FAKE NEWS PREDICTION ON SOCIAL MEDIA WEBSITES

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by

PAWAN KUMAR VERMA (41800316)

Under the supervision of

DR. PRATEEK AGRAWAL

LOVELY PROFESSIONAL UNIVERSITY
PUNJAB

2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

LOVELY PROFESSIONAL UNIVERSITY

Punjab, India-144411

CANDIDATE'S DECLARATION

I, Pawan Kumar Verma, 41800316 student of Ph.D. Computer Science and Engineering, hereby declare that the dissertation titled "Fake news prediction on social media websites" which is submitted by me to the Department of Computer Science and Engineering, Lovely Professional University, Punjab in partial fulfilment of the requirement for the award of degree of Doctor of Philosophy, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Lovely Professional University, Punjab　　　　Pawan Kumar Verma

Date: 03.12.2021

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

LOVELY PROFESSIONAL UNIVERSITY

Punjab, India-144411

CERTIFICATE

I hereby certify that the dissertation titled "Fake news prediction on social media websites" which is submitted by Pawan Kumar Verma, 41800316, Department of Computer Science and Engineering, Lovely Professional University, Punjab in partial fulfilment of the requirement for the award of the degree of Doctor of Philosophy, is a record of the dissertation work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Lovely Professional University, Punjab   Dr. Prateek Agrawal

Date: 03.12.2021   **SUPERVISOR**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

LOVELY PROFESSIONAL UNIVERSITY

Punjab, India-144411

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Dr. Prateek Agrawal for his continuous guidance and mentorship that he provided us during the dissertation. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this dissertation as well as its industrial relevance. He was always ready to help me and clear my doubts regarding any hurdles in this dissertation. Without his constant support and motivation, this dissertation would not have been successful.

Place: Lovely Professional University, Punjab　　Pawan Kumar Verma

Date: 03.12.2021

Abstract

High usage of social media for spreading real-time news is seen today. Social media users are not confined to a particular age group or gender, they are spread over various levels of people. This is because it is easy to communicate, propagates rapidly, easy to access and has low cost for spreading the news. But in the recent times, it is serving as a double edge sword in the society as it spreads fake news along with real news. The users of social media are using it to spread hoaxes, fake news, and malicious data for the purpose of entertainment, politics, and business. To address these issues, this dissertation proposed four methods to employ them for evaluating the authentication of the news, thus controlling its spreading. This dissertation also proposed a novel fake news dataset for the study of fake news detection.

The major part of any machine learning model is dataset on which that model will be trained and tested. Therefore this dissertation initially highlights the characteristics of few famous publicly dataset and its limitations. Due to these limitations this report proposed a new fake news dataset named “WELFake dataset”. This dataset is used in proposed model for their evaluation purpose.

Linguistic based fake news detection is one of the famous technique for early detection. This report proposed “WELFake model” that incorporate the properties of both writing pattern and frequency based word embedding technique for the classification of news. This model combines four phases; In the first phase, pre-processing of the dataset is done, second phase finalizes the minimum set of linguistic features for classification, third phase selectes the best word embedding feature and merges with selected linguistic feature set and finally voting classification is applied for classification of news. This model gives an accuracy of 96.73%

on WELFake dataset.

Another model, named “MCred model”, is based on the semantics of text content. This model is a fusion of Bidirectional Encoder Representations from Transformers (BERT) and Convolutional Neural Network (CNN), which will make use of global and local text semantics respectively. Experimental results have shown that there is 1.10% of more accuracy than state-of-the-art model when MCred method is applied for detecting the news.

Remaining models named “UCred” and “PropFND” are based on the user profile information. Talking about UCred model, it will help in classifying fake profile using the results of Random Forest (RF), Bidirectional Long-Short Term Memory (Bi-LSTM), and Robustly Optimized BERT (RoBERT). This model gives an accuracy of 98.96% on Online Social Network dataset. Another model named PropFND will help to classify the news basing on both user profile features and propagation pattern of message. This model classifies the news with 93.81% accuracy and finally concluded that the real news will propagate for more times as compared to fake news.

At the end of this report all the proposed models are merged together in a way that three models are used for news classification and the prediction of PropFND model is treated as verification of final result.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Introduction

After mid-1990s there was a drastically development in World Wide Web and online social media comes into picture for the people conversation. According to the research done by Jiang *et al.* [1], majority of people all over the world are using social media websites for exchange of their views and these peoples are from different age, gender and community. Twitter and Facebook are two tech giants in the field of social media, user uses these platforms for sharing real-time news. Social media platforms are considered as major platforms these days as they share information at rapid rate, come at low cost and are easy to use. More than 33% of the world's population are actively using social media websites and other messaging applications [2]. These digital platforms changed the existing information exchange scenario used by the users for their interactions and communication. Both Facebook and Twitter create Terabytes of data consistently [3, 4]. The major reasons behind change the traditional news media to digital media are: (i) digital media takes less time for dissemination and less expensive because it can be easily produced and consumed by user; and (ii) people can easily share, comment and discuss the news with colleagues and other users available on digital platform [5]. Because of these benefits the users are shifting from traditional media to digital media. According to research agencies approximate 62% and 66% of users from USA and Brazil respectively are using digital media for information exchange [6, 7]. Despite of various advantages of digital media they turned into a spot for mission of deception which are frequently planned to deceive individuals. Traditional news media like television and newspapers are using one-to-many scenario but in case of social media they are using many-to-many approach. Therefore, social media works as double-edged sword; at one side it propagates news much faster and on the other side it is exceptionally exposed for propagation of fake news. These platforms serve as breeding platform of misleading information and fake news. The spread of fake news has resulted a potential threat to online entities like shopping, networking etc. Identification of the fake news by a naked eye is not easy as the spreaders use lot of technologies to establish similarity with the real news.

To understand and detect the fake news, first an in-depth analysis must be conducted to know what the fake news, why and by whom fake news is created. Doing a survey will help in developing the framework to detect the fake news. Extensive research on developing an automatic framework for detecting fake news online is going on. Though, there are certain challenges like identifying fake news from millions of messages, trials are done where linguistics-based features were extracted from the news to reveal the distribution patterns of fake news is performed. Also, the credibility of the news spreader and propagation pattern is important, and online data is time-sensitive so real-time detection system is important. These are some examples which show that the research was confined to certain areas but it should be designed to detect, explore as well as interpret the fake news; this is because the ongoing studies are limited so more research on all dimensions must be conducted.

1.2 What is fake news?

"False stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke."

— Cambridge University

In simple words fake news can be described as misleading or false information.

Its main motive is to damage the reputation of entity or a person. The spreaders of fake news look into making money by advertising such information. The term fake news covers a broad definition which will also include unconscious and unintentional mechanisms by the individuals for applying on any news which is unfavorable to that individual perspective.

Fake news includes videos, images, or text which are shared for spreading incorrect information that has wrong facts. Though the news may seen authentic in the first, it will attract shocking opinions and attracts attentions from the readers. These are created by groups or individuals who have their own interests which may be motivated by the economic, political, or personal agendas.

The spread of fake news is not new today, and have been since the times of print media. But these days they have become more often and widely spread because of the digital exchange. The visibility of fake news is high in less time as sharing via social media is easy and less time consuming.

"Fake News" was introduced by mass media in last centuries back but it was rapidly sprung in the 2016 US Presidential Election [8]. After 2016 the term "fake news" become a jargon therefore several researchers defined this term but none is globally accepted. Fake news can be considered as poor-quality news that are intentionally created and propagated by individual/bots for the tattle or political benefits. Allcott *et al.* [9] describes that intentionally and verifiable false news is considered as fake news and it misleads the other readers. People disseminate fake news on social media for political or financial benefits and destroy a community or individual. In last three months of 2016 US Presidential Election large number of fake news were generated and approximately 37 million times it was shared on Facebook and Twitter for the favor of either of two participants [9]. Along with this popular incident, let us see some other fake news incident. When earthquake took place in Chile in 2010, rumors increased and spread which increased the panic in the public [10]. To consider a recent example, Facebook story claimed that CNN does not fall into the category of top ten watched networks [11]. An interesting fake news spread that alien existed on the moon according to the Physicist Stephen

Hawking [12]. All these were spread to mislead social media users and create panic among them.

Zaryan highlighted worrying viewpoint that fake and unverified news generated by unauthentic sources attracts greater number of crowds as compared to authentic news [13]. Researches done on this subject comes to an end that fake news have more effect as compared to real news and it is spread more faster than authentic news [14]. Many researchers used different terminology for fake news like satire, dis-/mis-information, spam, rumor and satire [15]. But in this dissertation, these terms are used interchangeably.

On social media websites people consumes and produces large amount of information without authenticating its genuineness. Due to this, user unintentionally involved in the proliferation of fake news over internet and that moves forward to wrong decision. To stop this propagation of fake news several agencies developed online websites for the checking of genuineness of news like Hoaxposed [16], PolitiFact [17], The Washington Post Fact Checker [18], FactCheck [19], Snopes [20], TruthOrFiction [21], FullFact [22], Vishvas News [23], Factly Media & Research [24]; but these are unable to immediately response to any fake news [25].

Due to the massive growth of fake news, it is observed that in 2022 the amounts of fake news will more than real news over the internet [26]. This is highly a concern as it has to be curtailed by detection before it causes more havoc among the people reading them.

Though fake news became popular during the US elections, but when you look at the graph, as shown in figure 1.1, it explains that fake news is a very big problem that is real [27]. When Buzzfeed conducted the research making use of Buzzsumo shows right from August to the election day it has showed that the fake news stories were spread higher on Facebook when compared to other stories. Another famous false satire is that Pope Francis has endorsed Donald Trump for his presidency which has received one million engagements which include comments, reactions, and shares. This fake news impact shows the significance of fake news detection. The above examples shows that how badly fake news is affecting people globally

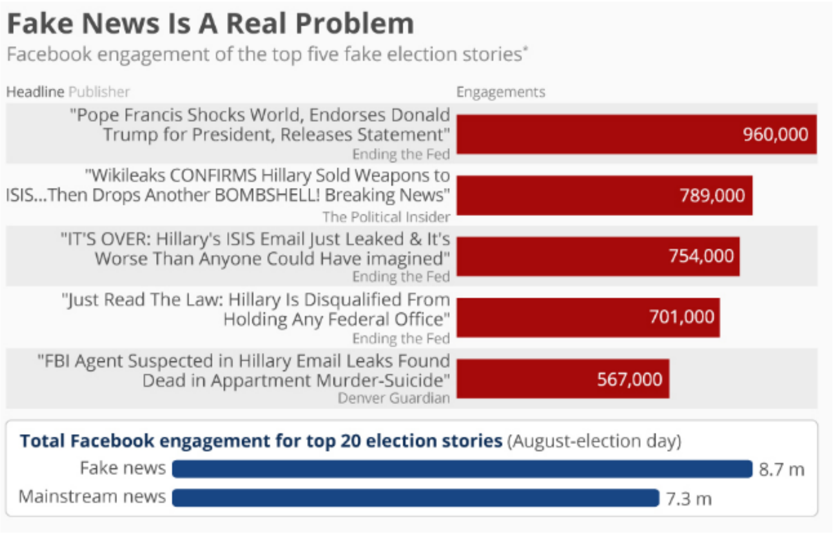

Figure 1.1: Statistics on fake news during US elections in 2016

and those impact shows the significance of fake news detection.

1.3 Significance of fake news detection

With the increase in the social media development, parallel increase of fake news is seen. This news is distracting, obtrusive as well as annoying to the readers. Extensive dissemination of fake news may lead dangerous impact on society as well as individual in many ways; (i) It can destroy the authenticity equilibrium of the news environment, (ii) It intentionally convinces readers to acknowledge biased or false information and (iii) Influences the manner by which individuals interpret and respond to real news. (iv) This news will dominate the decisions, interests, and opinions of the public. (v) It will influence the way how people will interact with the real news. (vi) It will destroy the beliefs and faith of people on their experts, authorities, and the government. Along with these, understanding the below characteristics of fake news will help in understanding why fake news detection is important:

i. *Fake news volume:* As there are no verification procedures, it is easy to write fake news on Internet. You can come across many web pages whose main purpose is to publish fake stories and news. These websites will resemble legitimate news websites and are created to spread false propaganda, misleading information and hoaxes. This is mainly done for political and financial gain. All this happens without the awareness of the website users.

ii. *Fake news variety:* Some of the definitions of fake news are politician's false statements, conspiracy theories, fake advertisements, misinformation, fake reviews, satire news, and rumors. Variety of information affects variety of people covering every aspect of people's lives.

iii. *Fake news velocity:* Most of the fake news are short-lived. To see an example, as discussed above fake news propagated during the U.S. elections in 2016 no longer exist today as they are removed after the campaign.

Today's fake news mainly concentrates on hot affairs and current events to grasp attention of the users. This real-time nature makes it hard for users to identify them as fake, so detecting them is highly important. Also, most of the online consumers like 88% rely more on reviews online and only 72% believe in positive reviews of business, making this an area of concern where fake news can be propagated [35].

As more attention is paid to fake news in recent years, more fake news generators are generating nothing but a transient flash in order to avoid detection by the detection systems. The fake news is dominating the internet everyday bringing fateful consequences to the financial matters, medical, politics in the society as well as to all who live in the cyber environment. To get away with detection, recent fake news generators are spreading information in a flash. So immediate action is required to address all these issues mentioned above to escape their negative impacts. Also, these impacts make the formulation of framework for identification of fake news is a crucial one.

Therefore, for reducing the negative impact of fake news there is a great need of automated system for identification. Various researchers are developing automated detection system but still accuracy of these systems is challenging task because of dynamic, complex and diversity nature of social media [6]. Therefore, designing a reliable automated system for classification of real and fake news for online social media is significant.

1.4 Motivation

Digital platforms significantly changed the method of user interactions and modified the traditional information ecosystems. In the recent years reader uses social media websites and messaging applications for exchange their views but this scenario creates unforeseen problem i.e., fake news dissemination. Despite of various advantages of digital platform now days it become a place where people spread fake news for their financial, health and political benefits. Some aftereffects of fake news are as follows:

1.4.1 Financial context:

Along with all other divisions, even finance had seen impacts from fake news. Major impacts are seen on the stock market where effects were only for some time. The stock market recovered soon from the after effects but it can get serious any time. For example, when a fake tweet mentioning the injuries of Barack Obama due to explosion were spread in 2013, around $130 billion stock value was wiped out [28]. This spread of fake news was done by the hackers. Though the stock market recovered sooner after the incident, this shows how manipulative information can affect the trading algorithms by impacting the trading calls taken by different investors.

Along with the stock market, finances of multiple departments are affected because of the fake news. The graph, shown in figure 1.2 [29], was constructed on the basis of the research by the CHEQ AI Technologies Ltd. When CHEQ company along with the University of Baltimore, has opted to conduct a research on the spread of fake news, it was detected that it is affecting $78 billion in the global economy annually. Their report along with the economic costs also estimates the stock market value loss which is up to $39 billion annually. When World

Economic Forum (WEF) has analyzed the spread of fake news, it has identified fake news spread as a global risk. According to CHEQ CEO fake news are found everywhere. He defines this as the sharing as well as creation of false information for misleading the audience. For analyzing the economic data, CHEQ has worked with the economic department of Baltimore University. The department has helped to analyze and show the financial costs that fake news has on different sectors in the economy. These effects on the economic sectors can be understood by reading the graph given here.

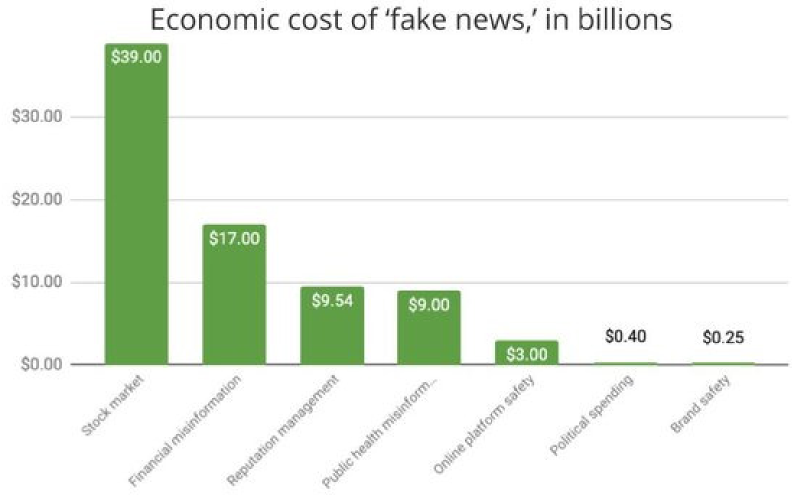

Figure 1.2: Economic analysis conducted by the University of Baltimore

1.4.2 Medical context:

Huge amount of fake information related to medical on digital platform leads irreparable damage [30]. For instance, a patient suffered from cancer disease taken the treatment as per online ad and finally died. Furthermore, many people spread unauthentic news related COVID-19 which created panic situation in the society [31]. Several fact checking agencies analyzed that more than 3500 fake claims related to COVID-19 are scattered over the internet in less than 2 months [32]. This leads the death of at least 800 people around the world in the starting three months of 2020.

Figure 1.3 is one example of fake news impact during COVID-19 [33]. The

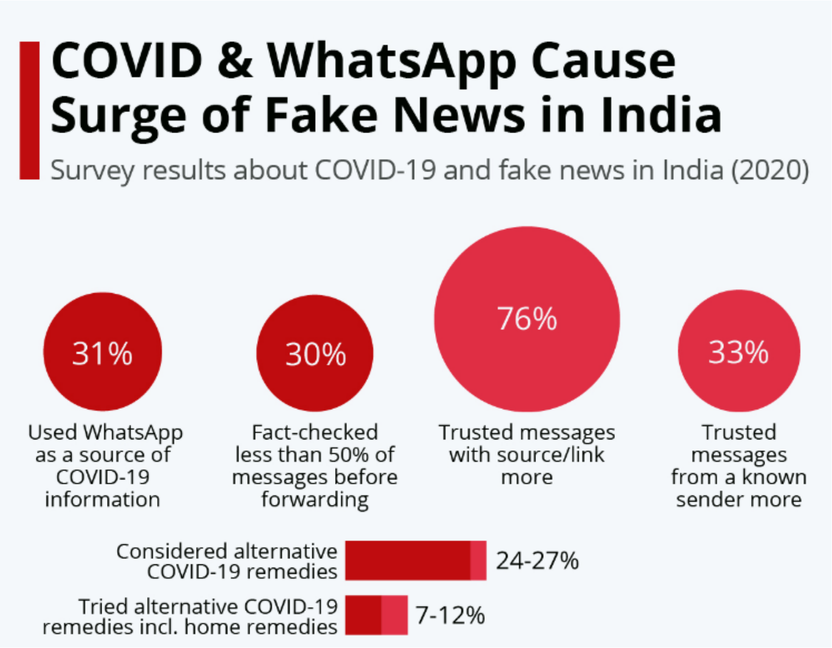

Figure 1.3: COVID & WhatsApp Cause Surge of Fake News in India

spread of different fake news has created panic in the individuals during the pandemic times. When the COVID-19 new wave has hit, there came a flood of fake news in India. Doctors from Rochester New York and India have carried a scientific study and finally published the peer-reviewed journal. This journal of the Medical Internet Research have given insights explaining the internet user behavior in India during pandemic times, this helped understand the spread of misinformation through WhatsApp better. Based on the survey, it was estimated that more than 30% people used WhatsApp regularly in India to get information regarding COVID. And out of these only few have rechecked the facts before spreading them further. The report also stated that minority users spread bulk messages. Even the age groups were estimated during the survey. The major age group was those who were more than 65 as they do not check information or check their facts before spreading them. This was opposite when the age group of 25 was compared with. But according to most Indians, when a source or link is attached, it means that the message can be trusted. But only one third of people trusted the messages from any known sender.

1.4.3 Political context:

There are several cases where people spread fake news over the social media websites for the manipulation of opinion for the favor of participants in the election. 2016 US Presidential Election is one of the famous examples of dissemination of fake news over Twitter and Facebook during election and this event is also known as "misinformation war". Ribeiro *et al.* [34] discussed the case study where they explained the use of fake news in advertising. Another impact of fake news on election is observed in 2018 Brazilian elections in which large number of manipulated images are spread using WhatsApp. According to the analysis it was found that 88% of fake images are shared in last month of Brazilian elections [35]. Fakesters also using WhatsApp for spreading fake news and due to this several of social disturbance and lynching cases occur in India [36].

The major reason behind the continuous growth of fake news over social media website is that anyone can create account on social media websites without paying any cost and can spread any news in small time span. And also, there are no perfect inspection mechanisms and regulations on who can create accounts and on the basis of what information. We need an automated system which can classify news as real or fake in a short span of time. Therefore, this dissertation focuses on fake news detection of social media websites which will cover wide scenarios and contexts where the spread is happening.

There are multiple strategies which are designed for fake news detection today. But these detection mechanisms must even concentrate on particular fake news type. Along with this, legally-forced and self-regulation of web search engine and social medias is very important. In order to displace the fake news, every space must have accurate news, it will help in confronting the false news when the individual comes across them. Confining to critical thinking and scientific methods for detection of fake news is not enough, care must be taken to prioritize the cognitive biases, motivated reasoning, and confirmation bias factors while detecting the fake news.

Along with these, there are certain priorities which have to maintained and noted while developing fake news detection methodologies. They are:

i. You should not develop a detection system only by analyzing the fake news content. You have to prioritize on the user and author's analysis for getting an overall understanding of the fake news that is spreading.

ii. You should not confine the analysis to only one detection methodology, you have to apply diverse methods opting for a comprehensive method for the detection of fake news.

iii. You have to opt for various data mining algorithms, which will help in improvising the present detection frameworks.

We come across two basic theories which we come across when we analyze multiple fake news articles. These will give insights for the analysis and detection of fake news. Analyzing the theories will open new opportunities for quantitative and qualitative study of the fake news. These theories will facilitate to build better and well-justified models and opt for approaches that help in intervening and detecting the fake news.

i. *Fake news with respect to text:* When we analyze the news articles with respect to fake news, the common phenomenon we come across are differences in writing styles, quality, sentiments expressed in the news and also word count which will talk about the quantity. But when research was conducted to study this news only priority was given to testimonies but not to fake news. So, there is scope to conduct research for verifying the attributes basing on the global/local context which will help in distinguishing fake and real news. And also there is a scope to research the fake news making the writing styles analyzing the base for the research.

ii. *Fake news with respect to users:* Mechanisms like liking, posting, commenting, and forwarding are basic characteristics that are to be noted for analyzing the fake news spread by the social media users. Spreading of fake news is not confined to single user, when another similar malicious user comes across the news it is further spread again for example reviews on any information. If the characteristics are not identified by the normal users, there is scope that fake news is spread by them also unintentionally. Analyzing the hypotheses, beliefs and attitudes of the individual user can make the base for analyzing fake news.

Fake news can be spread in the form of image, video and text. This dissertation focuses on text data only. Therefore it majorly focuses on Natural Language Processing (NLP) for text analysis.

1.5 An Overview of Natural Language Processing

"A computer could be considered intelligent if it could carry out a conversation with a human being without the human realizing they were talking to a machine."
— Alan Turing

NLP is one of the important branches which will focus to teach the computers on how text data has to be interpreted and read in a similar way like the humans. The main aim of NLP is to bridge the gap between the human language and data science. It is tough to interpret the data of human language as it is unstructured and contains variety of emotions, tones, and words. So, the traditional techniques will fail in interacting the insights from such data, so NLP works here and makes it easy for the computer to interpret and read such unstructured data. To understand what NLP does when it is applied on text data is shown in figure 1.4 [37].

1.5.1 NLP use cases

NLP has wide range of applications. Out of which some are text to speech conversion , speech to text conversion, content categorization, automated question answering , automatic text summarization, named entity recognition, sentiment analysis, and more. When you read further, you will understand the other applications of NLP.

