A Node2vec Based Fake News Detection Model Using News Source URL

A thesis presented for the degree

of

Bachelor of Science in Engineering in

Computer Science and Engineering

Department of Computer Science and Engineering,

University of Chittagong,

by

Mosharat Jahan

Student ID : 18701040

Session : 2017-2018

Under the Supervision of

Mr. Rezaul Karim

Associate Professor

Department of Computer Science and Engineering

University of Chittagong

August 01, 2023

Report Code:

University of Chittagong

Department of Computer Science and Engineering

8th Semester B.Sc. Engineering Examination 2021

Course No: CSE 800

Title: A Node2vec Based Fake News Detection Model Using News Source URL

Report Code:

University of Chittagong

Department of Computer Science and Engineering

8th Semester B.Sc. Engineering Examination 2021

Course No: CSE 800

Title: A Node2vec Based Fake News Detection Model Using News Source URL

Student Name: Mosharat Jahan

Student ID: 18701040

Session: 2017-2018

Hall: Shamsun Nahar

Signature of Student:

Submission Date: 01 August, 2023

Approval for submission

This thesis entitled as A Node2vec Based Fake News Detection Model Using News Source URL, by Mosharat Jahan, Student ID:18701040, Session: 2017-2018, has been approved for submission to the Department of Computer Science and Engineering, University of Chittagong, in partial fulfillment of the requirements for the Bachelor of Science in Engineering in Computer Science and Engineering.

Mr. Rezaul Karim
Associate Professor
Department of Computer Science and Engineering University of Chittagong
Chittagong-4331, Bangladesh.

Abstract

The rapid expansion of digital media and the internet has facilitated the spread of fake news, posing serious threats to information integrity and societal harmony. Now is the time of internet where every incident becomes a news and every news is seen all over the world in mere seconds. Which makes the possibility of spreading a false news more prominent. In a country like ours where people tend to be very excited over a simple news, spreading of fake news can be very impactful in many ways. There are numerous examples where a fraud and falsified news has done severe harm to people. So with time many researches are being conducted to detect fake news in different languages. Though most of the researches are conducted regarding high resource languages but the impact of fake news is not confined into language. This thesis presents a novel approach to fake news detection, leveraging web search results and harnessing the power of the node2vec algorithm. The objective of this proposed model is to detect fake news from various Bengali websites as there isn't many researches conducted regarding this language.

Acknowledgement

This is to certify that the thesis entitled **A Node2vec Based Fake News Detection Model using News Source URL**, submitted in full result.

I am grateful to my supervisor, **Mr. Rezaul Karim**, Associate Professor of Department of Computer Science and Engineering, University of Chittagong, for his kind word of support, encouragement and valuable time.

I wish to extend my profound sense of gratitude to my parents for all the sacrifices they made during my work and also providing me with moral support and encouragement whenever required.

Mosharat Jahan

Contents

List of tables

List of figures

Chapter 1

Introduction

The propagation of false news has grown to be a significant concern in today's quickly expanding digital era, posing substantial dangers to information integrity and public dialogue. It is essential to create strong and effective solutions to address this problem since the ease with which information can be shared on social media and the size of the internet have accelerated the spread of disinformation. Nowadays, news sharing amongst people is similar to an unofficial competition. As absurd as it may sound, it is a truth that individuals now evaluate one another primarily on the number of likes, comments, and views they receive on social media profiles. But some people often use unethical methods to gain access to these beliefs and preferences. The most popular method is to post erroneous or unverified news in order to periodically excite visitors and garner the desired likes, views, and shares. Even while it might not seem harmful at first, not all news has an impact on everyone. News that seems unimportant to us may be of great importance to someone else. People also utilize fake news for their personal gain, which is quite evident when we look at political parties in any nation. Political parties frequently broadcast misleading information about themselves or about their competitor parties in an effort to attract support and win elections. People's faith in online news is now seriously threatened by fake news as well as phoney product or other promotions. Sometimes the news may not be false, but the headline may be embellished to draw attention to something unrelated to the news it contains. The need to identify bogus news on the internet has become unavoidable as consumers these days prefer social media-based news to traditional news printed media. Some websites, such as www.politifact.com, www.factcheck.org, and www.jaachai.com, where the administrators manually update erroneous reports published on the internet with logical and factual justification, can be used to discover false news [1]. However, these websites are unable to operate in real-time and are not equipped to react swiftly to any breaking news. Numerous academics use a variety of techniques to effectively identify bogus news. Researchers and policymakers alike have shown a great deal of interest in fake news identification because the

effects of disinformation can result in societal discord, political instability, and a decline in public confidence in the media. In order to find linguistic patterns and textual discrepancies, natural language processing (NLP) techniques are typically used in traditional approaches of fake news identification. While these measures have had encouraging results, they frequently have problems coping with the constantly changing nature of fake news, as offenders constantly modify their evasion tactics. This thesis suggests a novel method for identifying fake news utilizing web search results and the potent node2vec algorithm to address these problems. We can extract rich semantic data from the network of linked web sites using the node2vec algorithm, a well-known node embedding method in graph analytics. This enables us to establish a more thorough picture of the information ecosystem surrounding a certain news story.

This study's main goal is to create a novel false news detection algorithm that goes beyond existing content-based approaches. We hope to use the collective wisdom of the internet to confirm or deny the veracity of a news piece by combining web search results into the study. This will provide us a distinctive viewpoint and enable us to cross-validate data from other sources, improving the accuracy and resilience of the model.

1.1 Background

News or article that are written with the intention of misleading and deceiving people with falsified or distorted content can be called fake news. The ease of access to internet and media has eventually increased peoples dependency on different internet platforms for any news. A news need not be fully human made for it to be called fake news. Even by changing a single date of a real news also makes the news equally false or just by changing a name of the place an incident took place. Rumors are also defined as fake news which is for an unknown reason very favorite of people now a days.

Every single day a new rumor circulates in the internet for gaining some public hype and attention. In order to mislead an occurrence, a story, or a target audience, fake news is typically circulated. Research demonstrates that people are affected for a long time by the news that

circulates online. Because of this, false information is spread to cause unwarranted anxiety in people. Prior to the internet, most people read news via printed sources because it was more difficult to access the news. Therefore, malicious individuals couldn't distribute fake information through the media as quickly as they do now. With the increased use of the internet and social media, people can now more easily get both true and fake news. Each fake news detection study chooses a concept that is relevant to the research question. We continue to refer to fake news in this study as "misinformation on the web" [2]. It consists of sarcastic fabrications, serious hoaxes, widespread hoaxes, rumors, social spam, etc. The concept also emphasizes other areas, including politics, entertainment, and society. In this age, the spread of fake news has become a worldwide issue that crosses linguistic and geographic boundaries. False information has also become an unavoidable problem in the Bengali language setting that affects public confidence in media and information sources. Even though techniques for spotting false news have been examined in depth in English and other widely spoken languages, tackling this issue in the unique setting of the Bengali language necessitates specialized and culturally aware solutions. Due to linguistic complexities, different script structures, and a dearth of language-specific datasets, traditional false news detection algorithms, while effective in some languages, may present special difficulties when used with Bengali. The lack of labeled false news data in Bengali makes it difficult to create precise and trustworthy models in this language.

This thesis suggests a novel false news detection methodology that is especially suited for Bengali language to address these issues. The method takes into account the peculiarities of the language and makes use of the collective wisdom of Bengali-speaking internet users while integrating web search results and utilizing the capabilities of the node2vec algorithm. We can access local comprehension and perception of news articles within the cultural and linguistic context by taking into account Bengali web search results. In addition to the conventional content-based analysis, this additional contextual information offers helpful indicators to confirm or deny the veracity of a news article. The Bengali adaptation of the node2vec technique creates embeddings for linked web pages that capture semantic relationships in the Bengali information network. This makes it possible to comprehend contextual linkages between news stories better and makes it easier to spot linguistically specific patterns of disinformation. The importance of this study rests

in its contributions to the creation of fake news detection algorithms that are uniquely adapted for Bengali. The suggested framework aims to increase media literacy and credibility among Bengali-speaking people by offering a cutting-edge and culturally appropriate method to addressing the problems of disinformation. The results of this study could ultimately help not only the Bengali-speaking community but also other languages with a dearth of labeled data for spotting bogus news. The knowledge gathered from this research can direct the creation of detection algorithms that are specific to a certain language and contribute to a larger global effort to combat false news across many linguistic environments.

1.2 Fake News

The spread of information through internet platforms and social media in the contemporary digital age has created both opportunities and difficulties. The rise of "fake news," a word that has attracted a lot of attention recently, is one of the most important issues. The purposeful dissemination of incorrect or misleading material, frequently disguising it as legitimate news with the intent to confuse readers or audiences, is referred to as fake news.

Fake news can take many different forms, making it a multidimensional and challenging problem to solve. It consists of completely made-up tales, deceptive headlines, doctored photos, facts taken out of context, and disinformation campaigns. False news can be produced and spread for a variety of reasons, such as sensationalism, financial gain, the promotion of particular ideologies, or swaying public opinion.

The emergence of digital media and social networking sites has greatly accelerated the presence and spread of fake news. Information may now be instantly shared and spread, quickly reaching a large worldwide audience. Sadly, this speed and accessibility also make it difficult for users to distinguish factual information from manufactured content, thus amplifying misinformation.

The effects of fake news may be widespread and harmful to society. It can aggravate societal differences, diminish public confidence in reliable news sources, and skew public conceptions of

reality. The impact of fake news can be particularly harmful in important circumstances, such as elections, public health emergencies, or geopolitical events, deceiving the public and affecting decision-making.

Fake news needs to be addressed from multiple angles, including media literacy, fact-checking programs, and the creation of efficient technical solutions. In recent years, a potential method for spotting false news has evolved through the use of cutting-edge machine learning and natural language processing algorithms.

This thesis proposes a novel method for detecting fake news in an effort to support ongoing efforts to prevent it. The suggested approach makes use of the node2vec algorithm, which builds node embeddings for news sources using web search results. By using such embeddings, we want to capture complex linkages between news sources based on their links in web graphs, potentially exposing patterns indicating the dissemination of bogus or real news. This thesis will also investigate and evaluate the effectiveness of current fake news detection techniques that depend on reading news content. Text-based techniques have shown potential in identifying false news based on linguistic patterns and semantic clues, such as those using Count Vectorization or TF-IDF for feature extraction.

This study aims to shed light on the efficacy of various methods for identifying false news through a thorough analysis and comparison of the node2vec-based model and content-based models. We hope to contribute to the creation of stronger, more precise tools to combat disinformation by improving our understanding of fake news detection techniques. This will eventually lead to the development of a society that is more informed, discerning, and resilient.

1.3 Impacts of Fake News

Not even a decade ago people used to wake up in the morning and first thing they did was to check weather today's news paper is delivered or not. Spotting people in road side tea stalls sipping their morning tea with a news paper in hand was a regular picture. This reminds us how

significant news papers are in our daily life to cope up with the world. But with time this scenario is rare now to even imagine. Now people do first thing in morning is to open their mobile phones and scroll through the news in internet. If we are talking about the impact of news then the impact of fake news also comes into light. Rumors and distorted news have a huge ability to change peoples view and behavior towards a particular incident.

There are many examples in Bangladesh of different tragic incidents due to rumors. In July 2019, five people were beaten to death and ten injured by mobs as a result of widespread news about the expected human sacrifice in the construction of the Padma Bridge [3]. At the time of 2016 US election, 25 percent of the Americans browsed a fake news website in the period of election which has been hypothesized as one of the issues that influenced the final results [4]. Like this there are vast amount of incidents which shows us the impact of falsified news so detection of fake news is unavoidable in every aspect.

1. The Degradation of Trust The loss of public confidence in media and information sources is one of the most important effects of fake news. A weakened foundation for informed decision making results from people's growing skepticism of the veracity of news sources and their propensity to discard even legitimate reports as fake material.

2. Secondly, social polarization Fake news frequently focuses on polarizing topics, capitalizing on feelings and pre-existing opinions to worsen societal polarization. Disseminating false information on purpose can aggravate societal divisions by encouraging people to retreat into echo bubbles and get into contentious arguments that are driven by false information.

3. Third, political influence Political environments can be strongly impacted by misinformation, with fake news items impacting voter attitudes and behavior. The integrity of democratic processes can be threatened, and national stability can be threatened by the deliberate deployment of fake news as a tool for political advantage.

4. Economic Implications The spread of false information may also have negative effects on the economy. False information about the economy, businesses, or financial markets can cause instability and panic, which can result in financial losses and decreased investor confidence.

5. Inappropriate Resource Allocation The spread of false information diverts resources and attention away from pressing problems and emergencies. When people and organizations react to incorrect information, important time and energy are diverted from solving urgent society problems to disprove falsehoods.

6. Public Security: Fake news occasionally poses dangers to the safety and security of the general public. For instance, inaccurate information on health crises, catastrophes, or disasters can hinder efficient response efforts and possibly endanger lives.

7. Negative effects on reputations Reputation of people, companies, and public personalities can suffer permanent harm by the dissemination of false and defamatory material. When bogus news spreads widely and quickly across digital networks, reputation management becomes more difficult.

The overall effects of false news highlight the importance of taking watchful and aggressive steps to stop it from spreading. To promote media literacy and protect the integrity of information in the digital age, it is crucial to establish reliable detection techniques and public awareness campaigns as misinformation grows more complex. The suggested node2vec-based fake news detection model, which is based on online search results and is designed for the Bengali language, aims to support these efforts by providing a novel strategy to address the problems caused by false information in a context that is unique to that language.

1.4 Motivation

Information and knowledge are just a click away in the world of internet. Even for tiny amount of query people tend to search it up on web. And there are thousands of answers of same question. But we need to remember not all of them are true or acceptable. Many people take advantage of internet and peoples innocence to distract them in a way that is a total waste of time for them. Some have pre decided intention for doing this and some just find it fun. Indeed what a ridiculous way to have fun! But what can one do, this is the way world is revolving now.

People use internet to keep track with the world news. But if you are keeping track of fake news then what is the motive of this total concept of keeping track with world. Most of the researches in this field revolves around English language mostly and to our knowledge there is not so many researches done to detect and tackle fake news spread in Bangla. So to outline and analyze a way to deal with fake news issue, this work is presented.

