produce a strong classifier.

Important characteristics:

- **Boosting**: Focusing on more difficult cases, iteratively modifies the weights of examples that were mistakenly classified. Decision stumps, or one-level decision trees, are commonly used by weak learners.
- **Weak Learners**:Typically uses decision stumps (one-level decision trees) as weak learners.
- **Adaptivity**: The algorithm enhances overall performance by adjusting to the mistakes made by the prior classifiers.

Algorithm Mechanism:

1. **Initialization:**
   - Assign equal weights to each training instance.
2. **Iterations in Training:**
   - Utilize the weighted training data to train a weak learner.
3. **Determine Error Rate:**
   - Determine the weak learners mistake rate using the training set.

4. **Update Weights:**
   - Increase the weights of misclassified instances.
   - Decrease the weights of correctly classified instances.
5. **Calculate Classifier Weight:**
   - Determine the weight of the weak learner based on its accuracy.
6. **Form Ensemble:**
   - Combine the weak learners to form a strong classifier.
7. **Final Model:**
   - Form the final strong classifier by taking the weighted sum of predictions from the weak classifiers.

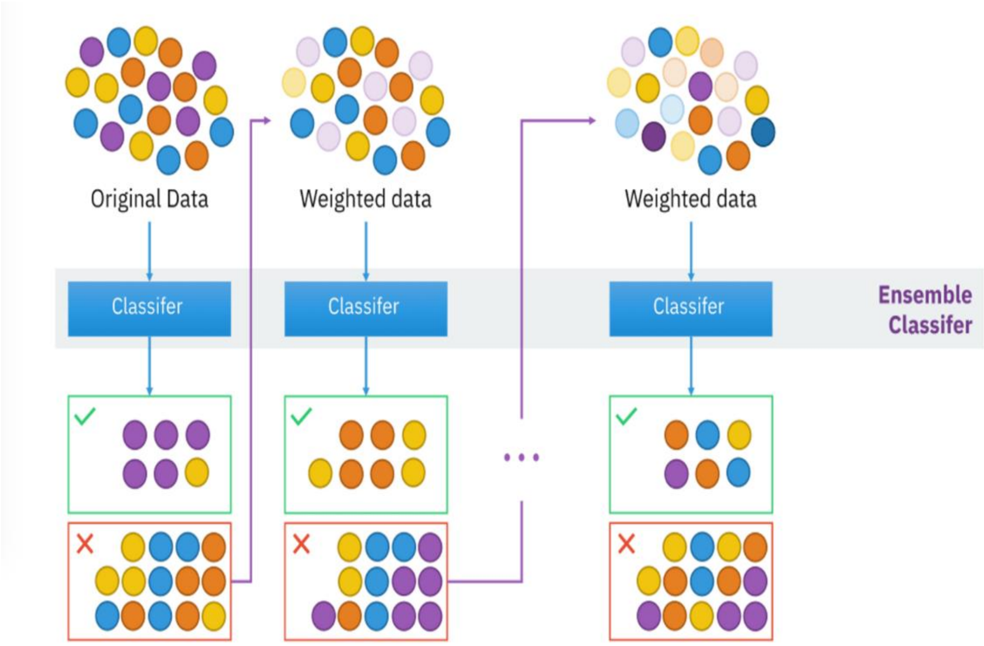

Figure 3.4 Working of AdaBoost Algorithm

3.3 Process Flow

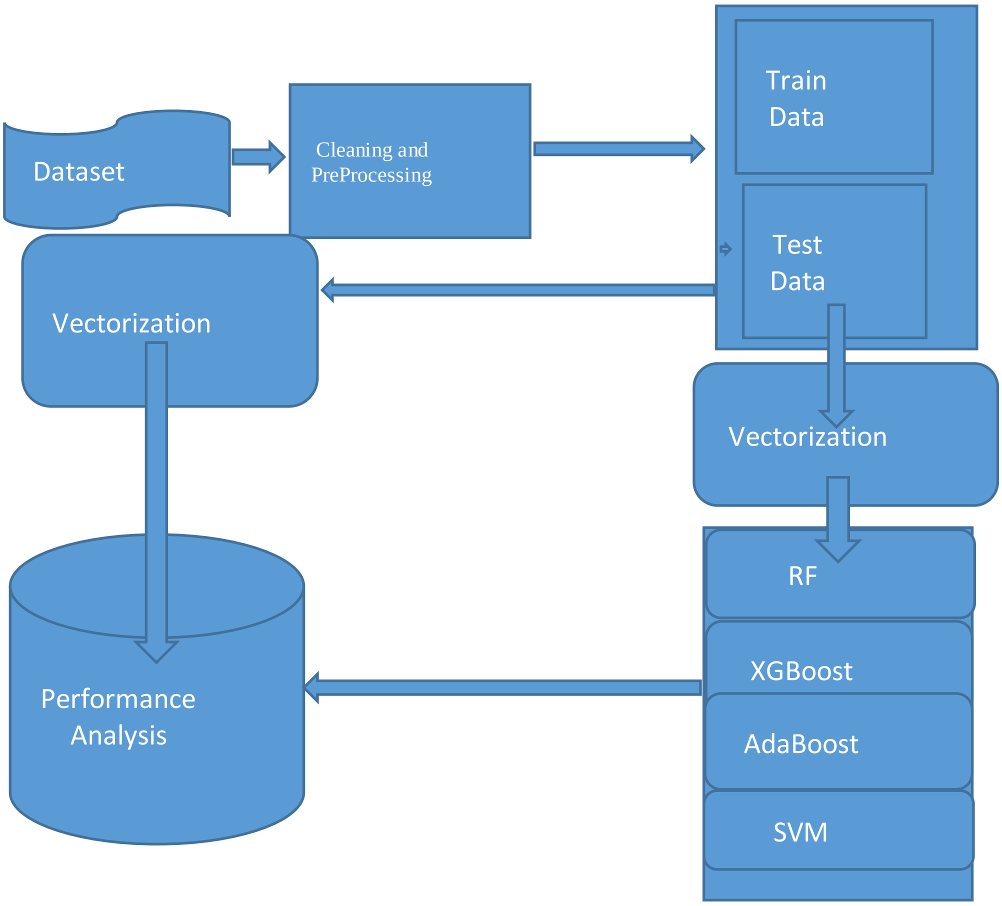

Figure 3.5 Process Flow Chart

3.3.1 Data Cleaning and Preprocessing

- First we are removing extra columns from the dataset named ‘Unnamed’. In some cases we are combining the title of the news article with the text but in other caseswe are simply removing it.
- There is also a need to shuffle the data before splitting it into training and testing set to remove any type of imbalance caused by the amount of data present for both the labels.
- We tried plotting the amount of data present for both the labels and these are thebar graphs we got:

