DESIGN AND DEVELOPMENT OF A TWITTER SPAM DETECTOR USING METAHEURISTIC APPROACHES

A

Thesis

Submitted for the Award of the Degree of

DOCTOR OF PHILOSOPHY

In

COMPUTER SCIENCE AND ENGINEERING

Submitted By

PINNAPUREDDY MANASA

41900524

Supervised By

Dr. Arun Malik (17442)

Computer Science and Engineering (Professor)

Lovely Professional University

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB

2024

DECLARATION

I, hereby declared that the presented work in the thesis entitled "Design and Development of a Twitter Spam Detector using Metaheuristic Approaches" in fulfilment of degree of **Doctor of Philosophy (Ph. D.)** is outcome of research work carried out by me under the supervision Dr. Arun Malik, working as Professor , in the Computer Science and Engineering of Lovely Professional University, Punjab, India. In keeping with general practice of reporting scientific observations, due acknowledgements have been made whenever work described here has been based on findings of other investigator. This work has not been submitted in part or full to any other University or Institute for the award of any degree.

(Signature of Scholar)

Name of the scholar: Pinnapureddy Manasa

Registration No.: 41900524

Department/school: Computer Science and Engineering

Lovely Professional University,

Punjab, India

CERTIFICATE

This is to certify that the work reported in the Ph. D. thesis entitled "Design and Development of a Twitter Spam Detector using Metaheuristic Approaches" submitted in fulfillment of the requirement for the reward of degree of **Doctor of Philosophy (Ph.D.)** in the Computer Science and Engineering is a research work carried out by Pinnapureddy Manasa, 41900524, is bonafide record of his/her original work carried out under my supervision and that no part of thesis has been submitted for any other degree, diploma or equivalent course.

(Signature of Supervisor)

Name of supervisor: Dr. Arun Malik

Designation: Professor

Department/school: Computer Science and Engineering

University: Lovely Professional University

Abstract

Twitter spam refers to unwanted or unsolicited content, often in the form of excessive, irrelevant, or deceptive messages, that is distributed on the Twitter platform. These spam messages can have various negative effects on users and the overall Twitter experience. Twitter spam has become a pervasive issue on the platform, affecting users and the overall Twitter experience. Spam refers to the distribution of unwanted or unsolicited content, often in the form of excessive, irrelevant, or deceptive messages. The effects of Twitter spam are far-reaching and can have negative implications for users and the platform as a whole.

Firstly, the user experience is significantly impacted by Twitter spam. Spam messages flood timelines with irrelevant or misleading content, making it harder for users to find meaningful and valuable information. Moreover, Twitter spam contributes to the spread of misinformation. Spam messages often contain false information, malicious links, or phishing attempts. This dissemination of misleading content can have serious consequences, such as the perpetuation of rumors, the amplification of harmful narratives, and potential security risks for users. For businesses and individuals using Twitter for promotion or marketing purposes, spam can tarnish their brand reputation. Being associated with spammy content can undermine their credibility and trustworthiness. It is crucial for individuals and organizations to maintain a clean and spam-free presence on Twitter to protect their reputation and maintain trust with their audience.

To combat the growing problem of Twitter spam, there is a pressing need for AI-based spam detection systems. Artificial Intelligence (AI) offers several advantages in effectively identifying and mitigating spam. AI-based systems can handle the immense scale and speed at which content is generated on Twitter. With millions of tweets being posted every day, AI algorithms can efficiently analyze and process this vast volume of data in real-time, enabling quick detection and mitigation of spam. AI-powered spam detection systems leverage advanced machine learning techniques to continuously learn and adapt to evolving spam patterns. They can identify subtle indicators and patterns that may go unnoticed by human moderators. By analyzing large datasets, AI models can uncover hidden connections and characteristics of spam

messages, enhancing the accuracy of detection. AI-based solutions automate the spam detection process, significantly reducing the manual effort required to combat spam. By automatically identifying and filtering out spam messages, these systems free up human moderators to focus on other important tasks, such as addressing user inquiries or handling content moderation that requires human judgment.