The urgent necessity to stop the spread of fake news among speakers of the Bangla language is what spurred this study's motivation. The Bangla-speaking community stands an increased risk of falling prey to false information because it is one of the most commonly spoken languages in the world. This can negatively affect societal cohesion, political processes, and public trust. Due to the quick development of the digital era, disinformation has found a home on social media and online forums that specifically targets Bengali speaking users. Bengali's linguistic peculiarities make it difficult to identify bogus news with any degree of accuracy. To address the particular traits of Bangla disinformation, which are complex sentence structures, linguistic variances among regions, and cultural nuances, specialist approaches are required. It is critical to modify current methodology and investigate cutting-edge methods specifically designed for the Bengali language in order to create robust detection models that can withstand the increasing strategies of false news dissemination.

Fake news has effects that go beyond one person's beliefs and perceptions. The democratic fabric of the area is seriously threatened since it has the capacity to affect political beliefs, electoral outcomes, and governmental decisions. The goal of this research is to provide the Bengali-speaking community with the tools necessary to evaluate the veracity of information critically, promoting media literacy and reasoned decision-making. To that end, a comprehensive fake news detection model for Bangla is being developed.

Furthermore, investor confidence and economic stability depend on the availability of trustworthy information. The spread of false information about financial markets, companies, or economic policies can result in financial losses and obstruct growth potential. These risks can be reduced, and a strong false news detection system can help create a more stable business environment. Beyond the digital sphere, this study's consequences are broad. The detection

model seeks to rebuild public trust in reliable news organizations and information sources by fostering media integrity and improving the accuracy of information. As a result, society may become more cohesive and united and people may feel empowered to have productive conversations based on fact-checked information. A language-specific false news detection model can also offer important support for policy measures catered to the Bangla-speaking community as governments and regulatory agencies work to address the challenges of misinformation. In order to combat fake news and maintain the integrity of information distribution, it presents a chance to coordinate regional initiatives with global efforts.

This study is motivated by the urgent need to battle false information in the Bangla language, promote media literacy, and enable the Bengali speaking people to distinguish between trustworthy information and misleading lies. This work seeks to contribute to a more knowledgeable, resilient, and digitally secure Bengali-speaking community, capable of navigating the information landscape with assurance and critical thinking. It does this by creating an effective and culturally relevant fake news detection algorithm.

1.5 Problem Statement

When reading a news from an article or a link not many people cares about the authenticity of the news. But the news they read does keeps a trace in their mind for some time or throughout the day. If someone reads a news about the weather he or she will take an umbrella in case the news shows that there is a chance of rain and will carry a water bottle in case the news shows it will be a very sunny day. Here the authenticity of the news is not too necessary but if the news turns out to be false and someone is carrying an umbrella throughout the day it will just be an extra baggage and nothing else. Though this is just an example but still because of a simple false news someone is facing a problem. So if there is a way that one can detect if the news is true or not even if it is a small and not harmful news one can relieve themselves from some extra hard work like carrying an umbrella. There have been many attempt made to detect fake news from internet like using social media platforms as domains but if we are going to use the social media

as a domain then we can only get the information an person or a news site has posted in their profile nothing more than that. If we are going to detect fake news by making a whitelist of the news found in different web portals then we will be needing human intervention for making that whitelist which is too much of a labor. There also have been many studies that uses linguistics features to identify which news is fake and which is true but as fake news are becoming more similar to the pattern of real news this process is not so useful now. Other than that the most thought about problem in this study is the language barrier. All this studies we are talking about were mostly done in English language or a high resource language like English but what about the low resource languages like Bangla? So keeping this in mind this study will focus on Bangla language to detect fake and falsified news from web. The main objective of this thesis is to create a node2vec-based model for fake news identification that makes use of web search results to find and combat false information in the Bangla language. The issues provided by fake news in the digital age, notably in the context of the Bengali-speaking population, are the subject of this study's concern.

1. The prevalence of fake news in the Bangla language causes serious problems for the accuracy of information and the trustworthiness of the media. The objective of the issue statement is to address the need for a context-sensitive, effective false news detection system that is specifically designed for Bengali.

2. Large and trustworthy labelled datasets are necessary for creating an accurate false news detection model. However, there aren't many of these datasets available in the Bangla language setting, which makes it difficult to create efficient detection models.

3. To verify the veracity of news stories, the suggested approach includes online search results as an additional source of data. Investigating how traditional content-based approaches can be complemented by web search-based verification can improve overall detection accuracy, according to the issue statement.

4. Several areas have shown that the node2vec approach is excellent in capturing semantic relationships in graph-structured data. Understanding node2vec's relevance to the context of

Bengali false news identification and its potential to enhance model performance are the main goals of the problem description.

5. Bengali's distinctive linguistic characteristics, regional variances, and cultural nuances provide particular difficulties in spotting bogus news. The goal of the issue statement is to develop a reliable and culturally appropriate detection model by resolving these contextual and linguistic difficulties.

6. The problem statement acknowledges the wider societal effects of fake news, such as its impact on political discourse, public trust, and social cohesion. The suggested methodology seeks to lessen these negative consequences by enabling Bengali speakers to make wise judgements and participate critically.

By addressing these aspects of the issue statement, a novel and efficient false news detection model will be created, promoting media literacy, a better information ecology, and the digital well-being of the Bangla-speaking community. In order to increase the credibility and trustworthiness of information distributed among Bengali-speaking people, the research aims to offer insightful knowledge about minimizing the spread of false information.

1.6 Significance of the Study

Due to its potential to revolutionize the field of fake news identification, this research is extremely important. We contribute to the development of a more reliable and accurate detection framework by expanding the present understanding of false news detection mechanisms and introducing a novel strategy that makes use of web search results and node2vec embeddings. In turn, this can promote an information ecosystem where trustworthy news sources are recognized from false ones, enabling people to make informed decisions and promoting a more democratic dialogue.

In addition, a variety of stakeholders, such as legislators, media outlets, fact-checking organizations, and technological businesses, will find the study's conclusions to be quite useful.

They can make use of the newfound knowledge to hone their tactics for battling fake news and protecting the accuracy of information published on various digital platforms.

1.7 Organization of the Thesis

- Chapter 1 of the thesis starts with an introduction of Fake News. It is followed by background of of the Fake News Detection, objectives and organization of the thesis.
- Chapter 2 examines the literature review current methods for identifying fake news, including content-based strategies, graph-based methods, and node2vec's uses in a variety of fields.
- In Chapter 3, the research methodology is presented including a description of the data collection procedure, the development of node2vec embeddings, and the application of the suggested false news detection model.
- In Chapter 4, a thorough examination of the experimental findings is provided ,which also compares the model's performance to more conventional techniques.
- In Chapter 5, the findings are thoroughly discussed, along with the results of using node2vec embeddings and online search results in the identification process. The ability of the model to spot previously undiscovered phoney news stories as well as the reduction of false positives and false negatives are both thoroughly examined in Chapter 5.
- Chapter 6 brings the thesis to a conclusion by outlining the research's major accomplishments, pointing out its shortcomings, and offering future prospects for improving false news identification and reducing the spread of disinformation in the digital age.

Chapter 2

Literature Review

As stated previously there have many attempts to detect fake news from internet by many researchers. And most of the attempts were made regarding content based features. The news content has many parts like image, text etc but as a huge part is consisted by text so main focus always remains on text based fake news detection. But the limitation in this way is not negligible because even though previously the pattern of text in fake news and real news could be separated, now it is not the case. So text based detection in loosing its credibility with time. To avoid loopholes of text based detection many studies focus on context based features which includes the idea of linguistic features of the news. In this approach user based features and network based features are focused on [5]. Which means the properties of news source and news link are studied to find out if the news is true or not. Just like different features every studies uses different methods for their corresponding work. Most of the previous studies used Support Vector Machine(SVM) [6]. It was intended to show that absurdity, punctuation marks and grammar are best for identifying poignant news by collecting a number of news articles from some news websites.

Links on the internet, a network-based feature, could potentially be a valuable resource for context data. The type of URL and host information are crucial factors in recognizing suspicious URLs, even though they are not related to the identification of fake news [7]. A research examined a number of link attributes to identify bogus news [8]. They created link-related features based on the presumption that connections to legitimate news will have a https prefix, .gov, .co, and.com domain extension. In addition, links to bogus news are sometimes excessively long, brief, complex numerical, or blog hosts. Additionally, they attempted to turn connections into one-hot vectors and utilize them as features for fake news identification, but when used in isolation, they produced no useful results. However, because all the features connected to the linkages were synthesized, they were combined into a single feature variable. By integrating this

variable in the model, the baseline model's performance was somewhat enhanced (2.44%), and accuracy of 53.28% was produced. But when compared to other factors like the text variable (64.35%) and the Twitter variable (57.22%), it was discovered that the conversion did not significantly boost performance. Use of "https" as a feature to assess the credibility of the URL had the opposite effect, with more "https"-enabled URLs turning up in fake news [9]. Further research is required because their study lacked sufficient links to support it. The two study instances listed above used URL-based source data often for false news detection research, but they were unable to demonstrate a discernible improvement in the performance of the models.

Among all the studies, the research done by Vishwakarma et al. is a noteworthy one. His study proposed URL search based result fake news detection using a set of reliable links. This research shows 85 percent accuracy and which shows the effectiveness of using web search links in detecting fake news. But the idea of the this whole study was not new as it used the concept of white list based phishing detecting [10]. Even after considering the effectiveness of this study there are still some drawbacks of this research which is it needs human participation to collect the set of links.

Another research by Perez Rosa et al. which includes previously mentioned SVM model used only linguistic features of text. They also collected a dataset of 240 legitimate news from different news websites in US and made another dataset consisting the fake version of news from the previous dataset. They used crowd sourcing to generate this fake version of news using Amazon Mechanical Turk(AMT) [11].Moreover as we can see here also human intervention is needed to collect the relevant dataset and also to generate false version of the news collected in the dataset. Moreover to extract more secretive characteristics of false news neural network based model are undoubtedly more reliable.

There are also some researches of clickbait detection which also uses linguistic features and neural network with dataset that normally contains some click baits from various news websites [12]. Not only English but there are some studies done on a couple of low resource languages like Indonesian using Naive Bayes classifier [13].

A noteworthy Research was done in Bangla language using SVM and MNB classifier that performs better than Nave Bayes [3]. Their study serves as an example of an experimental investigation into spotting fake news on Bangla social media, an issue that still needs a lot of attention. To identify bogus news in Bangla, they have used two supervised machine learning approaches, Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) classifiers. Inverse Document Frequency - Term Frequency As far as feature extraction goes, Vectorizer and CountVectorizer have been employed. The proposed method distinguishes bogus news based on the polarity of the connected post. Finally, the research shows that SVM with a linear kernel outperforms MNB with a 96.64 percent accuracy compared to MNB's 93.32 percent accuracy.

Another mentionable work in Bangla fake news detection was the creation of a first publicly accessible news dataset with the aim of collecting 50k data [1]. An annotated dataset of 50K news stories was provided in this study that can be utilized to create automated false news detection systems for a low-resource language like Bangla. Additionally, they offered a dataset analysis and created a benchmark system using cutting-edge NLP algorithms to detect bogus news in Bangla. They researched conventional linguistic aspects and neural network-based techniques to develop this system.

A new approach to detect fake news was initiated using Link2vec mechanism, which is a self supervised learning model in order to differentiate between fake and real news gained from web search result using Word2vec [5]. The result of this study was very appreciable as they showed that this way is far better than whitelisting method and doesn’t need human intervention. They used a composition pattern of web links containing news content as a new source of information for fake news detection. To properly vectorize the composition pattern of web links, This paper suggested a novel embedding method called link2vec, an extension of word2vec, to properly vectorize the composition pattern of online links. They deployed their link2vec-based model to two real-world fake news datasets in different languages (English and Korean) to assess its efficacy and language independence. They also used the traditional text-based model and a hybrid model that integrated text and whitelist-based link information suggested by a previous study as comparison models. The link2vec-based detection models outperformed all the

comparison models with statistical significance in the datasets in two languages, according to the results.

When news content and social context variables are combined, a novel machine learning (ML) false news detection strategy that outperforms other methods in the literature and boosts accuracy to 78.8% was initially proposed by Marco L. Della Vedova et al [14]. Second, they applied their technique within a Facebook Messenger Chabot and tested it in a real-world setting, achieving an 81.7% accuracy in detecting fake news. They first discussed the datasets they utilized for their test, then showed the content-based approach they employed and the way they suggested to combine it with a social-based approach that was previously proposed in the literature. Their objective was to categorize a news item as reliable or fraudulent. The 15,500 posts in the resulting dataset have 2,300,00+ likes, while 8,923 of them are fake and 6577 are real.

The traditional strategy is one that relies on content since it may be used in traditional news outlets and, more broadly, in any situation where there is a lack of social information. These techniques have historically been applied to web pages [15] and email communications [16] to detect spam. They have also been used in recent years to identify bogus news: Using convolutional neural networks with text and additional metadata on a large political fake news dataset, [17] used syntactic and semantic features for classifying between real articles and articles generated by sampling a trigram language model, obtaining a 91.5% accuracy; [18] in the context of a recent fake news challenge, [19] demonstrated the effectiveness of using a relatively simple approach based on term frequency (TF) and term frequency-inverse document frequency (TF-IDF).

Current research on clickbait detection uses deep neural networks [20] and hand-crafted linguistic features [21] with datasets that mostly contain clickbaits (headline and article of clickbait news) from various newspapers. Additionally, a knowledge base-based fact-checking system has been suggested by [22]. Additionally, [23] have suggested a fully automated fact-checking process using outside sources. Their dataset contains 761 claims from snopes.com that cover a variety of topics, including politics, local news, and interesting facts. Each assertion is

classified as either factually accurate (34%), a false rumor (66%), or both. There are extremely few works available for other low resource languages including Indonesian, Bangla, and Hindi, despite the fact that research focusing on the English language has made tremendous progress. 250 pages of fake and real news articles were compiled into a dataset by [24], who also suggested an Indonesian-language hoax detection model utilizing the Naive Bayes classifier. [25] developed a real-time news certification system for the Chinese microblogging platform Sina Weibo using a dataset of 50K news items.

To detect breaking news rumors of emerging topics in social media a research was done based on twitter only using both Word2vec and LSTM-RNN model, which was capable to detect breaking news rumors [26]. The majority of research in the subject of fake news detection focuses on the idea of examining and identifying hoaxes on their primary distribution channel: social media. Examples of the above are [27] and [28], which use traditional machine learning techniques (classification trees, SVM,...) to determine the likelihood that a given post is false based on its own features like likes, follows, shares, etc. The best results from using these kinds of approximations may be seen in [21], where click-bait news is identified with an accuracy of 93%. Other studies, like [29], analyze the relationships between people who disseminate news and the direction that the content takes in order to stop it and reduce its possible deceptive consequences. Another article [30] presents a straightforward approach to detecting fake news that is based on the Naive Bayes classifier, one of the AI algorithms. The study's objectives are to investigate the effectiveness of this strategy for a given problem using a manually labelled (fake or real) dataset and to provide evidence in favor of the use of machine learning to identify fake news. This article differs from others on related subjects in that it heavily relies on a Naive Bayes classifier that is used to distinguish between fake news and real news. Additionally, the developed system was tested on a recent data set, giving the opportunity to compare its performance to the most recent data.