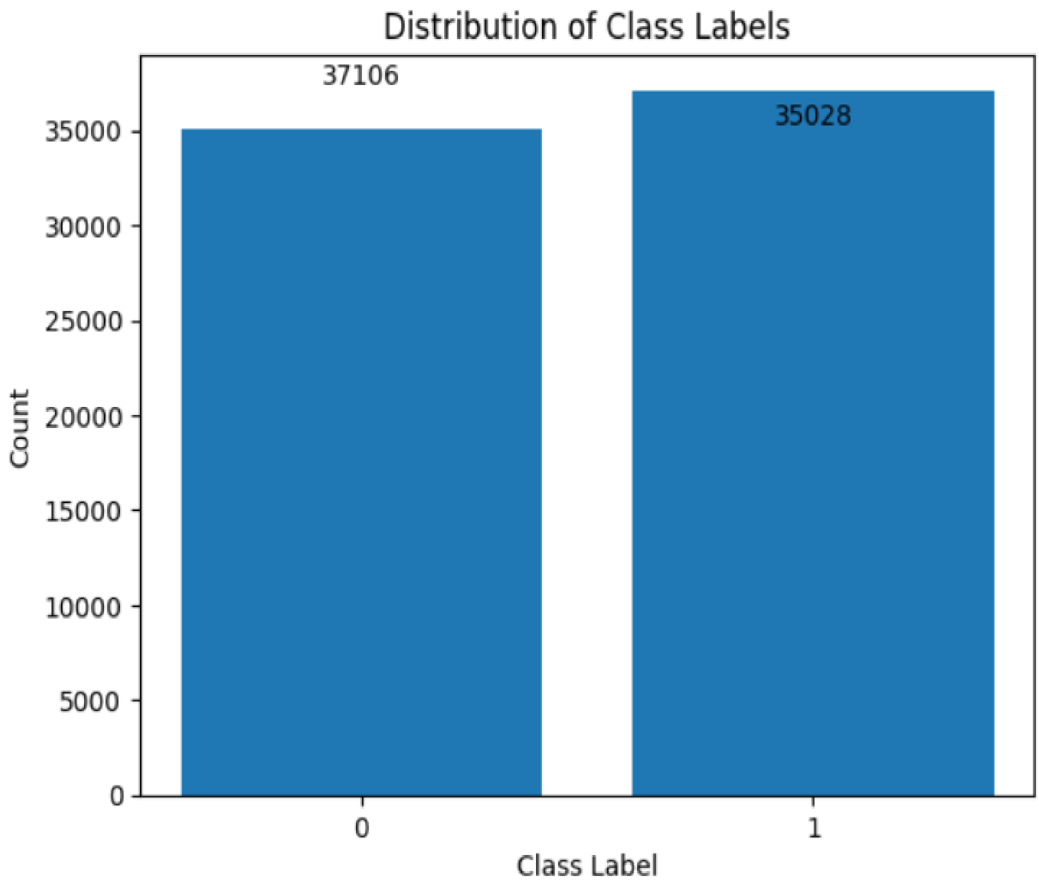

Figure 3.6 Distribution of Class Labels

- After that we are applying regular expression functions to remove things like: links, punctuation marks, brackets etc. which can deprive the performance of different classifiers.
- After applying regular expression functions we have to make the text data ready to feed it to the corresponding machine learning algorithms and for that we have used three to four vectorization techniques.
-

3.3.2 Vectorization Technique Used

In our implementation TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique has been used .

- **TF – IDF Vectorizer:** It stand for Term Frequency – Inverse Document Frequency. It simply tells us the relevance of a word in a particular text or the corpus. **Term frequency** of a word is calculated by simply dividing the frequency of that word in the document by total number of words in that document.[28]

Inverse Document Frequency can be calculated by dividing the total number of documents with the number of documents which contains that particular word. In the end the tf-idf term is calculated by multiplying the term frequency with the log of inverse document frequency. To form the vector we simply write the tf-idf value of the word for each document in the vector as we did in the countvectorizer. To use it on text first we need to import it from the scikit-learn library and simply applying it to the training and testing data.

3.3.3 Results for Basic Approach

We have achieved the following accuracies for different Machine Learning classifiers when applied with TF-IDF vectorization technique:

Table 3.1: Results for Basic Approach.

Classifiers/Vectorizers	**TF-IDF**
AdaBoost	92.93%
SVM	95.91%
Random Forest	93.80%
XGBoost	96.36%

As we see that all the ML algorithms AdaBoost, SVM, Random Forest, XGBoost are performing best with the TF-IDF vectorization technique. So we'll be using these four algorithms as an input to the ensemble models. It is because fusing the models with highest accuracy using ensemble models increases the accuracy of the resulting model.

3.3.4 Ensemble Learning

Using the useful method of ensemble learning, a number of independent models—also known as base models or weak learners—are combined to create a prediction model that is more accurate and dependable. The ensemble model then synthesises the predictions from each base model to arrive at a conclusion. Group learning often outperforms a single model by using the diversity and combined intelligence of the members of the ensemble.[29]

There are several techniques to group learning, each with special qualities and advantages. Let's take a closer look at these tactics:

Bagging:

Often referred to as "bagging," bootstrap aggregating is a popular ensemble technique where different base models are trained separately utilising different subsets of the training data. To create the subgroups, random samples are taken with replacement from the original training set using a process called bootstrapping. Every base model is trained using a separate bootstrap sample in order to bring variation to the training process. Usually, averaging or voting is used to aggregate the projections from several base models into a single final prediction. It is commonly known that bagging algorithms have the ability to reduce overfitting, which reduces variance and increases generalisation. Extra Trees and Random Forest are two examples.[30]

Boosting:

Base models are trained successively in the iterative ensemble process called "boosting," with each subsequent model trying to correct the mistakes made by the models that came before it. During training, more attention is paid to the data that the previous models misidentified, allowing the ensemble to learn from its mistakes and gradually improve its performance. Boosting methods such as AdaBoost, Gradient Boosting Machines (GBM), and XGBoost combine the predictions of multiple weak models in an effort to generate a final strong model. Boosting is particularly useful for handling intricate interactions and spotting minute patterns in the data.

