International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Combating Fake News: The Role of Effective Pre-Processing Techniques

Combating Fake News

Pummy Dhiman[1], Amandeep Kaur[2], Yasir Hamid[3] and Joseph Henry Anajemba[3]

[1]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*
[2]*Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India*
[3]*Department of Information Security Engineering Technology, Abu Dhabi Polytechnic, Abu Dhabi 111499, United Arab Emirates*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: This study highlights the importance of detecting fake news and the necessity of effective pre-processing techniques to clean and merge datasets. By utilizing NLP techniques and Python programming, this study successfully merged two Indian datasets and performed pre-processing tasks to improve the quality of the data. The experiment section provides detailed insights into the process of merging and pre-processing, including code snippets and graphs to replicate the results. While this study contributes to the development of effective pre-processing techniques for fake news detection, there is still much work to be done in improving the accuracy of fake news detection. Future research can explore the use of deep learning classifiers and multimodality in fake news detection to enhance our ability to detect fake news and promote a more informed society.

Keywords: Fake News, India, Internet Access, Dataset, Natural Language Processing

1. Introduction

In recent years, the dissemination of fake news has become a pressing issue worldwide, with real and fake content generators alike using persuasive tactics to sway public opinion[1]. One such model used to understand how information can inform attitudinal changes is the Elaboration Likelihood Model (ELM), where persuasion can occur via two routes: central routes requiring high cognitive effort and peripheral routes requiring low cognitive effort[2]. It is common for fake news sites to go down the peripheral route by presenting less information overall and relying on negative emotional cues to influence their readers[3]. This has resulted in a growing concern about the impact of fake news on public opinion and the need for effective strategies to combat its dissemination (Figure 1).

Fake news is not a new phenomenon and has existed for centuries. In ancient Rome, there were instances of false news being spread to discredit political opponents or influence public opinion[4]. However, the term ”fake news” gained prominence in the 2016 US presidential election when false stories were widely circulated on social media to influence the election outcome[5]. Fake news is defined as information that is false or distorted and deliberately spread in order to deceive people[6]. In spite of the fact that false news has existed since the dawn of writing, the term ”fake news” has only recently gained attention due to its global implications and wide-spread concern. Fake news can have serious consequences, such as inciting violence, polarizing communities, fluctuations in stock prices[7], damaging reputations, and eroding trust in institutions and media. It can also undermine democracy by distorting public opinion and manipulating electoral outcomes. Digitalization and easy internet use in India have made it easier for people to access news and information from a variety of sources. As positive as this is, it has also contributed to the propagation of false information.

A. Reason behind Fake News dissemination

There are various reasons why people spread fake news, including:
Sensationalism: People may spread fake news because it is more sensational and attention-grabbing than real news. Sensational news tends to get more clicks, likes, and shares on social media, which can be appealing to people seeking attention or looking to go viral.

Confirmation bias: People may also spread fake news because it confirms their existing beliefs or biases. There is a tendency for people to ignore information that challenges their beliefs[8] and seek out information that confirms them. Fake news that supports their beliefs can be appealing, and

E-mail address: pummy.dhiman@chitkara.edu.in,amandeep@chitkara.edu.in, yasir.hamid@adpoly.ac.ae, sana

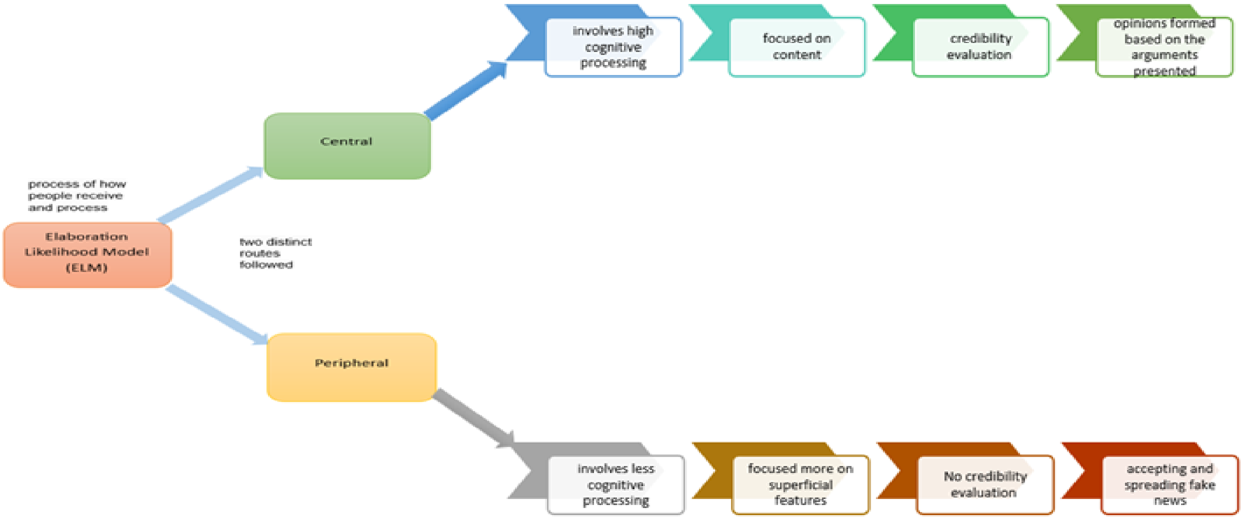

involves high cognitive processing
focused on content
credibility evaluation
opinions formed based on the arguments presented
Central
process of how people receive and process
Elaboration Likelihood Model (ELM)
two distinct routes followed
Peripheral
involves less cognitive processing
focused more on superficial features
No credibility evaluation
accepting and spreading fake news

Figure 1. ELM for Information route

people may share it without verifying its accuracy.

