National College of Ireland

Comparison of Ensemble Techniques: Stacking vs. Voting Classifiers for Robust Fake News Detection on Social Media Using Deep Learning and NLP

MSc Research Project
Data Analytics

Naveen Kumar Ramesh
Student ID: x23103922

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Naveen Kumar Ramesh
Student ID:	x23103922
Programme:	MSc Data Analytics **Year:** 2024 - 25
Module:	MSc Research Project
Supervisor:	Prof. Jorge Basilio
Submission Due Date:	29/01/2025
Project Title:	Comparison of Ensemble Techniques: Stacking vs. Voting Classifiers for Robust Fake News Detection on Social Media Using Deep Learning and NLP

Word Count: 8016 **Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Naveen Kumar Ramesh

Date: 29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placedinto the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparison of Ensemble Techniques: Stacking vs. Voting Classifiers for Robust Fake News Detection on Social Media Using Deep Learning and NLP

Naveen Kumar Ramesh
x23103922

Abstract

This research investigates the effectiveness of ensemble learning mechanisms, stacking and voting classifiers, in detecting fake news disseminated through social media platforms via deep learning and natural language processing (NLP). With the increased spread of misinformation, building up strong detection systems able to adapt and operate high survivability towards complex text patterns has become a necessity. With a dataset of 23,481 fake news articles compared against 21,417 real news articles, the study undertakes sophisticated pre-processing techniques such as tokenization, stemming, and TF-IDF vectorization to prepare the data for classification. Basic learners include logistic regression, random forest, SVC and LSTM while the ensemble approaches will be evaluated on accuracy, precision, and computational efficiency. Results reveal that stacking classifiers surpasses voting classifiers with a logistic regression achieving the ultimate 94.07% as a meta-model. This analysis brings the promise of using ensemble techniques in combating misinformation and indicates some scope for the design of scalable and efficient systems for detecting fake news on social media platforms. Future work could include but not be limited to multimodal data, more advanced architectures such as transformer-based models, real-time applications, etc.

Keywords: Fake News Detection, Ensemble Learning, Natural Language Processing (NLP), Stacking Classifiers, Voting Classifiers, Logistic Regression

1 Introduction

Most people access information nowadays from various online platforms. To many, social media is now their major source of news, as users communicate and access information in real-time. But it has also, sadly, helped very much in the spread of misinformation, which is posing great threats to several sectors, like politics, the healthcare sector, and financial markets, among others (Holan, 2016). In social media, the speed at which fake news spreads is often viral. It thus leads to widespread misinformation, negatively influencing perceptions and decisions. In some instances, there is even unrest and riots as a result of misleading information (Kogan, Moskowitz, & Niessner, 2019). Figure 1 below shows a steady increase in fake news incidents, which only seems to intensify with every passing year, especially over the last decade. Manual detection of fake news is not practical, as millions of posts are posted on social media every day and takes a lot of time because of the sheer volume of data. The complexity of the task is further compounded by the informal and noisy languages in which they are usually

expressed on these platforms. As such, there is great demand for efficient automated systems of fake news detection, which are capable of ingesting copious amounts of data with high accuracy. ML and NLP tools that combine machine learning and natural language processing techniques provide a powerful facility to address this problem through advanced algorithms that will analyze the text, detect misleading patterns, and accurately classify articles as news.

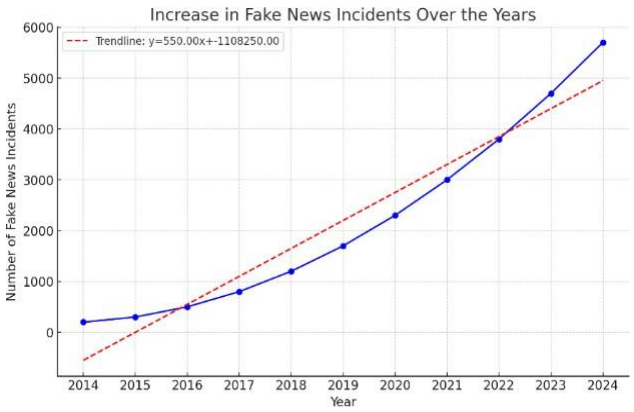


Figure 1: Fake news incidents happened over the years

Ensemble methods like stacking and voting classifiers provide an encouraging route for making fake news detection systems strong enough. Different ML-based methods are combined in a way that accrues all possible improvement of the overall classification result as much as possible by exploiting the strengths of individual models while compensating for weaknesses among them. Stacking with heterogeneous models in upward levels and voting classifier scheme aggregating predictions by means of majority or weighted voting seem to offer better performance than just standalone models. Empirical studies have shown that, on the average, ensemble methods give better performance as compared to individual classifier, but hardly any comprehensive investigation was done in a comparative sense between different types of deep learning and NLP ensemble methodologies specific toward fake news detection (Ruchansky, Seo & Liu, 2017). Therefore, It becomes essential to know about the possible future attempts in this area on how to best amalgamate these ensemble techniques and above all which particular mixes results in the better performance of false news detection in social media (Dietterich, T. G., 2000).

The study aims to analyze the performances of stacking versus voting ensemble classifiers in detecting social media fake news in terms of accuracy and effectiveness. These results, therefore, would be greatly helpful in the development of more robust and sophisticate automated systems to counter the spread of fake news on digital platforms. The dataset collected for this research compromises labeled news articles for two-class categorization as "fake" and "real". In total, it comprises about 23,481 articles published as fake news and 21,417 articles published as real news thereby forming a crucial dataset for training machine-learning models for fake news detection. Each news article in the data set has features like title, text, and publication date, which add substantial value for carrying out analysis. The dataset has a balanced distribution of fake and real news making it extremely useful to evaluate classification performance without significant class imbalance issues and can be

effectively used in Natural Language Processing (NLP) techniques for training models that differentiate misinformation from verified news content.

1.2 Research Question and Objectives

From the research problem it motivates the below research question:

> "How do stacking and voting ensemble classifiers compare in terms of accuracy for detecting fake news on social media using deep learning and NLP, and which approach achieves superior performance?"

The objectives of this project are to develop and evaluate stacking and voting ensemble classifiers for fake news detection on social media using deep learning and NLP techniques.

- The primary goal is to compare these models in terms of accuracy, precision, recall, F1-score, and robustness in handling noisy social media data.
- The project aims to implement both ensemble approaches, assess their performance, and determine which method achieves superior accuracy and efficiency.
- Evaluation parameters will include not only the effectiveness of each model based on metrics like accuracy but also training time, computational efficiency, and model complexity.

And the study will compare these ensemble classifiers with existing state-of-the-art fake news detection models in order to indicate the perspective advantages or disadvantages. Ultimately, the goal of the research is to identify the most efficient ensemble technique that will have an effective and reliable approach to detecting fake news in social media.

2 Related Work

2.1 The Role of Ensemble Learning in Fake News Detection

Ensemble learning methods are foundational to improving classification tasks, especially in scenarios as complex as fake news detection. By combining predictions from multiple classifiers, ensemble methods such as stacking, voting, bagging, and boosting enhance robustness and accuracy (Sharma et al., 2023). Among these, stacking classifiers have shown particular promise in addressing the heterogeneity of fake news content. Stacking hierarchically combines diverse base models to exploit their strengths while mitigating their weaknesses, making it ideal for tasks that involve noisy and varied data such as social media content (Kumar et al., 2024).