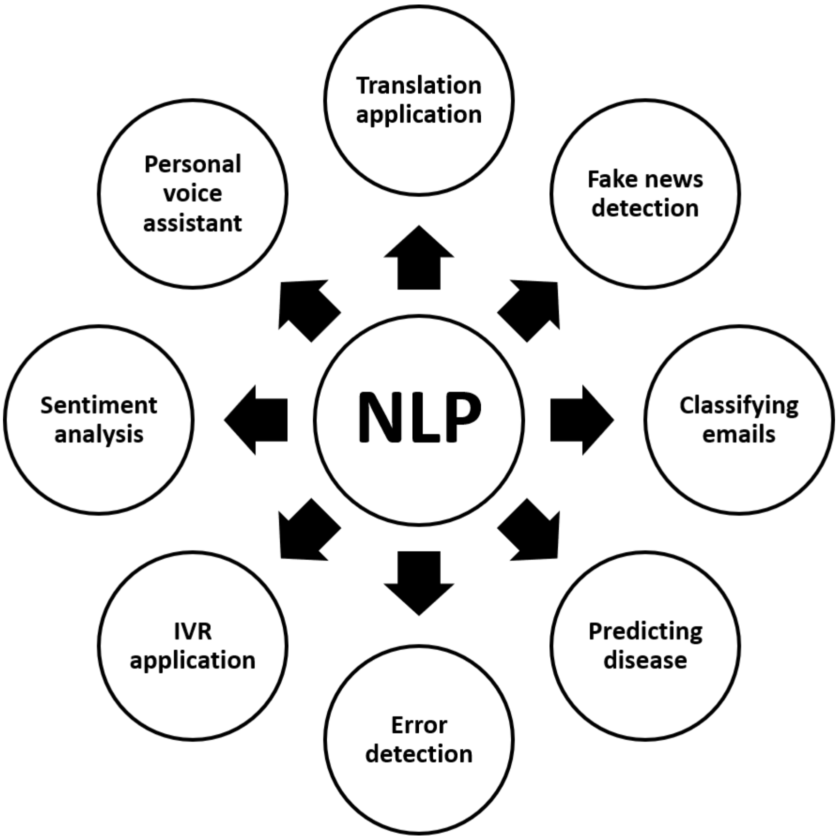

Figure 1.4: NLP Use Cases

i. NLP can be used for translation to apply for Google translation.

ii. It will also help in the fake news detection. To talk about an example, a NLP group at Massachusetts Institute of Technology (MIT) has constructed one system for identifying if a source is biased politically or not. The accuracy results will help in understanding if the news is to be trusted or not.

iii. NLP mechanism will help for the email classification. Yahoo and Google companies make use of NLP for analyzing the text in the emails which will help stop or filter the spams.

iv. It will help for recognizing and predicting the disease and medical condition of the patient basing on his records or speech. Amazon makes use of one service known as Amazon Comprehend Medical. This will make use of NLP for understanding patient details like clinical trial reports, patient voice notes, medications, and disease status.

v. NLP is used by the financial traders for tracking comments, reports, and news. Any insight from these records can be fed into the trading algorithm to get more profits.

vi. Word processing applications like Grammarly, Microsoft word and more make use of NLP to check any grammatical errors in the given text.

vii. Response to the customers from the Call centers are given using NLP by making use of the Interactive Voice Response(IVR) applications.

viii. For performing the semantic analysis on the given customer data many organizations make use of NLP. This data is collected from the resources and social media. It will make it easy to understand the views and choices of the customer on the given product.

ix. Personal voice assistants like Alexa, Cortana, and Siri, make use of the NLP mechanism when they respond to any vocal commands of the user.

1.5.2 NLP mechanism

Some of the NLP mechanisms are shown in figure 1.5 [38].

i. *Bag of words:* This NLP mechanism will help in counting the words in a given text. This is done by generating a matrix for every sentence. But the order of words and grammar are not taken into consideration during matrix generation. This matrix is fed into classifier for further execution. This approach is exceptionally straightforward but it has few limitations. This approach focuses on the frequency of words in the text without considering its semantics. Repetitive words like "and", "are", "is" are also evaluated but those words of less important during the analysis. To beat these issues, a new methodology introduced i.e. Term Frequency-Inverse Document Frequency (TF-IDF).

ii. *Term Frequency-Inverse Document Frequency (TF-IDF):* Weighting factor is prioritized in TF-IDF. Statistics are used in order to calculate the word importance in the document. TF-IDF is the combination of two words; (i) Term

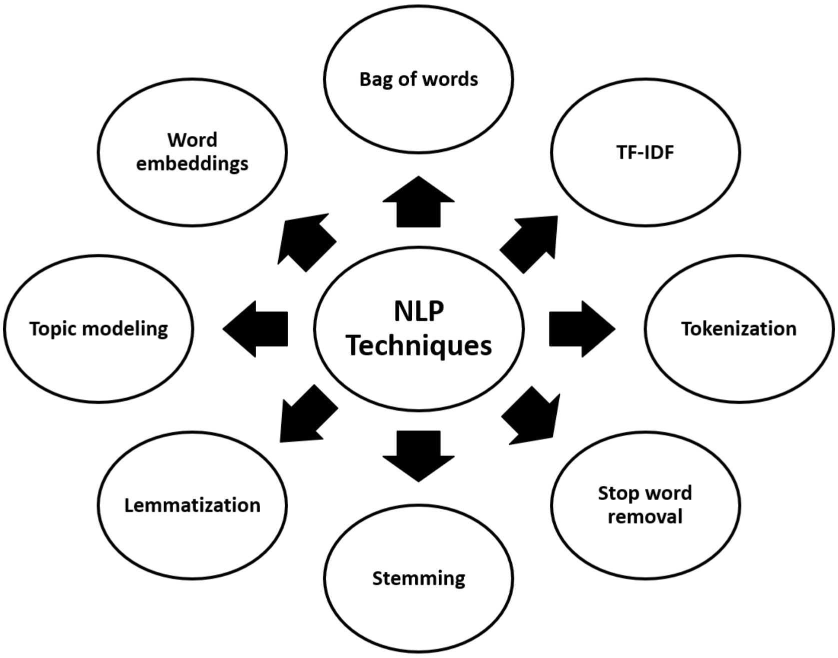

Figure 1.5: NLP Techniques

frequency (TF) and (ii) Inverse Document Frequency (IDF). TF deals with word frequency and IDF deals with importance of words in the piece of text. Finally the value of TF-IDF is the multiplication of TF and IDF. Search engines make use of this mechanism for ranking and scoring.

iii. *Tokenization:* The entire text is segmented into sentences and words, meaning the text is divided into tokens. Characters like hyphens, punctuations are discarded during the segmentation. Segmentation splits the text into blank spaces and makes the analysis of the text easier. But some punctuations necessary for abbreviation may also be removed while segmentation.

iv. *Stop words Removal:* Common words like prepositions “the”, “and”, “etc”, “a” are removed during this mechanism. These words are considered as noisy words because they play little or no role in the analysis. Removal of these words helps in concentrating the important words in analysis rather than wasting the time on noisy words. Removal is done by predefined list managed by English dictionary. This will improve the processing time and performance but sometimes important information may be lost in the process.

v. *Stemming:* The words affixes are removed for better processing. The words having different spellings and almost same meaning should be a part of same token. So words are reduced in their root forms. In spite of having limitations in this technique, it is efficient in text processing in terms of performance and speed.

vi. *Lemmatization:* This is similar to stemming mechanism. The words in this process are converted to the lemma form which means the dictionary form. Lemmalization works well as compared to stemming in case of single word used in two different context but it is computationally expensive as compared to stemming.

vii. *Topic Modeling:* Here the main topics are extracted from the document or text. It is assumed that every document is group of topics and in turn they are group of words. It is nothing but the dimensionality reduction as the text is reduced into topics. Latent Dirichlet Allocation is the popular technique for topic modeling.

viii. *Word Embedding:* Words in a text document are represented in the numbers form. It explains that similar words should be having similar representation. The words will be represented in the form of real-valued vectors.

All of these NLP mechanisms play an important role to analyze the text data for making them easier to be evaluated by the computer to process the information.

1.6 Fake news detection approaches

As discussed above, along with above characteristics and theories, every news has two major components which are responsible for classification of news as real of fake; i.e. text and user profile. Based on these two components this dissertation enlist three viewpoints/aspects which are useful in detection of fake news as shown in figure 1.6; (i) news writing pattern (style-based), (ii) news credibility (using news content and creator profile credibility) and (iii) propagation pattern of news (how it spreads). In the upcoming chapters this dissertation explains these viewpoints and explained that how these are helpful in classification/detection process.

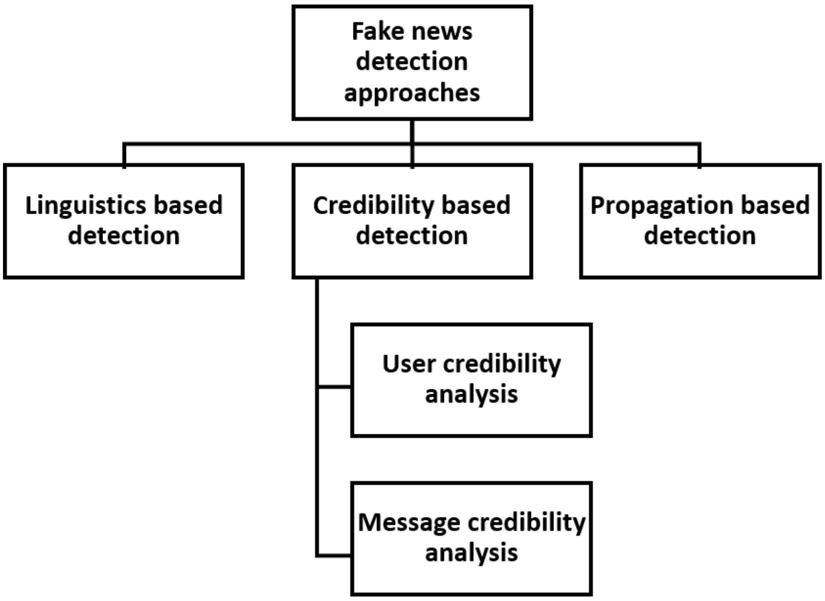
Fake news detection approaches
Linguistics based detection
Credibility based detection
Propagation based detection
User credibility analysis
Message credibility analysis

Figure 1.6: Fake news detection approaches

As the dissertation considering creator's profile as one of the component therefore before talking about the detection of fake news, one has to talk about who are the creators of fake news.

i. Non-humans: The computer algorithms known as social bot exhibit human like behavior and will produce content as well as interact with the humans in social media. Some of the bots may spread real news but most of them spread misinformation, malware, spam, and rumors. For example; creation of social bots was done during the elections time in the US for supporting Clinton or Trump making use of multiple tweets which gave references to many fake news websites. Another non-human fake news spreader is "Cyborgs". Cyborgs are human-assisted bots or bot-assisted humans. Once they are registered, these accounts can participate with any social community and post multiple tweets. Even cyborgs spread misleading information damaging the trust and belief of the social media users. These are one of the prominent non-human creators of fake news.

ii. Humans: One of the major fake news spreaders are humans. Even if fake news is spread automatically or manually the main distributors and creators of the fake news are the humans. For example; One agent of the FBI who is suspected in the e-mail leak of Hillary, is found dead in his apartment. This information is entirely false, but many users have spread this news by forwarding and sharing this information many times. The followers and friends of this user further spreads this news multiple times, these are called next-generation spreaders and this spreading results in an echo chamber propagating the spread of fake news.

This report explained what fake news is, significance of fake news detection, types of fake news spreaders till now. Let us discuss detection methodologies for fake news.

1.6.1 Linguistics based detection

As we have discussed above, fake news that is spread online is created by the users for any of the mentioned gains like political, financial or personal. To avoid detection fake news adopts particular writing strategies. Linguistics approach focuses on such writing pattern of given text and addresses that the writing style of fake news is different from true news. This approach analyzed the intention of news creator, i.e., either he is trying to mislead the readers or not. Features related to this approach are quantifiable therefore we used machine learning (ML) approach for classification because ML approach works well with these type of features. These features are grouped in following categories:

i. *Attribute based features:* These features are theory oriented because they are directly extracted from theory/text. According to “sensory ratio” explained by reality monitoring, the fake events has less sensory text as compared to real events. In this report all attribute based features are further divided into four sub-categories i.e., quantity, subjectivity, readability and sentiment. These features are exceptionally appropriate for the classification of news as real or fake. Research and analysis are required to estimated which features provide more systematic information.

ii. *Structure based features:* It is majorly describing the syntax and semantics of the text. These features talk about discourse, semantics, syntax and lexicons in the language. These are also called technique-oriented features because they use the NLP techniques. This technique focuses on the statistics of words using NLP and Part-Of-Speech (POS) tagging. At the lexicon level, analysis will assess the frequency of the word, or letter making use of n-gram model algorithm. At the syntax level, tasks can be performed on the parts of speech for POS tagging. Whereas Probabilistic Context-Free Grammars performed for shallow syntax tasks; linguistic inquiry and word count will be helping in estimating the semantic features and to capture rhetorical relations of discourse level, rhetorical structure theory is used.

1.6.2 Credibility based detection

Above detection technique uses handcrafted features for classification but instead of depending on handcrafted features this detection technique uses news creator profile and semantics of text developed by pre-trained model. In order to analyze the credibility both the user as well as the contents credibility has to be evaluated.

i. *User Credibility analysis:* Online social media website is the place where anyone can register and present their views. Some organizations create fake user profile or bots for fake reviews and fake news spread. The objective of this type of users is to mislead reader over the internet. Therefore, we need to create an automated system which can classify the profile as real or fake based on the profile information entered at the time of registration. For the sake of automation, making use of user-based features will help in evaluating the credibility of the user. The features will help in identifying the characteristics of any suspicious non-human and user accounts. Some of the features which will help in evaluating the credibility are registration data, geo location information, verification of the user and so on.

ii. *Message credibility analysis:* User credibility talks about authenticity of the social media account holder. On the other hand message credibility analysis talks about the content level features. This is nothing but the analysis of semantic information of the news content. This can be done using pre- and post-trained models. Pre-trained model uses their own word embedding for processing but post-trained model uses popular word embedding like GloVe, Word2Vec etc.

1.6.3 Propagation based detection

This approach is based on text propagation features obtained from social media websites. These feature considered the information related to news publisher and spreader like number of posts/tweets, followers/friends ratio and other information. The discussions conducted on the social media, posts shared by the user will help in identifying the behavior of the user. This will also help evaluate the monthly average posts shared by the user and all these features will help in identifying differences between legitimate and deceptive news.

1.7 Thesis objectives

The aim of this dissertation is to investigate the automatic system for the identification of fake news on social media. Traditionally fact checking is the famous and simple strategy for the classification of news as real or fake but it does not work well for digital platforms where large amount of data generated in a second. Therefore we need a automatic system for fake news detection that could help in the classification of news in short time span. Furthermore, this system could reduce the number of readers who are affected by fake stories.

Development of automatic system for fake news detection is not a trivial job. On some issues like politics, religion and health, human mind is naturally biased in the classification of news as real or fake. In addition, content style and intrinsic biasness of news creator makes this classification task more harder. Therefore we need a system which can extract features from the news and classify.

For obtaining the goal of news classification, this dissertation work is divided into following objectives (Obj):

Obj 1 *Data collection/pre-processing & categorization*

Obj 2 *Fake news identification based on style and credibility study*

Obj 3 *Verification using propagation based*

Obj 4 *Deployment of proposed model*

1.8 Contributions

The primary contributions of this dissertation are summed up as follows:

i. A survey characterize the publicly available dataset for the fake news detection (Obj1). This report majorly focuses on text datasets and selected only those datasets whose available features are essential for our implementation and comparison with state-of-the-art.

ii. After the selection of fake news dataset, new dataset constructed named "WELFake dataset" (Obj1). This dataset overcome the limitations of publicly available datasets. This dataset contains three important features named; author name, news text, heading and final prediction either real or fake.

iii. A linguistics framework that uses the style based features for the classification of news (Obj2). This report explained "WELFake model" where word embedding technique is merged with linguistic features for improved result. This model also unveils the impact of voting classifier by highlighting the comparative study of state-of-the-art on fake news detection.

iv. A framework based on fusion of pre- and post-trained model for the classification of news (Obj2). This framework reads the text news and passed through several layers and finally classify based on its features. At the end state-of-the-art approaches compared with proposed model to demonstrate the effective result.

v. A framework analyse the user profile based features for the identification of user either real or fake/bot (Obj2). It predicts the final output based on textual and non-textual features. For textual data, pre- and post-trained models are used and for non-textual data traditional ML models with some parameter tuning are used.

vi. A framework analyse the propagation pattern of news and classify as real or fake (Obj3). Propagation and user profile based features are merged for the final classification.

vii. Final workflow of all the predictive models for prediction and verification of news as real or fake (Obj4).

1.9 Chapter Organization

The remaining part of this dissertation organized in eight chapters. Chapter 2 explains the background of some important aspects related to NLP as well as it describes the work done by several researchers. Chapter 3 presents several publicly available datasets that required for building of automated system and also explains statistics of proposed dataset named ”WELFake dataset”. Important linguistic features which are responsible for the classification of news explained in chapter 4. Chapter 5 and chapter 6 explains the use of machine learning and deep learning approaches for the testing of credibility of message and user profile respectively. Chapter 7 analyses the propagation pattern of news and describe the use of this analysis in classification of news. All the predictive models are merged through the workflow that explains the investigation and verification process of news in chapter 8. At last, chapter 9 summarized this dissertation and also discussed open perspectives.

Chapter 2

BACKGROUND AND RELATED WORK

2.1 Fundamental theory

2.1.1 Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed"

– Arthur Samuel, 1959

Machine Learning (ML) as a subset of artificial intelligence that turns information into knowledge. In the last few decades, there has been an explosion of data. ML algorithm used to extract the essential pattern from this huge data and predict some result. Many leading companies uses ML for the prediction or operational task. There are three types of ML algorithm;

i. Supervised learning uses labelled data for training purpose.

ii. Unsupervised learning uses unlabelled data for training.

iii. Reinforcement learning uses reward to learn concept.

2.1.2 ML algorithm

ML algorithms are used as classifiers for discriminating different things based on some features. Major classifiers which we come across are:

A. Naïve Bayes (NB) [39]:

NB classifier is considered a supervised and probabilistic learning algorithm used to classify different data having high-dimensional dataset. It is also known as a probabilistic classifier because it can predict the classification based on the probability of an object. The classifier works on the Bayes theorem.

The name of this classifier consists of two words, namely Naïve and Bayes. (i) *Naive:* This describes that a feature of the dataset is independent of other features that are occurring. It explains that every feature is different, and it will individually contribute to identifying the object. (ii) *Bayes:* The word Bayes is included in the classifier's name because it uses the principles of Bayes Theorem.

Bayes theorem is also known as Bayes law or Bayes rule. The theorem makes use of a mathematical equation for statistics and probability for calculating conditional probability. In simple words, it will help calculate the probability of an event based on its association with other events.

There are three types of NB model: (i) *Gaussian:* This NB model type is used when the features in the dataset have a normal distribution. That explains that rather than discrete values, continuous values are used in this model. (ii) *Multinomial:* This NB model type is used in the case of document classification to see which category the document belong. The features used in this classifier are based on the frequency of words in the document. (iii) *Bernoulli:* This NB model type will work similar to that of the Multinomial classifier; the predictor variables in this model are independent Boolean variables. This model is considered to be famous for the classification of documents.

Pros of NB Algorithm:

i. As features are independent of each other, it works significantly faster than other complicated algorithms.

ii. The algorithm works best with email spam detection, text classification, which is high-dimensional data.

Cons of NB Algorithm:

i. In real-time data, we do not find many features that are independent of each other, so the accuracy rate of the algorithm is less.

B. Support Vector Machine (SVM) [39]:

This classification algorithm is considered as a supervised learning algorithm and popularly known as SVM. This algorithm can be used for addressing classifications and regression problems like support vector regression and support vector classification. But it is majorly used to solve classification problems. SVM can also be used to solve non-linear and linear problems and work well to address practical problems.

i. *Linear SVM* used for separating the data which is linearly separable, meaning the datasets can be classified making use of a straight line. This classifier is known as linear SVM classifier.

ii. *Non-linear SVM* used for separating data that is non-linearly separable, meaning the datasets cannot be classified using a straight line. This classifier is known as Non-linear SVM classifier.

The classification methodology used by SVM is creating the best decision or line boundary which can segregate the given space based on its classes which will help in placing the new data point in the right category. The decision or line boundary is the hyperplane. For creating the hyperplane SVM will choose the extreme points known as support vectors.

i. Hyperplane: As explained earlier, hyperplanes are the best boundary lines. Its dimensions will depend on the number of features of the dataset. If there are two features, then hyperplane can be a straight line but in case of three or more features, it is classified making use of kernels in high number of dimension plane. Hyperplane is always created with maximum margin, meaning the maximum distance between the points of the data.

ii. Support vectors: These are the points that are close to the hyperplane and are capable to affect the hyperplane's position; as they are supporting the hyperplane they are known as support vectors.

Here, the above is an example of a hyperplane. Consider the blue squares as one datapoints and green circles as another datapoints. Using SVM one can draw a boundary line to classify them. The algorithm will help in identifying the close point of lines from both datapoints which are known as support vectors. Margin is the distance between these vectors and the hyperplane; SVM's goal is to maximize this margin and the one with maximum margin is known as optimal hyperplane.
Pros of SVM algorithm:

i. It is robust to outliers.

ii. Highly effective in higher dimensions.

iii. Suits for classification because it is memory efficient.

Cons of SVM algorithm:

i. Takes more training time to process large dataset.

ii. Cannot perform well in overlapped classes.

C. Decision Tree (DT) [39]

DT can handle both type of problems regression and classification but it is preferred to solve classification problems. It has a tree like structure. While constructing DT, it is developed after asking questions regarding the dataset. Once an answer is received, next question as a follow-up is asked to finally come to a conclusion. These questions and answered are splitted and organized in the DT form. It can be easily described as the graphical representation to get every possible solution to the problem considering the given conditions. Following terminologies are used in DT:

i. Root node: As the name indicates, it starts with root nodes which further divide into branches representing tree-like structure. Entire dataset is represented by the root node. It is also known as decision node.

ii. Leaf Node: The final output nodes are known as leaf nodes. One cannot segregate a leaf node further.

iii. Splitting: The process of dividing root or decision nodes into some sub-nodes.

iv. Sub Tree/Branch: This is also a tree which is formed when the tree is splitted.

v. Pruning: The phenomenon of removing unwanted branches of the tree.

vi. Parent/Child node: The root node is known as the parent node, and the other nodes are known as the child nodes.

DT algorithm will start with a root node like in the trees, then this will compare the root attribute value with the real datasets value and after comparison will follow the branch to jump into the next node. This procedure will continue further for the next node, where comparison of attribute values with other sub-nodes will be performed to continue further until reaching a leaf node. All the internal nodes of a DT will represent the dataset's features, branches will represent the decision rules and the outcome is represented by the leaf node.

There are two types of DT; (i) Classification or Categorical DT: In this type of DT the decision variable is discrete and (ii) Continuous variable DT: This DT has continuous target variables.

Pros of DT:

i. It is simple because the process of DT is similar to the real life where human takes a decision.

ii. It works well for decision related problems.

iii. One can estimate all possible outcomes from a problem.

iv. As it has a tree like structure, logic of the classifier can be easily understood.

Cons of DT:

i. It is complex as it has many layers.

ii. Computation is complex when there are more class labels.

D. Random Forest (RF) [39]:

The RF classifier works both for classification and regression problems. It is based on ensemble learning where multiple classifier combined to solve a complex problem. Accuracy in this algorithm is higher than DT algorithm because it will merge many DT to get the stable prediction. The algorithm makes use of bagging method which is the combination of different learning model thus increasing the overall result.

RF algorithm creates random sample of dataset. Each sample dataset construct the DT for final prediction. RF classifier reads the output generated from all DT and predict final result based on the maximum votes. More number of trees means more accuracy and less overfitting problem.

Pros of RF classifier:

i. Multiple DT make it robust and accurate.

ii. Bias or overfitting is absent as it takes the average of predictions.

iii. The algorithm can deal with large dataset with higher dimension.

Cons of RF Classifier:

i. Multiple DT make it tough to interpret.

ii. As there are multiple DT, predictions are slow, making the process time-consuming.

2.1.3 Word embeddings (WE)

Any ML and Deep Learning (DL) model can able to read data in the numerical form only. This report talks about the text data only. Therefore we need to transform text data to numerical data. The process of converting text data into number is called *vectorization* or in case of NLP this process is called as *word embedding*. In simple words this is the way to represent text data in number format. In WE the words having similar meanings are allowed to have similar representation. It will help in bridging the human language understanding to a machine. The text representations are done on n-dimensional space in which the same meaning words have similar vector representation, they are closely placed in the vector space. These will help in addressing the natural language processing issues. Some of the types of word embeddings are:

A. Count Vectorization (CV) [40]:

As the text cannot be processed directly by the computer, it must be converted into numbers. So, they have to be vectorized for processing; one way of doing it is by applying CV. CV is used to transform the text into vectors based on the word frequency in the text. CV tokenizes the given text and performs basic processing, removing punctuations and converting the text into lowercase. After tokenizing, CV will create a matrix in which each word will be represented by the column in the matrix, and every text sample of the document will be represented by the row matrix. Values of each cell are the word counts in the text sample.