Political or ideological motivations: Some people may spread fake news to advance a political or ideological agenda. This could involve spreading misinformation to discredit political opponents or promote a particular ideology.

Financial gain: Some people may spread fake news to make money. Fake news websites generate revenue through advertising, and the more traffic they get, the more money they can make. The author described how spreading false information on the internet allows people to profit by preying on other people's emotions[9]. As a result, some people may create and spread fake news stories to drive traffic to their websites.

B. Various ways to disseminate fake news

With digital technology on the rise, fake news is easier than ever to spread today. Hence, it is important to understand the variety of ways in which fake news can spread, such as through social media, e-mail, fake news websites, and even traditional media (Figure 2).

Social media: The use of social media is a popular way for people to share news stories and information. Unfortunately, these platforms can also be used to spread fake news quickly and widely[10].

Email: Some people still use email to share news stories and information[11]. Fake news stories can be circulated via email, which can make them seem more credible to people who receive them.

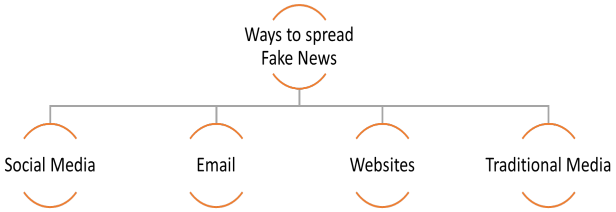


Figure 2. Ways for Fake News Propagation

Fake news websites: Some websites are dedicated to creating and sharing fake news stories. These websites can look professional and legitimate, making it difficult for people to distinguish between real and fake news.

Traditional media: In some cases, traditional media outlets may inadvertently spread fake news. If a fake news story is picked up by a reputable news outlet, it can give it more credibility and make it more likely to be shared by others.

C. The Importance of Detecting Fake News

The spread of fake news has become a growing concern in India[12], with the potential to cause harm to individuals, communities, and the democratic process. In recent years, fake news has been known to incite violence, mislead citizens during elections, and create communal disharmony. As a result, detecting fake news has become crucial to protect democracy, reduce violence, preserve communal harmony, and maintain the credibility of the news media in India. Following is the significance of detecting fake news in India:

Protecting democracy: Fake news can spread misinformation and propaganda that can mislead citizens during elections, potentially leading to biased voting outcomes[13], [14]. By detecting fake news, we can ensure that the public is informed and can make decisions based on accurate information. The Brexit referendum 2016 in the United Kingdom was also influenced by fake news[15]. There were numerous false claims made about the costs and benefits of leaving the European Union, and these claims were often shared widely on social media platforms. Some experts believe that the spread of fake news played a significant role in the outcome of the referendum.

Reducing violence: Fake news has been known to incite violence in India. The spread of fake news during the COVID-19 pandemic led to attacks on healthcare workers and the spreading of false information about the Citizenship Amendment Act (CAA), the Delhi riots of 2019 led to violent protests[16]. Detecting fake news can prevent such incidents from happening in the future.

Preserving communal harmony: India is a diverse country with a range of religions, castes, and cultures. Fake news has the potential to create communal disharmony by spreading misinformation and inciting hatred between different communities. For example, Fake news stories related to cow slaughter and beef consumption often exaggerate the issue or present it in a biased manner, leading to communal tensions and a polarized society. In 2015, a mob in Dadri, Uttar Pradesh, India lynched a Muslim man named Mohammad Akhlaq, after rumours spread that he had slaughtered a cow and stored its meat in his house. The rumours were based on fake messages circulated on WhatsApp and other social media platforms[10]. Detecting fake news can help preserve communal harmony and ensure that everyone feels safe[17].

Maintaining credibility: The news media has a significant impact on public opinion. Fake news erodes the credibility of the news media, making it difficult for people to trust the information they receive. By detecting fake news, we can maintain the credibility of the news media and ensure that people trust the information they receive. Overall, detecting fake news is essential to protect democracy, reduce violence, preserve communal harmony, and maintain the credibility of the news media in India. In this paper, we present our work on using deep learning to detect fake news in the Indian context, by merging two existing datasets and pre-processing them using NLP techniques in Python.

D. Objective and Research Questions

The objective of this paper is to perform Natural Language Processing (NLP) pre-processing on a dataset of Indian news articles to explore the linguistic and semantic features that are unique to this domain. The pre-processing techniques will involve text cleaning, tokenization, part-of-speech tagging, and other standard NLP techniques to prepare the dataset for further analysis. This research aims to identify the linguistic and semantic patterns that can be used to improve the accuracy of fake news detection models for Indian news articles. The research questions are:

- What are the linguistic and semantic features unique to Indian news articles that can be extracted using NLP pre-processing techniques?
- How can the identified linguistic and semantic features be used to improve the accuracy of fake news detection models for Indian news articles?

E. Study Organization

This article is organized as follows: Section 2 follows the introduction and discusses the motivational energies that drive our study in this area. Section 3 provides an overview of the background and work of researchers on this topic. Section 4 discusses the function of natural language processing in detecting bogus news. Section 5 provides information on the dataset used in this study. Section 6 describes the work done in the Python experiment and the findings produced. Section 7 finishes with recommendations for more research into recognizing bogus news.