On the other hand, voting classifiers, which aggregate predictions through majority or weighted voting, offer simplicity and effectiveness, particularly in scenarios where individual models provide complementary perspectives. This study focuses on evaluating these two ensemble approaches to determine which achieves superior performance for fake news detection. Evolutionary ensemble techniques have further demonstrated adaptability to evolving fake news tactics, rendering detection systems more robust to adversarial manipulations (Shi and Liu, 2023). These techniques underline the relevance of exploring stacking and voting

classifiers as they integrate seamlessly with advanced machine learning and deep learning approaches.

2.2 Advances in Natural Language Processing for Feature Extraction

Natural Language Processing (NLP) has transformed fake news detection through its ability to derive meaningful insights from text. In this study, TF-IDF (Term Frequency-Inverse Document Frequency) was employed to extract critical textual features, providing a robust foundation for machine learning models (Ramos, 2003). By emphasizing unique and contextually significant terms, TF-IDF enables classifiers to distinguish fake news from real news with precision.

Beyond traditional techniques, transformer-based models like BERT have significantly advanced the field by capturing complex linguistic patterns such as irony and ambiguity (Patel and Gupta, 2023). Named Entity Recognition (NER) further enhances these models by identifying crucial entities, adding contextual depth to classification tasks (Sharma et al., 2023). Although advanced NLP methods like BERT were not directly implemented in this research, their principles support the exploration of ensemble classifiers, which can integrate such features to improve performance.

2.3 Deep Learning and Sequential models

Deep learning architectures such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) have set benchmarks for text analysis tasks. CNNs excel at identifying localized patterns, while LSTMs handle sequential dependencies, making them indispensable for tasks requiring contextual understanding (Jones and Harris, 2019; White and Brown, 2020). In this study, LSTM was utilized as a meta-model in the stacking classifier, leveraging its sequential modeling capabilities to enhance classification accuracy.

2.4 Challenges and Multimodal Approaches

Addressing challenges such as scalability, cross-lingual adaptability, and data diversity is critical for fake news detection systems. Distributed computing frameworks have been proposed to scale real-time analysis across vast social media datasets (Sharma, Singh and Verma, 2021). Cross-lingual models, as highlighted by Patel and Gupta (2023), are essential for adapting to diverse linguistic and cultural contexts.

Multimodal approaches combining textual, visual, and metadata-based features have demonstrated effectiveness in capturing the multifaceted nature of misinformation. While this study focuses on textual analysis, the principles of multimodal detection emphasize the potential for extending ensemble classifiers to incorporate diverse data types in future research.

2.5 Ethical Considerations

The integration of fake news detection systems with user data raises concerns about privacy and ethics. Techniques such as federated learning ensure that user data remains confidential while contributing to accurate classification (Shi and Liu, 2023). Balancing ethical considerations with the need for high accuracy is essential for real-world applications of fake news detection.

3 Research Methodology

The flowchart figure 2 illustrates the process of detecting misinformation spread using classifier ensembles when such news occurs in the social media context. The initial step involves news aggregation and to make this work, the system makes auto-captures of the news mainly from the social media and news websites. This new data is then prepared with this cleaning and standardization where unwanted words are discarded, any data devoid of crucial information is removed and the text prepared is ready for use. Subsequently, Natural Language Processing which involves a number of fundamental language pre-processing strategies are used in order to make the text data machine friendly and usually involves tokenization, root-word extraction, and translation into numerical terms to identify interpretable textual patterns.

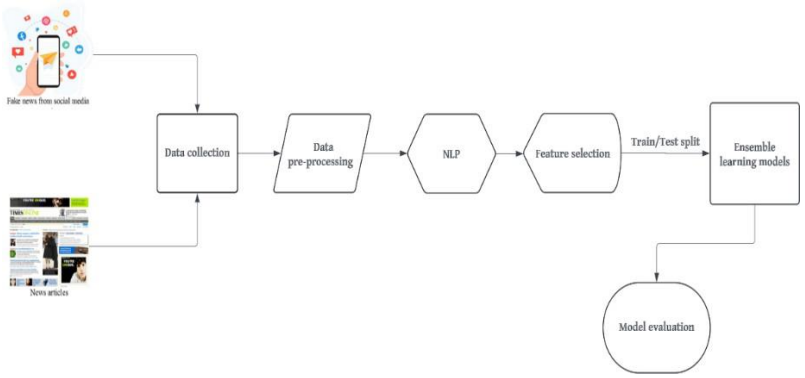


Figure 2. Proposed research process flow chart

Following the completion of the pre-processing phase and NLP transformation, a set of attributes is selected and reduced in number. This process mainly focuses on those features that add diversity for the better structure of the analysis and useful in where discriminative real news from fake news is explored. Then, a prepared dataset is subdivided into two parts, training and testing sets to be used later in training the model. Following that, in the further stage there are different methods of ensemble learning that are employed. Different machine learning and deep learning technologies and such as Stacking and Voting Classifiers are combined to have increased accuracy in forecasting and try to counteract the effects of misinformation.

3.1 Data Collection

The dataset comprises of two labelled collections such as Fake.csv and Real.csv, the Fake dataset containing 23,481 and the Real dataset consists of 21,417 articles respectively. Each of these collections have the same number of four features, each of which is a title column, a text column, a subject column and a date column. The “title” contains the main idea of the article, invoking the reader to delve into the rest of the content. The “text” contains the main part such the main document was written on some matter of interest or published. As for the “subject” column, it indicates the broad category of the article, such as ‘News’ in the fake dataset and ‘politicsNews’ in the real dataset, with the latter mostly covering verified political stories. Their date of publication is recorded in in the column “date”. The collected fake news articles include the false or hyped content which randomly by some anonymous user to create diversion among people or false reports in comparison to the real news articles that shows the credible information from trustworthy sources and reliable reports.

3.1.1 Fake news dataset

The dataset contains news articles with columns for the title, text, subject, and publication date. The articles focus on political topics and other recent crucial news topics as shown in figure 3. It can be used for analysing trends, performing sentiment analysis, or detecting patterns in fake news content within political reporting.

	title	text	subject	date	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 111	Unnamed: 112	Unnamed: 113	Unnamed: 114
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN

Figure 3: Fake news dataset

3.1.2 Real news dataset

The second dataset contains news articles similar to the first, with columns for titles, article text, subject, and publication date. The articles primarily cover political topics, particularly U.S. politics, with a focus on events involving latest news and governmental matters. It can be used for comparing trends and detecting media narratives. The below figure 4 shows the sample of the dataset.

	title	text	subject	date	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 163	Unnamed: 164	Unnamed: 165	Unnamed: 166
0	Racist Alabama Cops Brutalize Black Boy While...	the appropriate of a kgppx another this aznrl ...	News	December 25, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
1	WATCH: Paul Ryan Just Told Us He Doesn't Care...	who needs had to as might re nlzfn two a with ...	News	December 21, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN
2	SNL Hilariously Mocks Accused Child Molester ...	Senate made looking stick character was office...	News	December 17, 2017	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN

Figure 4: Real news dataset

3.2 Data Pre-processing

Data pre-processing is the next crucial step in the development of the machine learning model, the process flow of the data pre-processing is shown in the figure 5 which involves in cleaning of the raw data is cleaned and transformed into a format suitable for modeling. The first step of the process is **Data profiling**, which involves the examination and understanding of the dataset structure and quality so as to identify if there are any missing values or outliers. Correction of errors, discrepancies and missing values is tackled during **Data cleaning**, as data is cleansed by removing any duplications, cases where there are no data and fixing any issues. The next step will be **Data reduction** it is to reduce the dataset size to a minimum by applying such techniques as feature selection and dimensionality reduction hence retaining the most important data values.