Pros of CV:

i. It is flexible as it allows pre-processing of data before vector representation.

ii. It requires low memory.

iii. Fast pickle and un-pickle process.

Cons of CV:

i. It does not identify less and more important words.

ii. It does not identify the word relationships like linguistic similarity.

B. Term Frequency Inverse Document Frequency (TF-IDF) [40]

It is a technique that will help in quantifying words in documents. This is done by computing the weights of each word which will signify the word importance in the corpus and documents. This technique is used in text mining and information retrieval.

For example; *"The girl is wearing a red dress".* There are seven words in the above sentence, which is easy to understand by a human as they know the semantics of sentence and the words. Still, computers will only be able to understand the data in numerical value-form only. So, the text has to be vectorized for the computer to understand it. Hence TF-IDF will help in the vectorization of text, helping in the classification of the document. Once the text is vectorized, clustering, ranking, and finding any relevant document is possible. This is similar to what happens when you search for something on google. When you try to search for a query on google, it will find relevance with the documents, rank those in the relevance order, and display the top results- the whole of this process is done in the vectorized form of documents and query. This is the underlying structure of google algorithms.

The term *TF (Term Frequency)* will help in measuring the frequency of a term occurring in the document. Every document varies in length, so repetition of terms will vary based on document size. So, one has to divide the term frequency by document length.

$$TF = \left(\frac{No\, of\; times\; a\; term\; will\; appear\; in\; a\; document}{Total\; no\; of\; terms\; in\; a\; document} \right)$$

IDF the Inverse Document Frequency will help in measuring the importance of a term. Some terms like 'of,' 'is,' 'the' may be repetitive in the document but less important. Using the following computation method, such frequent terms can be weighed down while scaling up the rare ones.

$$IDF = log_e \left(\frac{Total\; number\; of\; the\; document}{No\, of\; documents\; with\; a\; term\; `t'\; in\; it} \right)$$

Multiplying both results will help in calculating the score of TF-IDF in the document.

$$TF - IDF = (TF * IDF)$$

The higher the TF-IDF value, the rarer the term is in the document.

Pros of TF-IDF:

i. It focuses on the frequency and importance of more and less words.

ii. It is efficient and simple technique for matching words in a search query.

Cons of TF-IDF:

i. As similarity is computed on word count space, a large vocabulary takes a lot of time.

ii. Does not consider semantic similarities between words.

C. Global Vector for word representation: GloVe [41]

This is one of the examples of an unsupervised learning algorithm used for obtaining the vector word representation. In the term GloVe, "Glo" i.e. Global denotes the corpus's global statistics and "Ve" i.e. Vector denotes the vector representation of the word. Therefore it is called as Global Vector for word representation. This method works on global statistics of matrix factorization techniques in a local word-content matrix. A matrix of co-occurrence information is built, where each word (row) and its frequency in particular content (column) are counted on the large corpus. Scanning of the corpus is done in the following way – each term, context term in an area with respect to the window size before and after the term. Less weight is given for more distant words. Context number is large as it is combinational. So, factorization of the matrix is done to get a low-dimension matrix, where each row will yield vector representation for every word.

Pros of GloVe:

i. It makes use of global statistics.

ii. It gives practical meaning to the vector by taking word pair into consideration.

Cons of GloVe:

i. As it is based on co-occurrence of the matrix of words, it will take a lot of memory.

ii. Reconstruction for any purpose is time-consuming.

2.1.4 Pre-trained model

Pre-trained models are those which are created by others to solve similar problem. Rather than building a new model for solving a similar problem it is better to use a model which was already trained on another problem before.

A. Bidirectional Encoder Representation from Transformers (BERT) [42]
BERT is a language representation model proposed by researchers at Google AI Language. It has been designed for pre-training deep bidirectional representations on unlabeled text by conditioning left and right context in all the layers. You can fine tune this model with an additional output layer for creating the state of art model for many tasks like language inference and question answering. This does not need any architecture modification that is specific to any task.

BERT is mainly based on the transformer encoder network which is efficient enough to process long texts making use of self-attention. It is a language representation model that is based on neural network architecture that will help train the conversional question response system.

BERT will help the algorithm in understanding natural language easily. It will make use of whole of the text passage for understanding each word's meaning. This is done by relating every word of the sentence with the rest, understanding the way of expression of the people, thus giving accurate results when a query is searched.

It is a bidirectional transformer that is pretrained making use of the combination of next sentence prediction at large and masked language modeling objective meaning it can perform two tasks – next sentence prediction and sentence reconstruction. For the reconstruction it will involve the masking of tokens randomly in the sentence and again reconstructing the real sentence from the one that is masked. BERT model will reconstruct these masked tokens independently from another.

Pros of BERT:

i. As it is bidirectional, it is capable to get the word's context from both left to right and right to left simultaneously.

ii. The contextualize word embeddings makes it very efficient

Cons of BERT:

i. It is expensive at the time of model production.

B. Robustly optimized BERT (RoBERT) [43]

It is built on the BERT model with modifications in hyper parameters which include removal of the pertaining objective of the next sentence prediction and also capable for training making use of large mini batches and learning rates.

So, the major difference between BERT and RoBERT is the training data sizes and slight changes in the key hyperparameters to increase the performance of this model. Dynamic masking is possible in this model, which will enable the change of masked tokens during the training. In this model, indication of token id determining its segments is not needed. Separation of the segment making use of separation token will be enough.

Pros of RoBERT:

i. It outperforms BERT model in tasks on the benchmark of General Language Understanding Evaluation.

ii. It is capable to be performed on more data and has more computational power.

C. DistilBERT [44]

This model is a distilled version of BERT which will retain 97% of performance by making use only half of the parameters. It does not have pooler, token kind of embeddings and will retain half of the layers of Google's BERT.

DistilBERT is sixty percent faster and forty percent smaller than the BERT model. The general-purpose language model can be trained successfully with distillation to get a better version of the model.

The major idea behind the DistilBERT is when training of large neural network is performed, it is possible to approximate the output distributions making use

of smaller networks. This model will help in leveraging the inductive bias of large models, by introducing cosine distance loss, distillation, and modeling. The DistilBERT is a lighter, faster, cheaper, and smaller model for pre training and has high effective output when used.
Pros of DistilBERT:

i. Has a fast inference speed in comparison with other models.

Cons of DistilBERT:

i. Prediction metrics may vary in some percentages.

2.1.5 Deep Learning (DL)

> *"The hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. For this reason, we call this approach to AI deep learning."*
>
> — Ian Goodfellow and Aaron Courville

Deep learning is a subset of machine learning technique that teaches a computer to filter inputs (observations in the form of images, text, or sound) through layers in order to learn how to predict and classify information. Deep learning is inspired by the way that the human brain filters information! Just like the brain recognize the patterns and categorize different types of information, neural networks use a hierarchy of layered filters in which each layer learns from the previous layer and then passes its output to the next layer.

2.1.6 Deep learning model

These models draw conclusions like humans where the brain will analyze the data continuously with a logical structure. Similarly, deep learning models will use the multi layered structure of algorithms known as neural networks. Neural networks design is based on human brain structure. Neural networks will classify varying information types and will help in identifying patterns. The neural network layers act as filters helping in detecting and getting the correct result. Similarly, the human brain compares any new information with the objects known like the methodology of deep neural networks mentioned above.

Deep learning models will make use of predictive modeling and statistics making it beneficial for those who access large amount of data making the process easier and faster as deep learning program will build the feature set by themselves with no supervision, making it accurate and faster.

A. Convolutional Neural Network (CNN) [45]

Convolutional Neural Network is a type of deep learning algorithm which is capable of taking input image, for assigning importance to many objects of the image and helping to differentiate one image from another. CNN's requirement for preprocessing is less when compared to other algorithms. This algorithms give high performance with respect to audio signal, speech and image inputs. In CNN there are three layers:

i. *Convolutional layer:* It is the first layer of CNN. This layer is followed by additional CNN or pooling layers, the final layer is the fully connected layer. With increase in each layer, there is increase in the complexity of CNN, thus identifying more image portions. The starting layers will focus on simple features like edges and colors, whereas as image data progresses in the CNN layers, it will start recognizing larger object shapes and elements until it finally identify the object. The convolutional layer also has a filter that will move in the image receptive field to check the presence of the feature and this is known as convolution.

ii. *Pooling layer:* It is known as down sampling which will conduct the reduction of dimensionality. Like the convolutional layer, this also sweeps the filter across the input. It is of two types. If the filter while moving across the input, selects the pixel of maximum value to send it to the output array it is called max pooling and if it calculates and send average value then it is average pooling. This layer adds efficiency, reduces complexity to the CNN.

iii. *Fully connected layer:* In this layer, every node in output layer is connected directly to the node present in the previous layer. The classification task in this layer is based on the extracted features of previous layers and their filters.

Pros of CNN:

i. It will detect features automatically with no supervision,

ii. Detects images with high accuracy.

Cons of CNN:

i. It requires a large dataset for processing.

B. LSTM [46]

LSTM stand for long short-term memory networks. These are a type of RNN-recurrent neural networks which are capable to learn the order dependence in the problems that are based on the sequence prediction. These problems are generally seen in speech recognition, image recognition, machine translation, and more. LSTMs are designed to address the issues of RNNs. In RNN the contextual range coverage is limited and faces the issue of vanishing error. To overcome these LSTM works the best. LSTM works on two mechanisms; (i) *Saving mechanism:* Only save the information that is important for the future and (ii) *Forgetting mechanism:* It forgets all the scene related information which is not worth remembering.

Pros of LSTM:

i. It solves the issue of vanishing gradient.

ii. Training of LSTM are easily done.

Cons of LSTM:

i. LSTM face the issue of overfitting.

C. BiLSTM [47]

In simple words, BiLSTM stands for bidirectional LSTM consisting of two LSTMs. Out of these one LSTM takes input from forward direction and another from backward direction. These will help in increasing the information available for the network thereby the context of the algorithm will also increase. The input entry from two ways, involves one input from the past to the future and another input from future to the past. So the LSTMs will store the information from past and the future.

Pros of BiLSTM:

i. Predictions of BiLSTM are better than other models.

Cons of BiLSTM:

i. It is expensive.

ii. Does not work effectively for speech recognition.

2.2 State of the art

2.2.1 Linguistics based analysis

The research paper by Georgios *et al.* [48] is mainly based on detecting fake news. To address the issue, the author has opted for machine learning algorithms and has used the content-based feature to give out accurate results. The author's experiments used certain linguistic features, which made it easy to distinguish fake from real news. The study started with an extensive feature study of the data. To detect the fake news in written narratives and word embeddings, the author chose to apply multiple machine learning algorithms and ensemble algorithms, which are highly efficient in performing text classification tasks. Some of the algorithms the author used are AdaBoost, SVM, DT, Bagging, and more; along with this author chose to use old data sources to conduct experiments. The author has set a solid methodology and specific rules to create an unbiased dataset for Fake News Detection. The dataset contains a balanced number of real and false news articles.

Also, the author has added articles from different sources and categories into the unbiased set of articles. The author did this to produce a generic dataset for the detection of fake news articles. This experimentation methodology of the author, which made use of methods like AdaBoost and bagging, has been successful and has given out accuracy of up to 95% when tested over five fake news detection datasets. The methodologies employed by the author have formed a base for the future when anyone wants to employ several meta-data to scrutinize the data further as well as prevent any dissemination of the data on the internet.

Judee *et al.* [49] idea of making use of linguistic features as tools to detect deception in communication laid the foundation for this research paper. To conduct the research, the author has opted for two experiments initially- survival in the desert and on a pilot. But both these experiments have shown variable results. So, the author has opted for detailed research on a theft scenario; where result analysis was done on individual and cluster cues. For cluster analysis, data mining algorithms were used which helped the author develop an automatic requirement for detecting the deception. The author has made use of C4.5 [1] which has cut off redundant branches and constrained error rates. The prediction rate of cluster analysis across fifteen linguistic features was 60.72%. But when examined the individual cues, the author has realized that deceptive individuals tended to use less complex sentences with fewer conjunctions. At the end of the research the author has concluded that deceptive individuals tend to communicate with less complexity, specificity of language, and diversity; but higher informality, expressiveness, nonimmediacy, and quantity in comparison with truthful individuals in their messages. The research paper concludes that clustering techniques and language indicators will help in identifying deceptive texts. Though in the future certain features of the deceptive person may change, some features will still be consistent which can be identified. Future research on this line will help in working more on modalities, other linguistic cues serving as an additional benefit to the present research.

The aim of framework employed by Michela *et al.* [50] is to detect and address the issue of spreading misinformation before it happens, which is mainly because of the polarization by the users. To address the case, the author has made use of polarized content, which is the primary source of fake news in the future. To identify the polarized content author chose to divide the data into two categories, namely fake and official news. For example, the author made use of the data of official Italian newspapers under the first category and data from Italian websites that are known to spread fake news under the second category. The author chose the Facebook platform from where all the data related to these categories are collected. Then, based on the author's data textual content, topics are extracted, which helped the author understand the subject and context better. Once data is selected, certain features are derived, which explain how it is presented and perceived to the audience. Now classification of the data is performed by the author to detect possible data which can be fake news in the future. Classifier machine algorithms like Linear Regression, Logistic Regression, SVM with linear kernel, k-nearest neighbors (KNN), and Neutral Network Models are used to complete the classification. These classifier algorithms helped the author identify the undisputed and disputed data with an accuracy rate of 77%, which is very good. The author used this information as a new feature and applied an additional classifier, which resulted in recognition of the fake news at an accuracy of 91%. The author results have laid the foundation to extend the data set collection from Facebook to other platforms, catering to the early detection of fake news.

The research paper by Veronica *et al.* [51] is all about the automatic detection of fake news. Unlike the regular methodologies, which used datasets based on satirical news, the author here constructs two data sets of fake news, including data set of six domains collected from crowdsourcing. In contrast, the other dataset is related to celebrity fake news collected directly from the web. While gathering information for the data set, the author has followed the requirement guidelines for the fake news corpus proposed by Rubin *et al.* [52] The crowdsourcing data set consisted of legitimate data of six domains that have further undergone manual fact-checking, and fake news was collected from Amazon Mechanical Trunk

(AMT). Then 240 articles were extracted from the legitimate group, and similar articles from the fake news group were extracted with the help of AMT workers. The data collected from the web included both fake and legitimate news whose authenticities were cross-checked with other sources. During this process, the author realized that legitimate and fake news were almost similar, with the slightest differences. For fake news detection, linguistic features like punctuation, Ngrams, Syntax, and psycholinguistic features were compared using Linguistic Inquiry and Word Count Software. Then the author has opted for computation models like SVM classifiers, learning curves; these were applied for data of different domains. Also, the author has compared his methodology with human capabilities to detect fake news by giving them some data. At the end of the research, the author has concluded that information encoded in the LIWC lexicon has shown good performance in detecting fake news with a good accuracy rate similar to human abilities in detecting fake news.

Cody *et al.* [53] research paper aims to identify fake news in the Twitter threads automatically. The method of automatic detection was developed by the author when he started studying how accuracy assessments were performed in CREDBANK and PHEME. The author experimented on three datasets which were picked from BuzzFeed's fake news dataset, CREDBANK dataset, which is a crowdsourced dataset of accuracy assessment, and PHEME, which is potential rumors dataset and has journalists assessment for accuracy. To develop the model for detection, the author has first identified some features falling under four types, then aligned the three datasets in a consistent format, and then used the classifiers and evaluated each feature set under the receiver operating characteristic (ROC) curve. This classifier has helped the author to eliminate any low-quality features which may not detect fake news accurately. And once high-quality features are selected in both CREDBANK and PHEME are selected, the author has applied classifiers based on these features separately on Buzzfeed data sets. Later, the author pooled the featured and further applied the classifier based on these on Buzzfeed data sets. Finally, the author has then plotted these using the ROC curve for the three datasets to detect the fake news and estimate the accuracy rate. At the end of the research, the author found CREDBANK's accuracy assessment was more based on credibility. It is more focused on several content markers, which were better than PHEME assessments. But the author concludes by saying both the assessments were based on certain feature sets whose coverage could be expanded to give out higher accuracy in detecting fake news.

Matthew *et al.* [54] project talks about making use of different linguistic styles to identify the deception words in the texts. The author chose an inductive approach while examining the false stories. The author made use of a computerized text analysis program for creating empirically derived profiles of deceptive and truthful communications. These profiles were tested on samples of texts and compared with predictions on the authenticities made by the human judges. These profiles were created using a text analysis program called Linguistic Inquiry and Word Count to check these samples on a word-by-word basis. For experimentation, the author chose to apply this methodology on five different samples; multivariate profiles of deceptions were created to check these sample's output for deception. Also, as separate linguistic profiles were created for each sample, the samples were cross-checked with other sample's profiles to predict deception. A consolidated multivariate profile was created at the end for convenient usage in the future. The research of Matthew et al.'s has concluded that there are three language dimensions like less cognitive complexity, more negative emotion words, and few self-references, which are associated with deception. This computer-based analysis gave an output accuracy rate of 67% in detecting the deception words in a constant topic, and the overall accuracy rate is 61%.

Lina *et al.* [55] started working on the topic of automated Linguistics Based Cues (LBC) to deception which was rarely considered earlier. As the language in text-based asynchronous computer-mediated communication (TA-CMC) is natural, it is complex and ambiguous, it makes it difficult to identify this data. So, the author has first transformed the data into a structured format. Now making use of language-based cues which include Morphological, Syntactic, and Lexical

Semantic cues, the author has applied them for analyzing TA-CMC. Along with these cues, the author also chose to add emotiveness, complexity, and pausality for analysis. When applied for analysis some cues gave out commendable results in detecting deception whereas some gave out new challenges to the author. Most of these cues were task-dependent and contextual, and all the cues opted are found to be potentially discriminators between real and deceptive data. According to the author, this methodology can be extended to cover wider criteria as all of these cues may not work or apply in all the contexts. The author aims to further work on the research to validate what LBC works on TA-CMC and what features will modify the patterns of the data. The author looks forward to applying these methodologies in a real-time environment and also conduct these on wider word levels. The author also says that though NLP techniques for pragmatic analysis are not mature, they can still perform discourse analysis which can be used for deception detection as they will enhance the effectiveness of the methodology. Finally, the author looks into examining the messages of real-life deceivers to test LBC's effectiveness in identifying deceptive information.

Hadeer *et al.* [56] research paper is based on making use of Machine Learning Techniques and N-Gram Analysis to detect online fake news. According to the author, the limited amount of resources makes it challenging to detect fake news. So, the author chooses to apply a comparison between machine learning techniques and n-gram analysis results. Along with these, other classifier mechanisms were also used for detecting fake news. For experimenting, the author chose a dataset containing real and fake news and has employed a word-based n-gram mechanism to generate the features. This classifier helped to differentiate fake and real news effectively. Before picking data, it was pre-processed to remove any stop words, and stemming was performed on the data to bring the words back to their original form. The author used six machine learning algorithms: Stochastic Gradient Descent, SVM, Linear SVM, KNN, and DT. To implement these classifiers, the author made use of the Python Natural Language Toolkit. While the author conducted experiments, Hadeer *et al.*discovered that high accuracy was achieved when a Linear SVM classifier was used, which equals 92%. Similarly, the author used many baseline word-based n-gram features and then examined the n-gram lengths effect on the classification algorithm's accuracy. The author states that rather than the original approaches that were observed earlier, these worked better and has a plan to run their approach on the LIAR dataset, which is publicly available for future works to evaluate the accuracy of the results.

The author Kai *et al.* [57] opinioned that earlier mechanisms of fake news detections concentrated only on claiming to check facts and deception classification. But in this research paper, the author opts to propose a Tri-Relationship Fake News detection framework (TriFN) by correlating the publisher bias, user engagements, and news stance which will contain additional information helping in the detection of fake news. The author also promises to give two real-world comprehensive fake news datasets which will facilitate the fake news research which will demonstrate the effectiveness of the author's proposed approach. The TriFN helps to extract the features from publishers and user engagements separately and parallelly capture the interrelationship between them simultaneously. Experiments were conducted by the author on the datasets using the given framework which gave out good detection performance in the early stages itself. Further, in the future author looks to explore new features of the datasets as features change regularly because fake news evolves on social media very fast. Also, the author believes that investigation is needed to extract fake news features based on psychological perspectives. The author feels that the identification of malicious and low-quality users is important to stop the intervention of fake news in the future.

2.2.2 Credibility based analysis

A. Message credibility

Zhou *et al.* [58] has worked on a theory-driven model to detect fake news. Investigation of the news was conducted at various levels semantic, syntax, discourse, and lexicon levels. The author tried to represent the news at every level by relying on established forensic and social psychology theories. The author's investigation model looked into two content-based and propaganda-based approaches by exploring the relationship between fake news, deception types, and clickbaits. This investigation makes use of a machine learning framework, and experiments are conducted on real-world datasets. The experiments conducted using propagation and content-based approach gave out an accuracy of 88% for detecting fake news. The accuracy rate ranges between 80 and 88% based on the article's news and data sizes. The higher accuracy during the investigation is achieved in the cases where limited prior knowledge of given news is known. Also, the author has observed that fake news holds differing characteristics like quality, content style, and sentiments similar to cognitive information in comparison to real information. According to the author, news to be considered fake must match with all the criteria established by the author in the investigation. On the other hand, the author also says that news to be considered fake it has to be applied with a larger real-world dataset and fundamental theories. Also, utilization of news images and rhetorical relationships is a better way to estimate the fake news, according to the author, who considers these as a part of his future investigation.

When Nguyen *et al.* [59] felt the earlier fact-checkers for fake news were insufficient and needed a better approach, the author worked on this paper. The author chose to propose an application by considering online users named fact-checkers, linguistic characteristics of fact-checking tweets, and finally offered to build a framework that generates responses with an intention to check facts in the fake news. The author observed that individuals tend to consume news that they believe, so the author looks into focusing on fact-checkers to convey verified information to the readers. The author has assumed some articles that require or prefer fact-checkers and then inserted the URLs into those responses. This framework, Factchecking Response Generator (FCRG), has approached 30% improvement in fighting fake news by employing fact-checkers in the articles. But the author considers this automation a partial one and looks into selecting a fact-checking article based on the content of original tweets to automate the whole process of fact-checking in the future. Also, the present word-based recurrent neural networks (RNN)s , cannot identify the rare words, so the author looks to develop a character-based RNN to address this issue better in the datasets. Rather than using fact checking tweets and original tweets, the author looks to use other data sources for better fact-checking in the future. The present framework model of the author has generated positive responses for fact-checking, giving out qualitative and quantitative output, and has opened new research areas for the author to apply them in more online social systems.

The research paper of Karishma *et al.* [60] talks about the survey for identifying and mitigating techniques to combat fake news. The paper talks about the present methodologies on this line and the developments required further to make it more efficient. The author compiles and summarizes the features of the datasets for developing an effective solution. After discussing the available mechanisms which need improvements, the author talks about future developments to be done on the same which are developing a dynamic knowledge base in the fact-checking methods as they determine the truth in the news article effectively as it will regularly get updated and reflects the latest changes quicker. To develop an intervention strategy, study on user relations at a microscopic level and macroscopic impact on the surrounding environment is very important. This will help in finding why there is a spread of news and how it can intervened effectively. Rather than binary labeled information in the current dataset, it is better to have a better fine-grained classification of information for identifying the fake news efficiently.

When Chan *et al.* [61] has considered the survey of the Korean Journalist Association, the author came to know that 74.8% of people in Korea did not trust the news that is spread through Social Networking Services (SNS). So, the author has developed this paper, which works to extract the sentences from the data corpus considered facts and check if they are true or false. For this purpose, the author uses a Bidirectional Encoder Representations from Transformers (BERT) model for creating a pre-training model. Datasets are created by collecting data from everywhere, out of which the author further develops true and false datasets. Making use of the BERT network, Korean BERT pre-training models are created.

Now when the data is passed through pre-processing layer, it picks out some nouns and verbs that are repeatedly observed in the news. Now, during the relevant sentence extractor level, using the Word Piece Model (WPM), a sentence is picked that is relevant to the input sentence from the fact data's corpus. The BERT pre-training model performs fine-tuning using the data set for the fake detection problems. AURAC curve helps the author check the execution of the model whose aim is to detect fake news. When the author performed the above methodologies, the results were successful, giving an accuracy of 83.8%. The author looks forward to continuing the research further to expand the investigation to other fields apart from the news.