Fake news detection on images has not been the subject of many studies. According to [31], real and fake news have different distribution patterns for the photographs. They offer a collection of visual features for news verification that are taken from visual picture material and expose the hidden properties of image distributions in news events. In addition to the aesthetic aspects, it

also suggests a number of statistical picture features that are utilized to condense image statistics. For news authentication, visual elements are combined with statistical information. [32] employs K-means clustering algorithm (based on publishing day) to obtain a broad overview of how the photographs were utilized over time and Google reverse image search to locate related images. Edge detection, scaling, and color conversion algorithms are used to detect the alteration in the image (assuming that the original image is present someplace on the internet).

Recent studies are attempting to use context-based detection approaches to automatically identify bogus news [33]. With the exception of bogus news information, our technique identifies and utilizes pertinent elements. The components of the social environment that facilitate the propagation of news are described. User-based features and network-based features are both present in this approach. User-based features refer to a variety of user traits that users may have when they share real or bogus news on social media. The amount of followers, postings, and the date the account was created are among them. The characteristics of the network where news is disseminated serve as the foundation for network-based features. The majority of earlier research on context-based fake news identification with user- or network-based characteristics has only been applied to social media platforms like Twitter, Reddit, and Weibo, where contextual information is simple to get.

All these studies and many others which were not mentioned here somehow plays a role in detecting fake news in internet using different techniques and methods and models, which have some advantage and disadvantage, success and drawbacks but each and every study is important for making a further better model or finding a better way to detect fake news than the already existing ones. This paper also aims to create a new dimension with a low resource language like Bangla to make the task of detecting fake and falsified news in the internet easy.

Chapter 3

Methodology

3.1 Introduction

This chapter outlines the research approach used to create the node2vec-based false news detection model for the Bangla language using web search results. The chapter starts with a review of the two primary methods for embedding word and node representations in a network, respectively, word2vec and node2vec.

Word embeddings, which are dense vector representations of words in a continuous vector space, are created using Word2Vec, a well-liked and effective natural language processing (NLP) technique. The name "word2vec" comes from its primary goal, which is to convert words from a vocabulary to vectors in order to capture their contextual and semantic links based on the distributional characteristics of words in a given corpus. The distributional hypothesis, which asserts that words with comparable meanings tend to appear in similar settings within a text, is the foundation of word2vec. By learning word representations using unsupervised learning from enormous volumes of text input, Word2Vec makes use of this presumption. Words with comparable meanings or contexts can be situated adjacent to one another in the vector space thanks to the embeddings created by word2vec, which encode semantic similarities. Continuous Bag of Words (CBOW) and Skip-gram are the two main architectures used in word2vec.

Continuous Bag of Words (CBOW), here the model is trained to predict a target word based on its context words in the CBOW architecture. The words that appear in a specified window size around the target word in a phrase are known as the context words. By taking into account the terms in the target word's context, the model learns to predict the target word—basically, it learns to comprehend a word based on its surroundings.

The Skip-gram architecture is the opposite of the CBOW architecture. Skip-gram predicts the context words given a target word rather than the target word from its context words. In other words, it develops the ability to foretell the words that will probably appear close to a specific target term. When unusual words are the main emphasis, skip-gram is very helpful because it can learn more about them by taking into account different contexts.

In the word2vec training process, the neural network weights are modified using stochastic gradient descent or a comparable optimization approach. The goal is to reduce the loss function, which assesses the discrepancy between real word representations of context or target words and anticipated word representations. Once trained, the word2vec model generates a set of dense vectors called word embeddings that, on average, have hundreds of dimensions and represent each word in the lexicon. Word2vec can encode linguistic knowledge, analogies, and semantic similarities between words thanks to these embeddings, which capture semantic links and contextual data. Many NLP tasks, including language modeling, part-of-speech tagging, sentiment analysis, and named entity recognition, have been transformed by Word2Vec. In addition, word embeddings produced by word2vec have proven to be extremely useful as pre-trained features for downstream tasks, greatly enhancing the performance of numerous natural language processing applications.

With the help of the Node2Vec methodology, nodes in a graph can produce low-dimensional embeddings. This is an expansion of the Word2Vec method. It is an effective learning approach for learning graph representations that captures the semantic relationships and structural details of a graph's nodes. Node2Vec makes it easier to perform different graph-based tasks including link prediction, node categorization, and community recognition by encoding nodes as dense vectors in a continuous vector space. The main concept behind Node2Vec is to use biased random walks to tour the graph before using the Word2Vec approach to extract node embeddings from the random walks. Random walks are collections of nodes that a "walker" traverses in a graph; each walker's path is controlled by transition probabilities that strike a balance between moving to new regions of the graph and remaining in the immediate vicinity of the current node. The following stages are involved in creating node embeddings using Node2Vec: The first step in representing a graph is to turn it into an adjacency matrix or list. The

cornerstone for the random walk process is this representation, which captures the relationships between nodes.

The second step known as Biased Random Walks Node2Vec runs random walks to investigate the topology of the graph and record the connections between nodes. Return and in-out are the two parameters used by the method. The in-out parameter defines how probable it is for the walker to visit a node that is either similar to the present node or in a different neighborhood, while the return parameter governs how likely it is for the walker to return to the node it came from.

Following the generation of the random walks, Node2Vec interprets each set of nodes as a sentence in the Word2Vec context. Then, using the Word2Vec algorithm, node embeddings are learned from these phrases, with each node being considered as a word. This can be accomplished by using Word2Vec's Skip-gram or CBOW architecture.

After that each node in the graph receives low-dimensional embeddings as a result of the Word2Vec training procedure. The structural and semantic characteristics of each node in the network are represented by these embeddings, which are dense vectors in a continuous vector space. With Node2Vec, learning embeddings may be adapted to various graph architectures and applications using a flexible framework. The technique can explore particular aspects of the graph or capture various conceptions of similarity by changing the parameters of biased random walks.

As characteristics for numerous downstream tasks, including node categorization, connection prediction, or visualization, Node2Vec's learnt node embeddings can be employed. They make it easier for researchers and practitioners to understand and anticipate the behavior of complicated graph-structured data by facilitating efficient and effective graph-based machine learning. Network analysis, recommendation systems, and social network research have all benefited from the use of Node2Vec as a useful tool for extracting meaningful representations from large-scale graphs.

3.2 Data Collection

The data gathering procedure for creating the fake news detection dataset in Bangla is described in this section. To ensure the collecting of trustworthy and varied news pieces, web scraping methods and the choice of dependable news sources are described. To prepare the text data for upcoming comparison model training and evaluation, the preprocessing steps, which include tokenization, stop-word removal, and stemming, are described in depth. The information should fit the criteria of false news because it has the story's content. 'Banfakenews Dataset' is an experimental dataset that we created for this study [1]. The linguistic characteristics of the data are theoretically immaterial because web search is accessible in practically all languages worldwide. To test whether web search results are consistently available as a legitimate input for false news identification, regardless of regional and linguistic factors, we thus employ datasets in Bangla language in this work. Table 1 provides the percentage of real and fake news in our dataset.

Table-1: Dataset Details Information

Type of Article	Total Data	Percentage
Real	14707	81.83%
Fake	3264	18.16%

We obtained 242 different categories from our dataset since various outlets categorize the news in different ways. We combined related categories from several news sources to make a single category in order to generalize it. Finally, we divide the dataset's news stories into the 12 categories (Table 2).

Table-2: Number of News in Each Category

Data Type	Fake	Real	Total	Fake %	Real %
Crime	72.0	310.0	382.0	18.848168	81.151832
Education	55.0	330.0	385.0	14.285714	85.714286

Entertainment	182.0	783.0	965.0	18.860104	81.139896
Finance	2.0	378.0	380.0	0.526316	99.473684
International	154.0	2166.0	2320.0	6.637931	93.362069
Lifestyle	183.0	280.0	463.0	39.524838	60.475162
Miscellaneous	1142.0	696.0	1838.0	62.132753	37.867247
National	168.0	5610.0	5778.0	2.907580	97.092420
Politics	166.0	886.0	1052.0	15.779468	84.220532
Sports	99.0	1959.0	2058.0	4.810496	95.189504
Technology	48.0	259.0	307.0	15.635179	84.364821

Here a representative sample from the dataset that was utilized in this study to illustrate how the proposed node2vec-based false news detection model can be applied to web search results in the Bangla language is provided. The dataset serves as the basis for our inquiry into identifying false information and comparing the effectiveness of various classification techniques. The dataset is made up of a carefully chosen selection of news stories in Bengali that were obtained from reliable websites and media portals. These articles illustrate the breadth of knowledge communicated across multiple sectors by covering a wide range of topics, including politics, health, the economics, entertainment, and more. We carefully chose news stories from widely respected news outlets that follow accepted journalistic standards in order to ensure the accuracy and dependability of the data. To produce a high-quality and reliable corpus, a thorough process of data preprocessing was used to exclude any duplicated, unnecessary, or misleading information from the dataset.

Every news item in the dataset is labeled as either "fake" or "real" based on a ground truth that has been expertly annotated. The "real" labeled data includes items that have been thoroughly fact-checked and confirmed as factual and reliable, whereas the "fake" data includes stories that have been validated as containing misinformation, disinformation, or deceptive content. Using a selection of the news items and their accompanying labels, we display a table of the dataset in this example. This example is being provided to show how the dataset's structure and composition support the succeeding steps of model training, evaluation, and comparison. In order to preserve user privacy, data confidentiality, and intellectual property rights, the dataset utilized in this study complies with stringent ethical standards. The dataset example provided here serves as a crucial point of reference for the subsequent analyses and findings, enabling a thorough comprehension of the efficacy and potential for mitigating false information within the Bengali language information ecosystem of the proposed node2vec-based fake news detection model.

Table-3: Dataset example (Real News)

Label	**Real**
Source	banglatribune.com
Published Date	9/19/2018 19:54
Category	Sports
Headline	সূচি অদল-বদলে বিরক্ত মাশরাফি
Content	গ্রুপ পর্বের খেলা শেষ হওয়ার আগেই সুপার ফোরের সূচি প্রকাশ করেছে এশিয়ান ক্রিকেট কাউন্সিল (এসিসি)। বুধবার ভারত-পাকিস্তান এবং বৃহস্পতিবার বাংলাদেশ-আফগানিস্তান ম্যাচের ফল অনুযায়ী সূচি হওয়ার কথা ছিল। কিন্তু ভারত 'বেসক্যাম্প' দুবাই ছেড়ে আবুধাবিতে খেলতে রাজি নয়। তাই সূচিতে ওলট-পালট হয়েছে, আফগানদের মুখোমুখি হওয়ার আগেই বাংলাদেশকে ধরা হচ্ছে 'বি' গ্রুপ রানার্স-আপ! অন্যদিকে পাকিস্তানের কাছে হারলেও 'এ' গ্রুপে ভারতকে ধরা হবে চ্যাম্পিয়ন। এশিয়া কাপের মাঝপথে এমন উদ্ভট সিদ্ধান্তে মাশরাফি মুর্তজা বিরক্ত। আফগানদের মুখোমুখি হওয়ার আগের দিন সংবাদ সম্মেলনে সুপার ফোরের সূচি নিয়ে সবচেয়ে বেশি কথা বলতে হলো বাংলাদেশ অধিনায়ককে। তার কথা, "এভাবে সূচি বদল অবশ্যই হতাশাজনক। গ্রুপ চ্যাম্পিয়ন হওয়ার লক্ষ্য নিয়ে এই টুর্নামেন্টে খেলতে নেমেছিলাম আমরা। গ্রুপ চ্যাম্পিয়ন হলে সুপার ফোরে 'এ' গ্রুপ রানার্স-আপ দলের সঙ্গে প্রথম ম্যাচ খেলবো। কিন্তু আজ সকালে জানতে পারলাম আফগানিস্তানের বিপক্ষে জিতলেও 'বি ২' বাংলাদেশ।" মাশরাফির পরের মন্তব্যে ফুটে উঠলো ক্ষোভ, 'পাগলও তো এভাবে সূচি বদল মেনে নেবে না। একটা আন্তর্জাতিক টুর্নামেন্টে গ্রুপ পর্ব শেষ হওয়ার আগেই আমরা কিনা গ্রুপ রানার্স-আপ! তাহলে আগামীকাল আফগানিস্তানের বিপক্ষে ম্যাচটি এখন শুধুই নিয়ম রক্ষার ম্যাচ! অবশ্যই আন্তর্জাতিক ম্যাচের মূল্য আছে। গ্রুপ ম্যাচ বলেন বা যা-ই বলেন, একটা নিয়ম থাকে টুর্নামেন্টের। কিন্তু আমরা নিয়মের বাইরে চলে যাচ্ছি, আর এটাই হতাশাজনক।'

Table-4: Dataset Example (Fake News)