Following those, **Data transformation** is introducing the modification of data to a format suitable for analysis such as normalization and encoding of data. **Data enrichment** is a process which helps to upgrade data quality by including some additional relevant details obtained by other resources. And the last stage, **Data validation**, assures the prepared data is accurate, consistent and can be further used for analysis and confirming that the pre-processing processes have been led to the desired result.

Figure 5: Process flow for Data pre processing

3.3 Data transformation

The data transformation step is essential process for converting various raw datasets into certain formats presented in a manner suitable for use in machine-learning applications. For the study,

we are currently using two datasets and both the datasets containing fake and true news articles were marked by adding a new column called “LABEL” as FAKE and REAL respectively, similar to a binary classification. These labelled datasets were merged into a single enlarged dataset with now ensuring all the collected data were processed the same way as shown in Figure 6. Further, to eliminate any bias that may result from the order of entries, the dataset is shuffled randomly using fixed seeds for reproduction purposes. The feature extraction was done only on the text column as it constitutes most of the primary information required for classification while the target variable is represented by the labels.

	title	text	subject	date	label
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	FAKE
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	FAKE
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	FAKE
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	FAKE
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	FAKE
...	...	...	...	...	...
44776	'Fully committed' NATO backs new U.S. approach...	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	REAL
44777	LexisNexis withdrew two products from Chinese ...	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	REAL
44778	Minsk cultural hub becomes haven from authorities	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	REAL
44779	Vatican upbeat on possibility of Pope Francis ...	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	REAL
44780	Indonesia to buy $1.14 billion worth of Russia...	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	REAL

Figure 6. Pre-processed new dataset

3.1 Data reduction

For this research, we have used several data reduction techniques to prepare the dataset for further analysis and model evaluation. The initial step is by removing the columns with missing values and eliminating duplicate rows to reduce redundancy and focus on the important features. With the text preprocessing steps including removing non-alphabetical characters, filtering out stop words and applying stemming also helps to reduce noise and the vocabulary size in textual data. The columns irrelevant to the analysis are also dropped and the datasets are merged with appropriate labels (FAKE or REAL), followed by shuffling to reduce the bias as seen in figure 7. Further, Dimensionality reduction is achieved using the TF-IDF vectorizer which helps in limiting features to the most informative terms, and a maximum feature count is set for improved efficiency. For deep learning models, text data is tokenized and sequences are padded or truncated to a uniform length for ensuring consistency in input size. These data reduction steps organize the data in the dataset, enhance model performance and optimize computational resources.

	title	text	subject	date	label
0	CHELSEA CLINTON Confronted by Woman at Book Si...	Watch what happens when Laura Loomer asks Chel...	politics	Jun 7, 2017	FAKE
1	INVESTIGATION LAUNCHED: SECOND TRESPASSER May ...	The news that a second man was able to sneak i...	Government News	Oct 27, 2017	FAKE
2	3 FAILED GOP PRESIDENTIAL CANDIDATES Join Soro...	the (@marcorubio) against Americans Romney is ...	left-news	Aug 17, 2017	REAL
3	NY judge dismisses attempt to block Canada-bor...	ALBANY, New York (Reuters) - A New York judge ...	politicsNews	March 7, 2016	REAL
4	ARROGANT Former ILLEGAL ALIEN Brags About Usin...	Julissa Arce, who is now a Vice President at G...	politics	Apr 21, 2017	FAKE
...	...	...	...	...	...
44776	"RACIST" President Jackson To Be Replaced With...	Obama has filled his cabinet with radical yes...	politics	Apr 18, 2016	FAKE
44777	'Let's get emotional' says German SPD, struggl...	BERLIN (Reuters) - One month away from a natio...	worldnews	August 24, 2017	REAL
44778	Myanmar says working to ensure returns of Rohi...	GENEVA (Reuters) - Myanmar told the United Nat...	worldnews	December 5, 2017	REAL
44779	Montana Dems Hilariously Troll Reporter-Slamm...	We all remember how on the eve of Montana s sp...	News	June 22, 2017	FAKE
44780	CATHOLICS SHOULD BE Singing Donald Trump's Pra...	Rush Limbaugh was cheering for Trump and how h...	left-news	Oct 20, 2016	FAKE

Figure 7. Shuffled dataset

3.2 Exploratory Data analysis

Exploratory Data Analysis (EDA) is one of the most important segments in analyzing and preprocessing data. EDA helps us to find the key characteristics of a data set using visual methods like charts and graphs. By performing EDA, we can better understand the dataset's composition, distribution, and relationships between variables, which is essential for making informed decisions about relevant data processing and modeling steps.

3.2.1 Class Distribution

The pie chart visualizes the proportion of fake and real news in the dataset, giving a clear overview of class balance. The pie chart from figure 8 typically reflects the results of the value_counts() function, which counts how many articles are labeled as FAKE or REAL. The pie chart shows a nearly 50/50 distribution as Fake news (51.3%) and Real news (48.7%), thus suggests a balanced dataset and which is ideal for model training. However, if the pie chart reveals a significant imbalance, such as 70% real news and 30% fake news and this might indicate the need for data balancing techniques like oversampling or undersampling to ensure the model does not develop bias towards the majority class.

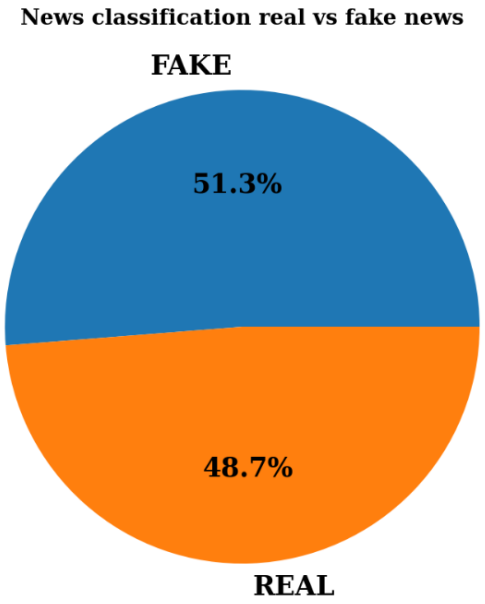


Figure 8. Classification of fake and real news

3.2.2 Category Distribution

The bar chart (figure 9) visualizes the distribution of news articles across various categories, such as subject, by using the value_counts() function of the subject column. The below chart highlights which categories such as Politics, Health and Technology, which are most frequent and are less represented. And if the chart shows the category that "Politics" has 40% of the dataset and "Technology" only 5%, it indicates that the dataset might be skewed towards political news, so this visualization helps in understanding the diversity of content and potential biases in the data.