2. Motivation

Fake news has become a significant challenge for media organizations and consumers worldwide, with the potential to misinform and mislead people on a massive scale. In India, this problem is especially acute, as the country is home to a diverse range of languages and cultures, which can make it more difficult to detect and combat fake news[18]. Although this fake news propagation gains attention worldwide, to the best of our knowledge, there is limited work done only in Indian News, which motivates us to work in this direction. To detect the authenticity of the news, one way is manual checking which frequently involves all of the methods and steps that can be utilized to validate the news. It might involve going to websites that conduct fact-checking. The accuracy of the given news is verified by subject matter experts acting as fact-checkers in manual false news identification. This method does not scale well to the vast amounts of data produced by social media use. Another method to tackle this problem is automatic fake news detection, these methods heavily rely on NLP[19], Data Mining[20], and Artificial Intelligence. Here, datasets play a crucial role in the development and training of fake news detection systems. For a country with diverse languages and news sources like India, collecting an extensive and trustworthy dataset requires considerable time and effort. However, researchers and practitioners have begun developing datasets specific to Indian news that can assist in detecting fake news. The dataset can be used to develop NLP models that can analyze the characteristics of fake news articles in the Indian context, identify patterns, and distinguish them from real news. However, before analyzing the dataset, it is necessary to pre-process it. Pre-processing the dataset involves cleaning and formatting the text data, removing stop words, stemming or lemmatizing the text, and performing other text pre-processing tasks.

This ensures that the data is standardized, and the NLP models can analyze the text effectively.

A huge and balanced dataset plays a vital role in achieving appreciable results in news classification. In this paper, we merged two datasets and apply pre-processing for Indian news for fake news detection datasets. Our approach focuses on specific features of Indian news, including the syntax, vocabulary, and structure of the news articles, which we believe are key to detecting fake news in this context. By developing effective NLP pre-processing techniques for Indian news data, we can take a step towards a more reliable and trustworthy media ecosystem in India, which is essential for maintaining the integrity of democratic institutions and ensuring the safety and well-being of citizens.

3. Background

As was mentioned before, false news is a problem that is getting worse in today’s culture, and it is getting more and harder to tell the difference between the two. To address this problem, machine learning and deep learning models are being developed to detect fake news using text datasets. However, these datasets require pre-processing to remove noise and improve feature extraction, which is crucial for the accuracy of the models. In this literature review, we will discuss the efforts made by researchers to pre-process text datasets for fake news detection.

The authors explore the use of NLP and sentiment analysis[21] in clinical analytics to improve healthcare outcomes. They highlight the potential of NLP and sentiment analysis to transform healthcare delivery and improve patient outcomes. However, there are challenges and limitations to using these technologies, such as data privacy concerns and the need for more reliable data sources.

An evaluation of different NLP techniques and a discussion of various datasets used for fake news detection is provided by the authors. They compare the features of these datasets and analyze the effectiveness of various NLP techniques, including supervised and unsupervised learning, feature engineering, neural networks, and deep learning[22].

Authors in[23] proposes a method for extracting financial sentiment information from online news, message boards, and microblogs using natural language processing. To reduce noise and improve feature extraction, the authors propose a six-step NLP processing approach integrating negation handling. They also introduce a three-class sentiment classification to improve sentiment analysis for financial texts.

The authors[24] describe the general approach to false news identification as well as the taxonomy of feature extraction, which is critical in obtaining maximum accuracy using various machine learning and NLP algorithms. The authors focus on pre-processing techniques for text datasets, such as stop word removal, stemming, removal of number values, punctuation and special symbols, and duplication removal. They report achieving high accuracy on Bi-LSTM models after pre-processing.

The authors used the PHEME dataset[25] to detect fake news, and pre-processing was done using the word2vec model. The data were then standardized and divided into train and test sets before being fed to an ML model for rumor prediction.

The authors utilized conventional techniques[26] for text cleaning and pre-processing, which involved using the NLTK to clean the text data, eliminating stop words, symbols, and special or unidentified characters, transforming all text to lowercase, and then dividing each sentence into tokens. They also eliminated void values from all data columns and accomplished remarkable accuracy on the Pymedia and Politifact datasets.

4. NLP for fake news detection

The majority of data in the world is unstructured, which is because human communication involves language rather than structured data formats[27]. This unstructured data is generated daily through various means such as tweets, emails, SMS, social media posts, and articles, among others. Text, being the most unstructured data type, can be difficult to understand[28]. To bridge this gap, NLP, a branch of AI, helps computers understand human language[22]. Before analysis, raw text data undergoes multiple pre-processing stages to convert it into a suitable format, which can be further used to identify fake news. The specific steps involved may vary depending on the task at hand and the specific tools and techniques being used but generally include the following (Figure 3):

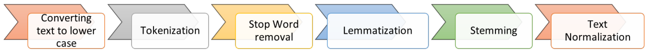


Figure 3. NLP Steps

- **Text Cleaning:**This process involves removing any extraneous or undesirable text data, including HTML elements, punctuation, and special characters[24].

- **Tokenization:**This involves breaking the text into individual words, phrases, or other meaningful units (tokens). Tokenization is a crucial step in NLP because it provides a basis for further analysis and processing of the text data. It basically converts sentence into words. Input: Hello How are you. Output: [“Hello”, “How”, “are”, “you”, “.”]

- **Stop Word Removal:** Stop words are common words that typically do not carry much meaning, such as ”the”, ”a”, ”an”, ”and”, ”in”, ”is”, ”of”, and ”to”. Removing these words can help to reduce noise in the text data and improve the accuracy of subsequent analyses. Input: ”The cat sat on the mat and looked at the dog next door.” In this sentence, the stop words are ”the”, ”on”, ”and”, ”at”, and ”next”. These words are used frequently in English, but they do not carry significant meaning and can be removed from the sentence without changing its overall message. After stop words removal, the sentence might look like this: Output: [”Cat sat mat looked dog door.”]

- **Stemming:** This is the process of reducing words to their root or fundamental form. This is often accomplished by eliminating word-ending suffixes (Figure 4).

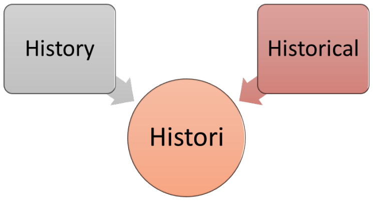


Figure 4. Stemming

These base words may not have any meaning.