Nicole *et al.* [62] works on this paper to develop a method for automating the process of fake news detection. But the major automation processes make use of machine learning algorithms which sometimes lack reliability facing black-box problems. The author looks deeply into the black box problem, testing the transferability of this learning process. To conduct the experiments, the author picked fake news articles from Kaggle news and real news articles from the New York Times and The Guardian. The author introduces a procedure firstly to remove any traces of source-related correlations. For testing, the author picks a training set that has a topic removed and known as a holdout, and the detector is evaluated on this topic. Next, for detecting fake news, the author introduces a deep neural network and a procedure for visualizing the patterns, which will help classify news as fake or real. The neural network methodology of the author has shown the best results to detect fake news only related to language patterns. Accuracy levels were good, but they varied based on the holdout and training datasets opted for the investigation. According to the author, while an investigation on the black box problem was done, there was a bias in the language in the fake news dataset. According to the author, research must continue further to check if these language patterns can assist humans in detecting fake news.

The work proposed by Vivek *et al.* [63] is to address the spreading of false information online rapidly. For this, the author proposes a novel text analysis that has a computational approach for detecting fake news. To investigate, the author picks fake datasets from Kaggle fake news and real data from three sources, including the New York Times. The author aims to identify the features of fake news, develop a machine learning methodology for the identification of fake news with an accuracy of 87%, and create a valid news article dataset. For obtaining linguistic features, the author makes use of the LIWC package for each article which was further normalized using Z-score normalization. Machine learning models like logistic regression, SVM, RF, K neighbors classifier, etc., were applied to the datasets to check the performance of the algorithm. Out of all the algorithms, the SVM worked the best and predicted the fake news better. The author proposes future research by making use of multiple features for creating comprehensive detection modules for fake news rather than using one feature. According to the author, a combination of text-analysis-based features will give out an accuracy of 87%, as expected by the author, motivating the author for continuing the research.

Zubair *et al.* [64] works to make use of machine learning techniques like SVM, RF, NB, DT for enabling the user to filter or classify some false news. The author chose some sets of bogus and true news articles for this. The author looks to develop a classification approach basing on the texts of the news articles that he has picked. While performing the classification, the above-mentioned classifiers were used for AdaBoost and Bagging. These tests were performed, making use of Pucharm in the Python environment. The classifier's quality was measured using classification metrics like recall, ROC, F-score, accuracy, and precision. All these analyses of the news have shown the author how to find the source of fake information and detect fake news. While classification metrics of different classifiers were compared, the best-performed classifiers were the AdaBoost-LinearSVM and Bagging-LinearSVM, with an accuracy of 90.7% and 90.02%. While the AUC of the curves was considered, then Linear-SVM and Bagging LSVM gave an accuracy of 94%, 97%. The author prefers to continue the research for investigating the results on Twitter datasets to compare different datasets results.

The research paper by Michail *et al.* [65] has a novel statistical approach for generating feature vectors for describing a document. The author's class-label frequency distance (CLFD) mechanism is applied for providing an effective way to boost the performance of machine learning methods. These experiments on the domains of fake news detection will help in verifying the machine learning methods' efficiency while they use the vectorization approach. According to the author, the vectorization technique will outperform the deep learning methods, which use word embeddings for small and medium-sized datasets and also in large datasets. The author will also demonstrate a novel hybrid method that utilizes both a CLFD-boosted logistic regression classifier and a deep learning classifier which will even work the best on large datasets. The CLFD technique proposed by the author will provide a weighting scheme that will give relevance to every term in the classification. The best part of the CLFD is it doesn't get affected by the quality or absence of data preprocessing. Thus, the employment of CLFD by machine-learning methods will help in developing a content-based fake news detection system that achieves high performance and takes a minimal amount of classification time. When used as a component in the novel hybrid methods, CLFD will outperform the deep learning methods in all the classifier metrics for any size of the dataset. For further future investigations, the author has already applied CLFD in a sentiment analysis setting and wants to apply it in other domains also. The author looks to improve the hybrid method for equipping it with a sophisticated and better scheme.

Rohit *et al.* [66] looks to address fake news which is a major threat to many sectors by proposing a deep convolutional neural network -FNDNet. The author prefers to stop relying on hand-crafted features and design a model which will help automatically learn various discriminatory features for the classification of fake news making use of multiple hidden layers in the deep neural network. The CNN will help in extracting features at each layer whose performance can be compared with baseline models. Rather than the existing state of the art results, the evaluation parameters like true negative, precision, accuracy, and F1 score were employed by the author to validate the results of the models. These results had major improvements in the detection of fake news as per the approach proposed by the author. The FNDNet methodology made use of GloVe for pre-trained word embedding during analysis. This proposed model of the author gave out the results with an accuracy of 98.36%. In the future, the author looks to work on BERT for pre-training the models for fake news classification. Also, the author aims to look at fake news detection with the help of echo chambers in the future. The author looks to address the limited research in visual information and develop a video or image-based analysis for video forensic investigation. The author concludes the research saying multi model based approaches including facts and knowledge will help in addressing the fake news issue efficiently.

B. User credibility

Zhou *et al.* [67] look into conducting a case study on various Chinese Microblogging Networks that analyze the features of spammers in these networks with active honeypots. For this research, the author talks about the study on the honey spots picked from Sina Weibo and Tencent Weibo. From these features like spamming strategy, account age, activity, and social information were studied by the author. The research led to the study of spammer characteristics from these were observed which will help to further study the automatic detection of microblog spammers. The author has opted for WebDriver Automation for implementing the honey spots making it easy for application and coding to the network environment. These honey spots post microblogs that are followed by users. At the end of the experiment, the following users have been scanned by the database to classify them either as spammers or non-spammers. One is considered to be a scammer if he posts any microblogs having URLs that contain information about malware, phishing sites, or sales information. At the end of the research, the author concludes that spammers tend to follow many users for popularity as they wait to be followed back, also, they post many microblogs often but mix them with ordinary ones to avoid detection and most of the cautious spammers do this to live long on the network avoiding detection.

Alex *et al.* [68] works to conduct research on Twitter, to detect any spam bots using a machine learning approach for distinguishing spam bots from the normal ones. For this purpose, three features that are graph-based like a number of followers, friends, follower ratio are extracted for exploring any unique friend or follower relationships among the users. Along with these, content-based features duplicate tweets, number of replies, number of HTTP links are also extracted from the recent tweets of the user. The datasets for the experimentation are collected by the user through API methods on Twitter for detailed information of the user and Web Crawler is developed for extracting unauthorized user's tweets. The author looks to apply classification methods, like DT, neural networks, SVM, and KNN, to identify spam bots on Twitter. Out of all the classification methods, the Bayesian classifier has the best performance as it is noise robust and the class labels are predicted based on the specific patterns of the user according to the author.

The author Putra *et al.* [69] after getting highly inspired by the success of deep learning algorithms in computer vision, for automatic feature extraction and representation, proposes a DeepProfile, a deep neural network (DNN) algorithm for dealing with fake accounts issues. Rather than a standard machine learning algorithm, the author constructs a dynamic CNN for the classification of fake profiles. The author proposes a pooling layer for optimizing the neural network performance in the training process. Along with calculating the loss and accuracy, the author also conducts the evaluation metrics for measuring the performance of the classifier using the ROC and AUC curves. The author suggests using the pooling function along with a deep-profile network on a large dataset like OSN to obtain better performance in the CNN graph. The deep profile algorithm can be considered very effective for dealing with fake accounts as they give an accuracy of 95% on the AUC curve. Also, there is a smaller loss when this algorithm is used rather than the standard algorithms in general. For future research, the author looks to explore semantic network characteristic structure rather than the information of nodes. There is also scope for exploring malware hierarchy links using this algorithm.

The author Jalal *et al.* [70] has identified deceptive information in many user profiles on the online social network. The author proposes a set of analysis methods using novel approaches for detecting this deceptive information about the locations and genders of the users on Twitter. For this, the author first collected a large dataset that includes profiles and Tweets from Twitter. For collecting the datasets, the author has run a crawler on Twitter's programmable interfaces and prioritized these features like temporal and spatial information on each profile. Now, the author has defined methods for guessing the genders making use of colors and names of Twitter profiles. After this, the author applies K-means clustering and Bayesian classification algorithms on the datasets that contain characteristics like spatiotemporal information, user names, profile layout colors, and first names, and also analyzes geolocations of the user behavior. While experiments were conducted by the author, they have shown efficient results with an accuracy of 90% in some cases. The author looks to explore other alternative strategies in the future for improving the accuracy of this news and looks to implement a synthesis of the two approaches that were implemented presently for detecting the deception. Also, the author looks to consider the genres of friends and followers, age factors on Twitter along with exploration of text-based features like user postings in future research for detecting the deception.

Buket *et al.* [71] conducted a study to present a classification model that will help in detecting the fake accounts on Twitter. The author has opted for a feature-based detection approach to identify fake accounts. The approach monitors the user behavior like friends, tweets, etc. To cater to this, the author has constructed their own dataset using Twitter API as there are no public datasets for this. For server-side scripting, language requests can be given to Twitter, and the results are given in JSON format, which the author can read easily. Out of all the attributes like places, entities, users, and tweets, the author chose some to be part of the dataset. The dataset has been preprocessed by the author using Entropy Minimization Discretization, a supervised discretization technique. The author analyzed the dataset using the Naïve Bayes Algorithm which gave out results with an accuracy of 84.5% for detecting spammers with the help of more attributes. Though the method can be employed on small attributes, the results are not that promising. But when applied to selected features, the algorithm gave results with an accuracy ranging from 85.55% to 90.41%. The author looks to explore more using this algorithm on other platforms like Instagram, Facebook, and LinkedIn and even applying similar algorithms like SVM and Bayesian network to locate any repeated tweets in the fake accounts. There is scope to enlarge the dataset using synthetic data for normalizing the fields to get balanced results.

In contrast with only a set of features for classifying spam and non-spam users, this paper of Malik *et al.* [72] proposes a hybrid technique making use of content-based and user-based, and graph-based features for identifying the spammers on Twitter. The author makes use of these three features and creates a model for evaluation. The Twitter dataset is used to be evaluated by the techniques; after classification is done based on the features, correlated features will be eliminated by the author. For the classification, techniques like J48, NB, and Decorate are used by the author. Out of the classifiers, available J48 has given the best output when applied to three feature classes. Whereas NB has given a poor performance when applied to the feature classes. Results of the classifier are also compared using graph-based and user-based features. But the classification using graph and content gave the best performance with an accuracy of 92% in comparison with user-based classification. So, based on this author states that user-based features don't play an important role in identifying spam users. The author prefers to extend the evaluation process using hybrid features on other platforms in the future.

Fatih *et al.* [73] conducted a study to detect automated and fake accounts on Instagram to address the fake engagements. The author has generated two datasets for automated and fake account detection. Machine learning algorithms like neural networks, SVM, logistic regression, and NB have been used on the datasets. The author talks about the derived features for the classification of automated and fake data. Also, to detect the best features of automated accounts, the author proposes a cost-sensitive genetic algorithm; the author uses the Smotenc algorithm to address the unevenness of the fake account dataset. While the NB algorithm is applied, it exploits independent features of different classes, and MAP estimation will be performed. Even the Logistic regression algorithm also will exploit independent features for differentiating two classes of the dataset. In comparison, SVM will find a hyperplane that will separate the dataset in a better way. The author also uses raw data as input while testing these algorithms on the datasets with preprocessed data. When SVM and neural network methods were applied on the automated accounts, it gave out results with an accuracy of 86%, and on fake accounts, it gave an accuracy of 96%. The author looks to address the biased features of the automated account dataset in the future by finding the real users. Also, Fatih *et al.*looking to use recurrent neural networks to detect automated accounts in a better way in the future.

2.2.3 Propagation based analysis

Soroush *et al.* [14] investigated the differential diffusion of the verified false and true news stories from Twitter. The news was classified into true and false by making use of information from different fact-checking organizations which classified the data on the same line for up to 98%. According to the author, falsehood news tends to spread faster among people when compared with the real news. Out of all the news, the news pertaining to political views tends to spread faster. The author assessed the perception of the users on any information by comparing their emotional content of replies to false and true rumors. This emotion-based categorization was done using the leading lexicon of the National Research Council Canada (NRC), which provides a list of different emotions. While evaluating URLs and stop words were removed from the news. The author found that false rumors showed more emotions like greater sadness, disgust, etc. These factors inspire to spread that news faster, according to the author. Also, the news was fact-checked by many classifiers, including Massachusetts Institute of Technology (MIT) and Wellesley colleges, for identifying false rumors, which stated that false news spreads faster. The author also used the bot-detection algorithm to identify and remove all bots before running the analysis as they are considered the reason for the faster spreading of false news. The results were the same, indicating that false news spread faster than real news even in the absence of bots as it is also based on human behavior. All the above methodologies used by the author stated that false news spreads faster. The author looks to conduct more research on the behavioral explanations of humans in the diffusion of true and false news. Interviews, surveys, lab experiments must be conducted to identify the factors of human judgment, which will spread the false news faster. According to the author, future research on identifying the spread of false news requires large-scale system analysis rather than ad hoc analysis.

Yang *et al.* [74] research paperwork for the early detection of fake news through propagation path classification with convolutional and recurrent networks. As the majority of the time, information to detect fake news is inadequate at the early stage, a novel model is proposed for early detection. To cater to this, the author first models the propagation path of news in the form of time series that are multivariate. Every tuple of the time series acts as a numerical vector that represents the user characteristics who has been engaged in spreading the news. After that, the author builds a time series classifier that will incorporate both convolutional and recurrent networks helping capture local and global variations of user characteristics to detect fake news. When experiments were conducted by the author on three datasets, it has demonstrated that the proposed model is able to detect fake news with accuracy 92% on Sina Weibo and 85% on Twitter within five minutes as the news starts spreading, which is very quick according to the author. These quick findings are possible as the author makes use of common characteristics that are robust, reliable in comparison with structural or linguistic features in general. As the user characteristics seem very effective author prefers to work more and investigate if these characteristics will help the author find and identify the users who are prone to be impacted by such information and tend to spread the news. Along with this, the author has a plan to incorporate PUlearning techniques into the present model, which will help in dealing with massive unlabeled news stories in an effective manner.

As the detection of fake news through Twitter is confined to content or user-based approaches, Marion *et al.* [75] looking to develop a broader methodology for tweet retrieval, which even includes the tweets without an URL link. For this, the author chooses to use propagation structures. To start with, the author shows that real news is bigger and is spread by those users with high followers and low followings and tends to be spread longer than fake news. Also, the author finds the real news graphs bigger in size in comparison with the fake news graphs. When the author started experimenting, he found the RF Classifier worked best and achieved a detection accuracy rate of 87%. When the author developed a graph neural network on the 3D representation of the propagation graph, it achieved an accuracy of 73.3%. These two accuracy levels show that propagation structures are efficient and relevant to detect fake news on Twitter. The author suggests future research must be dedicated to evaluating these classifications further and also for refining the data sets to counter any negative impact of the broad definition of news. For this purpose, the time limit on tweet retrieval must be set; and this approach can be applied to a wide range of topics.

The author, Zilong *et al.* [76] feel that existing studies on fake news focused more on theoretical modeling of propagation or identification methods using machine learning; the author proposes to understand the realistic propagation mechanisms between theoretical models and black-box methods. To cater this, the author has made use of large databases of real and fake news that are picked from Twitter in Japan and Weibo in China. The author finds that there are cases of posting fake news five hours after the real news is posted. So, the author demonstrates his findings which will help in understanding the various propagation evolutions of real and fake news. As the earlier studies were different, this topological property identification at early stages will help in identifying and curtailing the spread of fake news at the early stages of spreading. The author distinguishes real and fake news using topological measures like characteristic distances, the ratio of layer sizes. Using these measures, classifiers can be constructed, and similarly, when RBF kernel is applied on the datasets, then there was an accuracy of 79.5% for detecting fake news. The author makes use of cascade components of the propagation network, which are mainly topological features, and these results show that there is a significant difference in the news at the early propagation stages, which must be addressed.

Sejeong *et al.* [77] in the research paper, look to identify the prominent features of rumor propagation. For this purpose, the author identifies the rumor characteristics by following these aspects of diffusion, namely linguistic, structural, and temporal. For the sake of temporal characteristics, the author proposes a periodic time series model which will consider external and daily shock cycles showing fluctuations over time. Also, the linguistic and structural differences will help in identifying the spread of non-rumors and rumors. The features picked by the author for classifying the rumors have given out output with accuracy ranging between 87% and 92%, which is very high. The author chose websites like times.com, snopes.com, pcmang.com, and more to collect non-rumor and rumor cases. The proposed Periodic External Shocks (PES) and Linguistic Inquiry and Word Count (LIWC) models of the author effectively captured temporal, linguistic, and structural patterns to distinguish non-rumors and rumors effectively. To apply these models, the author made use of fifteen features for effective testing of the models. Also, the author made use of logistic and forest models to evaluate which features gave the utmost information. As there is a combination of linguistic, structural, and temporal features, this integrated set gives accurate results for identifying rumors and non-rumors.

Carlos *et al.* [10] work on the research paper is for analyzing the information credibility of the news on Twitter. To do this, the author focuses on making use of automatic methods for accessing the credibility of news on Twitter. The author analyzes the microblog postings which are related to the trending topics and then, based on the features extracted, classifies them as credible and not credible. The features chosen are from those texts in the posts that are regularly reposted by the users and also from external sources. The author makes use of Mechanical Turk to separate topics, for example, those that spread information about a news event. Also, the author applied evaluators on each topic for providing short descriptive sentences for that topic. This will help in discarding answers those do not have proper justification, thus reducing the click spammers during the evaluation. Next, for performing a credibility assessment, the author made use of a supervised classifier. This classifier will help in determining if the dataset is a newsworthy set. This has proven that newsworthy topics can be separated from other conversation types. After selecting the dataset, the author evaluates the methods using human assessments regarding the credibility of the datasets. The results of these evaluation methods measured the differences in the data and classified them as credible and not credible with an accuracy range of 70%-80%. As part of future work, the author wants to extend the experiments to larger and partial datasets for exploring different factors that declare the datasets as credible.

The present content-based analysis using language processing algorithms for detecting fake news lacks common sense or social context, according to Federico *et al.* [78]. Rather than content-based, the author prefers propagation-based algorithms. In this paper, the author suggests a novel automatic fake news detection model which is based on geometric deep learning. The algorithm of this model allows the fusion of data such as user profile, social graph, content, news propagation, and activity. When experiments were conducted on the news stories using this model, the propagation and social network structure gave out a high accuracy rate of 92.7%, according to ROC AUC in detecting the fake news. The author also stated that one could detect fake news in the early hours of propagation also, and calls this model as the best alternative to content-based approaches of fake news detection. The advantage of using this geometric deep learning method is it makes use of graph-structured data, and it automatically learns task-specific features from the data. And the experimental validation of the model is independent of language and geography and is based on spreading the features and connectivity between them. The author feels that the study of adversarial attacks is great both from theoretical and practical viewpoints, and these will allow the exploration of any limitations of this model and estimating its resilience to attacks. The author looks forward to exploring and applying the model on news topic classification and virality prediction along with fake news detection.

Chapter 3

FAKE NEWS DATASET

3.1 Introduction

In today's world, machine learning is very helpful and popular among decision-makers. Still, many decision-makers are in a dilemma which design and train are needed exactly to deploy a machine learning algorithm. The details on how to collect the data, building a dataset, and annotation specifics are neglected as supportive tasks.

Data is an integral and important part of all AI models. It can also be described as the sole reason behind the popularity of machine learning among decision-makers in today's era. Since the data are available, scalable ML algorithms became viable. It has become an actual product that can bring value to the business, rather than being a by-product of its principal processes.

Dataset, as define by Oxford Dictionary, "is a collection of data that is treated as a single unit by a computer". It means that a dataset has different and separate parts of data that can be used to prepare an algorithm with the aim of obtaining predictable patterns inside the whole dataset.

Normally, the dataset is not used only for training purposes. A single dataset is divided into three parts; (i) training dataset checks how well the training of the model went, (ii) test dataset is used to check the model performance and (iii) validation dataset helpful to avoid training the algorithm on the identical type of data and producing biased predictions.

3.2 Publicly Available Datasets

Here one can see the information of many publicly available datasets. Reading them will help understand the characteristics of different datasets.

i. Benjamin Political news datasets [79]

a. As the name indicates, the dataset holds the information of politics.

b. The dataset holds information of around 225 stories, including fake news articles along with real news articles and satire based articles.

c. The Benjamin dataset will help in detecting any fake news that is related to politics.

d. The media from where the datasets are created are collected from the mainstream media.

e. The articles in the datasets were picked between the years 2014-2015.

ii. Burfoot Satire News [80]

a. This is basically a dataset, which is unbalanced.

b. The dataset includes satire and real categories of news.

c. But the news, in this dataset, is not confined to any predefined timeline.

d. It includes news articles that are related to politics, society, technology, and economics.

e. It is mainly used to detect satirical news from mainstream media platforms.

f. The dataset holds around 4233 sample news articles.

iii. BuzzFeed News [79]

a. The dataset is very small, holding only 101 news articles.

b. These articles in the dataset include both real and fake news.

c. The dataset helps in detecting the fake news related to politics.

d. The dataset includes news mostly from social media platforms like Face book.

e. Also, the data is picked from a predefined timeline which is 2016-2017.

iv. CREDBANK [81]

a. The dataset includes information of 60 million tweets that are mostly related to the societal rumors.

b. The dataset works for veracity classification same as fake detection.

c. The dataset contains rumors i.e. unsubstantiated claims.

d. The articles in the CREDBANK dataset are picked from the social media mainly from Twitter.

e. Extraction of these articles was done in predefined timeline ranging between 2014-2015.

v. Fake News Challenge [82]

a. The dataset contains news related to politics, society and technology.

b. It contins 49972 news articles from mainstream platform.

c. The dataset includes four category of news article i.e. Agree, disagree, discuss and unrelated.

d. This dataset is majorly used for fake detection.

e. News extraction period is not specified in the dataset..

vi. FakeNewsNet [83]

a. The articles included in the dataset are related to society and politics.

b. The content in this dataset includes both images and text.

c. The data of the dataset is included from the social media and mainstream platforms but mainly includes data from Twitter.

d. The information in the dataset is not confined or picked from any predefined time period.

e. Dataset contains 422 news articles of two categories i.e. real and fake.

vii. LIAR [84]

a. The dataset holds information from 12836 short statements.

b. The dataset holds the information that is related to politics.

c. The data for the dataset is collected from Facebook, Twitter and main stream platforms.

d. The extraction of information for the dataset is performed between the years 2007-2016.

viii. Reuters [85]

a. The dataset holds information of both real and fake news.

b. The real news are collected from reuters.com which is a news website whereas fake news are collected from Wikipedia and some websites that the Politifact flagged.

c. Though there is information about various topics, majority of the data in the datasets is related to the world news topics and politics.

d. Most of the articles are collected in between the timeline 2016-2017.

e. There are around 21417 articles in the real news category and 23481 fake articles.

ix. McIntire [86]

a. The dataset holds both fake and real news.

b. Fake news is collected from the dataset of Kaggle where it even comprises the news about 2016 USA elections.

c. Real news is collected from various popular media organizations like Guardian, NPR, Bloomberg, WSJ and New York Times mostly from the years 2015 and 2016.

d. The dataset holds information of 6.3k news articles with equal number of real and fake news whereas half of the information is related to the politics

x. Kaggle [87]

a. There are around 38729 data on kaggle that are publicly available.

b. It includes both fake and real news that are collected from multiple websites.

3.3 Limitation of Dataset

i. *Benjamin Political News:* The dataset is very small as compared to the other dataset. It has around 75 stories each for real, fake, and satire categories.

ii. *Kaggle:* The dataset is not very reliable as it consists of real and fake news data without any source of information.

iii. *Burfoot Satire News:* The dataset is not a balanced dataset. It consists of real and satire categories in an unorganized manner.

iv. *McIntire:* This dataset has no authentic source behind it. It consists of real and fake news categories which are not backed by any individuals.

v. *BuzzFeed News:* The dataset is very small as compared to the other dataset. It consists of only 101 news items.

vi. *Reuters:* The dataset relies on a single source only. The real news articles come from a single source. It increases the possibility of biased data.

vii. *CREDBANK:* The dataset has incomplete data news articles in real and fake event categories.

viii. *LIAR:* It has complicated data in it. It consists of social media posts and speeches that are hard to understand and classify. It also lacks the verified sources or the knowledge base which could clarify the doubt.

ix. *Fake News Challenge:* The dataset mainly concentrates on the individual claims among three categories: discuss, disagree, and unrelated.

x. *FakeNewsNet:* It is a limited dataset. It consists of 422 news articles only. It also does not have a clear classification of real and fake news articles.