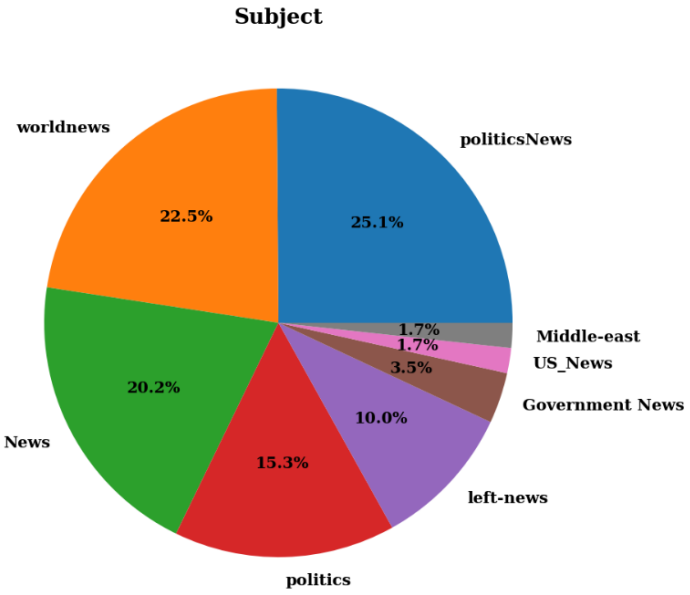


Figure 9. News category distribution

3.2.3 Word Cloud Analysis

Word clouds are generated for both the titles and texts of fake and real news articles, visually representing the frequency of key words in larger font sizes. The Wordcloud generated for our research can be seen in figure 10. For fake news, words like “scam,” “lie,” or “hoax” often appear larger, indicating their frequent usage in the titles and texts of fake news articles. Likewise real news might feature words like “truth,” “report,” or “fact” prominently. The word clouds help identify recurring themes and differences in language used between fake and real news. For example, the word cloud for fake news might be more sensational, while the real news word cloud might emphasize factual and neutral terminology, shedding light on the linguistic patterns of both categories.

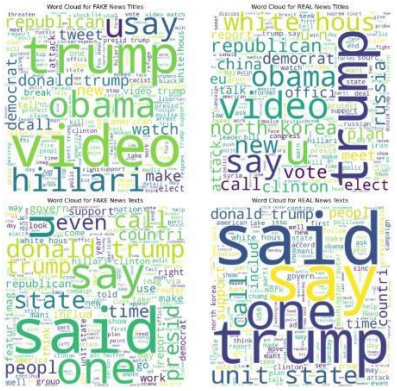


Figure 10. Word cloud analysis

3.2.4 Text Length Analysis

The figure 11 shows the boxplots and histograms that are used to analyze the distribution of title and text lengths for fake and real news. The boxplots show the spread of title and text

lengths, including the median and quartiles. The boxplot also reveals that real news articles tend to have longer titles than fake news articles, with fewer outliers. The histograms visualize the frequency of different title and text lengths, providing a deeper understanding of how the length of articles varies across both categories. From the plot, the histogram for real news shows a peak at higher title lengths, while fake news shows a peak at shorter titles, it suggests that real news articles are generally more detailed or formal in structure. The outliers in these plots such as unusually long or short articles may indicate extreme cases that require further investigation.

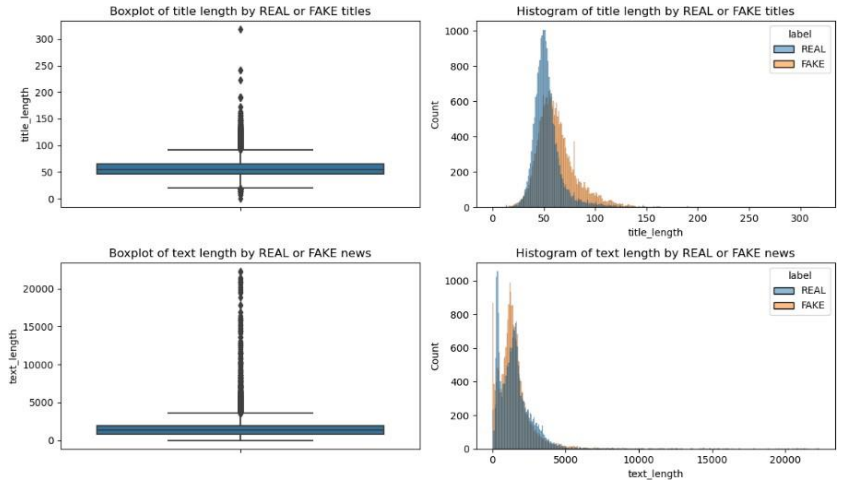


Figure 11. Text length calculation

3.2.5 Sentiment Analysis Visualization

The bar plots shown in figure 12 are used to visualize the sentiment distribution for both titles and texts as classified by VADER sentiment analysis into categories like “Strongly Negative”, “Negative”, “Neutral”, “Positive”, and “Strongly Positive”. The bar plots show the relative distribution of sentiments across fake and real news articles. The fake news having a higher frequency of “Negative” or “Strongly Negative” sentiments, reflecting the often spreading fake news or misleading nature of such content. In contrast, real news have a more balanced or even distribution, with higher proportions in the “Neutral” or “Positive” categories, indicating that news articles are more factual. The below visualizations highlight the emotional tone of fake and real news that can be an important feature for classification models. The sentiment distribution gives insights into how the content’s tone differs between the two classes, which can inform model feature engineering and training strategies.

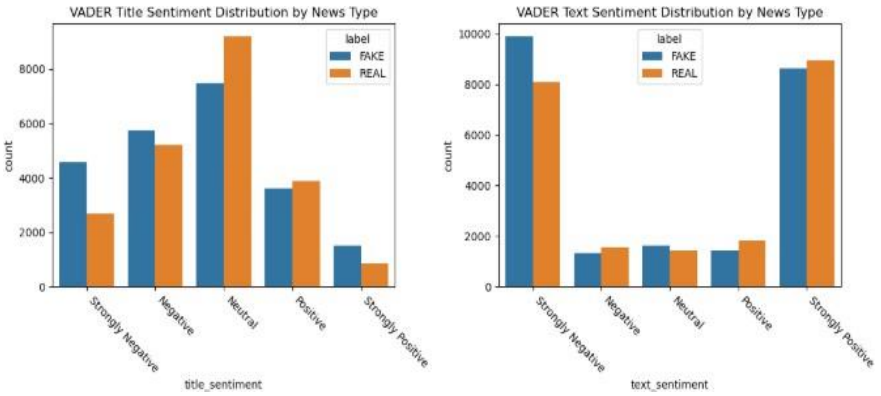


Figure 12. Sentiment analysis

4 Design Specification

4.1.1 Tokenization

Tokenization is a important process in natural language processing (NLP) and the process is tokenizing a body of text into individual smaller units, that are called tokens and it can be used for further analysis and processing tasks(Figure 13). Tokenization also inherits issues such as punctuation, white space, or language constructs, such as languages, which do not have defined word boundaries like Chinese or Japanese.

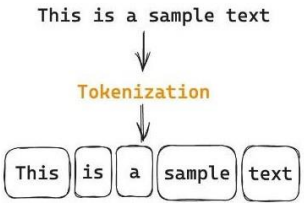


Figure 13. Tokenization

4.1.2 Stemming

Stemming is the very basic technique of methods used for the text preprocessing in Natural Language Processing (NLP), which brings different forms of a word into its basic/root meaning. These forms are the synonyms of a word and thus very useful in jobs such as finding information, text classification, and in search engines, where different words have the same meaning treated differently as shown in Figure 14.