- **Lemmatization:** This is similar to stemming, except that instead of eliminating suffixes, it reduces words to their basic or dictionary form (lemma)[16] (Figure 5). Because it incorporates the context of the word and its part of speech, this is typically more accurate than stemming. It is slow process as compare to Stemming.

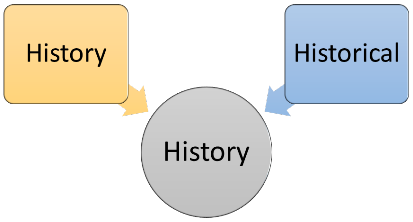


Figure 5. Lemmatization

- **Part-of-Speech Tagging:** POS tagging is the process of assigning a label to every word in a sentence that indicates its grammatical function, such as noun, verb, adjective, adverb, preposition, and so on. POS tagging is important in NLP because it can help to identify the grammatical structure of a sentence, which is useful for tasks like information extraction, text classification, and machine translation. For example, consider the sentence ”The cat sat on the mat.” A POS tagger would label ”The” as a determiner, ”cat” as a noun, ”sat” as a verb, ”on” as a preposition, and so on.

- **Named Entity Recognition:** It involves recognizing and categorizing named entities present in a text corpus, including individuals, locations, institutions, and specific dates. NER has diverse applications in NLP, including information retrieval, question answering, and sentiment analysis. For example, consider the sentence ”Barack Obama was born in Hawaii in 1961.” A NER system would identify ”Barack Obama” as a person, ”Hawaii” as a place, and ”1961” as a date.

- **Vectorization:** Vectorization is a process of converting text data into numerical vectors that machine learning algorithms can understand. Here are some examples of vectorization techniques used in NLP:
  Bag-of-Words (BoW) Vectorization: In this technique, the frequency of words in a document is used to create a vector. Each vector dimension represents the frequency of a particular word in the document.

  Term Frequency-Inverse Document Frequency (TF-IDF) Vectorization: This method is comparable to BoW, however, it considers not only the presence but also the frequency of a word in the entire collection of documents. It assigns a score to each word in a document based on both its frequency within that document and its frequency across the corpus.

  Word Embeddings: This technique represents words as dense vectors of continuous values, rather than as sparse vectors of 0s and 1s. Word embeddings capture semantic relationships between words, such as synonyms and antonyms, by placing similar words closer together in vector space. Popular algorithms for creating word embeddings include Word2Vec and GloVe.

5. Dataset

The dataset is a collection of data samples that are used to train, validate, and test ML, and DL models. In the case of fake news detection, the dataset consists of articles or news pieces that are used for recognition purposes as either real or fake.
The importance of the dataset in fake news detection lies in the fact that machine learning models can only learn from the data they are trained on. Therefore, the quality and diversity of the dataset play a crucial role in the performance of the machine learning model (Figure 6). A well-prepared dataset can improve the accuracy and reliability of the model, while a poorly prepared dataset can result in biased or inaccurate predictions[29].

A comprehensive analysis of the two Indian datasets used in this study is presented in this section. These datasets have been collected from reliable sources and contain a large amount of text data related to news articles. The first dataset contains news articles from various online news portals, while the second dataset consists of news articles collected from social media platforms.

Dataset 1: The process of detecting fake news requires the development or selection of an appropriate dataset. Despite the availability of various public datasets, there has been a lack of a comprehensive dataset solely dedicated to Indian news. To address this gap, the IFND (Indian Fake News Dataset) was created by[30], comprising 37,809 authentic news and 19,059 false news headlines from 2013 to 2021. Additionally, each headline is accompanied by an image. It is worth mentioning that the dataset used in this

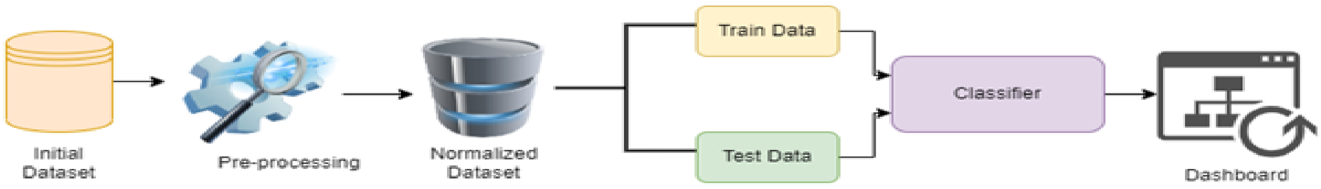


Figure 6. Role of Dataset in fake news detection

study does not cover news posts from social networking sites. However, the significance of these platforms in the propagation of false news cannot be ignored. Due to their vast user base and the speed at which information is shared, social networking sites have become a primary source for the spread of misinformation, propaganda, and fake news. The lack of fact-checking mechanisms, coupled with the ease of creating and sharing content, makes it challenging to differentiate between real and fake news. Therefore, including data from social networking sites in fake news detection research is imperative. Analyzing the characteristics and patterns of fake news on these platforms can provide valuable insights into how false information spreads and help in designing effective strategies to combat it.

Dataset 2: Over the past few years, the increase in the dissemination of false information through social media has raised alarm, since it has the potential to adversely affect both individuals and the entire society. To better understand the effects of fake news, researchers often use datasets that contain only fake news stories from social media. Using such a dataset for impact assessment can provide valuable insights into the potential harm caused by fake news, including its impact on public opinion, political decision-making, and social cohesion. The authors of this dataset[31] gather 4,803 incidents of fake news from six well-known fact-checking websites from June 2016 to December 2019, including 5,031 Twitter links and 866 YouTube video links. The dataset aims to create a benchmark for bogus news incidents in India and contains a collection methodology as well as an impact evaluation for social media. The dataset is accessible for use by the research community.