Label	**Fake**
Source	earki.com
Published Date	26/09/2018
Category	Politics
Headline	যে ১০টি সম্ভাব্য কারণে সাহারা খাতুন অ্যাম্বুলেন্সে নির্বাচনী প্রচারণা চালিয়েছেন
Content	গত ২৫ ডিসেম্বর ফেসবুকে একটি ভিডিওতে দেখা যায়, সংসদ নির্বাচন উপলক্ষ্যে ঢাকা ১৮ আসনের প্রার্থী সাহারা খাতুনের পোস্টার সংবলিত ৮টির বেশি অ্যাম্বুলেন্সের একটি বহর চলছে উত্তরার একটি রাস্তায়। প্রচারণার কাজে ঘোড়ার গাড়ি থেকে শুরু করে ট্রাক পর্যন্ত ব্যবহার করার নজির দেখা গেলেও অ্যাম্বুলেন্সের ব্যবহার সম্ভবত এই প্রথম। রোগী পরিবহনের পরিবর্তে অ্যাম্বুলেন্সে প্রার্থীর প্রচারণার কার্যক্রম কেন পরিবাহিত হচ্ছে, তা জানতে মাঠে নামে পড়ে eআরকি যানবাহন গবেষণা দল।এ ব্যাপারে অত্র এলাকার একজন মুমূর্ষু রোগীর অনুভূতি জানতে চাইলে তিনি বলেন, 'ভাইজান, আমগো চিকিৎসা লাগবো না। নৌকা মার্কায় ভোট পরলেই আমরা সুস্থ হইয়া যামু!' এতটুকু বলার পর তিনি হুট করে কাশতে কাশতে প্রচন্ড অসুস্থ হয়ে পড়ায় আমরা তার বাকি বক্তব্য জানতে পারিনি।সুতরাং আমরাই ভেবে বের করেছি কিছু সম্ভাব্য কারণ, যে সব কারণে সাহারা খাতুনের নির্বাচনী প্রচারণার কাজ চালাতে হয়েছে অ্যাম্বুলেন্সেই!১# এগুলো আসলে অ্যাম্বুলেন্স না। এগুলো হলো মাইক্রোবাস। গাড়িদের একটা যেমন খুশি তেমন সাজো প্রতিযোগিতায় এই মাইক্রোটা অ্যাম্বুলেন্স সেজেছে। অবশ্য এক বিএনপি কর্মী এ সম্পর্কে মনে মনে বলেন (ভয়ে শব্দ করে বলেননি), 'দেশের অবস্থা এত মারাত্মক যে মানুষ এখন যেকোনো মুহূর্তে হামলার ভয়ে আছে। তাই কেউ মাইক্রো, প্রাইভেটকার না কিনে সবাই অ্যাম্বুলেন্স কিনছে।' অথবা হতে পারে, ঢাকা ১৮ আসনে অন্য কোনো গাড়ি নেই। এই এলাকায় সব কিছুই এমার্জেন্সি, তাই সবই অ্যাম্বুলেন্স। বাধ্য হয়েই তাই প্রচারণাও চালাতে হচ্ছে অ্যাম্বুলেন্সেই।'২# সাহারা খাতুন অ্যাম্বুলেন্সে প্রচারণা চালিয়ে বোঝাতে চাচ্ছেন, তার আসনে স্বাস্থ্যখাতে এতটাই উন্নতি হয়েছে যে এখানে কোনো রোগীই নাই। এখানকার মানুষ অসুস্থ হওয়া বাদ দিয়ে দিয়েছে। ভবিষ্যতেও কেউ অসুস্থ হবে না। তাই অ্যাম্বুলেন্সগুলো বেকার পড়ে আছে। এই বেকার অ্যাম্বুলেন্সগুলো তাই কাজে লাগানো হচ্ছে নির্বাচনী প্রচারণায়।৩# হয়তো টানা অনেকদিন চলতে থাকায় এই প্রচারণাটি হঠাৎ করে অসুস্থ হয়ে পড়েছে। সিজনটা এমনিতেই বেশি ভালো না, চারিদিকে নানা ধরণের রোগবালাইয়ের প্রকোপ্ বেড়েছে। তাই হয়তো দেরি না করে তড়িঘড়ি কিছু অ্যাম্বুলেন্স জোগাড় করে প্রচারণাটিকে নিয়ে ইমার্জেন্সি সেবা গ্রহণের উদ্দেশ্যে হাসপাতালে নিয়ে যাওয়া হচ্ছিল!৪# মনোনয়ন পাননি, তবু মানিক ভাই বলে গেছেন, 'দিজ ইলেকশন, ভেরি ইম্পর্টেন্ট ইলেকশন।' সাহারা খাতুন সেই কথাকেই সিরিয়াসলি নিয়ে ফেলেছেন। নির্বাচনের গুরুত্ব বোঝাতে প্রচারণা চালাচ্ছেন অ্যাম্বুলেন্সে করে।৫# সাহারা খাতুনের জনপ্রিয়তা এতই বেশি যে মুমূর্ষু রোগীরাও ঘরে না বসে থেকে তার নির্বাচনী প্রচারণায় নেমে পড়েছে। কারো হার্ট অ্যাটাক হলে দ্রুত আইসিইউতে যাওয়ার পথে রাস্তায় দশজনের কাছে ভোট চাইতে চাইতে যাচ্ছে। কেউ এক্সিডেন্ট করলে সেই অবস্থাতেই চালাচ্ছে নির্বাচনী প্রচারণা। আর এসব সিরিয়াস রোগীদের জন্যই অ্যাম্বুলেন্সে করে প্রচারণার সুব্যবস্থা করেছেন সাহারা খাতুন।৬# স্বরাষ্ট্রমন্ত্রী থাকাকালে সাহারা খাতুন বলেছিলেন, কারো বেডরুমে নিরাপত্তা দেয়া সম্ভব না। মন্ত্রীত্ব চলে যাওয়ার পর তিনি তার ধারণা পাল্টেছেন। সর্বক্ষেত্রে নিরাপত্তা নিশ্চিত করার জন্যই প্রচারণা চালাচ্ছেন অ্যাম্বুলেন্সভিত্তিক। প্রচারণার সময় কোথাও কারো উপর হামলা হওয়া মাত্রই যাতে হাসপাতালে নেয়া যায় তাই এই ব্যবস্থা।৭# উনি আসলে এই প্রচারণার মাধ্যমে বিরোধী দলকে হুমকি দিতে চাচ্ছেন। বলতে চাচ্ছেন, 'দেখে নাও, অ্যাম্বুলেন্স সাথেই আছে।

3.3 Data Preprocessing For Proposed Node2vec Model

In this step each title of fake and real news from the collection is now typed as a search query into an online search engine. The first links from each search result are also gathered when a lot of links are indexed as a consequence of a web search. In this study, Google was the web search engine we used to put our suggested research model into practice. Google was the most popular platform and search engine. Since 2009, Google has maintained a commanding 90% market share in the search engine industry, according to StatCounter. It presently processes more than 40,000 search requests every second on average, which equates to about 3.5 billion and 1.2 trillion worldwide searches per day and year, respectively1. A ranking mechanism called "PageRank" was created by Larry Page and is used in part to determine the order of search results returned by Google. According to Google2, "PageRank" calculates a rough evaluation of a website's significance by counting the quantity and calibre of links pointing to a page. The essential premise is that more significant websites will probably get more links from less important websites. Google's search is made to prioritize websites with high PageRank content and high relevance to search phrases at the top of the search results page through a number of technologies, including PageRank. After collecting the links, preprocessing these links is applied. Unnecessary sub-paths from the links are deleted, and the domains are extracted and identified. Consequently, each link is collected and stored in the dataset.

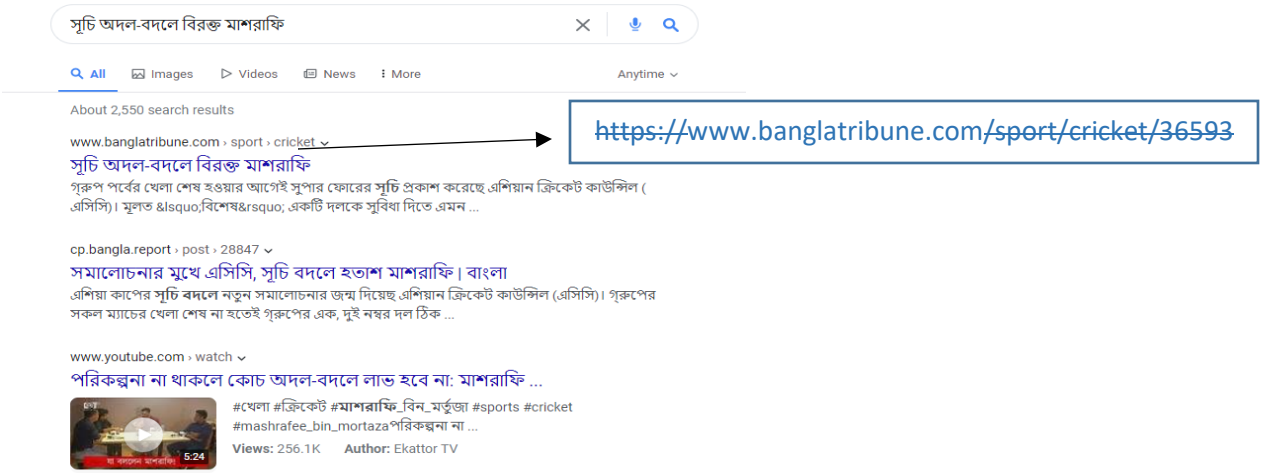

Fig 1: Link collection process according to news title

Links, sometimes referred to as edges in the node2vec model, are vectorized using node embeddings produced by the random walk and Word2Vec adaption process. Learning low-dimensional representations (embeddings) for nodes in a graph that accurately capture their structural and semantic links is the main goal of node2vec. These embeddings can be used to vectorize links between nodes in the graph once they have been obtained. The node2vec model's vectorization of links can be explained as follows:

Graph Representation: The node2vec algorithm receives a graph as input that is represented as either an adjacency list or an adjacency matrix. This graph representation, which serves as the foundation for the random walk process, records the relationships between nodes.

Fig 2: Web Graph for Node2vec Model

Biased Random Walks: Node2Vec conducts biassed random walks to investigate the topology of the graph and provide node sequences for each walk. Biassed random walks establish a balance between discovering more extensive graph topologies and exploring nearby neighbourhoods. The chances of the walker returning to the node they came from and the likelihood of exploring nodes in the same neighbourhood or in different regions of the graph, respectively, are controlled by the two parameters "return" and "in-out," which are used to do this.

Word2Vec Adaptation: In the context of the Word2Vec method, biassed random walk-generated node sequences are viewed as sentences. Assuming that each node in the graph represents a "word," Word2Vec is used to extract embeddings from these sequences. The goal of Word2Vec is to guarantee that co-occurring nodes in these random walks have similar embeddings.

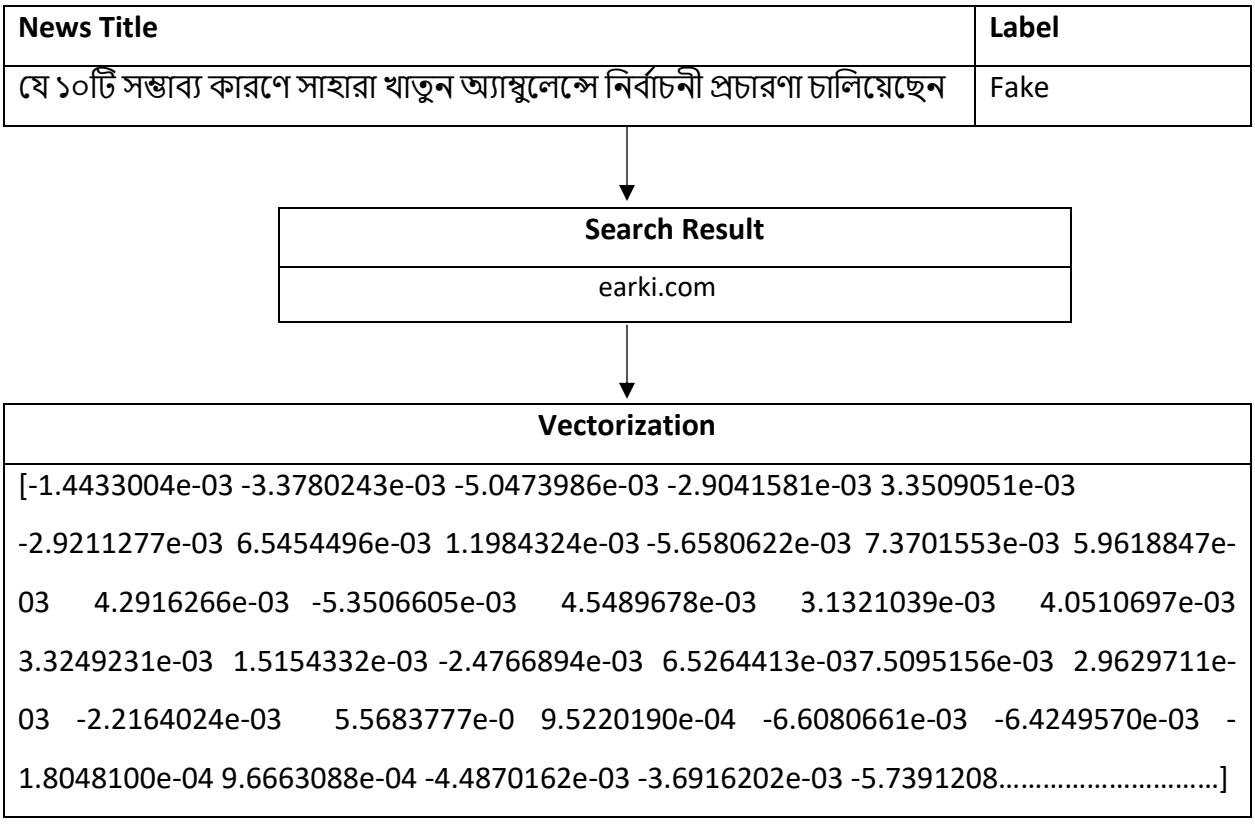

Fig 3: Link Vectorization for Feature Extraction

Embedding Generation: Low-dimensional embeddings are created for each node in the graph after the Word2Vec conversion is finished. These embeddings are dense vectors that generally have hundreds of dimensions and exist in a continuous vector space. They reflect each node's connections to other nodes in the graph and its structural and semantic features.