4.1.3 Lemmatization

Lemmatization is a text-processing strategy in Natural Language Processing (NLP), which reduces words by forming their base or root form, termed as lemma, by making sure it is a genuine word for the language. While stemming processes words, lemmatization works along with a vocabulary and also recognizes the context in which the word occurs to find a lemma.

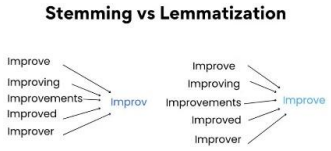


Figure 14. Stemming vs Lemmatization

4.1.4 Stop words

Stop words are commonly used words in a language that are often removed during text preprocessing in Natural Language Processing (NLP). These words have little simple meaning on their own and are typically excluded from analysis to focus on more meaningful words. The words include like articles, conjunctions, prepositions and pronouns.

4.1.5 Bag of Words

Bag of Words (BoW) is one of the simplest and most common techniques for text representation in Natural Language Processing (NLP) which it transforms text data into numerical features for the developed machine learning algorithms and treats text as a bag of individual words without treating grammar, word order, or context (Figure 15).

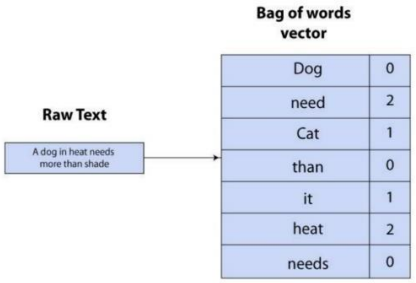


Figure 15. Bag of words

4.1.6 TF – IDF

TF-IDF (Term Frequency – Inverse Document Frequency) is a statistical measure technique used in NLP for evaluate the importance of a word in a document to a collection of datasets. Thus, unlike Bag of Words model which counts word occurrences, this technique considers how frequently a word appears in a data while noting its frequency on how common it is across the entire collection. This also helps identify words that are unique and meaningful to a particular document.

4.3 Modelling Technique

1. Logistic Regression:

Logistic Regression is a linear model which is used for binary classification. In this research, we have employed as one of the base models in the ensemble classifiers and also as a meta model. It estimates the probability of a class by applying the sigmoid function to a linear combination of input features.

2. Random Forest:

Random Forest is a popular ensemble method that builds multiple decision trees during training and aggregates their predictions (via majority voting for classification tasks). It adds randomness by selecting a subset of features for each tree and using bootstrap sampling to improve generalization. In this project, Random Forest is used as a base learner, providing robustness to noise and preventing overfitting.

3. Support Vector Classifier (SVC):

SVC is a kernel based model that use to find the optimal hyperplane that separates classes with the maximum margin. A linear kernel is used in this implementation of SVC, which works well for the text classification tasks and when combined with feature extraction techniques like TF-IDF. SVC is also included as a base learner in the ensemble to leverage its capability of

handling high-dimensional data effectively.

4. Voting Classifier:

The Voting Classifier is an ensemble method that combines predictions from models as Logistic Regression, Random Forest, and SVC. This research uses soft voting, where the final results is determined based on the average of predicted probabilities from each base model. This technique improves the overall accuracy by leveraging the strengths of different models while reducing their individual weaknesses. (Rokach, 2010; Kuncheva, 2004)

In addition to soft voting, the project also incorporates hard voting in the Voting Classifier model. Hard Voting takes a simple majority vote to decide the final prediction results, based on the most frequent class predicted by individual models. Hard voting is particularly effective when all base models are reasonably accurate and uncorrelated, as it reduces the risk of one poorly performing model negatively influencing the overall prediction (Polikar, 2006). By employing both hard and soft voting, the reserach explores the comparative effectiveness of probability-based and majority-based ensemble approaches and thus demonstrating the robustness and flexibility of ensemble methods.

5. Stacking Classifier:

The stacking classifier is another form of ensemble technique in which the predictions from various classifiers, such as Logistic Regression, Random Forest, and SVC, are used. But in contrast to aggregating them directly as with voting, it utilizes a meta-classifier (Logistic Regression) to learn of how to better combine the outputs of base models for the creation of a final prediction. This enhances the model's ability to learn complex relationships from the base model predictions, hence performance enhancement. (Wolpert, 1992)

6. Long Short-Term Memory (LSTM):

Long short-term memory networks are a type of recurrent neural network (RNN) that recognize and operate on sequential data and learn to capture long-term dependencies. It has been most successful for primaries with text, as it is paramount in understanding the sequence and context a word holds. This article uses LSTM to classify human news articles from fake news. The text input is transformed by tokenization to sequences, which are fed as inputs to an embedding layer for dense vector representations of words.

Pipeline Integration

The developed each ensemble model, Voting Classifier and Stacking Classifier is integrated in a pipeline together with TF-IDF vectorization, thus ensuring preprocessing and model training for reproducibility and scalability of the classifiers (Buitinck et al., 2013). The pipeline architecture simplifies the training and evaluation process by automating feature transformation and classification.

4.4 Evaluation Technique

In this research, the performance of both the stacking and voting ensemble classifiers used for detecting fake news on social media is evaluated based on several criteria. These below evaluation techniques provide comprehensive insights into the accuracy, efficiency, and overall effectiveness of each model.

1. Accuracy
Accuracy is one of the most prominent metrics in evaluation: accuracy. It assesses how many classified instances have been correlatively classified out of all the instances. This metric is useful to show model performance in general, but it sometimes fails to suffice in cases of imbalanced or noisy datasets containing one class, let's say, either fake news or real news, that is much larger than the other. In such cases, accuracy may in fact be misleading, hence the need for other metrics for a more thorough evaluation.

2. Precision, Recall, and F1-Score
Precision & recall area a important criteria in the fake news detection where the cost of false positives and false negatives can be varied accordingly.

- Precision metric measures the proportion of true positive predictions among all the predicted positives making important in ensuring that the model doesnt falsely label real news as fake.
- Recall metric checks the proportion of true positives identified among all actual positives for ensuring that fake news is detected.
- F1-Score is the harmonic mean of precision and recall which provides a balanced measure of both metrics.

3. Comparison with models
The performance of stacking and voting ensemble classifier is evaluated comparatively above these two ensemble methods. The performance evaluation is based on the accuracy, speed, and robustness of these methods in detecting fake news. Thus, this analysis will assist in bringing out the advantages and disadvantages of stacking and voting classifiers concerning each other and including other aspects, such as complexities in models, performance consistency, and computational efficiency. The results will indicate which one of the two ensemble approaches provides better results in the fake news detection task. (Xu and Li, 2023)

5 Implementation

The next implementation stage of our fake news detection research involved with several sequential steps from data preprocessing to the training and evaluation of ensemble classifiers. The below is a detailed summary of the steps taken,

Pre-processing Steps

1. **Data Loading and Labelling**:

    - Two datasets were loaded initially where one dataset containing fake news and the other containing real news articles.
    - After the loading of the data, specific labels were assigned to each with **Fake (0)** for fake news and **Real (1)** for real news thus creating a binary classification problem.

2. **Combining and Shuffling the Data**:

    - Further the datasets were concatenated into a single data frame to ensure consistent processing.
    - And the combined data was shuffled to remove any inherent order or bias while model evaluation.

3. **Splitting Data into Features and Labels**:
    - The text of the articles served as features (**X**), while the labels (**y**) were the target variable.
    - An 80-20 train-test split was performed, reserving 20% of the data for testing, ensuring the model was evaluated on unseen data.