By analyzing the datasets separately, we will be able to identify any differences or similarities between them, which will be useful for combining them into a single dataset. This analysis will help us understand the characteristics of the datasets and how they can be used effectively for fake news detection using NLP and ML, DL.

Overall, this section will provide a detailed analysis of the datasets, which will serve as the foundation for our fake news detection models.

6. Experimental and Results

The experiment was performed on a Lenovo system with a 12th Gen Intel(R) Core(TM) i5-1235U processor, clocked at 1.30 GHz, running the Windows 11 operating system, and implemented using Google Colab.

After importing basic libraries or modules, Dataset1 is loaded which is in csv format, with Latin-1 encoding.

```
df=pd.read_csv('Dataset1.csv',encoding='latin-1')
```

This dataset contains text news from 15 web sources, which is visualized in below Figure 7.

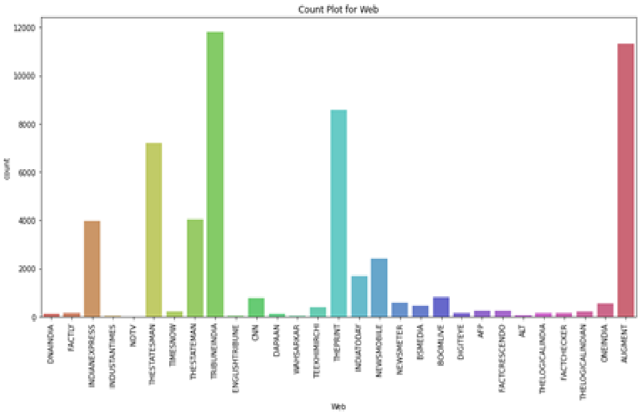


Figure 7. Count Plot for web sources

This dataset contains news of the following unique categories, as shown below(Figure 8).

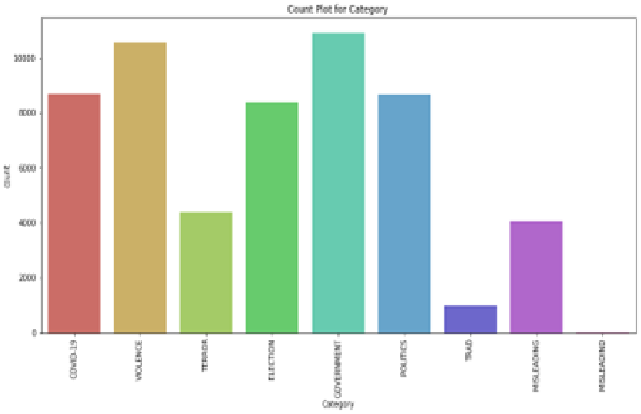


Figure 8. Count Plot for News Categories

A total 66.7% of news falls under the True category and the remaining fall under Fake(Figure 9).

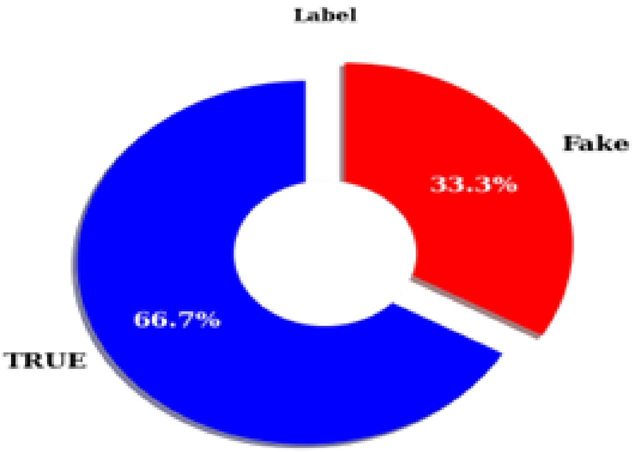


Figure 9. News Count of Dataset 1

Dataset 2 is having all fake news incidents from social network platforms.
After the removal of duplicate and null values, we got 4650 unique entries from various fact-checking websites, as shown in Figure 10 below.

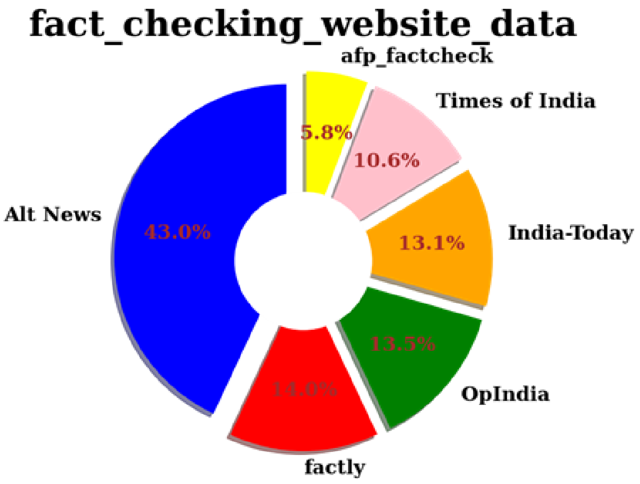


Figure 10. News Count of Dataset 1

After this, the next step is to concatenate these two datasets namely Dataset1 and Dataset2.

```
df_new_1 = [df_new,data_new]
result = pd.concat(df_new_1)
result.reset_index(inplace = True)
result
result = result.drop('index', axis = 1)
```

Table I and Figure 11 display the statistics of the merged dataset.