Voting:

To create the final forecast, voting is a straightforward ensemble procedure that combines the forecasts of different base models. There are several names for voting, such as ensemble or majority voting. The class label that receives the majority of votes is used to make the final forecast, with each base model receiving an equal amount of votes. There are two ways to cast a vote: hard voting and soft voting. Hard voting just considers the final decision made by each base model, but soft voting also considers the expected probability or confidence ratings that the base models assigned to each class. Soft voting often results in forecasts that are more accurate because it considers the degree of certainty or uncertainty in the predictions given by the base models.

Weighted Ensemble:

In a weighted ensemble, the predictions from each base model are given a distinct weight. The relative importance or effectiveness of each model is represented by these weights. The weighted predictions of the base models are added together to create the final forecast. When specific models are predicted to perform better or have a greater impact on the out- come, weighted ensembles can be helpful. The ensemble can capitalise on the advantages of various models and raise overall prediction accuracy by applying the proper weights.[24]

The cleaning and preprocessing as well as the vectorization parts are same here except the fact that we are using one more vectorization technique here which is Word2Vec. Also in case of majority voting ensemble we are using both soft and hard voting. In case of weighted ensemble it makes no sense to use hard voting because it kind of gives preference to the model having more weight.

3.4 Final Methodology:

As we discussed earlier we are combining four best ML classifiers that we have used in the basic approach to improve its accuracy. The flow diagram for the methodology is:

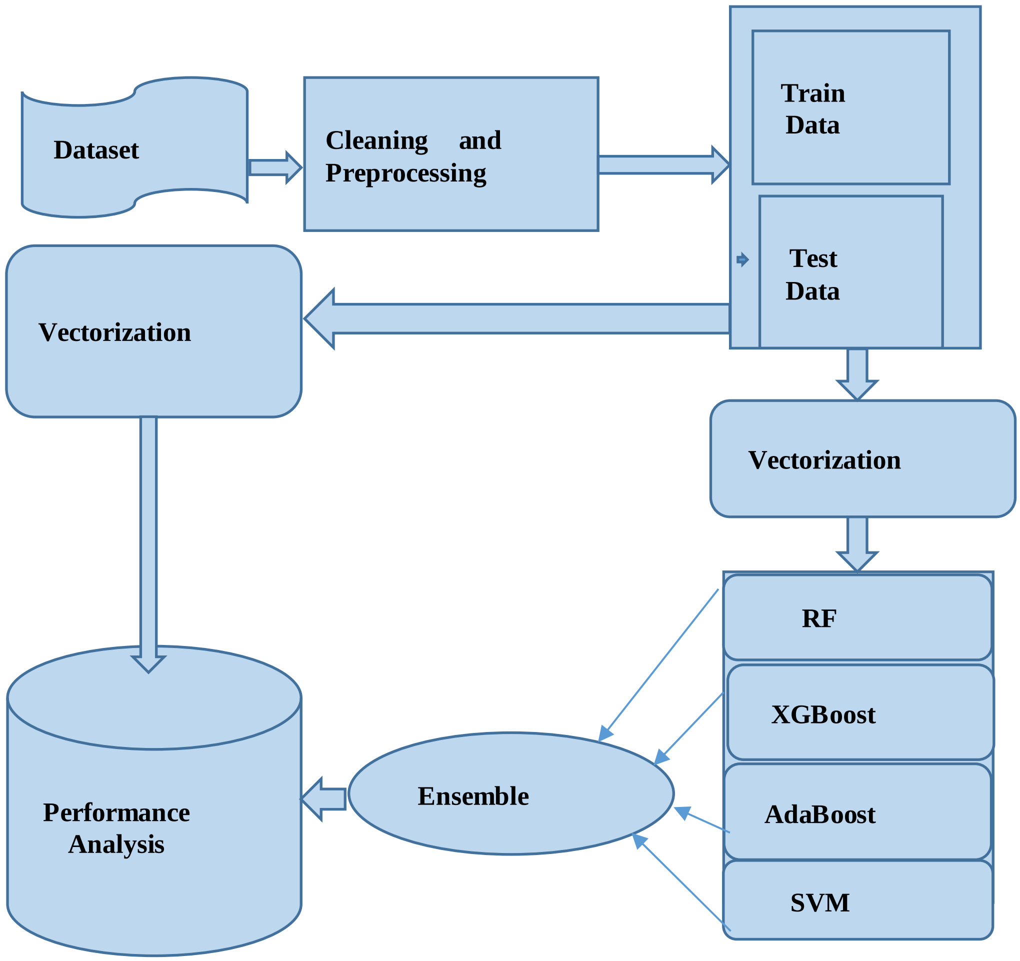

Figure 3.7 Final Methodology using Ensemble Methods

CHAPTER 4

RESULTS AND DISCUSSION

We have evaluated the impact of Ensemble models on the accuracy of state-of-the-art machine learning algorithms. The accuracy of the formerly employed algorithms significantly improved as a result of the ensemble learning. Soft voting method is used to calculate the effect of majority voting ensemble.

Table 4.1: Results for Accuracy Score

Classifiers/Vectorizers	**TF-IDF**
AdaBoost	92.93%
SVM	95.91%
Random Forest	93.80%
XGBoost	96.36%
Voting Classifier	97.03%

Table 4.2 Table for Various Metrics

	Precision	Recall	F1-score
AdaBoost	0.9208	0.9409	0.9307
SVM	0.9532	0.9665	0.9598
Random Forest	0.9313	0.9471	0.9392
XGBoost	0.9515	0.9777	0.9644
Voting Classifier	0.9628	0.9790	0.9708

With an accuracy of 97.03% in categorizing news articles as true or fraudulent, the Voting Classifier showed remarkable performance that is indicative of its resilience and trustworthiness. This model has a 96.28% precision rate, meaning that almost all the articles correctly predicted are authentic and highly important in minimizing false positives. The model, which has a recall of 97.90%, indicates the way it will most effectively capture accurate positive items and, hence, ensures that most of the actual articles can be caught. Therefore, it is found to give consistent results in identifying trustworthy news and avoiders of false negatives by giving the harmonic mean of precision and recall (97.08% F1 score). Therefore, when taken together, the above measures jointly lay out the strong aptitude and high accuracy of the model in correctly identifying news.