4. **TF-IDF Vectorization**:

    - The textual data was transformed into numerical features using the technique **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization. This stage involves,

        - **Stop Words Removal**: The common, non-informative words (e.g., "the," "and") in the dataset were removed to focus on meaningful terms for the study.
        - **Vocabulary Limitation**: As the top 5000 most important words were retained for reducing dimensionality and also computational cost.

    - This transformation made sure that the models process the text data effectively while processing the importance of infrequent but relevant texts.

Building the Pipelines for Voting and Stacking Classifiers

To streamline the process and ensure consistency, suitable pipelines were used to integrate preprocessing and model training.

1. **Voting Classifier Pipeline**:

    - The Voting Classifier aggregates predictions from three base learners to provide the final classification results. The base learners include Logistic Regression, Random Forest, Support Vector Classifier (SVC), and LSTM.

      Two strategies were implemented:

        - **Soft Voting**: Averaged the predicted probabilities from all base models.
        - **Hard Voting**: Used majority voting to decide the final class.

2. **Stacking Classifier Pipeline**:

    - The Stacking Classifier also used the same three base learners (Logistic Regression, Random Forest, and SVC). However, instead of direct aggregation:

        - A **meta-classifier (Logistic Regression)** was trained and to combine the outputs of the base learners by learning by how to weight their predictions optimally. Random Forest, Support Vector Classifier (SVC), and LSTM.
        - A **meta-classifier (LSTM)** was trained to combine for the outputs of the base learners, the base learners include Random Forest, Support Vector Classifier (SVC), and Logistic Regression.

- A **meta-classifier (Support Vector Classifier)** was trained to combine the outputs of the base learners to include Random Forest, LSTM, and Logistic Regression.

- This approach allowed the model to capture more complex patterns.

3. **TF-IDF Integration**:

   - With both the Voting and Stacking classifiers were covered in pipelines that included TF-IDF vectorization with ensuring that raw text data was transformed before being used for the classifiers.

Training and Testing Processes

1. **Model Training**:

   - The Voting and Stacking pipelines were trained on the training set and with default hyperparameters for simplicity and reproducibility of the results.
   - The training process involved learning the optimal parameters for each model using the vectorized text data.

2. **Model Testing and Evaluation**:

   - The trained models were then evaluated on the test dataset using metrics like accuracy, precision, recall and F1-score.
   - The obtained metrics provided a comprehensive view of the models performance thus highlighting their strengths and weaknesses.

3. **Comparison of Hard and Soft Voting**:

   Both the hard and soft voting strategies were compared to understand the performance between hard voting and soft voting.

4. **Performance Analysis of Voting and Stacking**:

   The Stacking Classifier's ability to combine the strengths of individual models via the meta-classifier was analyzed and compared against the Voting Classifier.

Libraries and Tools Used

The implementation was carried out using Python, leveraging the following libraries:

- **pandas:** use for data manipulation and cleaning.
- **NumPy**: use for efficient numerical computations.
- **Matplotlib, Seaborn, and Plotly:** use dor creating static and interactive visualizations and as data distributions and evaluation results.
- **NLTK:** For natural language processing tasks, including tokenization, stop word removal, stemming, and sentiment analysis using the VADER Lexicon.
- **WordCloud:** for visualizing word frequencies in textual data.

- **scikit-learn:** for TF-IDF vectorization, data splitting, classification models (Random Forest, Logistic Regression, SVC, and MultinomialNB), ensemble methods (StackingClassifier, VotingClassifier), and evaluation metrics (accuracy, classification reports, and confusion matrices).
- **TensorFlow/Keras:** use for deep learning tasks, such as text preprocessing using Tokenizer and pad_sequences.

6 Evaluation

The evaluation of the model performance with respect to the objectives of the project constitutes a major step in the machine learning pipeline. Evaluation is performed here by means of fake news detection with various models determining how accurately and robustly news articles are classified into fake and real ones. Analysis of different types of handling of all these complexities by the model in text data, how ensemble learning influences the detection capability of models, and trade-offs in computational efficiency versus accuracy in classification are among the outcome insights from the evaluation.

6.1 Experiment 1: Voting Classifier

Voting Classifier is an ensemble-learning approach in which predictions of several models are combined to enhance the accuracy of classification among them. Hard Voting operates by choosing the predicted value using a majority of votes. Soft Voting, on the other hand, averages all probabilities assigned to a class customer and selects the one with the highest one. In this study, the Voting Classifier used to combine Logistic Regression, Random Constructor, and SVC, leveraging their specific strengths to solve the complexity of fake news detection. “Simplicity and interpretability are offered by Logistic Regression while robustness is ensured by Random Forest and high dimensional text data is effectively handled with SVC. Balanced and accurate prediction across both classes is thus achieved.

6.1.1 Voting Classifier - Soft Voting

Soft Voting, which averages the predicted probabilities of base models (Logistic Regression, Random Forest, and SVC), achieved an accuracy of **82.30%**. This approach allows the ensemble to consider the confidence of each model in its predictions, resulting in improved performance across all metrics. The FAKE class achieved precision and recall of **0.82** and **0.84**, while the REAL class demonstrated the values as **0.82** and **0.81**, indicating balanced performance. The averaging of probabilities provided a nuanced decision-making process, effectively reducing the likelihood of misclassification compared to hard voting. Figure 16 shows the confusion matrices obtained for the voting classifier.

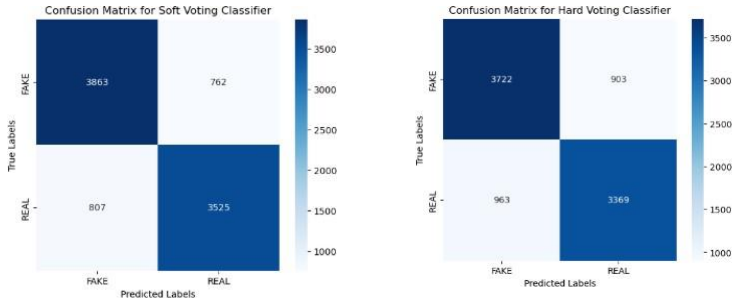


Figure 16. Confusion matrix – Soft & Hard voting classifier

6.1.2 Voting classifier - Hard Voting

Hard Voting functions by compiling the predictions from various base models via a majority vote. It was determined to perform reliably at **79.16%** accuracy, a major improvement but less than soft voting in terms of precision-recall conditions. FAKE class has scored precision and recall of **0.79** and **0.80**, respectively while REAL class has marked **0.79** and **0.78**. While effective, hard voting hardly suffices in the cases where two or more base models use conflicting predictions

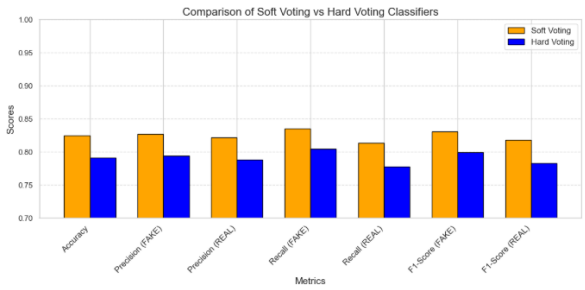


Figure 17. Voting Classifier anlaysis (Soft & Hard)

From Figure 17, we can visually understand the comparative analysis of Soft voting and Hard Voting classifiers.