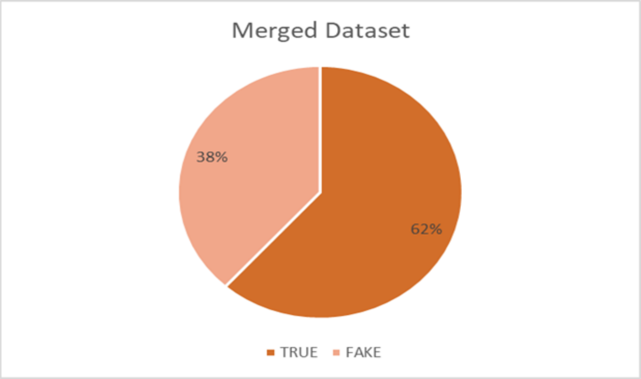


Figure 11. Merged Dataset proportion

Following Figure 12 shows the 5 top and bottom news statements.

Now pre-processing will take place as follows. **Data Cleaning:** Data cleaning is an important step in NLP pre-processing[32], as it helps to standardize the input text and remove any noise or irrelevant information that may affect the accuracy of the analysis. Here we used various methods as described below.

- text.lower(): This operation converts all the characters in the input text to lowercase. This helps in ensuring that text data is consistent, as sometimes the same word can appear in different cases, such as ”Hello” and ”hello”. By converting all the text to lowercase, we eliminate this variability, making it easier to compare and analyze the data.

- text.strip(): This operation removes any leading and trailing whitespace from the input text. Whitespace can include spaces, tabs, and line breaks, and removing them helps to ensure that the text is uniform and consistent.
  Now we proceed in removing the punctuations and special characters. `remove_punctuation` function is the same as before and is defined to take a string of text as input and remove all punctuation.
  This can help to simplify the text data and make it easier to analyze.
  After this, tokenization is performed, as shown below:

  ```
  import re
  def tokenization(text):
  tokens = re.split('W+',text)
  return tokens
  df_new['Statement']= df_new['Statement'].
  apply(lambda x:tokenization(x))
  ```

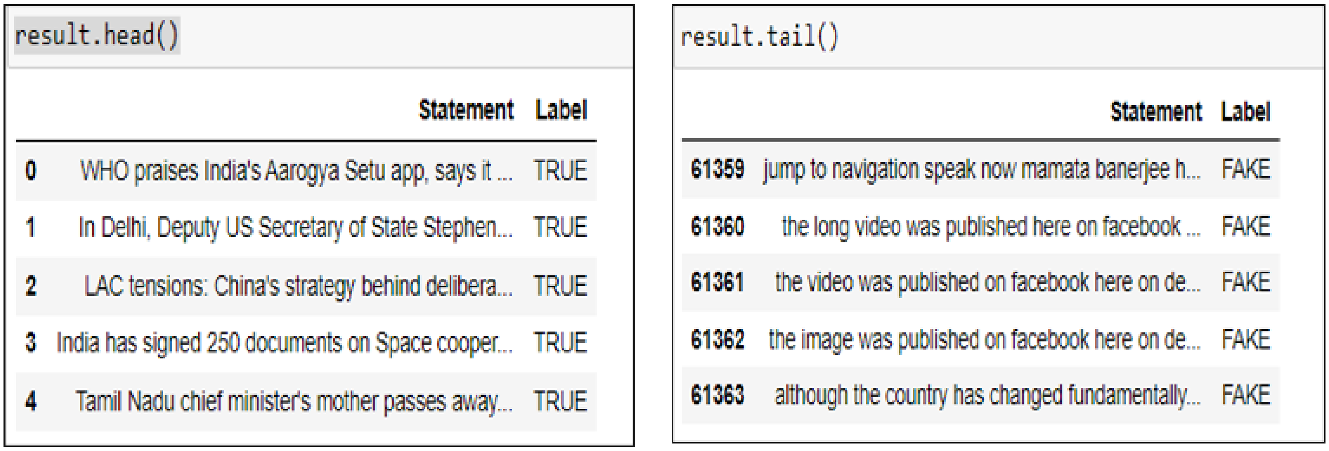
  The purpose of the function is to split the input text into individual tokens (words) and return them as a list.

- The re.split() method is used to split the input text into tokens. The regular expression pattern 'W+' is

TABLE I. Dataset Statistic

Dataset	Label	Statistics	Total

Merged Dataset	TRUE
	FALSE
		37800(62%)
		23564(38%)
			61364



```
result.head()

                                          Statement  Label
0   WHO praises India's Aarogya Setu app, says it ...  TRUE
1   In Delhi, Deputy US Secretary of State Stephen...  TRUE
2   LAC tensions: China's strategy behind delibera...  TRUE
3   India has signed 250 documents on Space cooper...  TRUE
4   Tamil Nadu chief minister's mother passes away...  TRUE
```

Top 5 rows

```
result.tail()

                                               Statement  Label
61359  jump to navigation speak now mamata banerjee h...  FAKE
61360     the long video was published here on facebook ...  FAKE
61361    the video was published on facebook here on de...  FAKE
61362    the image was published on facebook here on de...  FAKE
61363   although the country has changed fundamentally...  FAKE
```

Bottom 5 rows

Figure 12. Preview of Merged Dataset

used as the delimiter to split the text into tokens. This pattern matches any non-alphanumeric character, which means that it will split the text at any point where a non-alphanumeric character is found. This allows the function to split the text into individual words while ignoring any punctuation or other non-alphanumeric characters. The function returns the list of tokens as the output.