They are more concerned with the model's accuracy in classes separating between "fake" and "real" news. They are all high in precision, recall, and F1-score, as stated by the classification report on classes. High values, such as 96.28 for precision, 97.90 for recall, and 97.08 for the F1 score, would imply that the built model can consistently show both classes. The power of the ensemble model outperformed the individual models, where Random Forest had a reward accuracy of 93.80%, a precision of 93.13%, and a recall of 94.71%. This was further enhanced to 94.89% for the Voting Classifier, meaning that the ensemble approach harnessed the ability of multiple classifiers—like the ensemble of Random Forest, XGBoost, AdaBoost, and SVM—to deliver balanced and accurate results.

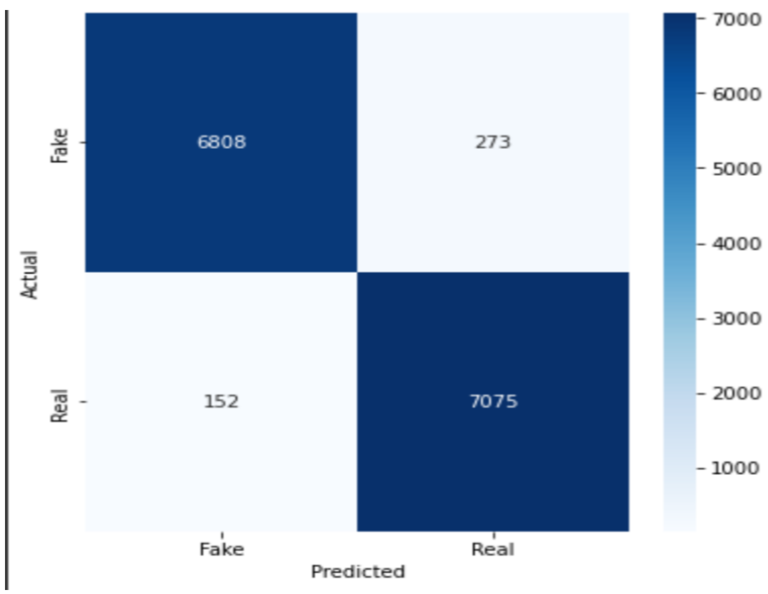

Figure 4.1 Voting Classifier Confusion Matrix

Confusion Matrix Analysis: The confusion matrix of the Voting Classifier is abundant on true positives, indicating: very well how it can recognize news articles; that is, that the high value in the matrix means the overall high accuracy of the model is extensively leveraged by the capability to detect big chunks of both genuine and bogus news pieces. It further appears that the model makes very few classification mistakes because the dataset only has 273 false positives and 152 false negatives. Although it did make identification mistakes occasionally, if it identified fake news as accurate news occasionally, and little accurate news was classified into fake news, these errors might also be influential about the model's usefulness. False negatives may further spread fraudulent information, while false positives erode the confidence in the trusted news sources. However, such low error rates do prove the model to be robust and reliable within real-world situations.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