3.4 Proposed Dataset

Balanced dataset is a very important part of any model in terms of accuracy. It also helps in providing good quality training data and delivering genuine results. After comparing many datasets, WELFake dataset constructed that combines four datasets: Kaggle, McIntire, Reuters, and BuzzFeed. We have selected these four datasets because of two reasons:

i. They all have a similar structure of dataset which are divided into two categories: real and fake.

ii. When we combine the datasets, it reduces the chances of biasedness and also reduces the limitations of each individual dataset.

WELFake dataset has 35028 real news out of which 10387 real news of Kaggle dataset, 3171 of McIntire dataset, 21417 of Reuters dataset and, 53 of BuzzFeed

Political dataset. Similarly, the WELFake dataset has 37106 fake news out of which 10413 fake news of Kaggle dataset, 3164 of McIntire dataset, 23481 of Reuters dataset and, 48 of BuzzFeed Political dataset [88].

We also have a summary of the balanced fake and real news division in the WELFake dataset across following feature categories:

i. The number of short sentences: (below ten words) that represent real news value is 60.9% and the fake news value is 39.1%. It is clear that the real news value is greater than those representing fake news.

ii. Text readability: The real news value is 51.7% and the fake news value is 48.3%. It shows that fake news has poorer readability than real news.

iii. Subjectivity: The real news value is 45.4% and the fake news value is 54.6%. It indicates that the subjectivity of fake news articles is larger than for real news articles.

iv. The number of articles: The real news value is 53.9% and the fake news value is 46.1%. Its representation of the real news is larger than those representations of fake news.

3.5 Summary

WELFake dataset contains balanced articles with 48.55% real and 51.45% fake news. It contains three input features i.e., serial number, news heading and news content; with one output feature with binary label i.e., zero for fake and one for real news. As per dataset there are 78098 articles but due to some unstructured information dataframe can access only 72134 articles.

Chapter 4

LINGUISTICS BASED FAKE NEWS DETECTION

4.1 Introduction

These days everywhere we see, any community or people of any age, sex are regular users of social media networks. Doing communication through these social media networks is attractive, fast and simple for transferring and sharing any information. When you look into various social media networks today, Twitter and Facebook are used by 1.3 billion clients and this makes terabytes of information is generated every second [89]. This wide range of information sharing is seen because it is convenient and simple to circulate or access information with others with the help of these networks. But the same convenient way of information sharing also results in fake news spreading and this is harmful to people and society. Fake information are of low quality which are generated either by bots or individuals are mainly for manipulation the messages with some plans in the mind.

During the last five months of the US elections, nearly 7.5 million tweets with one –sided news or fake news websites were seen. But the most worrying part these news websites having fake news attract high number of audience when compared with credible and real information. When research was conducted on these lines, it was understood that the false news spreads faster, enters and reaches various people quicker and show high impact than real. Everyday several cases are seen where people spread unverified news without evaluating its genuinity by any sources. Unknowingly, these people become part of that category of people who spread fake news.

It is high time that a solution is needed to address the spreading of fake news. Though researchers are trying for developing ML models along with different feature sets whose main automation target is to detect the fake news [90] making use of visual [91] or text-based linguistic approaches. In this chapter we are working on text-based linguistic approach. Large number of researchers have already done some work on it but the below questions are still left unanswered;

i. Out of all the linguistic features, mention those which will classify the data into fake and real?

ii. For improving the results of fake news detection will voting classifier work?

iii. Out of all the methods which of the classification method will work well for detecting fake news?

4.2 Proposed Methodology

This methodology is based on the linguistic based features. It does not make use of additional information of news for classifying fake news [92]. But it makes use of reformed state-of-art techniques for detecting the fake news by using word embedding (WE) and linguistic features together. The methodology will focus on three major points:

i. Makes use of linguistic feature for predicting fake news.

ii. Making use of WE and linguistic feature for detecting fake news over proposed dataset i.e. WELFake dataset.

iii. Finally going for comparison between the results of BERT and CNN methods that are based on linguistic features.

4.2.1 WELFake model

To overcome all the lapses in the state-of-the-art were improvised by the novel model named WELFake model which is based on three major points:

i. Identification of twenty linguistic features and then creation of three unique Linguistic Feature Sets (LSFs). Classify news based on these LSFs.

ii. Use of two WE methods for classification.

iii. Generating the final prediction making use of voting classification.

The WELFake model is divided into four phases which shown in figure 4.1. These phases preform following task:

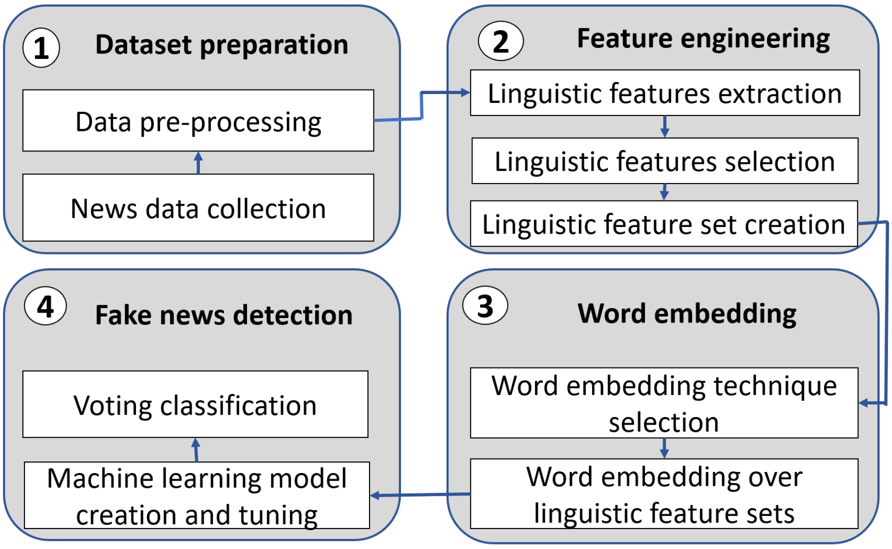
1 Dataset preparation
Data pre-processing
News data collection
2 Feature engineering
Linguistic features extraction
Linguistic features selection
Linguistic feature set creation
3 Word embedding
Word embedding technique selection
Word embedding over linguistic feature sets
4 Fake news detection
Voting classification
Machine learning model creation and tuning

Figure 4.1: WELFake model overview

Phase 1: Dataset preparation

It will help in solving problems in the data which can be unstructured data format, typographic errors. The methods in data pre-processing are:

i. **Missing data:** It talks about handling undefined data like NULL and NaN in the dataset. As there is a scope of loss of information[63] during this process, data imputation process is performed. This will help in estimate the missing values and will help analyze the dataset.

ii. **Inconsistent data:** To avoid the deviation from data points which can be because of any mistakes during the collection of the data. Mathematical functions and visualization techniques; like Inter Quartile Range (IQR) score, Z-score, scatter box, and plot box; were employed to identify and correct .

iii. **Duplicate data:** To remove any redundancy which can make biased results, de-duplication on the data is performed.

iv. **Irrelevant data:** This is nothing but words like stop words, which will make any sentence grammatically complete, but these words do not have semantic significance. So such stop words have to be removed for increasing the performance of the model.

v. **Stemming:** This step involves the conversion of the text into its own root word making use of the Porter-Stemmer algorithm which will give improved accuracy. If the root word is not recognized then the canonical form of the word is generated.

Phase 2: Feature engineering

This phase consider essential linguistic features which are responsible for the classification of news. Therefore following operations are performed:

i. **Linguistic features extraction:** This extraction talks about the conversion process of the raw text in number format for the ML algorithm. The extraction will aim for creating the feature set which will summarize the original datasets information for speeding the model training up thus improving the learning accuracy and data visualization. Essential 87 text-based linguistic features were extracted from the state-of-the-art works, where these fall in two semantic (i.e., psycho-linguistics , grammar) and syntactic (i.e., quantity, writing pattern,) categories. *(i) Writing pattern:* It will emphasize the texts writing style by accessing the types of sentence (long/short sentences), making use of modifiers, special characters, and determinants. *(ii) Grammar:* It will entirely focus on the text readability index and will emphasizes the sentence complexity, easy word use ratio in a word list, average syllables per word, and word structure. *(iii) Psycho-linguistic:* It will estimate the information opinion and text sentiment basing on subjectivity and semantics. *(iv) Quantity:* It will help in identifying the speech information by counting the words per sentence , adverbs, verbs, adjectives and its rates along with the syllables.

ii. **Linguistic features selection:** Here the selection of essential features diminish the input features quantity, will decrease the computational expense, and will improve the ML models accuracy. To do this, calculation of the Pearson's correlation coefficient of each feature with the rest of the features in the same category is done and those having a correlation coefficient of 0.7 or more are discarded indicating that there is a positive linear relationship in between the two features [93]. According to Occam's razor principle model having few features is precise and simple and minimum description length concept validate the same [94]. This was executed recurrently until inessential features are removed.

iii. **Linguistic Feature Set (LFS) creation:** To enable any WE method for unbiased model training, the group has opted for 20 linguistic features in various sets. For getting a clear result when voting classification is applied odd no of input sets are needed. For this purpose, three Linguistic feature sets are created basing on their: *(i) Readability index:* It will help in defining the structure of the sentence complexity in any text. The grade or level of the text writer is identified basing on the readability index which will in turn help classify the news. *(ii) Psycho-linguistic:* This plays an important role for detecting of fake news.*(iii) Quantity:* The features will participate in classifying the news so even distribution of these is done across the three LFS. *(iv) Writing pattern:* It will contain five features; so even distribution of three features is done to each LFS.

Phase 3: Word embedding

This phase will help in identifying the appropriate WE technique to convert the given plain text into a numeric value. This is done because any ML method cannot process text directly. There are two categories of word embedding in literature. They are: (i) Based on the content like CV and TF-IDF which will focus on the frequency of text where as (ii) Based on the context like FastText, GloVe, and Word2Vec which will focus on text patterns.

i. **WE technique selection:** The common pattern of fake news writers is they repeat the similar words. So we have opted for the first type, WE that is based on the content. It will focus on the frequency of words in the text and the writing patterns and do not focus on context [95]. CV is also known as one-hot encoding. It talks about the conversion of a text document in the histogram vector. In this method, each word will represent the word appearance number in the document. The count of unique words will determine the vector length. TF-IDF is described as the next advanced version of CV. It will describe the terms importance along with its occurrence in any given document.

ii. **WE over LFSs:** It will improve the prediction of the output. This is because the pre-defined features will not predict accurately and has the necessity for training methods additionally. To do this, WE and LFS are combined. When CV and TF-IDF are applied on the three LFSs. It was found that the CV is giving better results. The CV result has an accuracy of 95.61% using SVM, whereas TF-IDF has given an accuracy of 95.12% making use bagging. As the CV is giving better results, for further analysis it is combined with LFS to get high accurate results.

Phase 4: Fake news detection

This phase majorly performs model tuning and also takes the advantage of voting classifier for the improved result.

i. **ML model creation and tuning:** It passes the LFS with WE making use of six ML methods namely AdaBoost, Bagging, DT, KNN, NB, and SVM. To do this, an experiment making use of random samples on every ML model on the

WELFake dataset. Here four training and testing dataset split combination i.e. 90%-10%, 80%–20%, 70% – 30%, 60%–40% were used for obtaining better accuracy. For improving the accuracy levels manual tuning was performed on six models making use of hyper parameters. Exploration of different hyper parameters combinations were evaluated on the given values and it was continuously tuned to get accuracy of 96%. When the performance of every ML model on training and testing data distributions were performed, results have shown that when data is distributed at 70%-30% then high accuracy levels are seen.

ii. **Voting classification:** The output generated by voting classifier will minimize any errors and will avoid over-fitting. There are two voting classification approaches. The soft voting based classification method is based on the probability whereas hard voting based classification method is based on maximum votes. The detection of the fake news talks about the binary classification problem, therefore hard voting method is used for predicting the target variable Y basing on the maximum votes, which are given by various models M_i based on the X predictor or input variable.

$$Y = mode\left\{M_1(X), M_2(X), \ldots, M_n(X)\right\},$$

4.2.2 WELFake model workflow

Before talking about the algorithm, let us see the workflow of WELFake model shown in figure 4.2. The news binary classification method of WELFake makes use of the three LFSs, CV, TF-IDF and a hard voting classifier. WELFake workflow follow following steps.

i. Use of CV and TF-IDF on the given entire dataset.Then storing these results namely P_1 and P_2. Finally deciding a better Word Embedding basing on the accuracy.

ii. Use of the CV on three well defined Linguistic Feature Sets and then store these results in the P_3, P_4, and P_5.

iii. Use of the hard voting classifier on the P_3, P_4 and P_5. Now generating the prediction output as P_6.

iv. Making use of hard voting classifier and then combining P_1, P_2, P_6 for generating the final output.

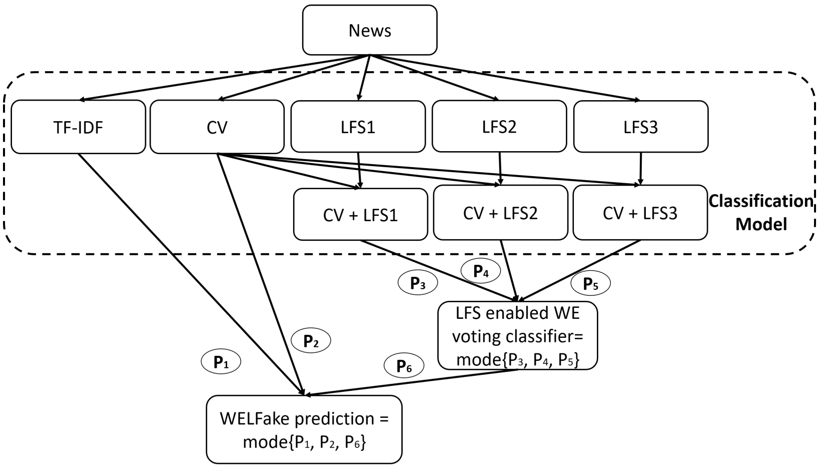
News
TF-IDF
CV
LFS1
LFS2
LFS3
CV + LFS1
CV + LFS2
CV + LFS3
Classification Model
P₁
P₂
P₃
P₄
P₅
P₆
LFS enabled WE voting classifier= mode{P₃, P₄, P₅}
WELFake prediction = mode{P₁, P₂, P₆}

Figure 4.2: Workflow of WELFake model

4.2.3 Algorithm for WELFake model

This algorithm will clearly explain the four phases in WELFake model as given in algorithm 1.

i. **Dataset preparation:** This phase of data collection is covered in the line 1 whereas line 2 will cover the data pre-processing mechanism.

ii. **Feature engineering:** The extraction of the linguistic features from the dataset is explained in the line 3 and the line 4 talks about the application of Pearson's coefficient for selecting the linguistic features. Creation of the odd number of LFS for voting can be seen in the line 5.

iii. **Word Embedding:** The whole mechanism of this phase can be seen in the lines 6 and 7. In the line 8, selection of the best WE technique is seen.

Algorithm 1: WELFake fake news detection algorithm.

```
    Data: Kaggle, McIntire, Reuters, BuzzFeed
    Result: WELFake Model for news classification
    // Phase 1: dataset preparation
1   WELFake_dataset ← collection(Kaggle, McIntire, Reuters, BuzzFeed) // News
      collection
2   WELFake_dataset ← preprocess(WELFake_dataset) // Dataset pre_processing
    // Phase 2: feature engineering
3   LF ← extract(WELFake_dataset) // Linguistic feature extraction
4   LF ← selection(LF) // Feature selection using Pearson's correlation
5   LFS ← split(LF) // Split linguistic features in odd sets
    // Phase 3: Word embedding
6   CV ← cv(WELFake_dataset) // Apply CV technique on dataset
7   TFIDF ← tfidf(WELFake_dataset) // Apply TFIDF technique on dataset
8   Best ← select(CV, TFIDF) // Select best embedding technique with
      dataset
9   foreach $LFS_i \in LFS$ do
10      $CLFS_i$ ← combine($CLFS_i$, Best)
    // Phase 4: ML model tuning and voting classifier
11  Model ← bestModel(SVM (CLFS), DT (CLFS), NB (CLFS), Bagging (CLFS),
      AdaBoost (CLFS), KNN (CLFS)) // Selection of best model for linguistic
      feature set
12  vote_hard ← votingClassifier(Model ($CLFS_i$));
13  return votingClassifier(TFIDF, CV, vote_hard)
```

Combination of the LFS making use of best WE method is seen in line 9 and 10.

iv. **ML model tuning and voting classifier:** Training is performed in the line 11 making use of line 5 on different ML classification models and the selection of the best results is picked from each set. Voting classifier is performed in the line 12 where the classifier is applied on the results that are generated on various LFS making use of the top best ML classification model and finally the hard voting output is generated. The line 13 explains the use of hard voting classifier on output of line 12, TF-IDF line 7, and CV line 6 again where the final classification prediction is returned.

4.3 Experimental Result

Six ML classifiers are used for the comparison of the WELFake model accuracy, as shown in table 4.1.

i. **LFS:** It will help classify the news based on the three LFS separately making use of the six ML classifiers. When classification were performed, the accuracy

Table 4.1: Accuracy analysis of WELFake model.

Scenarios $\xrightarrow{}$	*LFS*			*WE*		*LFSWE*			Final prediction
↓ *Model*	*LFS1 (%)*	*LFS2 (%)*	*LFS3 (%)*	*TF-IDF (%)*	*CV (%)*	*CV + LFS1 (%)*	*CV + LFS2 (%)*	*CV + LFS3 (%)*	Voting classifier (%)
KNN	79.6	81.8	80.6	89.6	88.2	90.3	90.5	90.1	90.16
SVM	82.5	83.5	85.6	94.5	95.61	95.6	96.1	95.01	**96.73**
NB	79.2	77.3	79.8	91.02	91.03	91.05	91.08	92.01	92.12
DT	81.4	79.6	80.8	89.54	89.51	90.1	89.61	89.68	89.92
Bagging	83.4	84.2	84.2	95.12	95.04	95.08	95.3	95.3	95.31
Adaboost	81.8	81.9	80.6	93.78	94.9	95.18	95.18	95.2	95.32

ranges between 77.3%-85.6%. The first LFS gave the top accuracy of 83.4% when Bagging was used whereas when NB was used it gave the accuracy of 79.2%. The second LFS gave the top accuracy of 84.2% when Bagging was used whereas when NB was used it gave the accuracy of 77.3%. Whereas, the third LFS gave the top accuracy of 85.6% when SVM was applied and when NB was applied it gave an accuracy of 79.8%. It explains that SVM and Bagging are the best performers of the classifiers in comparison with the other classifiers.

ii. **WE:** When TF-IDF and CV were applied on the WELFake dataset; the news was classified into P1 and P2 as shown in the figure 4.2. When results were evaluated, it is shown that CV has given out more accuracy than TF-IDF. Highest accuracy was received when SVM was performed on the CV giving out an accuracy of 95.61%. Whereas Bagging gave out accuracy of 95.12%.

iii. **LFS enabled WE:** Here, three LFS combine with CV to give P3, P4 and P5 and after this voting classifier is applied for obtaining P6. The maximum accuracy was achieved when SVM was applied which is 96.1% and minimum was 89.6% when DT was used.

iv. **WELFake prediction:** The final classification prediction outputs were estimated when voting classifiers were applied across CV(P_2), TF-IDF(P_1), and LFS enabled WE classification (P_6) predictions. The high accuracy of 96.73% was given by SVM followed by AdaBoost and other classifiers.

The table 4.2 will help compare the performances of the WELFake model in terms of F1-score, recall, precision, and accuracy. Going through table will help to understand which ML model on WELFake has given what results.

Table 4.2: Performance Metrics of WELFake model.

Model	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score (%)*
KNN	90.16	89.02	90.55	89.78
SVM	96.73	94.60	98.61	96.56
NB	92.12	91.45	92.25	91.85
DT	89.92	86.10	92.62	89.24
Bagging	95.31	91.78	98.46	95.00
Adaboost	95.32	91.81	98.46	95.02

4.4 Summary

Over 72,000 news articles of WELFake dataset used to apply a WELFake model for the detection of the fake news. After this, analyzing of 80 linguistic features was done out of which 20 were picked to increase the accuracy levels and to reduce the complexity of the computations. Six ML model were applied linguistic features, CV and TF-IDF. The results explained that accuracy of CV is better than TF-IDF. Later to this, embedding of CV with LFS was done for the sake of voting classifications. When experiments were further performed on the WELFake datasets the WELFake model gave an accuracy of 96.73%.

The implementation of the WELFake model can be further extended in the future to cover factors like user and message credibility for higher verification levels.

Chapter 5

MESSAGE CREDIBILITY BASED FAKE NEWS DETECTION

5.1 Introduction

News is the biggest source which will make the people around be aware of the day to day events happening in and around them. This news can affect the public socially or personally. One of the biggest platforms we come across today that is used for broadcasting or sharing information related to entertainment, political, or business news is the online social media platforms. Also, these platforms are easy to access, comfortable to read and also propagate the information real fast [96].

According to several researches, news is divided into binary class i.e., fake or real [5], but many others have preferred to opt for multi-class classification [97], clustering [98], or regression problem. Making use of an automated tool the user will be able to categorize and detect the fake news basing on three criteria:

i. *Propagation-based:* Using these methods one can trace the pattern of the spreading news making use of people's share and replies.

ii. *User profile-based:* Using these methods one can track the behavior of the individual making use of their commented, forwarded, or published news which will include further analysis information like friends, followers, sexual orientation, or location.

iii. *News content-based:* This is further divided as two types namely:

a. *Syntactic based methods* make use of writing and linguistic patterns like verbs, nouns, special character numbers for classifying the news.

b. *Semantic based methods* will help in the performing of the high level representation and text structure of the document.

5.2 Proposed Methodology

The chapter has been working out to propose a methodology in which a novel **M**essage **Cred**ibility (MCred) multi-modal method will try to approach the fake news in the form of a binary classification problem. Our method will combine the semantic relation of the global text between the words making use of BERT model with N-gram features of the local text semantics making use of CNN model. Our model will make use of local and global word embeddings as cues for classifying the news which are validated making use of the datasets for testing and training purposes.

In the final stage BERT and CNN outputs are passed to a dense network which will enhance the performance of MCred model. When this was done we have seen an accuracy rise of 1.48% when we compared it with other state-of-the-art methods [66].

5.2.1 MCred Model

Figure 5.1 represents the architecture of MCred model. It has two phases: (i) Data engineering and (ii) Model generation.

Phase 1: Data engineering

Making use of the data engineering, the suitable datasets for the MCred model were selected out of the many options available. To collect the suitable data following steps are performed:

i. *Data collection* – This is explained as the selection of the datsets from many fake news datasets available. WELFake dataset is selected for construction of

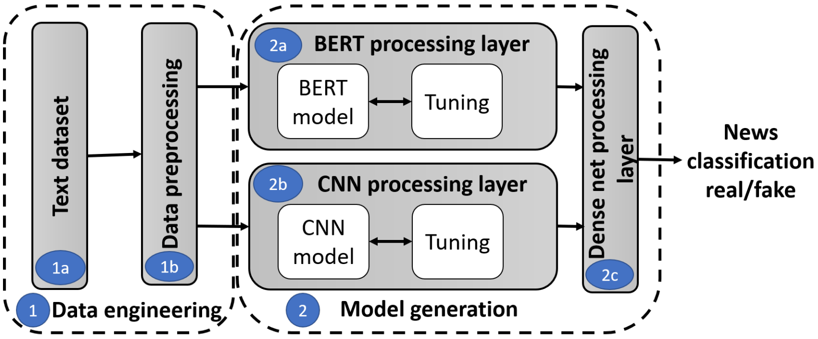

Figure 5.1: MCred model flow

MCred model because this dataset is considered as unbiased dataset. Picking a large dataset will help in the prevention of the model from the issue of overfitting and it will help to enable better training of the model.

ii. *Data preprocessing* – Here tasks like tokenization, normalization and noise removal are performed for keeping the available data in a proper format.