- In the next step, Natural Language Toolkit (nltk) library is used to download a list of English stop words, which are common words that are often excluded from text analysis as they are not informative.
- In the next step, the 'stemming' function, takes a text string as input and applies stemming to each word in the text using the 'stem' method of the PorterStemmer instance. The result is a stemmed version of the input text, where each word has been reduced to its root form.
- Porter stemmer is imported from the 'nltk.stem.porter' module, which is used to perform stemming on the text. The practice of reducing words to their base or root form, such as transforming 'running' to 'run,' is known as stemming. The code defines the 'stemming' function, which takes a text input and applies stemming to each word in the text. The function returns the modified text with stemmed words.
- Lemmatization is the process of reducing a word to its lemma. For instance, the words "cats" and "ran" would be shortened to "cat" and "run," respectively. This can be useful for text analysis and natural language processing tasks that require a common base form for words that have similar meanings. The function uses the wordnet_lemmatizer from the Natural Language Toolkit (nltk) library to lemmatize each word in the input text.
- Now the removal of text in square brackets and words containing numbers is removed.
- To remove URLs from a given text input (vTEXT) a Python function that uses the regular expression library re is used. It searches for patterns that match a typical URL structure, starting with "http://" or "https://" and ending with various possible characters such as letters, numbers, slashes, question marks, equals signs, and percentage signs.
  Next, remove any sequence of digits from the input text using regular expressions, and return the cleaned text.
- Removal of Emojis Emojis are non-standard textual symbols that can convey a wide range of emotions and meanings, and they can often be ambiguous or context-dependent. Removing emojis can help simplify the data and make it easier to analyze or classify using machine learning models. Moreover, emojis can add noise to text data, especially when working with datasets that contain user-generated content, such as social media posts, comments, and

reviews. In these cases, emojis can make it more challenging to extract meaningful insights from the text and can lead to bias or inaccuracies in the analysis.

- **Word Cloud** Word clouds are a popular and visually appealing way to represent text data in NLP. They are a type of data visualization that allows us to quickly and easily identify the most common words and phrases in a corpus of text. Figure 13 depicts the word cloud generated from the combined, pre-processed dataset.

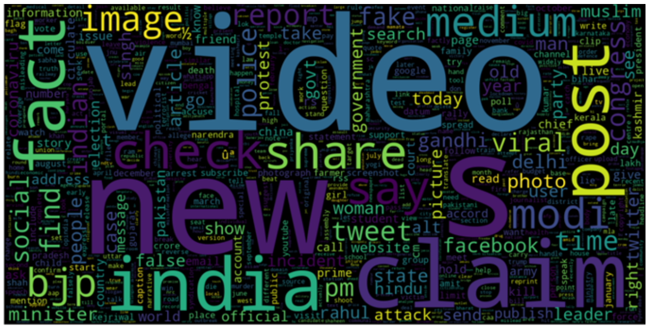


Figure 13. Visualizing Top Words in Combined Text Data

7. Conclusion and Future Prospects

In conclusion, this study has demonstrated the importance of detecting fake news and the need for effective preprocessing techniques to merge and clean datasets for fake news detection. Through the use of NLP techniques and Python programming, this study has successfully merged two Indian datasets and performed pre-processing tasks such as text normalization, stop word removal, and stemming to improve the quality of the data. The experiment section has provided detailed insights into the process of merging and pre-processing, including code snippets and graphs to aid in replicating the results. Overall, this study has contributed to the development of effective preprocessing techniques for fake news detection, which can be further refined and applied in future studies.

Despite the advancements made by this study in preprocessing techniques for detecting fake news, there is still room for improvement in the accuracy of identifying false information. One promising direction for future research is the utilization of deep learning (DL) classifiers to enhance the performance of fake news detection models. DL classifiers have exhibited great potential in several natural language processing tasks and can potentially augment the effectiveness of fake news detection models. Additionally, exploring the use of multimodality in detecting fake news is another area for future investigation. Multimodality refers to incorporating multiple modalities like text, image, and video to develop a more complete understanding of the news article. By integrating various modalities, fake news detection models can potentially identify more subtle nuances that may not be apparent in a single modality. Overall, the incorporation of DL classifiers and multimodality in detecting fake news offers exciting prospects for future research in this domain. By continually refining and improving these techniques, we can increase our ability to detect fake news, thereby promoting a more knowledgeable society.

References

[1] T. Janevatchararuk, “Fake news situations: Comparison, factors analysis, and model,” *Journal of Social Sciences and Humanities*, vol. 45, no. 1, pp. 119–140, 2019.

[2] B. Osatuyi and J. Hughes, “A tale of two internet news platforms-real vs. fake: An elaboration likelihood model perspective,” 2018.

[3] C.-Y. Chen, M. Kearney, and S.-L. Chang, “Belief in or identification of false news according to the elaboration likelihood model.” *International Journal of Communication (19328036)*, vol. 15, 2021.

[4] J. Posetti and A. Matthews, “A short guide to the history of ‘fake news’ and disinformation,” *International Center for Journalists*, vol. 7, no. 2018, pp. 2018–07, 2018.

[5] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[6] P. Dhiman, A. Kaur, and A. Bonkra, “Fake information detection using deep learning methods: A survey,” in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*. IEEE, 2023, pp. 858–863.

[7] K. Khan and M. O. Koti, “Impact of rumors and fake news on stock market,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 14, pp. 6094–6097, 2021.

[8] A. Kumar Sharma, S. Chaurasia, and D. Kumar Srivastava, “Deep sentiment approaches for rigorous analysis of social media content & its investigation,” *International Journal Of Computing and Digital System*, pp. 171–185, 2021.

[9] G. Kaur, P. ., A. Kaur, and M. Khurana, “A review of opinion mining techniques.” *ECS Transactions*, 2022.

[10] S. Banaji, R. Bhat, A. Agarwal, N. Passanha, and M. Sadhana Pravin, “Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india,” 2019.

[11] N. Dhamani, P. Azunre, J. L. Gleason, C. Corcoran, G. Honke, S. Kramer, and J. Morgan, “Using deep networks and transfer learning to address disinformation,” *arXiv preprint arXiv:1905.10412*, 2019.

[12] S. Shafi and M. Ravikumar, “Dynamics of fake news dissemination: a case study in the indian context,” *Media Watch*, vol. 9, no. 1, pp. 131–140, 2018.