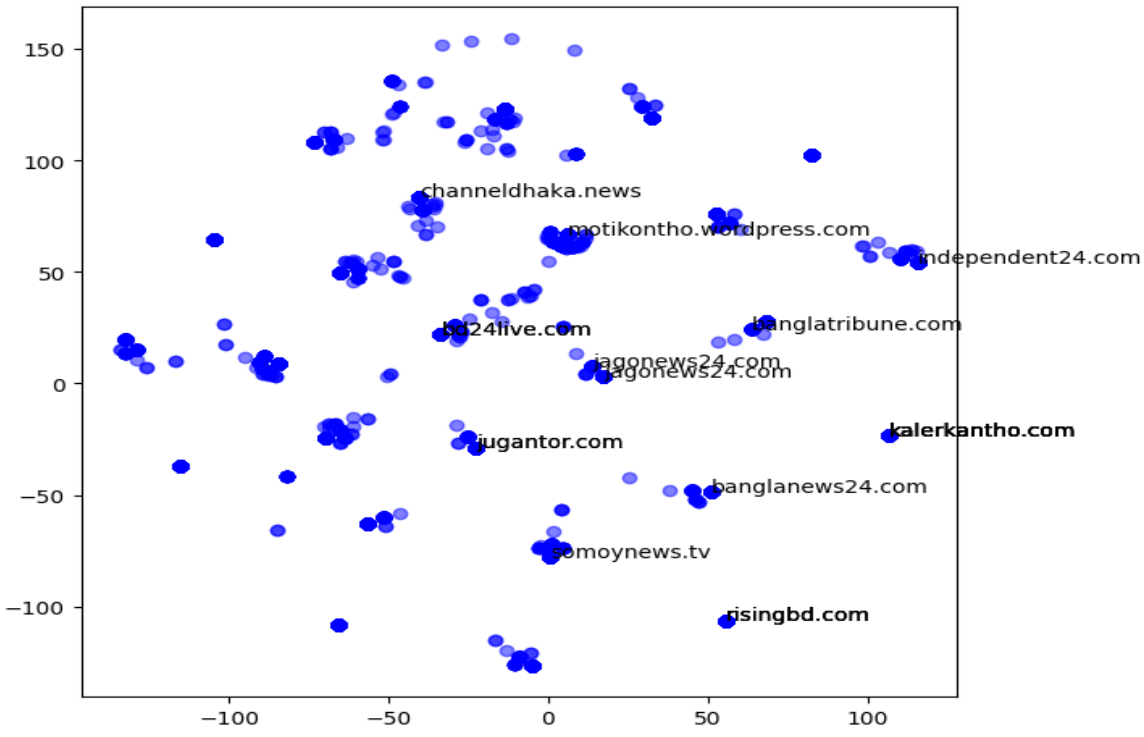

Fig 4: Embedded Nodes in Graph

As characteristics for numerous downstream tasks, including node categorization, connection prediction, or visualisation, Node2Vec's learnt node embeddings can be employed. They make it easier for researchers and practitioners to understand and anticipate the behaviour of complicated graph-structured data by facilitating efficient and effective graph-based machine learning. Network analysis, recommendation systems, and social network research have all benefited from the use of Node2Vec as a useful tool for extracting meaningful representations from large-scale graphs.

3.4 Data Preprocessing for Text-based Comparison Model

Before supplying the raw text data to the classifier, some preprocessing must be applied. Unnecessary symbols and other elements that are unimportant to our classification may be included in a raw text. Different emoticons, such as :D and ;), may be useful for sentiment analysis, but not in this situation. Additionally, we deleted special characters like @, #,!, etc. from our text. These components have the potential to lower or weaken the classifier's performance. Our dataset is ready for the classifier algorithms after the number values, punctuation marks, and special symbols have been removed. The table below shows characters that were removed in preprocessing.

Table-5: Characters considered removing in preprocessing

Category	**Characters**
Special Characters	@, #, $, %, —, ,, ...
Bangla & English Digits	1, 2, 3, 4, ... 0
English Alphabets	A, B, C, ... Z; a, b, c,.....z
Emoticons	:), :D, :(, :o, ...

Before feeding our text into the classification algorithms, we employ Count Vectorizer and TF-IDF Vectorizer (term frequency inverse document frequency) to extract characteristics from the text.

The count vectorizer creates a vector with the same number of dimensions as the particular word in the corpus. Every word has a distinct dimension and contains 1 in that dimension while 0 is present in all other dimensions, maintaining the frequency of every word. Instead of just providing a count, TF-IDF vectorizer provides numerical representations of the words whether they are present or not. The frequency of words is calculated by multiplying them by their inverse document frequency. Simply said, words that come frequently yet everywhere should be given very little relevance or weight. Words in the Bangla language like the ones below

ও, আর, এবং, আরও.....

don't hold a lot of interest. When a term only occasionally or infrequently appears, it is actually more important and should be carefully considered as such. This would result in better categorization performance. It is a technique used to explain the importance of a keyword within a document. The formula [1] is

$$tf\ idf(t, d, D) = tf(t, d)\ idf(t, D)$$

if the term is indicated as "t," a particular document is denoted as "d," and the complete document is denoted as "D."

The frequency of 't' in 'd' is indicated above by the symbol tf(t, d).

idf(t, D) measures how frequently or seldom 't' occurs across 'D'

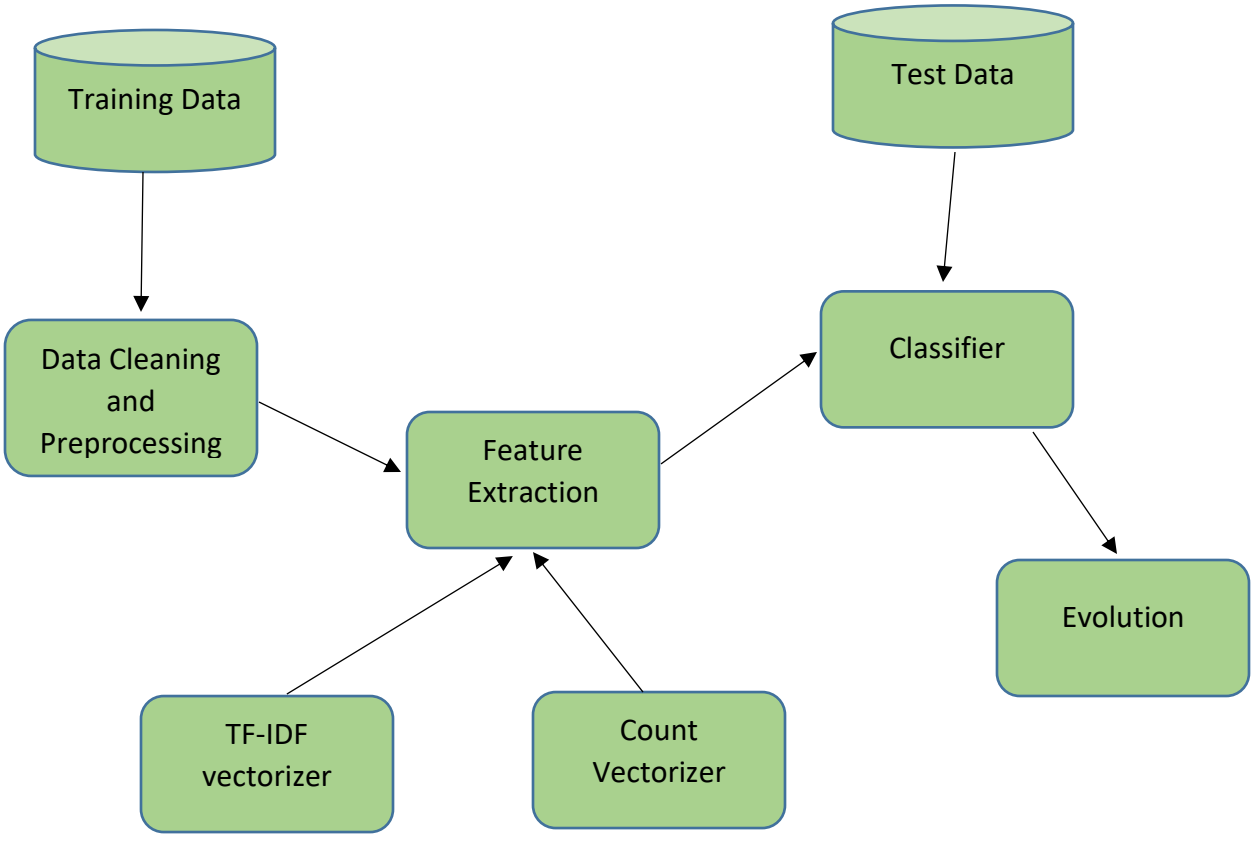

Fig 5: System Flow Diagram of Text Based Comparison Model

3.5 Classification Models

The Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Logistic Regression classification algorithms are used to identify bogus news. The architecture, training procedure, and hyperparameters of each model are presented in detail, with an emphasis on their advantages and disadvantages when used to identify fake news in Bengali.

SVM: A supervised machine learning approach called the Support Vector Machine (SVM) is generally used for classification tasks, while it can also be applied to regression issues. SVM is especially effective when dealing with high-dimensional data and scenarios where the classes cannot be separated linearly. In the 1990s, Vapnik and his associates created the algorithm. Finding an ideal hyperplane that maximally separates the data points of distinct classes is the main goal of SVM. A hyperplane is a subspace in a multidimensional space that has one less dimension than the dimensionality of the feature space. The hyperplane is a flat (d-1)-dimensional surface for a binary classification issue, where d is the quantity of features. The hyperplane that best distinguishes the data points from two different classes, in your example fraudulent and authentic news, is referred to as the decision boundary.

The hyperplane that maximizes the margin between the two classes is what SVM seeks to identify. The margin is the separation between the nearest data points from each class, referred to as support vectors, and the hyperplane. The crucial data items that establish the decision border are these support vectors. A hyperplane can properly divide the data points into distinct classes if they are linearly separable, which is the ideal case. Real-world data, however, frequently cannot be separated linearly. The "kernel trick," a method used by SVM to solve this problem, indirectly translates the data into a higher-dimensional feature space where it can be linearly separated. The polynomial kernel and the radial basis function (RBF) kernel are frequent SVM kernels. In real-world situations, obtaining a perfect separation with a wide margin could result in overfitting, particularly if the data are noisy or overlap. To manage the trade-off between maximizing the margin and allowing some misclassifications, SVM incorporates a regularization parameter (C). A larger C value enforces a stricter margin and may cause more misclassifications,

whereas a smaller C value enforces a looser margin and results in a bigger margin. Finding the ideal balance between overfitting and underfitting requires tuning C. According to my thesis, news sources are categorized as legitimate or fraudulent using SVM based on the attributes gleaned from two different approaches:

Node2Vec-based Features: The news sources are represented in a high-dimensional feature space by the node embeddings produced by the Node2Vec technique. Based on these embeddings, SVM finds the best hyperplane to distinguish between fraudulent and legitimate news sources. In this feature space, SVM will locate a decision boundary that maximizes the difference between fraudulent and legitimate news sources.

Text-based Features: To convert the news information into numerical feature representations for the text-based features, you can use Count Vectorizer or TF-IDF. Based on the text content of the sources, SVM will operate in this feature space to identify the decision boundary that distinguishes false from legitimate news sources.

CNN: Deep learning algorithms like Convolutional Neural Network (CNN) are frequently employed for image identification jobs. It has, however, also been successfully used for problems involving natural language processing, such as text classification. CNNs are well suited for jobs involving grid-like data structures, such as images and sequential data, like text, since they are particularly good at capturing local patterns and hierarchical representations in data.

Learning hierarchical feature representations from the incoming data is the main goal of a CNN. Convolutional layers, pooling layers, and fully linked layers are just a few of the many layers that make up the model. As the data travels through the network, each layer applies a particular operation to the incoming data, enabling the CNN to acquire increasingly abstract and complicated properties. The fundamental component of a CNN is the convolutional layer. It processes the input data through a number of learnable filters (also known as kernels). Every filter recognizes particular motifs or characteristics in the input, such as edges, corners, or more intricate structures. The CNN gains the ability to modify the filter settings to find features that are pertinent to the task at hand. Local patterns in the text data are captured by convolutional layers.

Following the convolutional layers, pooling layers are used to down sample the feature maps produced by the convolutions. By pooling the data, vital information is preserved while the spatial dimensions are reduced. Max-pooling is a typical pooling operation where the largest value inside a window (for example, 2x2) is taken, hence minimizing the size of the feature map. The model becomes more resilient to minor fluctuations in the input data thanks to pooling, which also aids in decreasing computing complexity. Following the convolutional and pooling layers are the fully connected layers, which resemble the layers in conventional feedforward neural networks. They use the previously learnt information from the underlying layers to generate predictions for the desired classes. There will be a single output node with a sigmoid activation function that forecasts the likelihood that the input text will belong to one of the classes in the case of binary classification (fake or true news detection).

CNN based on a Node2Vec model:

Node2Vec-based Features: The node embeddings produced by the Node2Vec technique serve as input features for the CNN in the node2vec-based false news detection model. In a high-dimensional feature space, each node embedding represents a different news source.

CNN Architecture: In this scenario, the node embeddings will be input into a layer on the CNN architecture. No embedding layer is required, unlike in the text-based model, because node embeddings are continuous-valued vectors.

Convolutional Layers: To identify regional patterns and hierarchical representations, the convolutional layers in the CNN will act on the node embeddings. To find pertinent features, the filters will move over the node embeddings.

Pooling Layers: Pooling layers downsample the feature maps produced by convolutions and follow the convolutional layers. The most essential features can be kept by using max-pooling, which is similar to the text-based model.

Fully Connected Layers: To create predictions, the fully connected layers will incorporate the learnt features from the convolutional and pooling layers. The likelihood score of each news

source's veracity will be provided by a single node with a sigmoid activation function in the output layer.

Evaluation: You will assess the CNN model's performance using common classification measures including accuracy, precision, recall, F1-score, and confusion matrices after training it with node2vec-based features.

CNN in Text-based Model:

Text preprocessing: Preprocess the news information by deleting stop words, special characters, and conducting tokenization before feeding the text input into CNN. Count Vectorizer or TF-IDF is then used to represent the text data as numerical vectors.

Architecture on CNN: An embedding layer is usually followed by a number of convolutional layers with activation functions (such as ReLU), pooling layers, and fully connected layers in the CNN architecture for text categorization.

Embedding Layer: The embedding layer creates dense continuous word embeddings from the sparse numerical vectors acquired from the Count Vectorizer or TF-IDF. These embeddings assist the model comprehend the contextual links between words by capturing the semantic meaning of each word.

Convolutional Layers: To identify pertinent patterns and local characteristics in the text input, convolutional layers process word embeddings. Convolutional layers' filters move over input embeddings to create feature maps that show the presence of particular patterns.

Pooling Layers: The pooling layers reduce the spatial dimensions of the feature maps while maintaining the critical data. In order to choose the most important features from an embedding window, max-pooling is frequently utilized.

Fully Connected Layers: To produce predictions, fully connected layers mix the learned data from convolutional and pooling layers. One node with a sigmoid activation function exists in the output layer and provides a likelihood score for the veracity of each news source.

Logistic Regression: For binary classification tasks, the supervised machine learning algorithm logistic regression is frequently utilized. When the output variable (target) is binary, as in your thesis topic of fake news identification, where the classes are denoted by 0 for fake and 1 for true news, it is very well suited for the situation at hand. The main goal of logistic regression is to calculate the likelihood that a given instance belongs to a given class. Contrary to linear regression, which makes predictions about continuous values, logistic regression uses the sigmoid function to estimate the likelihood of a binary result. The final classification choice is then based on the expected likelihood. A representative model In logistic regression, the output is the likelihood that the instance belongs to the positive class (actual news), while the input data is represented by feature vectors. The logistic function, also known as the sigmoid function, converts the linear combination of feature values and model coefficients to a number between 0 and 1.