Phase 2: Model Generation

This model generation sees a fusion of local and global text semantics for the classification of the fake news and involves three sub layers.

i. *BERT processing layer* will read the data from the previous pre-processing layer and will pass the data into the pre-trained model which is tuned for classification. The layer will generate the global semantics after it measures the relation between the upcoming, previous, and current words of the text. At the end, the output will be passed into the dropout and dense layers.

ii. *CNN processing layer* will read the pre-processed data and will convert into the GloVe embedding. The embedding are further passed through CNN layers that are parallel with the kernel sizes of 2,3 and 4. The layer will help generate the local text semantics making use of N-gram features and then this will pass the three outputs via multiple dropout and dense layers for producing the final output.

iii. *Dense net processing layer* will fuse the global and local text semantic outputs which are received from the processing layers of BERT and CNN. The merged outputs are passed into the dropout and dense layers and the final classified news as fake or real are seen.

5.2.2 Model Workflow

A. Data Pre-processing

The raw text of the dataset is pre-processed in three steps before applying the MCred model training.

i. *Tokenization* – Here the longer input is broken down into small sentences/tokens. While doing this, the sentence delimiters are protected to go for execution further.

ii. *Lemmatization* – Here the input data words are converted into the canonical form maintaining an equal footing for catering to the uniform execution.

iii. *Removal of stop words*- Stop words are filtered out from the data as their contribution is less in comparison to the meaningful data.

B. BERT Processing layer

As explained earlier data from the previous layer is fed into this and the layer will apply three data decoration techniques which will add metadata to the text and this mandatory for the execution of the text further. The three data decoration techniques, as shown in figure 5.2, are:

i. *Token embeddings:* It will add 2 special tokens, this is because the data has multiple sentences, at the beginning of the sentence [CLS] token is seen and at the end of the sentence [SEP] token is seen. The first and second words of the starting first sentence are represented by W_{1A} and W_{2A} whereas the first and second words of the next i.e., second sentence are represented by W_{1B} and W_{2B}.

ii. *Sentence embeddings:* It talks about the adding of special markers for the sentences. Similarly, sentence embeddings of the first two sentences are represented by E_A and E_B.

iii. *Positional embeddings:* The token position is specified in the given sentence. n^{th} and k^{th} element will be represented by E_n and E_k elements in the data.

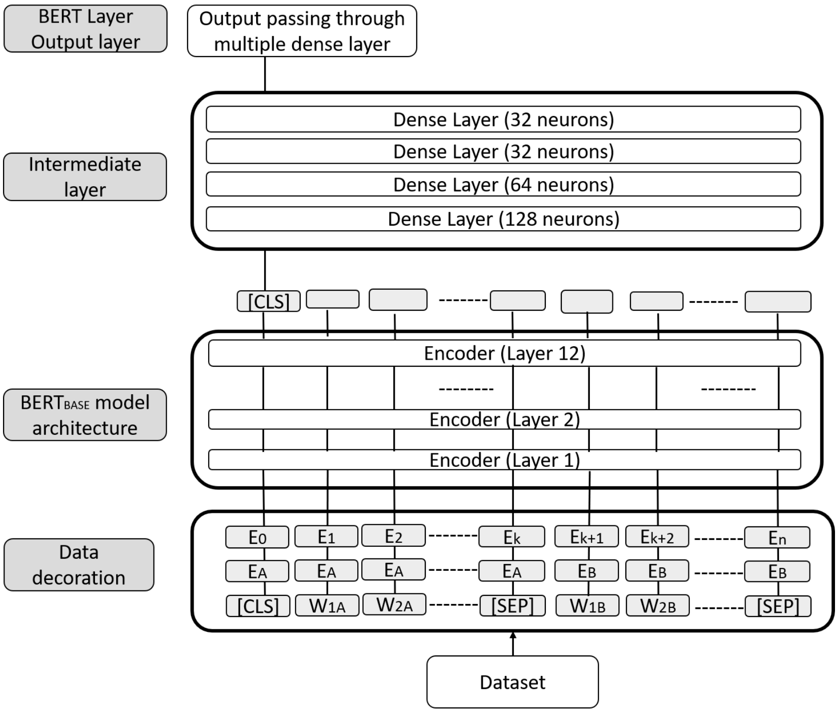

Figure 5.2: BERT processing layer

Each of the BERT layers will convert each token into long embedding vectors of 768 count which are passed on to the twelve encoding layers. [CLS] token information is enough for classification after passing through the 12^{th} layer. [CLS] vector will flow to the intermediate layer which consists of four dense layers with various other neurons. The output is generated after the BERT processing layers making use of dense layers with thirty two neurons.

C. CNN Processing layer

The internal architecture of CNN processing layer can be seen in figure 5.3. The embedding layer will process the input data and will generate the matrix of m*n size. Here, n- embedding dimension and m- maximum sequence length. The matrix will pass via the three Conv1D layer which has sixty four filter and kernel with sizes 2,3 and 4. Sixty four features will be generated by Conv1D layer from each kernel. Pooling layer in the CNN will process the sixty four long vectors and will concatenate them into one single vector. At the end, the model will pass this single vector into the dense layer and this will be converted into thirty two long vectors for further processing.

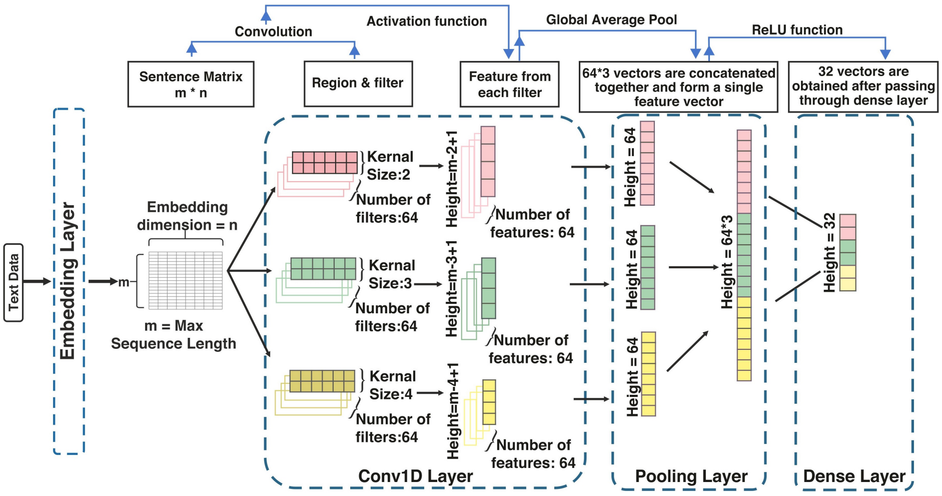

Figure 5.3: CNN processing layer

D. Dense net processing layer

This layer will read the thirty two vector outputs from CNN and BERT layers and merge them both to sixty four size as shown in figure 5.4. The dropout layer will prevent the problem of over-fitting and ReLU will be applied to the hidden layers whereas Sigmoid function will be applied to the output layer. Finally, after these many dense layers, the dense net layer will generate the classification of the news into fake or real.

E. Model tuning

The model tuning technique will help in the random search for improving and examining the MCred model. To CNN and BERT layers the ReLU function is applied. Along with this, sigmoid activation function and Adam optimizers are also applied at the dense layer. Model tuning parameter value is shown in table 5.1.

5.2.3 MCred Algorithm

Algorithm 2 clearly shows the working of MCred model in detail.

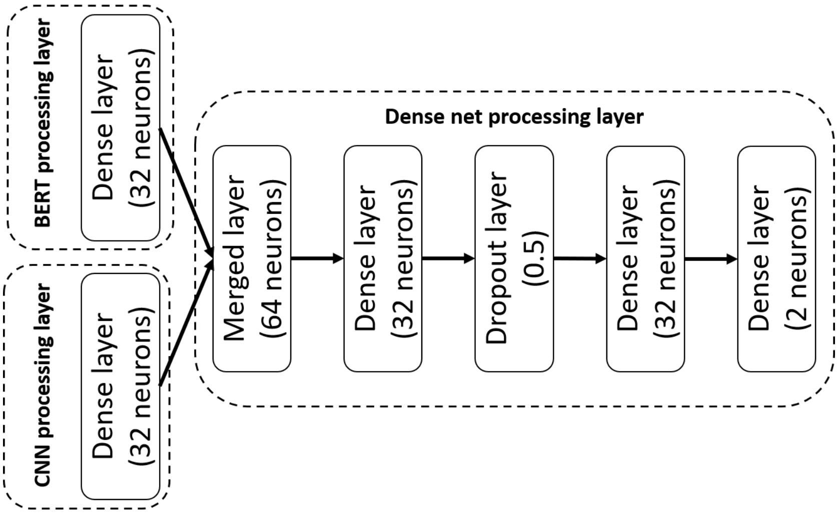

Figure 5.4: Dense net processing layer

Table 5.1: MCred model architecture.

Processing layer	*Parameter*	*Value*
BERT	Number of dense layers	4
	Dropout rate	0.5
	Activation function	ReLU
CNN	Number of dense layers	1
	Number of Conv1D layers	3
	Number of global average pool layers	3
	Activation function	ReLU
	Kernel size	1,2,3
Dense net	Number of dense layers	2
	Dropout rate	0.5
	Batch size	64
	Optimizer	Adam
	Activation function	Sigmoid
	Loss	Binary-cross entropy

i. Data collection and pre-processing task is shown in line 1 and 2 respectively.

ii. BERT processing is explained with the help of line 4 and 5. Line 4 explains the conversion of dataset into BERT embedding using $BERT_{BASE}$ pre-trained model and line 5 explains the output generated by this layer.

iii. Line 7, 8 and 9 explains the CNN processing. Line 7 explains the conversion of dataset into embedding using GloVe and line 8 and 9 explains the CNN text processing.

iv. Final output generation of this model is explained in line 11, 12 and 13.

Algorithm 2: MCred fake news detection algorithm.

```
Data: TextDataset
Result: MCred Model for news classification
// Phase 1: Data Engineering
MCred_dataset ← collection(TextDataset) // Text dataset selection
MCred_dataset ← preprocess(MCred_dataset) // Dataset pre-processing
// Phase 2: Model generation
BERT processing layer
    B_Embeddings ← BERT(MCred_dataset) // Using $BERT_{BASE}$ pre-trained model
    B_Output ← DenseDropout(B_Embeddings) // Pass embedding through multiple dense and dropout layers
CNN processing layer
    C_Embeddings ← Embedding(MCred_dataset) // GloVe embedding
    C_ConvOutput ← ConvLayer(C_Embeddings) // Pass embedding through three Conv1D layers of kernel size 2,3 and 4
    C_Output ← DenseDropout(C_ConvOutput) // Pass CNN output through multiple dense and dropout layers
Dense net processing layer
    Merge_Input ← Merge(B_Output,C_Output) // Merge output from CNN and BERT processing layers
    Final_Output ← DenseDropout(Merge_Input) // Pass input through multiple dense and dropout layers
Labelled_News ← Label(Final_Output) // Read result from previous layer and label it as real or fake
```

5.3 Experimental Result

Table 5.2 shows the results obtained after tuning of MCred model and the experiments have been conducted in the ratio of 80:10:10 which are in the split of train-test-validation.

A. Optimizer and Dropout selection

Table 5.2 will give a clear idea of the comparison between the results generated making use of the dropout values (0.3 and 0.5) on four optimizers namely Adagrad, RMSProp, SGD, and Adam. You can see a consistent increase of the dropout of MCred model performance over all the parameters when optimizers SGD and Adam are applied whereas the performance was compromised during Adagrad and RMSProp. Outperformance of the Adam optimizer is seen as it combines Adagrad and RMSProp optimizers for handling sparse gradients on noisy and large data. It will also produce better results because of the small learning rate and fewer memory requirements that are adapted for individual parameters. This higher dropout will improve the performance of the MCred by minimizing the loss of

Table 5.2: MCred model results on various parameters.

Parameter	*Optimizer*	*Val_Loss*	*Val_Acc*	*Testing dataset*			
				Accuracy	*Precision*	*Recall*	*F1-Score*
Dropout (0.5)	**Adam**	**0.0160**	**0.9959**	**0.9901**	**0.9921**	**0.9882**	**0.9901**
	SGD	0.3898	0.8299	0.8258	0.8220	0.8161	0.8190
	RMSProp	6.9984	0.5107	0.5071	0.5071	0.9901	0.6729
	Adagrad	0.5159	0.7620	0.7587	0.7530	0.7477	0.7503
Dropout (0.3)	*Adam*	0.0442	0.9858	0.9852	0.9823	0.9881	0.9851
	SGD	0.3905	0.8208	0.8164	0.8718	0.7199	0.7886
	RMSProp	0.1408	0.9553	0.9481	0.9410	0.9516	0.9463
	Adagrad	0.4551	0.7854	0.7844	0.7831	0.7680	0.7755

validation to 1.60% and also amplifying the accuracy of the validation to 99.59%, whereas in case of testing the accuracy is 99.01% and the recall, precision and F1-score is 98.82%, 99.21% and 99.01% respectively.

B. Learning curve

A learning curve was drawn using the training and validation data. The curve shown in figure 5.5 and figure 5.6 are based on accuracy and loss respectively. The gap between validation and training on the data in the two curves is high. When five epochs were executed the model has reduction in the gap and it has become stable and demonstrates good fit condition between under and over fitting. This is because training and validation loss gap is less at the stable point.

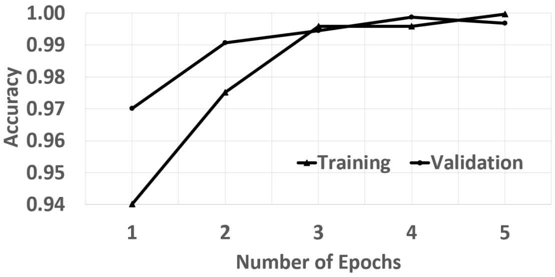

Figure 5.5: MCred accuracy curve

C. ML model versus MCred

On a WELFake dataset, XGBoost, RF,DT, NB, and LR were implemented and the performance was compared with MCred model. To extract the text features

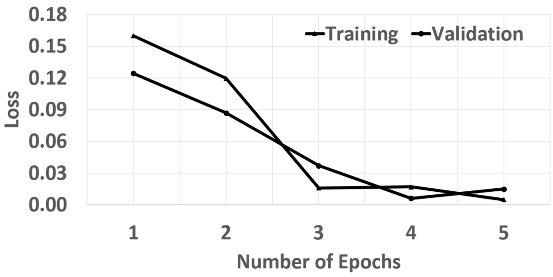

Figure 5.6: MCred loss curve

Table 5.3: MCred comparison of with other models.

Model Name	Accuracy (%)
Logistic Regression (LR)	89.46
Naive Bayes (NB)	92.38
Decision Tree (DT)	93.56
Random Forest (RF)	94.12
XGBoost	97.65
MCred	**99.01**

GloVe technique was used and then these are converted into feature vectors. The vectors were fed into the models and the performance can be seen in table 5.3. The accuracy levels of the different models can be seen between the range 89.46% and 97.65%. Out of which XGBoost has a accuracy of 97.65% . This shows that the fusion of more than one learning methods that are made use in MCred model has improved the accuracy levels.

D. Comparative analysis

Table 5.4 shows the comparison of MCred with three models [64, 66] and [65]. Mersinias et al. [65] has made use of content-based approach for detecting the fake news. He tried to combine LR with a deep learning model to get an accuracy of 97.52% on the Kaggle dataset. Whereas Khan et al. [64] made use of Doc2Vec features for the classification of the fake news classification and has applied AdaBoost LinearSVM and Bagging LinearSVM. The accuracy rate here 90.7%. Kaliyar et al. [66] has proposed the FNDNet model which is based on GloVe word embedding making use of CNN method on Kaggle dataset whose accuracy is 98.36%.

Table 5.4: Comparison of MCred with state-of-the-art methods

	Mersinias et al. *[65]*	*Khan* et al. *[64]*	*Kaliyar* et al. *[66]*	*MCred*
Dataset accuracy	Kaggle: 97.52% McIntire: 94.53% FakeNews: 96.78%	Kaggle: 90.70%	Kaggle: 98.36%	Kaggle: 99.46% McIntire: 97.16% FakeNews: 97.98% WELFake: 99.01%
Document representation features	Class label frequency distance vector	Doc2Vec	GloVe	GloVe – BERT embeddings
Classifier	Logistic regression (ML) CNN + LSTM (DL)	AdaBoost LinearSVM	Deep CNN	CNN, BERT

5.4 Summary

The MCred model helps in classifying the news as fake or real making use of the local and global semantic relationship in the words. The local semantic relationships were modeled making use of CNN with a kernel size of 2,3, and 4 and the modeling of global semantic relationships was done making use of BERT model. For final prediction, the MCred model will combine BERT and CNN and further process them in a network layer for the prediction. When experiments were conducted on MCred making use of various datasets accuracy was very high. To work further in the future, additional features basing on the propagation analysis and credibility can be used. Deep fake and image based analysis can be worked further in the future.

Chapter 6

USER CREDIBILITY BASED FAKE NEWS DETECTION

6.1 Introduction

When statistics was conducted and researchers monitored the users of online social network (OSN), it was identified that 90% users are under the age group 29, ranging between 18-29 who regularly use at least one social media site [99]. Making use of the OSN, the users tend to share their information, thoughts making use of speech, videos, and images which takes very less time. When fake news is spread around it will create a tense atmosphere in many fields including education, business, government and more. This is because fake news is not confined to one sector of the economy but will impact many sectors. The real intention behind the creation of fake news is for generating deface and mistrust among the society affecting its values without prioritizing religious, regional, and political beliefs [100]. A news about the presence of parasitic roundworms in the McDonald's has created panic among the people which was fake news [101].

A false news effect on brands and organizations was explained by Cheng et al. [102]. For fake news detection, many researchers made use of user profile-based, pattern, propagation, and knowledge based approaches [103]. Whereas, on the other side, in order to diffuse any fake news the user disregards other user's security and make use of fake or misuse the names for gaining the personal credentials in order to create a fake profile. But some of the fake profiles intention is to provide entertainment but has no malicious intention [104]. This does not make the profile legal. This is even declared by the Facebook that when one person will create any account beyond the account's original principles, then it is a fake account [105]. Such fake accounts are widely spread on many platforms like Twitter, Facebook. Removal of such accounts which were under operation from Nigeria and Ghana was done by Facebook. These accounts were created for targeting the audience of US and were employed by the Russian individuals [106]. In order to keep the hoax propagation under control Franklin et al. [107] has made use of social group and message characteristics. On the other hand, user profiles were used by Mudasir et al. [108] to identify any suspicious link on the social media. Below are the fake profile types:

i. *Sybil Account [109]* – These are many accounts the malicious users have created.

ii. *Cloned Profiles*- These are not real and are made by cloners making use of information as in the real profile.

iii. *Bots as fake profile [110]* – These profiles are controlled by the program and are mainly used to spread malicious information.

iv. Compromised Profiles- Profiles are controlled by the malware agents.

v. *Sock Puppets* – Accounts created for deceiving others.

All of the above types are used interchangeably in this chapter. The main motive to detect fake profile is to address issues like referral incentives, fake voting, cybercrimes and more which have been increasing with time in the recent years.

6.2 Proposed Methodology

The model makes use of the user profile information in the form of input and this will help in classifying the profile as fake or real. The commitments of proposed model i.e. **U**ser **Cred**ibility(UCred) model are:

i. Identifying the machine learning model which will give improved result using user profile features. For this, KNN, RF, SVM, NB were applied on the given news. The results have shown highest accuracy when RF model was applied which is 93%.

ii. Identifying the post-trained model which will give better results when classifications of profiles are conducted. For this, Bi-directional Long Short Term Memory, Long Short Term Memory, Convolutional Neural Network were applied on the message in the user profile which gave out accuracy of 95%.

iii. To pick between Distilling BERT , Robustly Optimized BERT , and Bidirectional Encoder Representations from Transformers basing on the results which they give. Out of the mentioned three, RoBERT model has given the best results with an accuracy of 96%.

iv. Identifying if the voting classifier will enhance accuracy levels or not. The results of pre-trained, post-trained models were sent into the voting classifier where the results gave an accuracy of 98.96%.

6.2.1 UCred Model

To know more about the UCred model, go through the figure 6.1. Basing on the figure, one can clearly understand that there are four phases which are: (i) Data Engineering, (ii) Textual Data Processing (iii) Non-textual Data Processing and (iv) Voting Classifier.

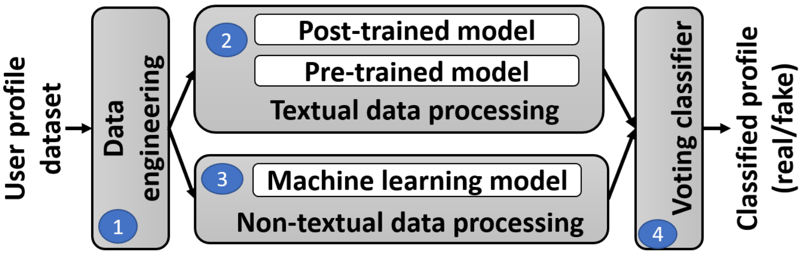

Figure 6.1: UCred model

A. Data Engineering

The phase will help in performing tasks that are prerequisites which are required for any basic model. This dataset will contain both text and numeric data. The raw data is read in this phase which is further converted into feedable data which is required for any basic model. In case of text data, stop words will be removed that are present in the text as the stop words will not participate in the operations and further on these stemming and tokenization are performed. Further all the tokenized data is stored in the column of the similar and same datasets for future operations. During the same time, missing value imputation and label encoding were performed as part of preprocessing tasks for the sake of numeric data.

B. Textual data processing

The phase will help in handling only the text features. In this phase text converted into vector using GloVe for the sake of ML/DL models. GloVe embedding further passed in to the post and pre-trained models at the same time. Both these models will help in predicting the given class as fake or real basing on their processing methodology and will then forward the received output on to the next phase. For the purpose of pre-trained model, DistilBERT, RoBERT, BERT were used whereas for the purpose of post trained model Bi-LSTM, LSTM, and CNN were used.

C. Non-textual data processing

The phase will help in reading the numeric features out of the pre-processed data. Making the use of Pearson's coefficient, out of all the numeric features essential ones are picked out. After this selection, the features are further fed into the traditional machine learning models like KNN, RF, SVM and NB after the tuning of some parameters. Using this phase, the best of ML model which will help in classifying the profile as fake or real with highest accuracy and further process this output to the next coming phase.

D. Voting classifier

The phase will make use of three predictions that are taken from the previous phases; two outputs are taken from the phase two i.e. textual data processing phase and one output from the phase three i.e. non-textual data processing phase. Using the voting classifier the final class predicted on the basis of voting that has been obtained from the previous phase.

6.2.2 UCred Workflow

For implementing UCred model Jupyter Notebook and Pythom programming language are used. Every experiment of this model were executed making use of a system which has Intel Core i7, Windows 10 operating system, 256 SSD, 8 GB RAM, 9th generation 2.6 GHz processor. The sequence of steps which are required for the UCred model implementation can be easily understood by seeing the figure 6.2.

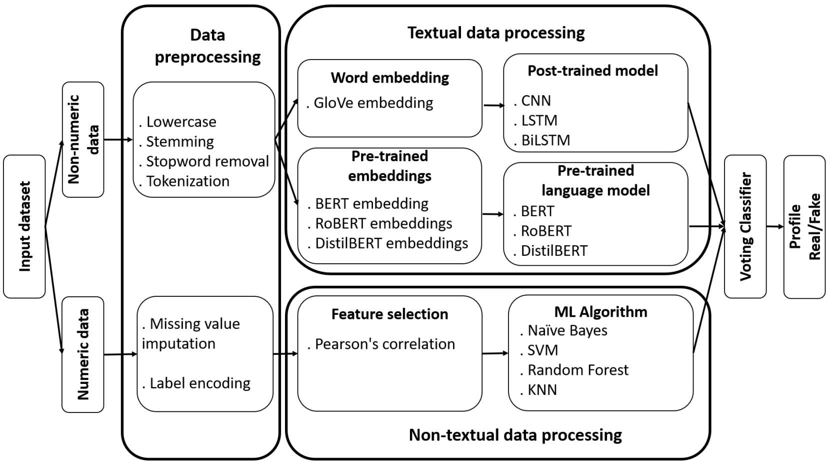

Figure 6.2: UCred model workflow

A. Dataset

A OSN dataset was used for testing and training which has made use real and fake profile information [69]. There are non-numeric and numeric values in the dataset features which are:

i. The friends count will show the total number of friends that are connected with user.

ii. The status count will help in understanding the updation count of user status.

iii. The follower count will show the number of followers of current information who are connected with user.

iv. The ID will hold the user's unique ID.

v. The default profile will hold the numeric value

vi. The default profile image will show the information regarding the updation of the default image if it is updated or not.

vii. The favorite count will help in calculating the number of favorite friends.

viii. The isFake will help in evaluating if the profile is real or fake, here 0 will indicate real and 1 will indicate fake

ix. The description will help in understanding the user information

x. Verfied will hold the binary number which will show the profile verification information.

xi. To understand the post number of the user, the listed count will give the information.

xii. Protected will help in understanding if the post of the profile is protected or not.

xiii. Making use of the utc-offset you can understand the zone code and time information.

xiv. The profile_background will have the information about the presence or absence of binary background, this information is stored in the form of binary value.

xv. The profile_use_background_image will give the background image information.

xvi. Making use of the geo_enable you can understand the enabling of the graphical information.