[13] G. O. I. M. O. LAW and J. L. DEPARTMENT, “FAKE NEWS DURING GENERAL ELECTIONS,” p. 5, 2019. [Online]. Available: http://164.100.24.220/loksabhaquestions/annex/171/AU5033.pdf

[14] J. C. Reis, P. Melo, K. Garimella, J. M. Almeida, D. Eckles, and F. Benevenuto, “A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections,” in *Proceedings*

of the international AAAI conference on web and social media, vol. 14, 2020, pp. 903–908.

[15] J. Mair, T. Clark, R. Snoddy, and R. Tait, *Brexit, Trump and the media*. Abramis academic publishing, 2017.

[16] P. Dhiman, A. Kaur, C. Iwendi, and S. K. Mohan, "A scientometric analysis of deep learning approaches for detecting fake news," *Electronics*, vol. 12, no. 4, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/4/948

[17] "2013 muzaffarnagar riots - wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/2013_Muzaffarnagar_riots

[18] S. Singhal, R. Kaushal, R. R. Shah, and P. Kumaraguru, "Fake news in india: scale, diversity, solution, and opportunities," *Communications of the ACM*, vol. 65, no. 11, pp. 80–81, 2022.

[19] S. K. Mohapatra, P. K. Sarangi, P. K. Sarangi, P. Sahu, and B. K. Sahoo, "Text classification using nlp based machine learning approach," *AIP Conference Proceedings*, vol. 2463, no. 1, p. 020006, 2022. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/5.0080301

[20] C. Kaushal, M. A. R. Refat, M. A. Amin, and M. K. Islam, "Comparative micro blogging news analysis on the covid-19 pandemic scenario," in *Proceedings of International Conference on Data Science and Applications: ICDSA 2021, Volume 2*. Springer, 2022, pp. 377–391.

[21] G. Kaur, A. Kaur, M. Khurana *et al.*, "A stem to stern sentiment analysis emotion detection," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2022, pp. 1–5.

[22] Y.-C. Ahn and C.-S. Jeong, "Natural language contents evaluation system for detecting fake news using deep learning," in *2019 16th international joint conference on computer science and software engineering (JCSSE)*. IEEE, 2019, pp. 289–292.

[23] F. Sun, A. Belatreche, S. Coleman, T. M. McGinnity, and Y. Li, "Pre-processing online financial text for sentiment classification: A natural language processing approach," in *2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*. IEEE, 2014, pp. 122–129.

[24] S. Rastogi and D. Bansal, "A review on fake news detection 3t's: typology, time of detection, taxonomies," *International Journal of Information Security*, pp. 1–36, 2022.

[25] H. M. Jabir, M. A. Naser, and S. O. Al-mamory, "Rumor detection on twitter using features extraction method," in *2020 1st. Information Technology To Enhance e-learning and Other Application (IT-ELA*. IEEE, 2020, pp. 115–120.

[26] W. Shishah, "Fake news detection using bert model with joint learning," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, pp. 9115–9127, 2021.

[27] M. K. Islam, M. A. Amin, M. R. Islam, M. N. I. Mahbub, M. I. H. Showrov, and C. Kaushal, "Spam-detection with comparative analysis and spamming words extractions," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1–9.

[28] S. A. Salloum, M. Al-Emran, and K. Shaalan, "Mining social media text: extracting knowledge from facebook," *International Journal of Computing and Digital Systems*, vol. 6, no. 02, pp. 73–81, 2017.

[29] S. S. Ashik, A. R. Apu, N. J. Marjana, M. S. Islam, and M. A. Hassan, "M82b at checkthat! 2021: Multiclass fake news detection using bilstm." in *CLEF (Working Notes)*, 2021, pp. 435–445.

[30] D. K. Sharma and S. Garg, "Ifnd: a benchmark dataset for fake news detection," *Complex & Intelligent Systems*, pp. 1–21, 2021.

[31] A. Dhawan, M. Bhalla, D. Arora, R. Kaushal, and P. Kumaraguru, "Fakenewsindia: A benchmark dataset of fake news incidents in india, collection methodology and impact assessment in social media," *Computer Communications*, vol. 185, pp. 130–141, 2022.

[32] A. Kathuria, A. Gupta, and R. Singla, "A review of tools and techniques for preprocessing of textual data," *Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 1*, pp. 407–422, 2021.

Pummy Dhiman is a PhD candidate in Computer Science and Engineering at Chitkara University in Punjab. She received her MTech degree from SDDGPI, Haryana. Her research interests include Machine Learning, Deep Learning, Natural Language Processing, and Data Mining.

Amandeep Kaur is presently working as a professor at Chitkara University, Punjab, India. She attained her doctorate degree from I.K. Gujral Panjab Technical University, Jalandhar, India. She has 23 years of experience. She has filed and published more than 80 patents. Her areas of research interest mainly include Medical Informatics, Machine Learning, Deep Learning, IoT, and Cloud Computing.

Yasir Hamid is an Assistant Professor at Abu Dhabi Polytechnic, where he teaches in the Department of Information Security Engineering Technology. He completed his Ph.D. in Computer Science and Engineering from Pondicherry University in 2019. He is an active member of many scientific societies and serves as an editorial board member of several journals. His research interests primarily focus on Machine Learning, Deep Learning, and Big Data Analysis.

Joseph Henry Anajemba (Member, IEEE) received the Ph.D. degree in Information and Communication Engineering from Hohai University, China, in 2021. He is currently an Assistant Professor with the Department of Information Security Engineering Technology, Abu Dhabi Polytechnic, UAE. He has authored or coauthored many top scientific researches. His research interests include the Internet of Things, Physical Layer Security (PHY), Cyber Security and Artificial Intelligence. He is an Associate Fellow of the Higher Education Academy (AFHEA), U.K. He has served as a reviewer and guest editor for several journals.