This paper examines majority voting ensemble approaches in terms of their potential to enhance detection accuracy in classifying information as fake or real based on text data. From the results obtained, the proposed ensemble solutions effectively solved issues that could be initiated from erroneous data. Ordinary solution machine learning solutions outperform the approaches in terms of accuracy.

In conclusion, the extensive testing and performance evaluation showed that ensemble models easily beat the individual models, at least in comfort. The primary voting ensemble model aggregation could pull up the best from the multiples of classifiers to achieve dependable performance.

The latter can more easily encompass the most common host of false information: textual data. The findings corroborate that ensemble methods are better performing in the area because they can leverage the merged wisdom of many classifiers to identify and differentiate true and false information more accurately.

While this study has clarified the usefulness of the majority and weighted voting ensemble techniques, there is a further area where improvement can be made in all the areas of concern based on:

Feature Engineering: Consider the impact on the behavior of an ensemble model with different approaches to text representation, either deep learning-based methods or word embeddings. Going further with other, even more, sophisticated elements, namely, the exploitation of linguistic patterns or contextual cues, will help these features be better engineered to detect fake content.

- **Ensemble Combination Strategies:** Innovation in the research of combining different ensembles—Stacking or Hybrid Ensembles—to realize the benefits of whatever ensemble techniques. This technique usually outperforms and is the most robust of all, especially in dynamic and complex false information.
- **User Perception and Behavioral Analysis:** Understanding how user perception and behavioral analysis employs the identification of deceptive material in textual visual cues. Works up to the level of conceptual settings and human biases shall disrupt the objective judgment of visual stimuli and drawing methods to incorporate user behavior and input in ensemble models to allow more precise decisions.[26]

• **Visual Data Integration:** Research is underway to discover how false data algorithms can integrate visual data that might include images and videos. Most of the false information is expressed through pictures, mainly on social media sites. Involuntarily or not, most of the false information is expressed through pictures, mainly on social media sites. We consider techniques that might strip relevant features of visual input and confer them to ensemble models to increase the robustness and accuracy of false information detection systems. [25]

In this way, our research proves how the majority voting ensemble method can enable improvement in accuracy for the detection of information based on text data and false information. We will continue to push the boundaries toward the advancement on the topic of the detection of false information and the development of more reliable and robust systems to detect and curtail the influence of false information.

LIST OF PUBLICATIONS

[1] Anup Kumar Srivastav, Virender Ranga and Dinesh Kumar Vishwakarma. *Enhancing Veracity:Leveraging Machine Learning Ensemble Methods for Fake News Detection.* Accepted and Registered for presentation at **1st International Conference on Advances in Computing, Communication and Networking, IEEE** (16-17 December, 2024), Greater Noida, India.

[2] Anup Kumar Srivastav, Virender Ranga and Dinesh Kumar Vishwakarma. *Machine Learning and Deep Learning in Fake News Detection: An In-Depth Review.* Accepted and Registered for presentation at **1st International Conference on Advances in Computing, Communication and Networking, IEEE** (16-17 December, 2024), Greater Noida, India.

REFERENCES

[1] Elyassami, Sanaa, et al. "Fake news detection using ensemble learning and machine learning algorithms." *Combating fake news with computational intelligence techniques* (2022): 149-162.

[2] Neela, K., and R. Deepa. "An Ensemble Learning Frame Work for Robust Fake News Detection."

[3] Hakak, Saqib, et al. "An ensemble machine learning approach through effective feature extraction to classify fake news." *Future Generation Computer Systems* 117 (2021): 47-58.

[4] Kumar, Nitish, and Nirmalya Kar. "Approaches towards Fake news detection using machine learning and deep learning." *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2023.