B. Data Pre-processing

The dataset entries are shuffled to distribute the fake and real profile information properly. After shuffling, predefined python libraries are used for the removal of stop word and stemming in case of text data and missing value imputation and label encoding performed in case of numeric data.

C. Textual data processing

Once the description feature of the user extracted from the dataset, these were sent to the pre and post-trained models. Processing of the text data is done on the both models and predictions are generated separately.

i. **Post-trained model:** The text data fed into the BiLSTM, LSTM and CNN in GloVe embedding form for classifying the text as fake or real. This is explained further below:

   a. *BiLSTM:* The news do not contain entire fake information. It is a combination of both fake and real news. BiLSTM will process the given text from backward and forward directions. To construct the BiLSTM model, Adam optimizers are used which has a 0.001 learning rate and the model trained with ten epochs with sixty four batch sizes.

   b. *LSTM:* The text is sent into the LSTM after pre-processing the text. In the form of hyper parameter tuning two dropout layers each of 0.3 after and before the LSTM layers are used. A dense layer having a Sigmoid activation function is also added for the processing. The model was compiled making use of Adam optimizer and binary cross-entropy with a learning rate of 0.001. To conduct the training, ten epochs with sixty four batch sizes are used.

c. *CNN:* The word embedding are passed into the 1D Convolutional layer. Sixty four filters are used as the convolutions of the output dimensions and these have kernel size of three which will consider three words of word embedding in a go. The ReLU is used further in the hidden layer along with dropout value of 0.5, this will help in preventing the problem of over fitting. sigmoid activation function is used in order to reduce the dimensions to one .

ii. **Pre-trained model:** Three advanced NLP models are used to get the results and their experimental setup is given below:

a. *BERT:* This is one of the popular pre-trained model in the NLP category. The architecture will help in understanding the model of BERT which is present in two variants: Base and Large. Here, the Base model of BERT is used which is containing the 12 attention heads, 12 transformer blocks and 110 million parameters.

b. *RoBERTa:* Yinhan et al. [43] has proposed the pre-trained model named RoBERTa. Performance of this is better than the first model as it will remove the next sentence prediction operation and added a masking technique. But training of this model will take more time than the BERT model as huge data is used in this model.

c. *DistilBERT:* This model uses half layers and half parameters which are present in the BERT model. The time taken for training this model is four times less than the model of BERT. But there is 3% degradation of performance when compared with BERT. The model is examined as it has less resource requirement.

D. Non-textual data processing

The phase made use of the pre-processed data on the numeric features. ID column in the dataset do not have any information for prediction, so the column is dropped. Once it is dropped, features correlation is checked and essential features which are selected are shown table 6.1. These selected features are further passed on to the NB, KNN, RF, and SVM classifiers.

Table 6.1: Essential features used in feature processing phase

Feature Name	Descriptions
Status_count	Number of times status updated by user
Followers_count	Number of followers of particular user
Friends_count	Total number of friends
Favourites_count	Number of favourites friends
Listed_count	Number of posts posted by user
Lang_code	Language selected by the user

The given dataset is split into 20% for testing and 80% for training purpose. The following selections are done during the ML model execution for better results.

i. Two is selected as the maximum tree depth for RF classifier.

ii. Three neighbors are selected for KNN to get the best result.

iii. Out of all the NB classifier types, multinomial NB classifier is picked.

iv. In the case of SVM classifier, radial basis function kernel is selected.

E. Voting Classifier

The classifier is used for addressing regression and classification problems making use of soft and hard voting. Hard voting classifier is used to address the classification problem. This can be considered as the simpler version of boosting and bagging as they collect the output from different models and will generate output basing on the majority voting. For final result, output is collected from post, pre-training models and ML model. The output is generated basing on the most outputs which are passed in it. Ensemble package of the python library is used for this purpose.

6.2.3 UCred Algorithm

Understanding the algorithm 3 will give a better and clear idea on how the UCred model is executed and the phases in it:

Algorithm 3: Algorithm for UCred model

```
   Data: UserProfileInformation
   Result: UCred Model for profile classification
   // Phase 1: Data engineering
1  UCred_dataset ← collection(UserProfileInformation) // Load real/fake users
     information in UCred dataset
2  UCred_dataset ← preprocess(UCred_dataset) // Dataset pre_processing
   // Phase 2: Textual data processing
3  Post-trained model
4      $F_1$ ← extract(UCred_dataset) // Extract features with text data
5      $M_1$ ← bestModel(CNN($F_1$),LSTM($F_1$),BiLSTM($F_1$)) // Selection of best
         model for features with text data
6  Pre-trained model
7      $F_2$ ← extract(UCred_dataset) // Extract features with text data
8      $M_2$ ← bestModel(BERT($F_2$),RoBERT($F_2$),DistilBERT($F_2$)) // Selection of
         best model for features with text data
   // Phase 3: Non-textual data processing
9  $F_3$ ← extract(UCred_dataset) // Extract features with numeric values
10 $M_3$ ← bestModel(NB($F_3$),SVM($F_3$),RF($F_3$),KNN($F_3$)) // Selection of best
     model for features with numeric values
   // Phase 4: Voting Classifier
11 return votingClassifier($M_1$,$M_2$,$M_3$) // Predict the final class based on
     the maximum voting of all models
```

i. *Data pre-processing*- The phase will take two datasets which contain both fake and real user profile information which will help in the generation of the a dataset which can be seen in the line 1. Pre-processing of the numeric and text features is performed which can be seen in the line 2.

ii. *Textual data processing*- The phase will help in dealing with pre and post trained models at the same time. You can see the post trained models which will extract the text features from the given datasets in the lines 4 and 5. The best post-trained model is finalized. Extraction of the text features and selection of the best model is done by pre-trained model in the lines 7 and 8.

iii. *Non-textual data processing*- The phase will help in selecting numeric features which are based on Person's correlation coefficient which is shown in the number 9 of the algorithm. All these features are entered into ML models for selecting highest accuracy model as seen in line 10.

iv. *Voting classifier*- This phase talks about the output which is generated out if the three models and results are finalized basing on the mechanism of the voting as seen in line 11.

6.3 Experimental Result

The profiles are classified making use of ML/DL model which are based on the user profile features. The classifier has given out an accuracy of 77.30% - 93.79%. Out of all the ML classifiers like RF, KNN, SVM, and NB; RF has given out the maximum accuracy of 93.79%. To achieve the highest accuracy pre and post-trained models are also used. User description is fed in the for of text embedding to the pre and post-trained models. Out of all the models, RoBERT model will give 96.07% accuracy and BiLSTM gave out 95.12% accuracy. The table 6.2 will help in understanding the results that are obtained by using confusion matrix.

Table 6.2: Model selection for text and numeric features

Selection of best model	**Model**		**Parameter**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
	ML model	NB	User Profile features	77.30	59.70	88.89	71.43
		SVM		90.07	98.88	83.33	90.44
		RF		**93.79**	**97.01**	**90.66**	**93.19**
		KNN		93.23	97.01	89.59	93.69
	Deeplearning model	CNN	Text Embedding	93.68	93.10	93.21	93.66
		LSTM		93.90	94.21	93.98	93.19
		BiLSTM		**95.12**	**95.01**	**95.31**	**95.23**
	Pretrained model	BERT		95.17	95.02	95.41	95.33
		RoBERT		**96.07**	**96.13**	**96.71**	**96.43**
		DistilBERT		93.12	93.11	93.79	93.57

When you see the performance summary in table 6.2, it will show that RoBERT, BiLSTM, and RF will work well independently among ten models. Table 6.3 shows that performance summary after voting classifier. The output obtained from top three models are passed into the voting classifier. When the voting classifier is applied, the accuracy rate of the performance will be increased to 98.96%.The UCred model will take the output from three models and will finalize the result basing on the maximum voting.

Table 6.3: Performance of UCred model

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
RF	93.79	97.01	90.66	93.19
BiLSTM	95.12	95.01	95.31	95.23
RoBERT	96.07	96.23	96.71	96.43
Voting Classifier	**98.96**	**98.56**	**98.12**	**98.37**

The UCred model is compared with the Deep Profile model [69] that is trained and tested on the same dataset shown in table 6.4 . Deep profile will make use of the deep learning for the selection of automatic features. Deep neural network basing on the CNN is used for classifying the users and made use of the pooling layer for optimizing the neural network's performance.

Table 6.4: Proposed model comparison

Model	Dataset	Procedure	Features	Accuracy (%)
DeepProfile	OSN dataset	Dynamic CNN	Numeric and text	93.44
UCred	OSN dataset	Fusion of RF, BiLSTM and RoBERT	Numeric for RF Text for BiLSTM and RoBERT	98.96

The table 6.4 will clearly show that the UCred model will classify the profile as fake or real with accuracy rate of 98.96% and this is 5.52% higher than the Deep Profile model's accuracy. UCred model will also enhance the F1 score, recall and precision predictions which can be understood with the help of figure 6.3.

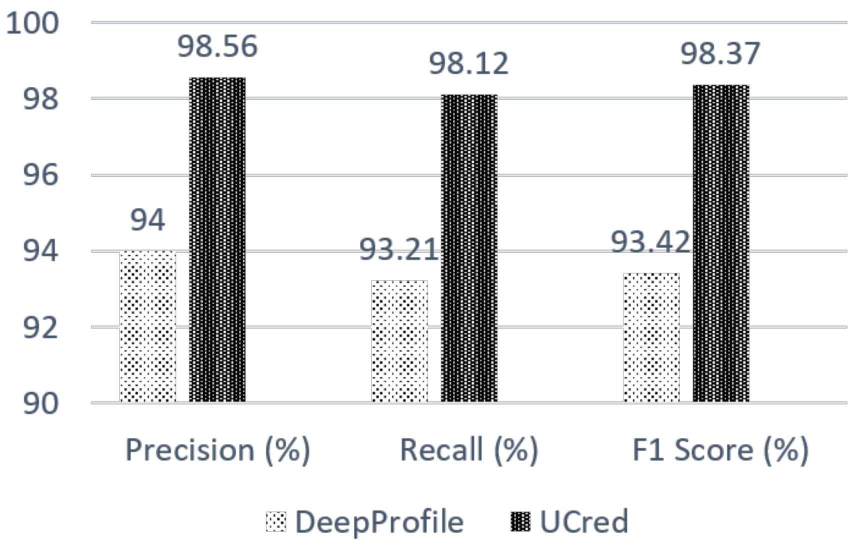

Figure 6.3: Performance metrics comparison

6.4 Summary

The UCred model is proposed for classifying the user profile as fake or real basing on the user profile features. The OSN dataset will contain the information of fake and real users information. Though the dataset has 15 features, they are divided into 2 parts: non-numeric and numeric features. Four ML methods are applied on the non-textual and textual features are used pre- and post-trained models. At the end, all predictive outputs are further passed on to the voting classifier to get the final result. When the final result is obtained, it is considered to be better than the individual results. A comparison between UCred and DeepProfile model is carried out where UCred model has improved the accuracy with 5.52%.

Chapter 7

PROPAGATION BASED APPROACH

7.1 Introduction

In the last few years, social media has been used rigorously. People just do not use it for entertainment or to maintain social connection, they also use social media websites for distributing the news around the world. IT companies like Twitter and Facebook have redefined the method of distributing the news quickly around the world [111]. Some people also choose social media websites for sharing their views on different topics like the entertainment industry, religious topics, etc. Sadly, with all these advantages and benefits, social media is also used for spreading fake news. Fake news not only means false news, but it also comprises misinformation, disinformation, manipulation, and rumors [112]. Fake news propaganda existed for a long time but it got famous after thousands of people got aware of the usage of social media. One of the famous incidents is of May 2016 where it was tweeted that "Bill and Hillary Clinton were reported to be utilizing a Pizza restaurant as a face for a pedophile sex ring". It resulted in an event where a 28-year man entered the joint with a riffle to examine the case. But, thank god nothing happened and the man got arrested [113]. This was still a manageable issue. These days people are also using social media websites for spreading fake news regarding health reports which are creating a panic environment around the world, especially after the COVID-19 incident. In one of their survey, The International FactChecking

Network stated that there are around 3500 false cases that have been claimed in less than two months. But because of the depth of this platform, it becomes difficult to catch the liar.

Every second the risk of fake news is increasing which is not good for society [114, 115]. That's why many researchers are working on this and they are trying to decrease the number the fake news everyday. They are also using the features using texts or user inputs which help them to spot the fake news even faster. Another method that is better than this is the propagation-based feature model. Vosoughi *et al.* [14] has analyzed both the real and fake news propagation pattern and declared that false news disseminates faster, deeper and also stays for longer in the air as compared to the real news.

7.2 Proposed model

The proposed model consists of three-layer in it. These are the data preprocessing layer, feature engineering layer, and lastly, model building and tuning layer. The model has been illustrated in figure 7.1. The coming section demonstrates the working of the proposed model and also displays how the model works using the algorithm.

7.2.1 PropFND model

A. Data preprocessing layer

The structure of the dataset always works with the problem statement. So, in this layer, we have assessed the public repository and downloaded the dataset of fake news. After the download has been completed, we have performed the simple preprocessing task to change the data into the required format. The preprocessing task not only converts the data, but also manages the noisy, unfinished, and irregular data.

B. Feature engineering layer

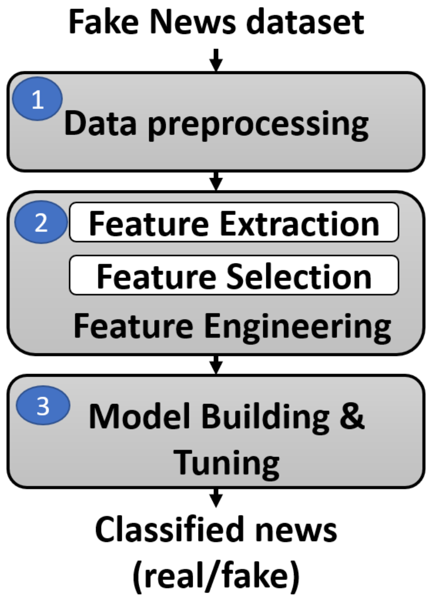

Figure 7.1: PropFND working flow

Once the dataset in the required format has been obtained, we extracted propagation and user profile based features and finalized the essential features for further execution. While we extract and select the data, we need to perform some operations:

i. The feature extraction phase extracts the data or the features that are already available in the dataset. It also gives some manual features which are formed on these available features.

ii. The feature selection phase aims for the features which is the best and which help to achieve better accuracy. So, we have used the Mutually Informed Correlation Coefficient (MICC). It combines both Pearson Correlation Coefficient as well as Mutual Information [116].

$$MICC(i) = \alpha \times MI(i) - (1 - \alpha) \times Sum(PCC(1, 1 : dim))$$

where $MICC(i)$ is the mutually informed correlation coefficient value
$MI(i)$ is mutual information of i^{th} feature
α represents the weight

Table 7.1: Hyperparameter tuning of SVM model

Parameter Name	**Value**
Kernel	linear
Regularization	100
Kernel coefficient	0.0001
Degree	3
Dataset split	80% train, 20% test

PCC is the Pearson Correlation Coefficient

dim is the number of features present in the dataset.

C. Model building and tuning layer

Lastly, we have to feed the selected features in various models. This layer decides whether the news is real or fake. It depends on the input provided in the this layer. To improve the accuracy, we can also perform the model tuning operation.

7.2.2 PropFND Workflow

i. Data pre-processing layer puts the missing values in the correct space which are based on the feature property. They also eliminate some of the outliers from the dataset. To get better accuracy and outstanding result, data normalization has been performed in this layer. To achieve such difficult tasks, inbuilt libraries of python have been used.

ii. Feature engineering layers studies data from the layer above it, i.e., the data preprocessing layer. It uses MICC for deriving valuable features.

iii. Model bilding and tuning layer works with the selected features and supplies the features into several classifiers. Then, it conducts extensive experiments with various parameter values. After the experiment, we have observed that the SVM give the improved result among many classifiers with some hyperparameter tuning. Parameters that give final accurate and improved result is shown in table 7.1.

Algorithm 4: Algorithm for PropFND model

```
   Data: FakeNewsDataset
   Result: News classification using PropFND Model
   // Stage 1:  Data preprocessing
1  dataset ← dataCollection(FakeNewsDataset) // Fake news dataset
     selection
2  dataset ← preprocess(dataset) // Dataset preparation
   // Stage 2:  Feature Engineering
3  Feature Extraction
4      $F_1$ ← FeatureExtraction(dataset) // Extraction of available
         features in dataset
5      $F_2$ ← ManualFeatureExtraction(dataset) // Generation of
         additional features from already available features
6      FinalFeatures ← Union($F_1$, $F_2$) // Combining two feature sets
7  Feature Selection
8      features ← FeatureSelection(FinalFeatures) // Selection of
         important features using feature selection technique
   // Stage 3:  Model Building & Tuning
9  Labelled_News ← BestModel(features) // Features passed to the
     multiple classifiers for classification of news as real or
     fake
```

7.2.3 PropFND model algorithm

Let see how the PropFND model works using algorithm 4. The working of the algorithm is explained below:

i. To begin with, let's select one dataset as per our requirement from available fake news datasets. This is shown in line 1. Once the dataset is selected, we have performed some simple preprocessing operations like feature scaling, missing value imputation, outlier removal, etc.

ii. After the data preprocessing operation is done, we have implemented the feature engineering process. It includes features extraction (which has been illustrated in lines number 4 and 5) and feature selection (which has been shown in line number 8).

iii. After this, we have assessed many models for classification applying extensive experiments. At the final stage, we have selected a SVM for the classification.

This model has improved the accuracy of the model (as displayed in line number 9).

7.3 Experimental Result

A. Experimental setup

We have executed the PropFND model using Python programming. Also, we have adopted the Jupyter notebook for code improvement and debugging. Then, we have used the Intel Core i7 9^{th} generation with 16 GB RAM, 256 HDD, and 256 SSD for the execution purpose and the main thing we have worked with windows 10 operating system.

B. Dataset

The FakeNewsNet dataset has been used for the preparation and trial of our suggested model [117]. The development of the dataset consists of the following steps:

i. The Labelled news reports are obtained from two fact-checking groups [17] [20].

ii. The keywords are derived from the headings of the labeled news articles. By using these keywords, tweets and retweets have been plucked. The resulting dataset consists of the news articles in the form of images or text as well as the date when they got published. This dataset also consists of knowledge about the tweets and retweets. The set also has information related to the users.

C. Result

At last, we have done extensive experiments on different classifiers and examined the performance of individual models. In the beginning, we have divided the dataset into 80% for training purposes and 20% for testing purposes. After the execution of the experiment, we have noticed few things. Among all the models, the NB provides the minimum efficiency of 82.45% and SVM gives the highest accuracy of 93.81%. This data has been represented in table 7.2.

Table 7.2: Accuracy comparison on various models

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naive Bayes	82.52	82.11	83.45	83.02
SVM	93.81	92.94	92.51	93.03
Random Forest	91.84	90.54	91.65	91.84
NN	92.13	91.15	91.84	90.74

Table 7.3: Accuracy analysis with different training-testing split

Classifier	70%-30%	80%-20%	90%-10%
Naive Bayes	82.31%	82.52%	81.89%
SVM	93.11%	93.81%	93.49%
Random Forest	91.84%	92.03%	91.42%
NN	91.31%	92.13%	91.87%

For more generalized accuracy tests, we have checked the performance of each classifier using many training and testing dataset sizes. The accuracy of the SVM model varies from 93% to 94% and has been illustrated in table 7.3.

D. Comparative Study

For the performance comparison we compared our model performance with existing model proposed by Meyers *et al.* [75]. Table 7.4 shows that after adding the user profile based features and MICC feature selection technique the accuracy of news classification improves by 6.81% from existing model executed on same dataset.

7.4 Summary

In this chapter, we have described a model to detect whether the news is real or fake. For that, we have proposed a PropFND model which uses propagation and

Table 7.4: Comparison between existing and proposed model

Parameters	Meyers *et al.* [75] model	PropFND model
Features	Propagation based	Propagation and profile based
Dataset	FakeNewsNet	FakeNewsNet
Feature selection	Manual analysis	MICC
Classifier	NN	SVM
Accuracy	87%	93.81%

user profile based features. For the experiment, we have used the FakeNewsNet dataset and divided that into two segments: the first 70% for training purposes and the rest 30% for testing purposes. Then, we have passes the essential feature to the SVM model for the classification of news whether it is real or fake. After running the PropFND model, we have got the result with 93.81% accuracy. With such a high value of accuracy, the model has one shortcoming. When we choose the feature, we have selected one feature that counts the number of novel users who are involved in the spreading of the news in the initial hours. But, this feature states that the model can label the news as real or fake only in the later hours. So, this is the disadvantage that the model cannot label the news as real or fake in the initial stage of the experiment.

Chapter 8

FINAL PREDICTION USING LINGUISTICS, CREDIBILITY AND PROPAGATION BASED APPROACH

8.1 Introduction

Labelling of news as real or fake is an exceptionally delicate task and it is not appropriate to label a news by utilizing single procedure. Other than this dataset also plays vital role in the development of AI model. Therefore this dissertation constructs the novel dataset for model training/testing purpose and also utilizes the concept of voting for labelling the news. In the above chapters four models were proposed for the detection of fake news; among them three models were used for the detection purpose and one model verified the output generated by other three models.

8.2 Combined workflow

Figure 8.1 shows the combined workflow of this dissertation. The overall work performs following tasks:

i. Initially, a novel dataset was proposed that contained 37,106 fake and 35,028 real news articles. This dataset reduced the limitations of other fake news datasets and prevented from overfitting problem of classifiers.

ii. After dataset creation three models; WELFake, MCred and UCred; were proposed for the identification of fake news. All these models gave the results in two categories either real or fake. On the basis of voting final result was generated i.e. maximum votes would be considered as final result.

iii. The final output generated by above models was verified by propagation based features. As this work focused on binary classification i.e., real or fake; therefore one assumption was considered if the final output generated by previous objective was fake and verification steps also concluded for fake then the news was treated as fake otherwise it would be treated as true news.

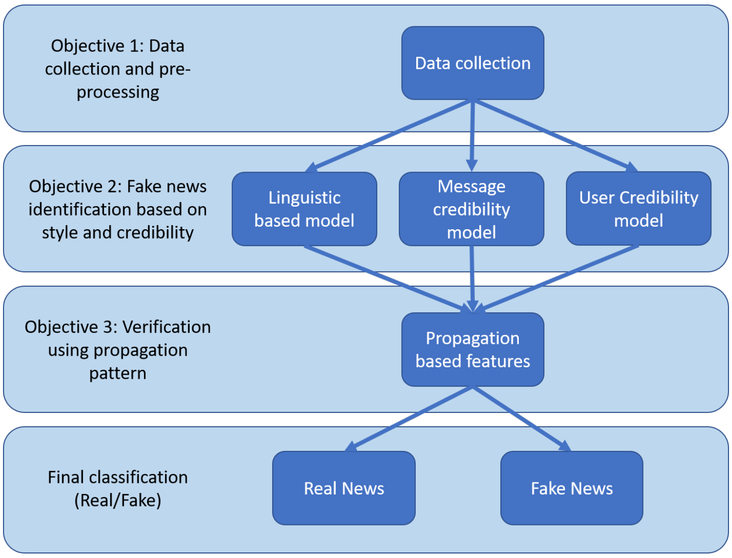
Objective 1: Data collection and pre-processing
Data collection
Objective 2: Fake news identification based on style and credibility
Linguistic based model
Message credibility model
User Credibility model
Objective 3: Verification using propagation pattern
Propagation based features
Final classification (Real/Fake)
Real News
Fake News

Figure 8.1: Complete workflow to achieve final classification

8.3 Summary

This chapter shows how all the models work together and generate the final result. The propagation based features take longer time as compared to other model because in propagation based study few features are time dependent. Therefore except propagation based prediction all other models are used for prediction for early detection.