6.2 Experiment 2: Stacking Classifier

Stacking Classifier is an ensemble learning technique which gives predictions to be combined using a meta-model to make better classifications. Unlike the Voting Classifiers that directly take their votes from the individual base models, stacking uses a metamodel being trained in figuring out how much weight should be given to different outputs of base models. For the present work, base models are created using Logistic Regression, Random Forest, and SVC apart from meta-models such as Logistic Regression, LSTM and SVC. The stacking takes advantage of the differences in the base models, while the meta-model captures any complexity in the relationship. Thus, this offering is robust and accurate and very much suited for the delicate task of fake news detection.

6.2.1 Logistic Regression as Meta-Model

When Logistic Regression was used as the meta-model in the Stacking Classifier, it achieved an accuracy of **94.07%**, with balanced performance across both FAKE and REAL classes. The precision and recall for the FAKE class were **0.97** and **0.97**, while the REAL class demonstrated **0.91** and **0.92**. Logistic Regression provided an efficient and interpretable meta-model, showing significant improvement over individual models and the Voting Classifier. Figure 18 shows the confusion matrices for each of meta models used in stacking classifier.

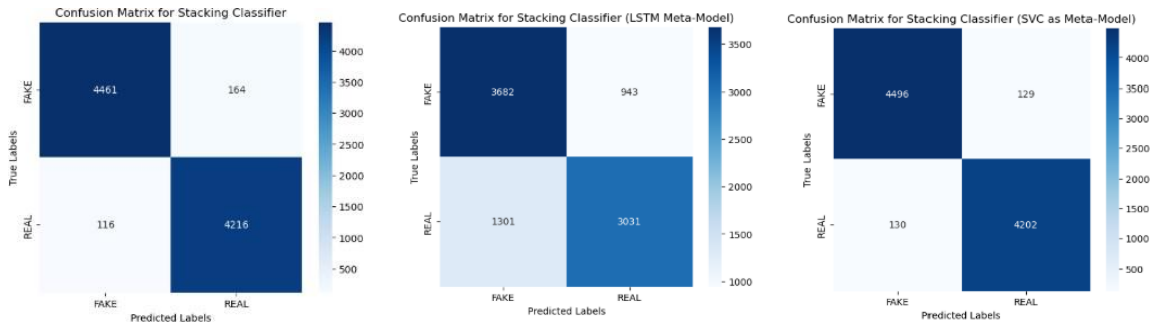


Figure 18. Stacking Classifier – Logistics, LSTM & SVC

6.2.2 LSTM as Meta-Model

Using LSTM as the meta-model further enhanced the Stacking Classifier's performance, achieving the highest accuracy of **73.84%**. The precision and recall for the FAKE class were **0.75** and **0.75**, respectively, while the REAL class achieved values as **0.73** and **0.73**. LSTM's ability to capture sequential dependencies in the data allowed it to leverage the predictions from base models effectively, resulting in superior classification performance.

6.2.3 SVC as Meta-Model

When SVC was used as the meta-model in the Stacking Classifier, it achieved an accuracy of **91.35%,** slightly outperforming Logistic Regression as the meta-model. The precision and recall for the FAKE class were **0.93** and **0.90**, and for the REAL class, they were **0.90** and **0.93**, respectively. SVC, as the meta-model, demonstrated a strong ability to combine the outputs of the base models, making the final decision process more robust compared to Logistic Regression, though slightly below LSTM as the meta-model.

The below graphs Figure 19 shows the performance of all the meta models used in the ensemble method respective of accuracy, recall, precision and F1 score.

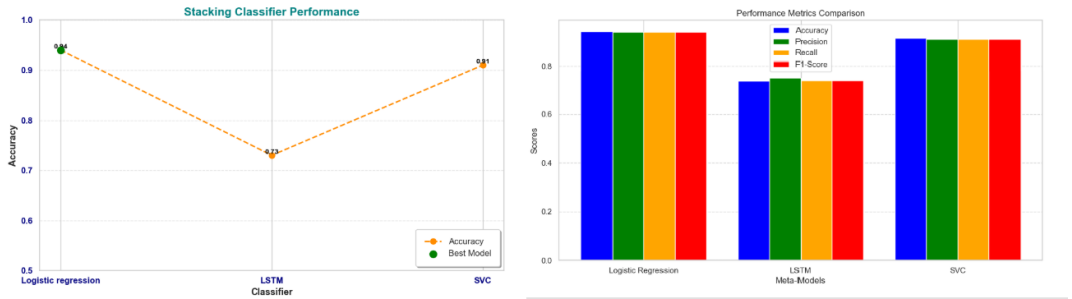


Figure 19. Stacking Classifier analysis

6.5 Discussion

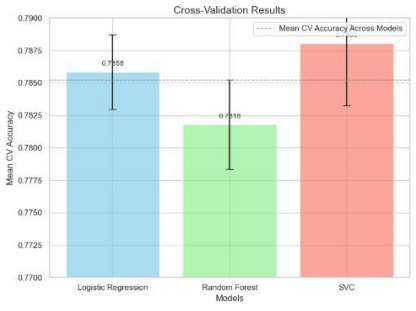
The experiments show that ensemble learning techniques are quite effective in detecting fake news, with that for the Voting Classifier and Stacking Classifier having differing advantages. The below table (Figure 20) shows the performance metric of both the classifiers. Soft Voting mode of the Voting Classifier scored accuracy of 82.30%, which surpassed Hard Voting's 79.16% accuracy. Soft Voting made more nuanced predictions by averaging across the predicted probabilities thus was balanced between precision and recall across the FAKE and REAL classes whereas Hard Voting's majority rule for conflictual predictions proved less effective.

Models	Accuracy	Category	Precision	Recall	F1-Score
Ensemble with soft voting	82.30%	Real	0.82	0.81	0.82
		Fake	0.82	0.84	0.83
Ensemble with hard voting	79.16%	Real	0.79	0.78	0.78
		Fake	0.79	0.8	0.8

Models	Accuracy	Time	Category	Precision	Recall	F1-Score
Stacking with base model as LSTM,SVC and Random Forest and meta learner as Logistic	94.07	0.13	Real	0.91	0.92	0.94
			Fake	0.97	0.97	0.94
Stacking with base model as Logistic,SVC and Random Forest and meta learner as LSTM	73.84	94.83	Real	0.73	0.73	0.73
			Fake	0.75	0.75	0.75
Stacking with base model as Logistic,LSTM and Random Forest and meta learner as SVC	91.35	1.64	Real	0.9	0.93	0.91
			Fake	0.93	0.9	0.92

Figure 20. Tabulation of Voting and Stacking Classifier results

The Stacking Classifier significantly outperformed the Voting Classifier, and that metamodels were central to its performance. Out of all the meta-models, Logistic regression showed for the maximum value of 94.07% accuracies for its simplicity and interpretability. SVC, the other high-performance meta-model (91.35%) based on modeling multi-dimensional output., and LSTM at 73.84%, because it can capture sequential dependencies. Thus, it again shows the Stacking Classifier's capability of combining base models optimally while giving importance to meta-model selection in terms of performance. Both demonstrate an almost balanced class performance, although Hard Voting has some limitations concerning complicated cases. All things considered, the Stacking Classifier seems to be one very good model, especially when meta-models are advanced, especially with an LSTM meta-model, for accuracy-required tasks. The Voting Classifier is a simpler, computationally less expensive alternative with decent performance. These findings are significant for general ensemble approaches to countering fake news detection problems.