[5] Singh, Gaurav, and Kamal Selva. "Detection of fake news using NLP and various single and ensemble learning classifiers." *Authorea Preprints* (2023).

[6] Pal, Aishika, and Moumita Pradhan. "Survey of fake news detection using machine intelligence approach." *Data & Knowledge Engineering* 144 (2023): 102118.

[7] Alarfaj, Fawaz Khaled, and Jawad Abbas Khan. "Deep Dive into Fake News Detection: Feature-Centric Classification with Ensemble and Deep Learning Methods." *Algorithms* 16.11 (2023): 507.

[8] Meel, Priyanka, and Dinesh Kumar Vishwakarma. "Fake news detection using semi-supervised graph convolutional network." *arXiv preprint arXiv:2109.13476* (2021).

[9] Sharma, Anu, M. K. Sharma, and Rakesh Kr Dwivedi. "Exploratory data analysis and deception detection in news articles on social media using machine learning classifiers." *Ain Shams Engineering Journal* 14.10 (2023): 102166.

[10] Agarwal, Arush, and Akhil Dixit. "Fake news detection: an ensemble learning approach." *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020.

[11] Khanam, Z., Alwasel, B., Sirafi, H., Rashid, M., 2021. Fake News Detection Using Machine Learning Approaches. IOP Conference Series: Materials Science and Engineering 1099, 012040.

[12] Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "Multiclass fake news detection using ensemble machine learning." *2019 IEEE 9th international conference on advanced computing (IACC)*. IEEE, 2019.

[13] Reddy, Harita, et al. "Text-mining-based fake news detection using ensemble methods." *International Journal of Automation and Computing* 17.2 (2020): 210-221.

[14] Vishwakarma, Dinesh Kumar, and Chhavi Jain. "Recent state-of-the-art of fake news detection: A review." *2020 international conference for emerging technology (INCET)*. IEEE, 2020.

[15] Jain, Akshay, and Amey Kasbe. "Fake news detection." *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE, 2018.

[16] Reis, Julio CS, et al. "Supervised learning for fake news detection." *IEEE Intelligent Systems* 34.2 (2019): 76-81.

[17] Yang, Yang, et al. "TI-CNN: Convolutional neural networks for fake news detection." *arXiv preprint arXiv:1806.00749* (2018).

[18] Castelo, Sonia, et al. "A topic-agnostic approach for identifying fake news pages." *Companion proceedings of the 2019 World Wide Web conference*. 2019.

[19] Ahmed, Sajjad, Knut Hinkelmann, and Flavio Corradini. "Combining machine learning with knowledge engineering to detect fake news in social networks-a survey." *arXiv preprint arXiv:2201.08032* (2022).

[20] Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. I. (2019, September). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)* (pp. 39-47). IEEE.

[21] Thota, Aswini, et al. "Fake news detection: a deep learning approach." *SMU Data Science Review* 1.3 (2018): 10.

[22] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. *arXiv preprint arXiv:1902.06673*.

[23] Sitaula, N., Mohan, C. K., Grygiel, J., Zhou, X., & Zafarani, R. (2020). Credibility-based fake news detection. *Disinformation, misinformation, and fake news in social media: Emerging research challenges and Opportunities*, 163-182.

[24] Giachanou, Anastasia, Guobiao Zhang, and Paolo Rosso. "Multimodal multi-image fake news detection." *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE, 2020.

[25] Vereshchaka, Alina, Seth Cosimini, and Wen Dong. "Analyzing and distinguishing fake and real news to mitigate the problem of

disinformation." *Computational and mathematical organization theory* 26 (2020): 350-364.

[26] Asghar, M.Z., Habib, A., Habib, A. *et al.* Exploring deep neural networks for rumor detection. *J Ambient Intell Human Comput* **12**, 4315–4333 (2021). https://doi.org/10.1007/s12652-019-01527-4

[27] Goyal, Priyanshi, Swapnesh Taterh, and Mr Ankit Saxena. "Fake News Detection using Machine Learning: A Review." *International Journal of Advanced Engineering, Management and Science* 7 (2021): 3.

[28] Park, Minjung, and Sangmi Chai. "Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques (Jan 2023)." *IEEE Access* (2023).

[29] Mahabub, A. (2020). A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. SN Applied Sciences, 2(4). https://doi.org/10.1007/S42452-020-2326-Y

[30] Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, *2020*, 1-11.