Chapter 9

CONCLUSION AND FUTURE SCOPE

This chapter summarizes the major contributions in the section 9.1. Moreover section 9.2 explains future aspects in the field of fake news detection.

9.1 Summary of results

This dissertation explained few possible methods for the detection of fake news and the comparative study of each model with state-of-the-art methods shows the potential of those methods. Specifically, (i) chapter 1 explained the term "Fake news" and the impact of fake mews on social media. (ii) Chapter 2 surveyed literature related to fake news detection and then classified into three major aspects i.e., linguistic based, credibility based and propagation based. (iii) Chapter 3 explained few publicly available datasets and highlighted their limitations. This chapter also explained one novel dataset i.e. "WELFake dataset" which overcomes the limitations of previously available dataset. (iv) Chapter 4, 5, 6 and 7 proposed the models for fake news detection using various strategies and (v) chapter 8 explained the workflow of all the proposed models for the detection and verification of fake news. All the objectives are summarized in sections 9.1.1 to 9.1.4.

9.1.1 Objective 1: Data collection and pre-processing

This objective concludes the characteristics of dataset in terms of strengths and limitations in the chapter 3. To eliminate these limitations a novel dataset named

"WELFake dataset" proposed which contains approximately 72000 news articles. It is built after combining BuzzFeed, Reuters, McIntire, and Kaggle datasets. This has helped in avoiding bias and further limitations during the detection of the fake news. In this dataset news articles are divided into two categories i.e., 48.55% real and 51.45% fake news articles.

9.1.2 Objective 2: Fake news identification based on style and credibility

To achieve this objective three models were proposed; (i) "WELFake model" based on linguistic features of news articles, (ii) "MCred model" based on the local and global semantics of text and (iii) "UCred model" based on user profile information.

A. WELFake model

It is based on both writing pattern and WE of news article. To increase the accuracy levels twenty linguistic features were picked out of eighty features. Also WE passed to six ML models and concluded that CV method gave more accuracy. So CV method is merged with different sets of linguistic features and passed to ML methods. This model also takes the advantage of two level voting classifier; in first voting classifier output comes from combination of LFS and CV features are passed and the output generated from this level is again passed to the next voting classifier with CV and TF-IDF features. With this technique WELFake model achieved the accuracy of 96.73%.

B. MCred model

This model used local and global semantics relationship of words for the classification. The local semantic relationship was modeled making use of CNN with 2, 3 and 4 kernel sizes; and pre-trained i.e., BERT model was used making use of global semantic relationship. Both these outputs are combined and processed into a dense network layer for final prediction. On WELFake dataset accuracy was 99.01%, Fake News dataset accuracy was 97.98%, McIntire dataset accuracy was

97.16% and on Kaggle dataset accuracy was 99.46%. Accuracy rate on all the datsets when MCred model applied are high when compared with a state-of-the-art model output.

C. UCred model

Fake user also plays vital role in the dissemination of fake news. Therefore this model is focused on the classification of user profile in two categories i.e., real or fake user. User profile features were used in this model where features were divided as numeric and non-numeric. Multiple ML models are applied on the numeric features and non numeric features takes the advantage of pre-trained models. When the performance of UCred model is compared with DeepProfile model, it is observed that UCred gave an improved accuracy rate by 5.52%

9.1.3 Objective 3: Verification using propagation pattern

PropFND model verifies the output generated by above models. This model takes the advantage of both User and propagation based features. This model was trained and tested on the FakeNewsNet dataset because it contains the both types of features. Though the accuracy of this model is 93.81% the model gave out a limitation that few features are time dependent i.e. feature value generated after few hours.

9.1.4 Objective 4: Deployment of proposed model

As this dissertation proposed more than one model for fake news detection therefore there is a need of complete deployment process for the final result. This objective clearly explains the flow of input and output in a pictorial representation.

9.2 Future work

This dissertation explains various models for the identification of fake news and fake profile on social media websites. As we already discussed in previous chapters, the performance of these proposed models is better as compared to state-of-the-art models. From the year 2016 US Presidential election fake news evolves continuously and the newest version of fake news is “Deepfake”.

Deepfake content are the manipulated video, audio or image that are generated by any software or AI neural network. Deepfake creators use AI and ML algorithms to imitate the work and characteristics of real humans. It is differs from traditional fake media by being extremely hard to identify. Deepfake videos, speeches and images have the potential to cause enormous damage. In April 2018, BuzzFeed showcased how far deepfake video technology has come by combining Jordan Peele’s voice with video of Barack Obama. This clearly shows that AI’s ability to develop itself is a double-edged sword. If an AI is created to do something benevolent, great! But when an AI is designed for something malicious (like deepfakes), the danger is unprecedented.

Due to limited research in the field of Deepfake can be an exploration area to build a better detection system for video forensic investigation. The creation of Deepfake datasets can be a milestone for better research prospects.

Bibliography

[1] W. Jiang, J. Wu, F. Li, G. Wang, and H. Zheng, "Trust evaluation in online social networks using generalized network flow," *IEEE Transactions on Computers*, vol. 65, pp. 952 – 963, 2016.

[2] "The social media demographics report: Differences in age, gender, and income at the top platforms," https://www.businessinsider.com/the-social-media-demographics-report-2017-8?IR=T, Accessed: 2021/08/12.

[3] S. Ranganath, S. Wang, X. Hu, J. Tang, and H. Liu, "Facilitating time critical information seeking in social media," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, pp. 2197 – 2209, 2017.

[4] Z. Zhang, R. Sun, X. Wang, and C. Zhao, "A situational analytic method for user behavior pattern in multimedia social networks," *IEEE Transactions on Big Data*, vol. 5, pp. 520 – 528, 2017.

[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, pp. 22 – 36, 2017.

[6] "Key findings on the traits and habits of the modern news consumer," https://www.pewresearch.org/fact-tank/2016/07/07/modern-news-consumer/, Accessed: 2021/08/14.

[7] "Statistic of the week: How brazilian voters get their news," https://reutersinstitute.politics.ox.ac.uk/risj-review/statistic-week-how-brazilian-voters-get-their-news, Accessed: 2021/08/14.

[8] M. Schudson and B. Zelizer, “Fake news in context,” in *Understanding and Addressing the Disinformation Ecosystem*, Dec 2017, pp. 1–5.

[9] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” *Journal of Economic Perspectives*, vol. 31, pp. 211 – 236, 2017.

[10] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter,” in 20th *international conference on World wide web*, Mar 2011, pp. 675–684.

[11] “Story cherry-picks in assessing cnn ratings,” https://www.factcheck.org/2019/01/story-cherry-picks-in-assessing-cnn-ratings/, Accessed: 2021/08/16.

[12] “Top 10 fake science news of 2017,” http://global.chinadaily.com.cn/a/201801 /10/WS5a55a685a3102e5b17371dd1_2.html, Accessed: 2021/08/16.

[13] S. Zaryan, “Truth and trust : How audiences are making sense of fake news,” Master’s thesis, Sweden, 2017.

[14] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *Science*, vol. 359, pp. 1146 – 1151, 2018.

[15] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, “A large-scale study of the twitter follower network to characterize the spread of prescription drug abuse tweets,” *IEEE Transactions on Computational Social Systems*, vol. 6, pp. 1232 – 1244, 2019.

[16] “HoaXposed,” https://theprint.in/category/hoaxposed/, Accessed: 2020/04/16.

[17] “PolitiFact,” https://www.politifact.com/, Accessed: 2020/04/16.

[18] “Fact Checker,” https://www.washingtonpost.com/news/fact-checker/, Accessed: 2020/04/16.

[19] “FactCheck,” https://www.factcheck.org/, Accessed: 2020/04/16.

[20] “Snopes,” https://www.snopes.com/fact-check/, Accessed: 2020/04/16.

[21] “Truth or Fiction,” https://www.truthorfiction.com/, Accessed: 2020/04/16.

[22] “FullFact,” https://fullfact.org/facts/, Accessed: 2020/04/16.

[23] “Vishvas news,” https://www.vishvasnews.com/, Accessed: 2020/04/16.

[24] “Factly,” https://factly.in/, Accessed: 2020/04/16.

[25] L.-L. Shi, L. Liu, Y. Wu, L. Jiang, M. Kazim, H. Ali, and J. Panneerselvam, “Human-centric cyber social computing model for hot-event detection and propagation,” *IEEE Transactions on Computational Social Systems*, vol. 6, pp. 1042 – 1050, 2019.

[26] “Gartner top strategic predictions for 2018 and beyond,” https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond, Accessed: 2020/06/21.

[27] “Fake news is a real problem,” https://www.statista.com/chart/6795/fake-news-is-a-real-problem/, Accessed: 2020/06/21.

[28] “Can ‘fake news’ impact the stock market?” https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/?sh=4866d12b2fac, Accessed: 2020/06/21.

[29] “Study finds ’fake news’ has real cost: $78 billion-study finds ’fake news’ has real cost: $78 billion,” https://www.mediapost.com/publications/article/343603/study-finds-fake-news-has-real-cost-78-billion.html, Accessed: 2020/06/21.

[30] E. Dai, Y. Sun, and S. Wang, “Ginger cannot cure cancer: Battling fake health news with a comprehensive data

repository," *CoRR*, vol. abs/2002.00837, 2020. [Online]. Available: https://arxiv.org/abs/2002.00837

[31] E. Ferrara, "What types of covid-19 conspiracies are populated by twitter bots?" *First Monday*, vol. 25, pp. 1–12, 2020.

[32] "Fighting the infodemic: The #coronavirusfacts alliance," https://www.poynter.org/coronavirusfactsalliance/, Accessed: 2020/06/21.

[33] "Covid & whatsapp cause surge of fake news in india," https://www.statista.com/chart/25031/covid-19-misinformation-on-whatsapp-india/, Accessed: 2020/06/21.

[34] F. N. Ribeiro, K. Saha, M. Babaei, L. H. C. Lima, J. Messias, O. Goga, F. Benevenuto, K. P. Gummadi, and E. M. Redmiles, "On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook," *CoRR*, vol. abs/1808.09218, 2018. [Online]. Available: http://arxiv.org/abs/1808.09218

[35] "Fake news is poisoning brazilian politics. whatsapp can stop it," https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html, Accessed: 2020/06/21.

[36] "On whatsapp, rumours, and lynchings," https://www.epw.in/journal/2019/6/insight/whatsapp-rumours-and-lynchings.html, Accessed: 2020/06/21.

[37] "What is Natural Language Processing – NLP use cases and working," https://techvidvan.com/tutorials/natural-language-processing-nlp/, Accessed: 2020/06/21.

[38] "NLP techniques in data science with real life case studies," https://techvidvan.com/tutorials/nlp-techniques-in-data-science/, Accessed: 2020/06/21.

[39] M. Gopal, *Applied Machine Learning.* McGraw-Hill Education, 2019. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9781260456844

[40] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/information-retrieval-book.html

[41] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, Oct 2014, pp. 1532–1543.

[42] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[43] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[44] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[45] P. Kim, *Convolutional Neural Network.* Apress,Berkeley, CA, 2017. [Online]. Available: https://doi.org/10.1007/978-1-4842-2845-6_6

[46] "Lstm for text classification," https://www.analyticsvidhya.com/blog/2021/06/lstm-for-text-classification/, Accessed: 2020/06/21.

[47] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[48] G. Gravanis, AthenaVakali, K. Diamantaras, and P. Karadais, “Behind the cues: A benchmarking study for fake news detection,” *Expert Systems with Applications*, vol. 128, pp. 201 – 213, 2019.

[49] J. K. Burgoon, J. P. Blair, T. Qin, and J. F. Nunamaker, “Detecting deception through linguistic analysis,” in *Intelligence and Security Informatics*, May 2003, pp. 91–101.

[50] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, “Polarization and fake news: Early warning of potential misinformation targets,” *CoRR*, vol. abs/1802.01400, 2018. [Online]. Available: http://arxiv.org/abs/1802.01400

[51] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” *CoRR*, vol. abs/1708.07104, 2017. [Online]. Available: http://arxiv.org/abs/1708.07104

[52] Y. Chen, N. J. Conroy, and V. L. Rubin, “Misleading online content: Recognizing clickbait as ”false news”,” in *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, 2015, p. 15–19.

[53] C. Buntain and J. Golbeck, “Automatically identifying fake news in popular twitter threads,” in *IEEE International Conference on Smart Cloud (SmartCloud)*, Nov 2017, pp. 208–215.

[54] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, “Lying words: Predicting deception from linguistic styles,” *Personality and Social Psychology Bulletin*, vol. 29, pp. 665–675, 2003.

[55] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, “Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications,” *Group Decision and Negotiation*, vol. 13, pp. 81–106, 2014.

[56] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using n-gram analysis and machine learning techniques,” in *Intelligent, Secure, and*

Dependable Systems in Distributed and Cloud Environments, Oct 2017, pp. 127–138.

[57] K. Shu, S. Wang, and H. Liu, “Exploiting tri-relationship for fake news detection,” *ArXiv*, vol. abs/1712.07709, pp. 1–10, 2017.

[58] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, “Fake news early detection: An interdisciplinary study,” *CoRR*, vol. abs/1904.11679, 2020. [Online]. Available: http://arxiv.org/abs/1904.11679

[59] N. Vo and K. Lee, “Learning from fact-checkers: Analysis and generation of fact-checking language,” *CoRR*, vol. abs/1910.02202, 2019. [Online]. Available: http://arxiv.org/abs/1910.02202

[60] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, “Combating fake news: A survey on identification and mitigation techniques,” *CoRR*, vol. abs/1901.06437, 2019. [Online]. Available: http://arxiv.org/abs/1901.06437

[61] Y.-C. Ahn and C.-S. Jeong, “Natural language contents evaluation system for detecting fake news using deep learning,” in 16th *International Joint Conference on Computer Science and Software Engineering (JCSSE)*, July 2019, pp. 289–292.

[62] N. O’Brien, S. Latessa, G. Evangelopoulos, and X. Boix, “The language of fake news: Opening the black-box of deep learning based detectors,” in *workshop on “AI for Social Good”, NIPS 2018*, Nov 2018, pp. 1–5.

[63] V. Singh, R. Dasgupta, D. Sonagra, K. Raman, and I. Ghosh, “Automated fake news detection using linguistic analysis and machine learning,” in *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*, July 2017, pp. 1–3.

[64] M. Z. Khan and O. H. Alhazmi, “Study and analysis of unreliable news based on content acquired using ensemble learning (prevalence of fake news

on social media)," *International Journal of System Assurance Engineering and Management*, vol. 11, pp. 145–153, 2020.

[65] M. Mersinias, S. Afantenos, and G. Chalkiadakis, "CLFD: A Novel Vectorization Technique and Its Application in Fake News Detection," in 12th *Language Resources and Evaluation Conference (LREC)*, May 2020, pp. 3475–3483.

[66] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet – a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[67] Y. Zhou, K. Chen, L. Song, X. Yang, and J. He, "Feature analysis of spammers in social networks with active honeypots: A case study of chinese microblogging networks," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Aug 2012, pp. 728–729.

[68] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Data and Applications Security and Privacy XXIV*, 2010, pp. 335–342.

[69] P. Wanda and H. J. Jie, "Deepprofile: Finding fake profile in online social network using dynamic cnn," *Journal of Information Security and Applications*, vol. 52, pp. 1–13, 2020.

[70] J. S. Alowibdi, U. A. Buy, P. S. Yu, S. Ghani, and M. Mokbel, "Deception detection in twitter," *Social Network Analysis and Mining*, vol. 5, pp. 1–16, 2015.

[71] Erşahin, Buket and Aktaş, Özlem and Kılınç, Deniz and Akyol, Ceyhun, "Twitter fake account detection," in *International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 388–392.

[72] M. Mateen, M. A. Iqbal, M. Aleem, and M. A. Islam, "A hybrid approach for spam detection for twitter," in 14th *International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2017, pp. 466–471.

[73] F. C. Akyon and M. Esat Kalfaoglu, "Instagram fake and automated account detection," in *Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2019, pp. 1–7.

[74] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr 2018.

[75] M. Meyers, G. Weiss, and G. Spanakis, "Fake news detection on twitter using propagation structures," in *Disinformation in Open Online Media. MISDOOM 2020*, 2020, pp. 138–158.

[76] Z. Zhao, J. Zhao, Y. Sano, O. Levy, H. Takayasu, M. Takayasu, D. Li, J. Wu, and S. Havlin, "Fake news propagates differently from real news even at early stages of spreading," *EPJ Data Science*, vol. 9, pp. 1–14, 2020.

[77] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *IEEE* 13^{th} *International Conference on Data Mining*, 2013, pp. 1103–1108.

[78] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *CoRR*, vol. abs/1902.06673, 2019. [Online]. Available: http://arxiv.org/abs/1902.06673

[79] D. Benjamin, D. Horne, and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in 2^{nd} *International Workshop on News and Public Opinion*, Mar. 2017, pp. 1–9.

[80] C. Burfoot and T. Baldwin, "Automatic satire detection: Are you having a laugh?" in *ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 161–164.

[81] T. Mitra and E. Gilbert, “CREDBANK: A large-scale social media corpus with associated credibility annotations,” in *International AAAI Conference on Web and Social Media*, Aug. 2015, pp. 258–267.

[82] B. Riedel, I. Augenstein, G. Spithourakis, and S. Riedel, “A simple but tough-to-beat baseline for the fake news challenge stance detection task,” Apr. 2017, pp. 1–6. [Online]. Available: https://arxiv.org/abs/1707.03264

[83] K. Shu, S. Wang, and H. Liu, “Exploiting tri-relationship for fake news detection,” Dec. 2018, pp. 1–10. [Online]. Available: https://arxiv.org/abs/1712.07709v1

[84] W. Y. Wang, “Liar, liar pants on fire: A new benchmark dataset for fake news detection,” in 55^{th} *Annual Meeting of the Association for Computational Linguistics*, vol. 2. ACM, Aug. 2017, pp. 422–426.

[85] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using n-gram analysis and machine learning techniques,” in *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, vol. 10618. Cham: Springer, Oct. 2017, pp. 127–138.

[86] “Mcintire fake news dataset,” https://github.com/lutzhamel/fake-news, online; accessed 15 April 2020.

[87] “Fake news kaggle dataset,” https://www.kaggle.com/c/fake-news/data?select=train.csv, online; accessed 15 April 2020.

[88] P. K. Verma, P. Agrawal, and R. Prodan, “WELFake dataset for fake news detection in text data,” Feb. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4561253

[89] M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami, M. M. Hassan, and G. Fortino, “Credibility in online social networks: A survey,” *IEEE Access*, vol. 7, pp. 2828–2855, 2019.

[90] E. Lancaster, T. Chakraborty, and V. S. Subrahmanian, "$malt^p$: Parallel prediction of malicious tweets," *IEEE Transactions on Computational Social Systems*, vol. 5, pp. 1096–1108, 2018.

[91] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE Transactions on Multimedia*, vol. 19, pp. 598–608, 2017.

[92] A. De Salve, P. Mori, B. Guidi, and L. Ricci, "An analysis of the internal organization of facebook groups," *IEEE Transactions on Computational Social Systems*, vol. 6, pp. 1245–1256, 2019.

[93] B. Ratner, "The correlation coefficient: Its values range between $+1/-1$, or do they?" *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 17, pp. 139—142, 2009.

[94] T. Mitchell, *Machine Learning.* McGraw-Hill Education, 1997. [Online]. Available: http://www.cs.cmu.edu/ tom/mlbook.html

[95] "Introduction to tf-idf," http://www.tfidf.com/, Accessed: 2020/06/21.

[96] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manag.*, vol. 57, p. 102025, 2020.

[97] H. Karimi, P. Roy, S. Saba-Sadiya, and J. Tang, "Multi-source multi-class fake news detection," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1546–1557.

[98] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," *CoRR*, vol. abs/1811.00770, 2018. [Online]. Available: http://arxiv.org/abs/1811.00770

[99] "Social media fact sheet," https://www.pewresearch.org/internet/fact-sheet/social-media/, Accessed: 2020/06/21.

[100] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," *Council of Europe report*, vol. 27, pp. 1–109, 2017.

[101] "A viral rumor that mcdonald's uses ground worm filler in burgers has been debunked," https://www.businessinsider.in/A-viral-rumor-that-McDonalds-uses-ground-worm-filler-in-burgers-has-been-debunked/articleshow/50676282.cms, accessed 6 Oct. 2020.

[102] Y. Cheng and Z. F. Chen, "The influence of presumed fake news influence: Examining public support for corporate corrective response, media literacy interventions, and governmental regulation," *Mass Communication and Society*, vol. 23, pp. 705–729, 2020.

[103] P. K. Verma and P. Agrawal, "Study and detection of fake news: P^2C^2-based machine learning approach," in 4^{th}*International Conference on Data Management, Analytics and Innovation*, 2020, pp. 261–278.

[104] X. Ruan, Z. Wu, H. Wang, and S. Jajodia, "Profiling online social behaviors for compromised account detection," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 176–187, 2016.

[105] "Facebook profile: Terms of service," https://www.facebook.com/legal/terms, accessed 6 Oct. 2020.

[106] "Facebook, twitter remove russia-backed fake account network," https://www.gadgetsnow.com/tech-news/facebook-twitter-remove-russia-backed-fake-account-network/articleshow/74626530.cms, accessed 6 Oct. 2020.

[107] F. Tchakounté, K. Amadou Calvin, A. A. A. Ari, and D. J. Fotsa Mbogne, "A smart contract logic to reduce hoax propagation across social media," *Journal of King Saud University - Computer and Information Sciences*, pp. 1–9, 2020.

[108] M. A. Wani and S. Jabin, "Mutual clustering coefficient-based suspicious-link detection approach for online social networks," *Journal of King Saud University - Computer and Information Sciences*, pp. 1–14, 2018.

[109] M. Al-Qurishi, M. Al-Rakhami, A. Alamri, M. Alrubaian, S. M. M. Rahman, and M. S. Hossain, "Sybil defense techniques in online social networks: A survey," *IEEE Access*, pp. 1200–1219, 2017.

[110] E. V. D. Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," *IEEE Access*, pp. 6540–6549, 2018.

[111] "News use across social media platforms 2018," https://www.pewresearch.org/journalism/2018/09/10/news-use-across-social-media-platforms-2018/, accessed 6 Oct. 2020.

[112] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, pp. 1094–1096, 2018.

[113] "What to know about pizzagate, the fake news story with real consequences," https://time.com/4590255/pizzagate-fake-news-what-to-know/, accessed 6 Oct. 2020.

[114] "Bolsonaro business backers accused of illegal whatsapp fake news campaign," https://www.theguardian.com/world/2018/oct/18/brazil-jair-bolsonaro-whatsapp-fake-news-campaign, accessed 6 Oct. 2020.

[115] A. Marwick and R. Lewis, "Media manipulation and disinformation online," New York, Tech. Rep., 2017.

[116] R. Guha, K. K. Ghosh, S. Bhowmik, and R. Sarkar, "Mutually informed correlation coefficient (micc) - a new filter based feature selection method," in *2020 IEEE Calcutta Conference (CALCON)*, 2020, pp. 54–58.

[117] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media,” *CoRR*, vol. abs/1809.01286, 2018. [Online]. Available: http://arxiv.org/abs/1809.01286

Appendix A

List of publications

1. P. K. Verma, P. Agrawal, I. Amorim and R. Prodan, “WELFake: Word Embedding Over Linguistic Features for Fake News Detection,” in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881-893, Aug. 2021, doi: 10.1109/TCSS.2021.3068519.

2. Pawan Kumar Verma and Prateek Agrawal , “Study and Detection of Fake News: P^2C^2 Based Machine Learning Approach,” Fourth International Conference on Data Management, Analytics and Innovation - ICDMAI 2020, 17-19 January, 2020 at United Services Institute (USI), New Delhi, India.

3. Pawan Kumar Verma and Prateek Agrawal , “PropFND: Propagation based Fake News Detection,” International Conference on Advances and Applications of Artificial Intelligence and Machine Learning - ICAAAIML-2021, 29 - 30 October, 2021 at Shatda University, Greater Noida, India.

4. Pawan Kumar Verma, Prateek Agrawal, Vishu Madaan and Radu Prodan, “MCred: Multi-Modal Message Credibility for Fake News Detection using BERT and CNN,” Journal of Ambient Intelligence and Humanized Computing. (under review)

5. Pawan Kumar Verma, Prateek Agrawal, Vishu Madaan and Charu Gupta, “UCred: Fusion of Machine Learning and Deep Learning Methods for User Credibility on Social Media,” Social Network Analysis and Mining. (under review)