Decision Boundary: To arrive at a final categorization choice, we establish a predicted probability threshold (for example, 0.5). Real news is classed as an instance if y exceeds the threshold; false news is classified as an instance if y is less than the threshold.

Training the Model: Using an optimization procedure (such as gradient descent) to minimize a loss function, such as the log loss (cross-entropy loss), the logistic regression model estimates the model coefficients (b0, b1, b2,..., bn) during training. To determine the best fit for the training data, the coefficients are adjusted iteratively. To avoid overfitting, logistic regression can be regularized. L1 regularization (Lasso) and L2 regularization (Ridge) are frequent regularization methods that modify the loss function by including penalty terms. Here, one of the classification models for both the node2vec-based and text content-based false news detection methods is logistic regression.

Node2Vec-based Model: Logistic regression uses the node embeddings as input characteristics for the node2vec-based model. To determine the likelihood that each news source is reliable, it learns the model coefficients.

Text Content-Based Model: Logistic regression uses the Count Vectorizer or TF-IDF numerical features as input for the text content-based model. To determine the likelihood that each news piece is authentic news, it estimates the model coefficients.

An entire graphic is shown below to provide a thorough and visually instructive overview of the false news detecting procedure. The comparison model, which makes use of text content-based features, is shown in the figure next to the suggested node2vec-based false news detection model. This graphical representation tries to provide a concise and understandable overview of the research approaches used.

The overall process graphic demonstrates the methodical technique used in the thesis to identify and categorize sources of fake news. The text content-based comparison model and the node2vec-based model are two unique methods of analysis that are displayed.

Data Gathering Beginning with the gathering of news sources' URLs from various online platforms and domains, the web graph is built. The gathered URLs are used to construct a web graph, where each node stands in for a news source and the edges signify the connections between them.

Node2Vec Embeddings: By using the node2vec technique on the web network, node embeddings are produced that accurately depict the connections and closeness between news sources in a high-dimensional feature space. Node embeddings serve as feature representations of the news sources, which makes the next classification work easier.

Classifier Training: To differentiate between phoney and legitimate news sources, support vector machine (SVM), CNN, Logistic Regression classifiers are trained on node2vec-based characteristics.

Model Evaluation: The node2vec-based model's performance is evaluated using a variety of evaluation criteria, allowing for an evaluation of how well it detects bogus news.

Text Comparison Model Path Based on Content:

News Content Gathering: To create the text-based dataset, news articles are gathered from various sources and domains.

Text Preprocessing: Tokenization, stop word removal, and normalization are among the text preprocessing stages that are applied to the gathered news items.

Feature Extraction: Count Vectorization or TF-IDF are used to convert the text data that has been processed into numerical features.

Support vector machine (SVM), CNN, Logistic Regression classifiers are trained on the text-based features to distinguish between bogus and legitimate news articles, much like the node2vec-based model.

Model Evaluation: Using common classification criteria, the performance of the text content-based model is evaluated in order to compare it to the node2vec-based model.

Comparing the performance indicators of the two models reveals the advantages and disadvantages of each strategy. This study intends to shed light on the possible benefits of using web graph embeddings for false news identification by comparing the node2vec-based model and the text content-based comparison model side by side.

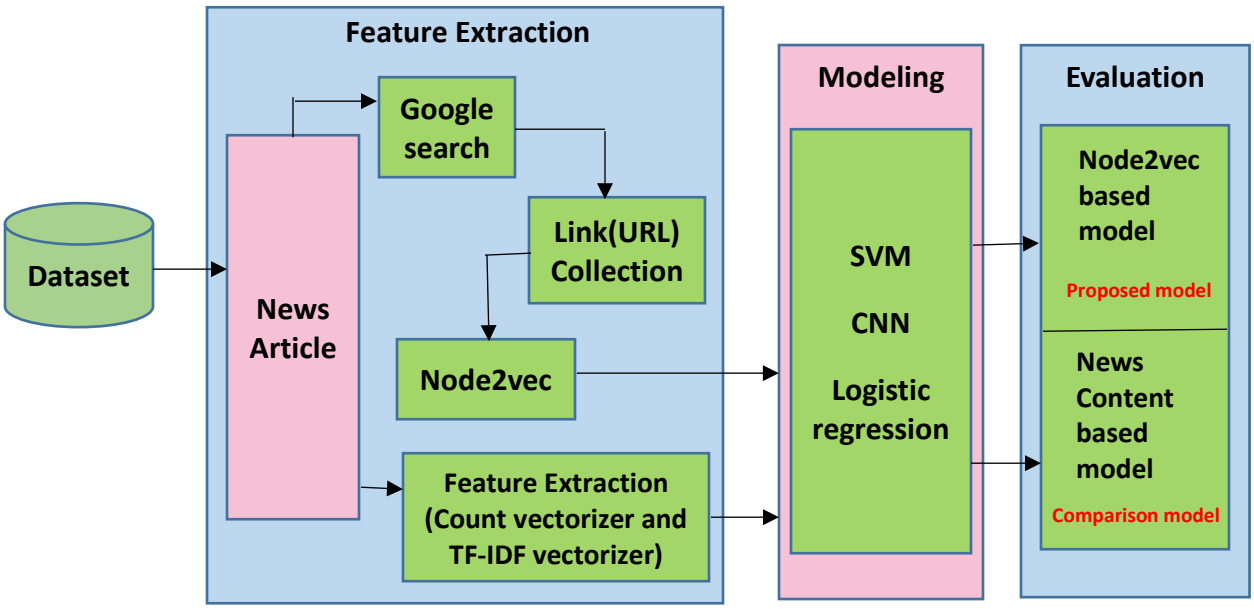

Fig 6: The total Process of Proposed and Comparison Model

Chapter 4

Experimental Result

The experimental findings from the application and assessment of the suggested node2vec-based false news detection model and the text content-based comparison model are presented in this chapter. A varied dataset of news articles and sources in the Bangla language was used for the studies. Each model's performance is evaluated using accepted assessment measures, which offers useful information about how well-suited each one is to identifying bogus news. We also evaluate and contrast the findings to comprehend the advantages and disadvantages of each strategies. We give a general summary of the dataset utilized in the assessment before getting into the experimental findings. The collection consists of news URLs and the news stories that go with them. Based on the actual facts, each entry is classified as either "fake" (0) or "real". To ensure a thorough review, the data covers a wide range of subjects and sources. The two divisions of the dataset were made at random. The classifiers are trained using 80% of the data in one section, and their effectiveness is tested using the remaining 20%.

4.1 Histogram Analysis for Real and Fake News Sources

The spread and consumption of reliable information have faced considerable difficulties in recent years as a result of the rise of false news. The deliberate dissemination of false or misleading material that passes for news has sparked questions about the veracity and dependability of media sources. This problem has impacted public opinion and decision-making processes in addition to affecting public confidence. In response to this urgent problem, our research looks at how legitimate and false news sources are distributed, adding to a greater understanding of the dynamics of disinformation. We have painstakingly collected a broad dataset that includes news pieces from multiple platforms and genres in order to get insights into how legitimate and

fraudulent news sources are distributed. Every news source has been painstakingly classified as "real" or "fake" based on real-world evidence.

In this context, our goal is to investigate the role that various sources play in the spread of accurate or inaccurate information. We have conducted a thorough study that clarifies the make-up of news providers in terms of their authenticity in an effort to comprehend the presence and distribution of actual and false news sources inside our dataset. The results of our analysis are presented in this chapter through two illuminating histograms, each of which shows the proportion of true and false news linked with various news sources. We have used histogram analysis, a potent visualisation approach, to depict the distribution of legitimate and phoney news sources in the search for data-driven insights. The histograms offer a thorough overview of the entire authenticity landscape by clearly displaying the percentage of actual and fraudulent news linked with each news source. We want readers to look for patterns and trends in the data as we present the histograms. We can learn a lot about the prevalence of false information in our dataset by analysing the distribution of true and fake news across various sources.

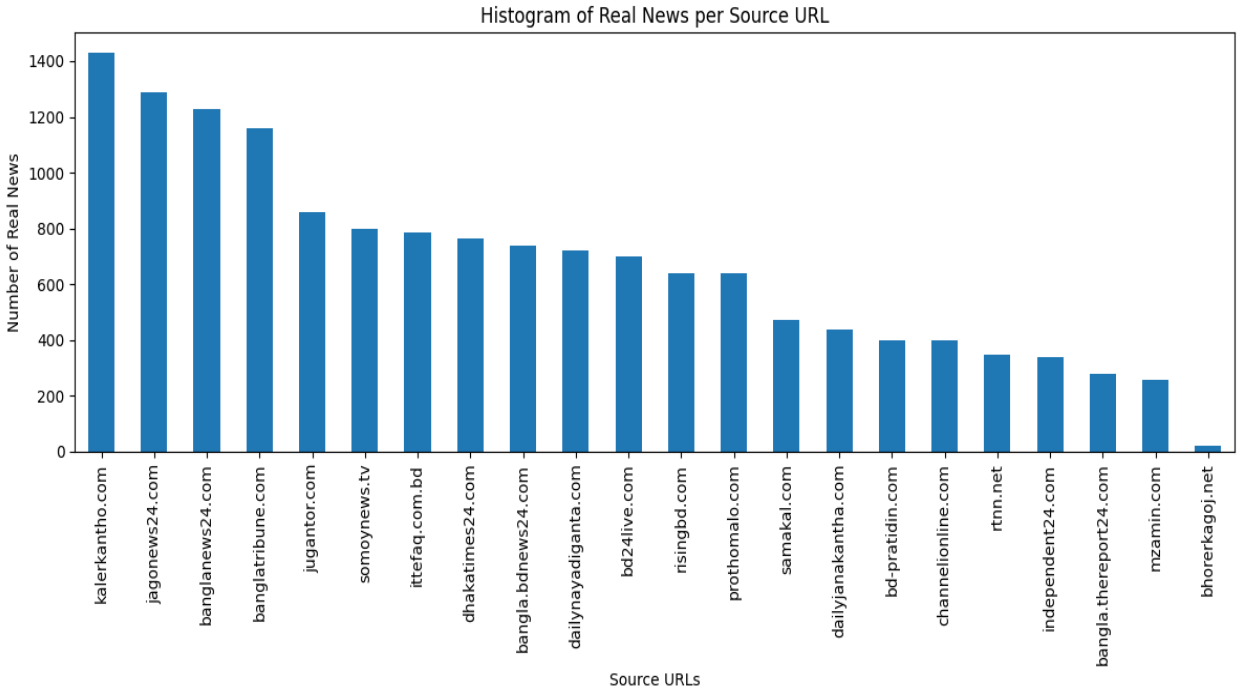

Fig 7: Histogram of Real News Per Source URL

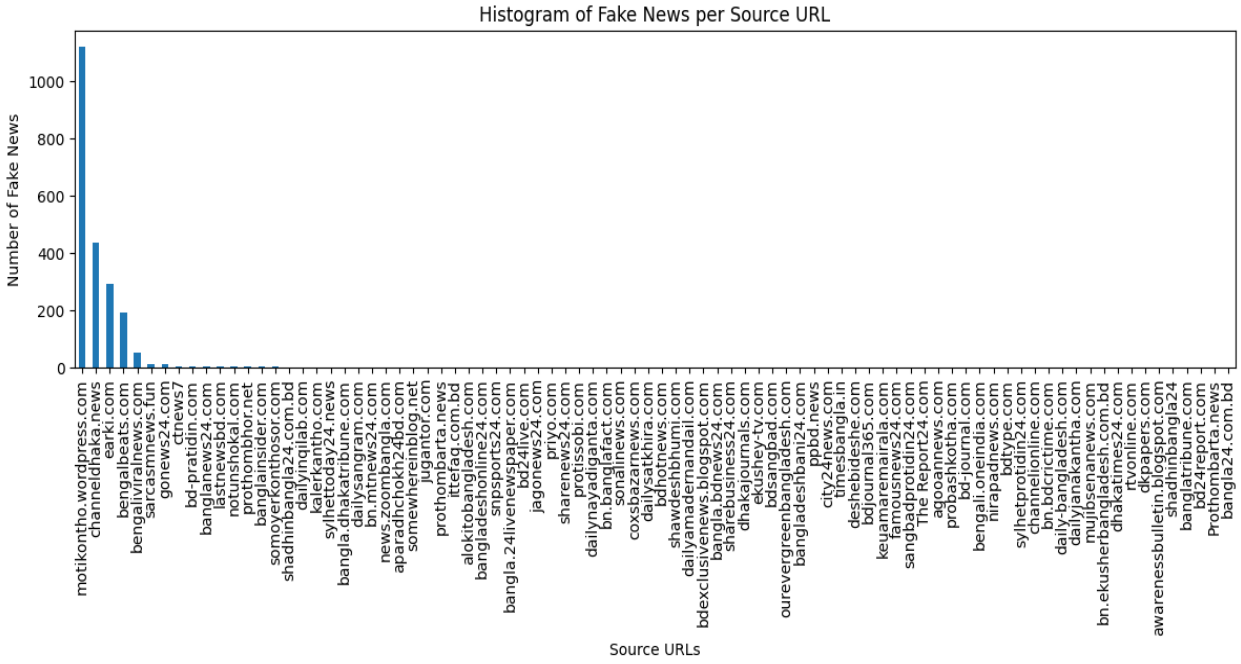

Fig 8: Histogram of Fake News per Source URL

The results of this histogram study have important ramifications for media literacy and the field of false news detection. The creation of focused interventions and tools to efficiently check information sources can be guided by an understanding of the distribution of reliable and unreliable news sources. Furthermore, this approach sets the door for future studies to investigate the elements that contribute to news sources' authenticity and their possible effects on how trustworthy people perceive the news to be.

4.2 Results

In this section, we give a tabular comparison of the accuracy results obtained by various models when it comes to the identification of fake news. The accuracy scores from our tests are shown in the table below, which gives a thorough analysis of how each model performs under various datasets or experimental circumstances. The Node2Vec-based Model and the Text-based

Model's accuracy numbers are displayed in the table. The table's rows, labelled Train and Test, relate to various experimental settings or datasets.