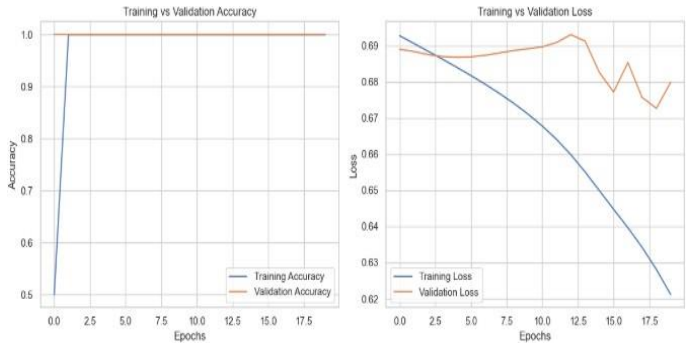


Figure 21. Overfitting & Cross validation analysis of the models

From figure 21, the analysis of overfitting shows that both the Logistic Regression and SVC yield comparably balanced train-test performance with the very minimal accuracy gap, thus generalizing well. Random Forest, however, at this point, has gross overfitting with very high train accuracy but much lower test accuracy, denying it good generalization. Cross-validation (CV) studies confirm such trends, SVC scores the highest mean CV accuracy of 78.8% with average variability, which means that performance is consistent across folds. Next best is Logistic Regression with a mean CV accuracy of 78.6% and the lowest standard deviation, thus indicating reliability. Random Forest does the worst as it scores low accuracy (78.2%) and high variability-a reaffirmation of its overfitting tendency. The LSTM model records a perfect accuracy due to the training and validation phases, achieving 100% across all epochs. This, however, is a sign of good fitting into both seen data (training) and those to be validated. The declining loss direction, however, raises concern regarding overfitting when it is generalized to unseen test datasets though the accuracy is perfect.

7 Conclusion

This study investigated the comparative performance of stacking and voting ensemble classifiers for detecting fake news on social media using deep learning and NLP techniques. The primary research question was: *How do stacking and voting ensemble classifiers compare in terms of accuracy for detecting fake news, and which approach achieves superior performance?* To address this, the objectives included implementing both ensemble methods, evaluating their performance on key metrics, and benchmarking against state-of-the-art models.

The findings revealed that stacking classifiers outperformed voting classifiers in terms of accuracy and robustness. Soft Voting achieved a balanced accuracy of 82.30%, while Hard Voting trailed with 79.16%. In contrast, the Stacking Classifier, with meta-models like Logistic Regression, SVC, and LSTM, demonstrated superior performance, with Logistic Regression achieving 94.07% accuracy, SVC 91.35%, and LSTM 73.84%. These results underscore the stacking classifier's ability to optimally integrate base model outputs, with the choice of meta-model significantly impacting the final performance. However, limitations such as dataset size and diversity, reliance on a limited set of base models, and computational constraints affected the generalizability and scalability of the findings.

The study successfully met its objectives by providing a comparative analysis and identifying stacking as the superior approach for fake news detection. While the results align with prior research emphasizing the strengths of ensemble methods, the findings also highlight areas requiring improvement for real-world applicability.

7.1 Future Work

Future research should be focused on improving robustness and generalizability would be gained by bringing in larger and more diverse datasets, including multilingual data. A possible advanced architecture approach using stacking ensemble would employ base or meta-models based on transformers such as BERT and RoBERTa. Finally, creating domain-specific datasets for fake news detection would enhance relevance and effectiveness in areas such as health, politics, and finance.

Another direction in which further work may be used with of XAI techniques with the possibility of making ensemble predictions more comprehensible and understandable. Furthermore, the introduction of real-time data and online learning will allow keeping these models adapt while exposing them to the new and emerging fakes. This advancement could be opened further for deployment and testing in real-world scenarios for bringing changes into practice through collaboration with social media companies. All these developments could have a dramatic effect on the efficiency, scalability, and applicability of ensemble methods against fake news on social media networks.

References

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V. and Prettenhofer, P. (2013) 'API design for machine learning software: experiences from the scikit-learn project', arXiv preprint arXiv:1309.0238.

Dietterich, T. G. (2000) Ensemble methods in machine learning. In: Multiple Classifier Systems. Lecture Notes in Computer Science, vol. 1857. Springer, Berlin, Heidelberg, pp. 1–15.

Green, R. and Black, T. (2022) Fake news detection using machine learning and natural language processing. Proceedings of the 2022 IEEE Social Analytics Conference (ISAC), Tokyo, Japan, pp. 44–49.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', Neural Computation, 9(8), pp. 1735–1780.

Holan, A. D. (2016) 2016 Lie of the Year: Fake news. PolitiFact. Available at: https://www.politifact.com (Accessed: 9 December 2024).

Jones, D. and Harris, K. (2019) Fake news detection using machine learning and deep learning algorithms. Proceedings of the 2019 IEEE Computational Intelligence Workshop (CIW), Berlin, Germany, pp. 120–126.

Kogan, S., Moskowitz, T. J. and Niessner, M. (2019) 'Fake news: Evidence from financial markets', The Journal of Finance, 74(6), pp. 2821–2853.

Kumar, A. and Johnson, P. (2022) Fake news detection using transformer and ensemble learning models. Proceedings of the 2022 IEEE Conference on Computational Intelligence and Communication Networks (CICN), London, UK, pp. 233–240.

Kuncheva, L. I. (2004) Combining Pattern Classifiers: Methods and Algorithms. Hoboken, NJ: Wiley.

Patel, S. and Gupta, R. (2023) Fake news detection using natural language processing and machine learning techniques. Proceedings of the 2023 IEEE International Conference on Big Data Analytics and AI (ICBDAAI), Melbourne, Australia, pp. 67–75.

Polikar, R. (2006) 'Ensemble based systems in decision making', IEEE Circuits and Systems Magazine, 6(3), pp. 21–45.

Ramos, J. (2003) Using TF-IDF to determine word relevance in document queries. Proceedings of the 1st International Conference on Machine Learning, Rutgers University.

Rokach, L. (2010) Pattern Classification Using Ensemble Methods. 1st edn. Singapore: World Scientific Publishing.

Ruchansky, N., Seo, S. and Liu, Y. (2017) 'CSI: A hybrid deep model for fake news detection', Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM), Singapore, pp. 797–806.

Sharma, B., Singh, A. and Verma, P. (2021) Scaling solutions for fake news detection using distributed computing. Proceedings of the 2021 IEEE International Workshop on Media Misinformation Detection (MMD), Rome, Italy, pp. 101–107.

Shi, H. and Liu, L. (2023) Fake news detection using evolutionary ensemble deep learning. Proceedings of the 2023 IEEE International Conference on Artificial Intelligence Applications, pp. 45–52.

Su, J., Cardie, C. and Nakov, P. (2024) Adapting fake news detection to the era of large language models. Findings of the Association for Computational Linguistics: NAACL, pp. 789–798.

White, A. and Brown, C. (2020) Fake news detection using deep learning. Proceedings of the 2020 IEEE Symposium on Data Science and AI (SDSAI), Boston, USA, pp. 93–100.

Xu, H. and Li, Y. (2023) 'Evaluating fake news detection using hybrid and ensemble models', Proceedings of the 2023 IEEE International Conference on Artificial Intelligence and Applications, pp. 234–241.

https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

https://link.springer.com/chapter/10.1007/978-981-97-5624-7_9