Table-6: Results for SVM Model

	Proposed Model Node2vec	**Comparison Model News Content Based**
Test Accuracy	0.99	0.96
Train Accuracy	0.99	0.98
Precision	0.99	0.97
Recall	1.0	0.98
F1 Score	0.99	0.98

The SVM model displays promising generalization performance by reaching high accuracy on both the training and test datasets in our proposed model which is 99% whereas the comparison model has accuracy of 96%. The model appears to be avoiding overfitting and is capable of producing accurate predictions on fresh, untested data, as seen by the little difference between training and test accuracy. After talking about the accuracy outcomes of the SVM model, we now move on to the section where we examine the effectiveness of the Convolutional Neural Network (CNN) model for fake news detection.

Table-7: Results for CNN Model

	Proposed Model Node2vec	**Comparison Model News Content Based**
Test Accuracy	0.87	0.86
Train Accuracy	0.86	0.86
Precision	0.87	0.86
Recall	1.0	1.0
F1 Score	0.93	0.92

The CNN model shows better accuracy in our proposed model that the traditional content based model. The accuracy for Node2vec model is 87%. The CNN model's balanced accuracy across actual and false news examples demonstrates its capacity to handle class imbalance, which is frequently present in fake news detection datasets. While the CNN model excels at catching complicated patterns, its computational costs may be higher than those of more established machine learning models like the SVM. For a deployment to be realistic, the trade-offs must be carefully considered.

Let's analyze the outcomes for the Logistic Regression model after looking at the accuracy tables for the SVM and CNN models.

Table-8: Results for Logistic Regression

	Proposed Model Node2vec	**Comparison Model News Content Based**
Test Accuracy	0.97	0.95
Train Accuracy	0.97	0.95
Precision	0.97	0.88
Recall	1.0	0.99
F1 Score	0.98	0.97

The competitive accuracy of the Logistic Regression model indicates its strength as an understandable and simple classifier. The model shows better accuracy than comparison model. And the accuracy for Node2vec model is 97%. But the accuracy for the comparison model which is a news content based traditional model is 95%. Its linear structure provides distinct decision limits, which contributes to its dependable performance in differentiating between instances of true and false news. As evidence of its capacity to manage class imbalance and provide accurate predictions for both classes, the Logistic Regression model displays balanced accuracy across instances of actual and fake news.

Now the confusion matrix findings from using Support Vector Machine (SVM) and Logistic Regression classifiers to both the Node2Vec-based and Text-based fake news detection models are presented and explained in this part. The confusion matrix is an effective tool that, by contrasting anticipated and actual labels, offers insightful information about the effectiveness of classifiers. We can learn more about how well the models can distinguish between legitimate and phoney news sources and articles by looking at the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values.

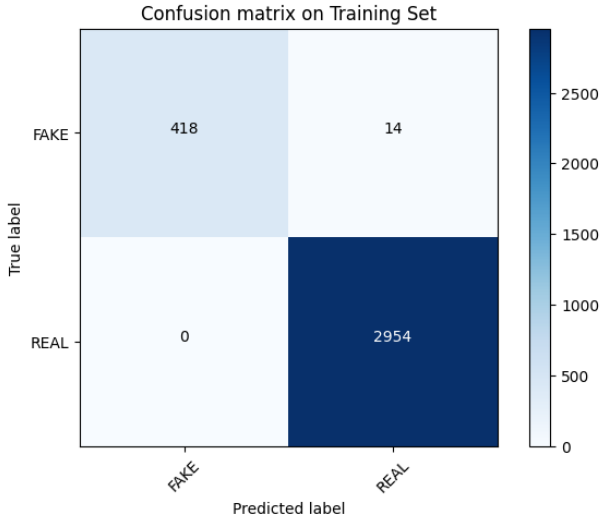

Fig 9: Confusion Matrix for SVM (Node2vec)

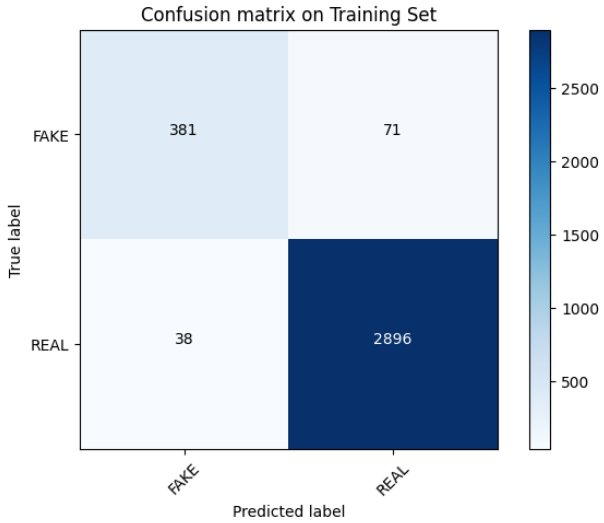

Fig 10: Confusion Matrix for SVM (Textbased)

The performance of the SVM-based Node2Vec model in categorizing news articles as authentic or fraudulent is shown in Figure 9 by the confusion matrix. The results of the confusion matrix are as follows:

2,954 news pieces were accurately identified as real news and were in fact authentic, which is known as a true positive (TP).

False Positives (FP): Fourteen news stories that were mistakenly labelled as real news but were fake news were examples of false positives.

True Negatives (TN): There were 418 news pieces that were projected to be fake news but turned out to be fake, totaling 418 TNs.

False Negatives (FN): There were no news stories that were expected to be fake news but turned out to be true, giving a FN count of zero.

The confusion matrix for the SVM-based Comparison Model (text-based) is shown in Figure 10. These are the outcomes:

2,896 news pieces were accurately identified as real news and were in fact real, which is known as a true positive (TP).

71 news articles that were mistakenly labelled as real news but were actually fake news were categorized as false positives (FP), suggesting instances of false positives. In comparison to the Node2Vec model, there are more false positives, which suggests that it is more likely that bogus news will be mistaken for the actual thing.

381 phoney news pieces were successfully identified as such, yielding 381 True Negatives (TN), which provided precise forecasts for fake news occurrences.

38 news stories that were mistakenly labelled as fake news but were actually legitimate reports are examples of false negatives (FN). False negatives suggest that some genuine news stories may have been incorrectly labelled as phoney.

Overall, the confusion matrix scenarios offer insightful information about the advantages and disadvantages of both methods. The Node2Vec SVM-based model performs admirably with few

false positives and negatives, demonstrating its promise as a reliable fake news detection model. The SVM-based Comparison Model, on the other hand, has a greater incidence of false positives and false negatives, indicating a need for more optimization to improve its accuracy and dependability. In order to ensure a more accurate classification of news stories and promote trust in media sources, the results from these confusion matrices can serve as a reference for future model upgrades and help in the building of more effective fake news detection systems.

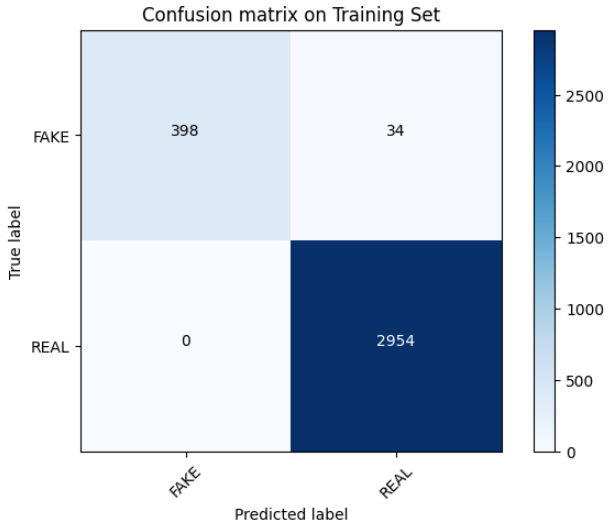

Fig 11: Confusion Matrix for Logistic Regression (Node2vec)

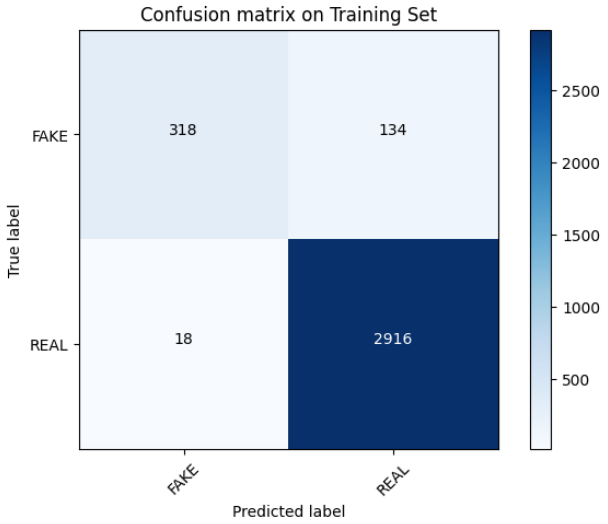

Fig 12: Confusion Matrix for Logistic Regression (Textbased)

The confusion matrix for the Logistic Regression model used with our suggested Node2Vec-based fake news detection model is shown in Figure 11. The confusion matrix shows how well the model performs in categorizing news articles as authentic or fraudulent using Node2Vec embeddings.

True Positives (TP): 2954 news stories were accurately classified by the model as being legitimate, and they were real news.

False Positives (FP): 34 news stories that were actually fake news were mistakenly identified as real. This indicates the instances of false positives where the algorithm mispredicted actual news.

True Negatives (TN): Since no news stories were flagged as phoney but turned out to be accurate, there are no true negatives.

False Negatives (FN): There were no news articles that were projected to be false but turned out to be true, giving rise to a false negative tally of zero.

The confusion matrix for the Logistic Regression model applied to our comparison text-based false news detection algorithm is shown in Figure 12 below. For feature extraction in this model, count vectorizer and TF-IDF are used.

True Positives (TP): 2916 news stories were accurately classified by the model as being real, and they were in fact real news.

False Positives (FP): 134 news pieces that were actually fake news were mistakenly identified as real news. This suggests that our suggested Node2Vec-based model has a higher rate of false positives.

True Negatives (TN): Although it is not mentioned, the true negatives count refers to the amount of fake news stories that were accurately detected as such.

False Negatives (FN): 18 news stories that were supposed to be fraudulent but turned out to be true news. Comparing this to the model we suggested, a larger false negative rate is evident.

Compared to the text-based model, which categorized 134 cases of false positives as actual news, our suggested Node2Vec-based model performed better, with only 34 instances. The text-based

model misses some actual news stories, as shown by the fact that our suggested approach has zero false negatives whereas the text-based model has 18.

According to the findings, our Node2Vec-based model makes more precise predictions and achieves a better balance between correctly categorizing authentic and fraudulent news. It can capture significant patterns and relationships in the data thanks to the efficient use of web graph embeddings, which improves accuracy and decreases false positives. However, because the text-based model relies solely on textual information, it may have trouble telling the difference between actual and false news.

The comparison of the confusion matrices reveals the advantages of our suggested Node2Vec-based fake news detection methodology and provides important insights into its potency in differentiating between true and fake news. These results support current initiatives to improve methods for identifying fake news and to promote media literacy in the fight against disinformation.

Chapter 5

Discussion

The experimental findings from comparing several false news detection models employing Node2Vec embeddings and text-based representations are thoroughly discussed in this chapter. In the broader context of false news identification, we examine the consequences of each model's performance analysis. With regard to both Node2Vec and text-based representations, the Support Vector Machine (SVM) model performed quite well. The model's ability to accurately discern between instances of true and fake news is demonstrated by the astonishingly high accuracy scores, which exceed 99%. The SVM model's accuracy, recall, and F1-score were all outstanding, demonstrating how well it could categorize a large number of legitimate news articles and identify every case of fake news. For both Node2Vec and text-based embeddings, the Convolutional Neural Network (CNN) model displayed competitive accuracy and performance. The CNN model's accuracy of 86%, but significantly lower than SVM, shows its effectiveness in distinguishing between authentic and false news articles. Additionally noteworthy were the model's precision, recall, and F1-score, which showed how well it could identify intricate patterns in textual material.

For both Node2Vec and text-based embeddings, the accuracy of the Logistic Regression model was commendable. The model is effective at identifying bogus news, as evidenced by accuracy scores that approach 97%. Its high precision, recall, and F1-score further demonstrated its propensity for making accurate predictions. When we compare the models' performances, we find that the SVM model had the highest accuracy of all the methods. With 99% or higher F1-scores, precision, recall, and robustness in detecting bogus news, it proved to be effective. Although slightly less effective than SVM, the CNN model nonetheless performed admirably, with an accuracy of more than 86%.

In terms of accuracy and false positive rates, our suggested Node2Vec-based model fared better than the text-based model. The Node2Vec model produced predictions with a better degree of

accuracy, demonstrating its efficacy in identifying significant linkages in the web graph. The text-based model, in contrast, showed a greater false positive rate, indicating its limits in identifying instances of actual and fake news. Our suggested Node2Vec-based model clearly outperforms text-based models when performance is compared. In comparison to the text-based model, it delivers higher accuracy and reduced false positive rates. By utilizing web graph embeddings, the Node2Vec model is able to capture significant relationships between news sources, producing predictions that are more precise. The text-based model, on the other hand, is purely dependent on text and may have trouble accurately differentiating between true and false news.

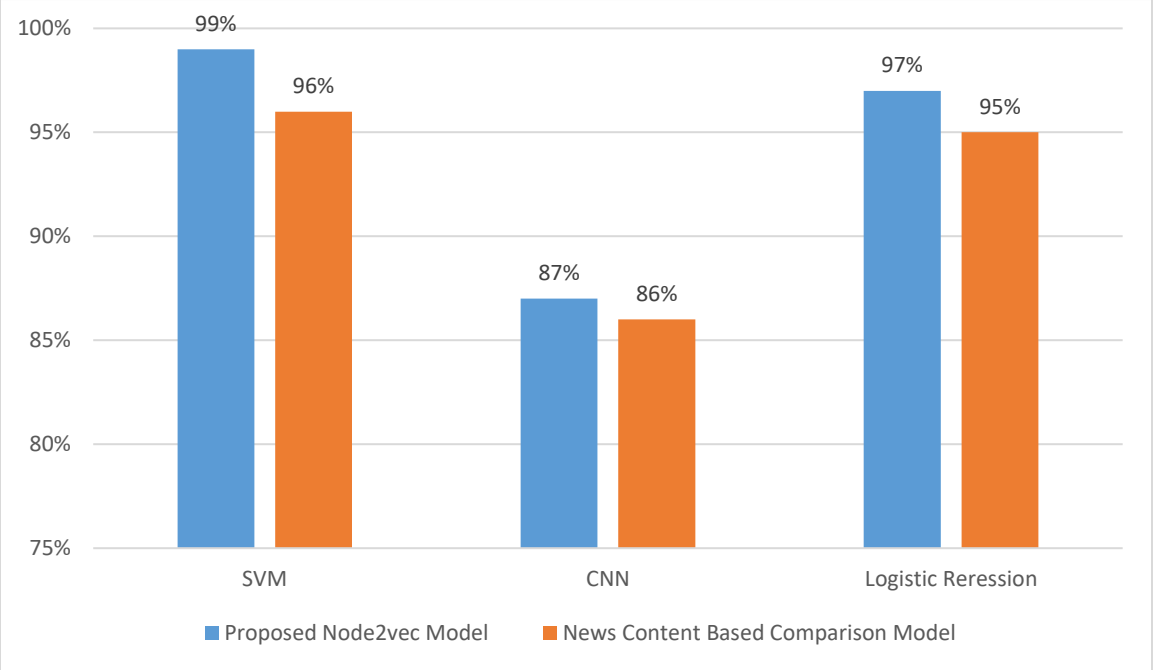

Fig 13: Classification Accuracy in Test Dataset

Chapter 6

Conclusion and Future Work

6.1 Conclusion

We summarize our research on identifying false news using text-based representations and Node2Vec-based embeddings in this chapter along with our main conclusions. We go over the benefits of our suggested model over the comparison model, as well as how well it can manage datasets that are unbalanced, and we also point out some of the Node2Vec algorithm's and our suggested approach's shortcomings.

The suggested Node2Vec-based fake news detection technique outperforms the conventional text-based comparison approach by a wide margin. First off, the use of graph embeddings considerably reduces the execution time. Our approach is ideally suited for real-time applications and large-scale datasets because of its improved efficiency, which enables prompt and precise predictions.

Second, our model demonstrates robustness in managing imbalanced datasets, a common problem in the identification of fake news. By utilizing Node2Vec embeddings, it effectively captures subtle patterns and significant relationships inside the online network, resulting in precise predictions even in the presence of a small number of fake news occurrences. Furthermore, by utilizing Node2Vec-based embeddings, the model is able to extract contextual data from the online graph, assisting in the identification of bogus news websites and creating links between them. The web graph representations help the machine recognize trends and classify bogus news sources, enabling more precise categorization.

Node2Vec has certain drawbacks even if it offers useful embeddings for our fake news detection model. Its sensitivity to the selection of hyperparameters, such as the walk length and the number of walks per node, is a severe constraint. To acquire the best outcomes, these factors

must be fine-tuned over a long period of experimentation. The accuracy and comprehensiveness of the web graph data may differ depending on the sources' accessibility, which can also have an impact on the algorithm's performance.

Although Google dominates the web search market, it is preferable to check node2vec's usability using additional search engines, like Bing and Yahoo. We intend to use a variety of search engines to further test and validate our fake news detection technology in upcoming studies.

Additionally, our suggested model has some restrictions. Its dependency on the availability of online graph data is one of its limitations. The model's performance might be impacted in situations when access to web sources is constrained. The quality and amount of the dataset also affect the performance of our suggested method, as is the case with any machine learning model. The accuracy and generalization abilities of the model could be further improved with a larger and more varied dataset.

6.2 Future Work

There are a number of intriguing directions for more study and improvements to our suggested false news detection algorithm going forward:

- Extension in several languages: At the moment, our model only supports Bangla. Its application to a wider range of scenarios and ability to work across several languages would both benefit from an expansion of its scope.
- Model diversification: Investigating additional deep learning and machine learning models could offer more information on how well they perform and contribute to an ensemble strategy for detecting fake news. The strengths of different models could be combined to provide forecasts that are more reliable and precise.
- Dataset augmentation would improve the model's ability to detect false information more successfully by addressing the dataset's class imbalance by increasing the number of examples of fake news. One approach to consider is the incorporation of synthetic data generating methods.
- Multi-URL Analysis: Increasing the model's scope to take into account numerous URLs for each news story would allow for a more thorough analysis of news sources. In order to provide a fuller representation of the news content, this would include combining embeddings from numerous URLs associated with the same news event.

In conclusion, our Node2Vec-based fake news detection model shows substantial promise for overcoming fake news detection difficulties. Its effectiveness is aided by its capacity to handle unbalanced datasets, quicker execution time, and web graph embeddings. Although the model shows promising results, there are still room for advancement, particularly in terms of overcoming Node2Vec algorithm restrictions and broadening the model's applicability to a wider range of languages and datasets. The future work described in this chapter paves the way for additional improvements in false news detection techniques as well as the encouragement of media literacy in the fight against disinformation.

References

[1] M. Z. a. R. M. A. a. I. M. S. a. K. S. Hossain, "Banfakenews: A dataset for detecting fake news in bangla," *arXiv preprint arXiv:2004.08789,* 2020.

[2] A. a. M. F. Bondielli, "A survey on fake news and rumour detection techniques," *Information Sciences,* vol. 497, pp. 38--55, 2019.

[3] M. G. a. H. M. R. a. R. M. a. P. J. a. A. H. S. Hussain, Detection of bangla fake news using mnb and svm classifier, IEEE, 2020, pp. 81--85.

[4] E. a. B. P. a. G. P. a. J. A. a. M. T. Grave, "Learning word vectors for 157 languages," 2018.

[5] J.-S. a. L. Y. a. A. H. Shim, "A link2vec-based fake news detection model using web search results," *Expert Systems with Applications,* vol. 184, p. 115491, 2021.

[6] V. L. a. C. N. a. C. Y. a. C. S. Rubin, Fake news or truth? using satirical cues to detect potentially misleading news, 2016.

[7] J. a. S. L. K. a. S. S. a. V. G. M. Ma, Beyond blacklists: learning to detect malicious web sites from suspicious URLs, 2009.

[8] R. a. K. G. a. A. D. a. G. J. a. N. P. Baly, "Predicting factuality of reporting and bias of news media sources," *arXiv preprint arXiv:1810.01765,* 2018.

[9] N. a. M. C. K. a. G. J. a. Z. X. a. Z. R. Sitaula, "Credibility-based fake news detection," *Disinformation, misinformation, and fake news in social media: Emerging research challenges and Opportunities,* pp. 163-182, 2020.

[10] L. a. B. E. a. H. M. a. O. S. Li, "Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate?," *Behaviour \& Information Technology,* vol. 33, no. 11, pp. 1136-1147, 2014.

[11] V. a. K. B. a. L. A. a. M. R. P{\'e}rez-Rosas, Automatic fake news detection, 2018.

[12] A. a. P. B. a. K. S. a. G. N. Chakraborty, 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, 2016.

[13] I. Y. R. a. A. R. A. a. R. F. Pratiwi, Study of hoax news detection using na{\"\i}ve bayes classifier in Indonesian language, IEEE, 2017.

[14] M. L. a. T. E. a. M. S. a. B. G. a. D. M. a. d. A. L. Della Vedova, Automatic online fake news detection combining content and social signals, IEEE, 2018.

[15] M. a. F. E. a. C. J. G. Sharifi, Detection of internet scam using logistic regression, IEEE, 2011.

[16] M. a. P. K. a. B. H. Vukovi{\'c}, An intelligent automatic hoax detection system, Springer, 2009.

[17] S. a. A. S. a. A. S. Badaskar, Identifying real or fake articles: Towards better language modeling, 2008.

[18] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648,* 2017.

[19] B. a. A. I. a. S. G. P. a. R. S. Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task," *arXiv preprint arXiv:1707.03264,* 2017.

[20] V. a. K. D. a. G. S. a. K. L. Y. a. V. V. Kumar, Identifying clickbait: A multi-strategy approach using neural networks, 2018.

[21] A. a. P. B. a. K. S. a. G. N. Chakraborty, Stop clickbait: Detecting and preventing clickbaits in online news media, IEEE, 2016.

[22] G. L. a. S. P. a. R. L. M. a. B. J. a. M. F. a. F. A. Ciampaglia, "Computational fact checking from knowledge networks," *PloS one,* vol. 10, no. 6, p. e0128193, 2015.

[23] G. a. N. P. a. M. L. a. B.-C. A. a. K. I. Karadzhov, "Fully automated fact checking using external sources," *arXiv preprint arXiv:1710.00341,* 2017.

[24] I. Y. R. a. A. R. A. a. R. F. Pratiwi, Study of hoax news detection using na{\"\i}ve bayes classifier in Indonesian language, IEEE, 2017.

[25] X. a. C. J. a. J. Z. a. X. F. a. S. Y. a. C. D. a. C. X. a. Z. J. Zhou, Real-time news cer tification system on sina weibo, 2015.

[26] S. A. a. D. S. H. a. F. B. C. a. L. J. Alkhodair, "Detecting breaking news rumors of emerging topics in social media," *Information Processing \& Management,* vol. 57, no. 2, p. 102018, 2020.

[27] E. a. B. G. a. D. V. M. L. a. M. S. a. D. A. L. Tacchini, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506,* 2017.

[28] N. a. L. K. Vo, The rise of guardians: Fact-checking url recommendation to combat fake news, 2018.

[29] K. a. B. H. R. a. L. H. Shu, "Studying fake news via network analysis: detection and mitigation," *Emerging research challenges and opportunities in computational social network analysis and mining,* pp. 43--65, 2019.

[30] M. G. a. V. Mesyura, "Fake news detection using naive Bayes classifier," *2019 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON),* pp. 900-903, 2019.

[31] Z. a. C. J. a. Z. Y. a. Z. J. a. T. Q. Jin, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia,* vol. 19, no. 3, pp. 598-608, 2016.

[32] S. a. D. A. a. A. A. a. B. S. S. Elkasrawi, What you see is what you get? Automatic Image Verification for Online News Content, IEEE, 2016.

[33] X. a. G. A. A. Zhang, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing \& Management,* vol. 57, no. 2, p. 102025, 2020.

[34] E. a. B. P. a. G. P. a. J. A. a. M. T. Grave, "Learning word vectors for 157 languages," *arXiv preprint arXiv:1802.06893,* 2018.

[35] K. a. S. A. a. W. S. a. T. J. a. L. H. Shu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter,* vol. 19, no. 1, pp. 22--36, 2017.

[36] D. K. a. V. D. a. Y. A. Vishwakarma, "Detection and veracity analysis of fake news via scrapping and authenticating the web search," *Cognitive Systems Research,* vol. 58, pp. 217--229, 2019.

Appendix A

Some Code Segments used for Node2vec and Machine Learning and Deep Learning Classification Models

A.1 Code for Generating Web Graph

```
web_graph = nx.Graph()
for source in df["Source"]:
    web_graph.add_node(source)


fig, ax = plt.subplots(figsize=(10, 10), facecolor='black')


pos = nx.spring_layout(web_graph)  # Define the layout algorithm
nx.draw(web_graph, pos, with_labels=True, node_size=50, node_color='blue',
edge_color='red', alpha=0.7, ax=ax)


plt.show()
```

A.2 Code for Node Embeddings using Node2vec

```
node2vec = Node2Vec(web_graph, dimensions=128, walk_length=30,
num_walks=200)

model = node2vec.fit(window=10, min_count=1)
```

A.3 Code for Extracting Features from URLs

```
features = []
for source in df["Source"]:
    features.append(model.wv[source])

features_array = np.array(features)
```

A.4 Code for Plotting the URLs in Graph After Feature Extraction Based on Their Similarities

```
tsne = TSNE(n_components=2, random_state=42)
embedding_2d = tsne.fit_transform(features_array)


df = df.reset_index(drop=True)


fig, ax = plt.subplots(figsize=(7, 7))


ax.scatter(embedding_2d[:, 0], embedding_2d[:, 1], c='blue', alpha=0.5)


num_points_to_annotate = min(len(df["Source"]), 20)

annotated_points = np.random.choice(len(df["Source"]), num_points_to_annotate,
replace=False)

for i in annotated_points:
    ax.annotate(df["Source"][i], (embedding_2d[i, 0], embedding_2d[i, 1]),
color='black')


plt.show()
```

A.5 Code for SVM Classifier

```
svm = SVC()
svm.fit(X_train, y_train)
svm_predictions = svm.predict(X_test)
t_pred = svm.predict(X_train)
svm_train_accuracy = accuracy_score(y_train, t_pred)
svm_accuracy = accuracy_score(y_test, svm_predictions)
print("SVM Test Accuracy:", svm_accuracy)
print("SVM Train Accuracy:", svm_train_accuracy)

ms_pre = precision_score(y_test, svm_predictions)
print("Precision :", ms_pre)

ms_rec = recall_score(y_test, svm_predictions)
print("Recall :", ms_rec)

ms_f3 = f1_score(y_test, svm_predictions)
print("F1 :", ms_f3)
```

A.6 Code for Logistic Regression Model

```
param_grid = {'C': [0.1, 1, 10]}
grid_search = GridSearchCV(LogisticRegression(), param_grid, cv=5)
grid_search.fit(X_train, y_train)
best_model = grid_search.best_estimator_


train_predictions = best_model.predict(X_train)
test_predictions = best_model.predict(X_test)

train_accuracy = accuracy_score(y_train, train_predictions)
test_accuracy = accuracy_score(y_test, test_predictions)
precision = precision_score(y_test, test_predictions)
recall = recall_score(y_test, test_predictions)
f1 = f1_score(y_test, test_predictions)

print("Logistic Regression Train Accuracy:", train_accuracy)
print("Logistic Regression Test Accuracy:", test_accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

A.7 Code for CNN Model

```
model = Sequential()
model.add(Embedding(input_dim=X.shape[0], output_dim=X.shape[1],
input_length=X.shape[1]))
model.add(Conv1D(filters=256, kernel_size=5, activation='relu'))
model.add(GlobalMaxPooling1D())
model.add(Dense(units=64, activation='relu'))
model.add(Dense(units=1, activation='sigmoid'))


model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

cnn_probabilities = model.predict(X_test).flatten()
cnn_predictions = np.where(cnn_probabilities >= 0.3, 1, 0)
cnn_accuracy = accuracy_score(y_test, cnn_predictions)
cnn_train_probabilities = model.predict(X_train).flatten()
cnn_train_predictions = np.where(cnn_train_probabilities >= 0.3, 1, 0)
cnn_train_accuracy = accuracy_score(y_train, cnn_train_predictions)
cnn_precision = precision_score(y_test, cnn_predictions)
cnn_recall = recall_score(y_test, cnn_predictions)
cnn_f1 = f1_score(y_test, cnn_predictions)

print("CNN Results:")
print("Accuracy:", cnn_accuracy)
print("Train Accuracy:", cnn_train_accuracy)
print("Precision:", cnn_precision)
print("Recall:", cnn_recall)
print("F1-score:", cnn_f1)
```