AI-based spam detection systems can adapt and evolve alongside changing spamming tactics. As spammers constantly modify their techniques, AI algorithms can be updated and trained to counter new and emerging forms of spam. This adaptability ensures a proactive approach in mitigating spam and staying ahead of spammers' strategies. A technique for detecting spam that uses a swarm optimization methodology is presented in this research. A dataset for the identification of spam tweets is used to train the machine learning model. The input features from the dataset serve as the foundation for the development of metaheuristic features. The appropriate properties are selected using the Whale swam Optimization Algorithm (WOA). The stochastic gradient descent (SGD) algorithm replaces the conventional objective function of the WOA to carry out the feature selection process. The Adaboost classifier is trained to recognize spam in tweets using the chosen subset of features. With WOA and SGD, the Adaboost classifier derived the best results.

Deep learning algorithms-based tweet spam detection is a useful method for locating and removing spammy content on Twitter. Deep learning models leverage the power of neural networks to learn intricate patterns and features from large volumes of data, enabling them to make accurate predictions. The proposed model processes a dataset consisting of tweets and additional metadata, such as follower count and user actions. The model is divided into two sections that operate on this dataset. In the initial step, the focus is on the tweet content, which is analyzed using Global Vectors for Word Representation (GloVe) language model to extract lexical features. These features are then input into a Long Short Term deep learning model to detect spam. In the second phase, a Convolutional Neural Network model is employed to classify the tweets, utilizing both the metadata within the tweets and additional meta-heuristic characteristics. These characteristics include factors like tweet length and the presence of question marks. Combining the outcomes from the Long Short Term Memory and

Convolutional Neural Network models into a single set of findings gives the final result.

Twitter spam detrimentally affects users, the spread of information, and brand reputations. The adoption of AI-based spam detection systems is vital in combating this problem effectively. AI's ability to handle the scale, speed, and complexity of Twitter data, coupled with its adaptability and automation capabilities, makes it a crucial tool in maintaining a spam-free and trustworthy Twitter environment for all users.

ACKNOWLEDGEMENT

With immense pleasure and deep sense of gratitude, I wish to express my sincere thanks to my supervisor Dr. Arun Malik, Professor, School of Computer Science and Engineering, Lovely Professional University. In my expedition towards this degree, I have found a teacher, an inspiration, a role model, and a pillar of support in my Guide. He has been there providing his incredible support and guidance at all times and has given me worthy counselling, suggestions and recommendations in my quest for knowledge. It was a great privilege and honour to work and study under his guidance. Without his able guidance, this thesis would not have been possible, and I shall infinitely be indebted to him for his overwhelming assistance. I am highly obliged for his unfathomable confidence in me and providing continuous encouragement at all times.

Special thanks go to examiners of term end reports and reviewers of the journals who vetted my submissions and gave valuable comments to improve the work further.

Finally, I wish to express my profound gratitude to my mother, father, brother, husband and all members of my family. Their incomprehensible faith in me, their sacrifices, unflappable strength, and encouragement at all times has been an inspiration and is solely responsible for refurbishing my life and profession. I pray to God and hope that He empowers me, to live up to their expectations all the time.

Date: 20/09/2023 Pinnapureddy Manasa

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Overview

Identification and removal of spam is becoming an increasingly crucial task as social media platforms like Twitter see rapid expansion, as both the site's dependability and the security of its users depend on it. "Twitter spam detection" is the process of finding and deleting junk accounts and content from Twitter. This is done in a number of ways, such as through machine learning as well as deep learning. It's important for Twitter to keep the user experience good by keeping trash from filling the site and making sure people can access and connect with relevant and useful content. This chapter introduces the functioning and issues in social media platforms and how important is twitter spam detection.

1.2 Social Media

The manner in which we interact with one another, communicate with one another, and consume information have all been radically changed by social media platforms, which already count more than 4 billion active users globally [1]. The impact of social media may be seen in many spheres of life, including business, entertainment, politics, and even personal relationships.

One of the best things about the rise of social media is that it has made it easier for people to talk to each other. A quick and simple approach to communicate with people worldwide is by using social media platforms like Facebook, Twitter, Instagram, and LinkedIn[2][3]. It has completely changed the way we interact by dismantling geographical borders and making it possible for us to maintain relationships with friends, family, and coworkers regardless of the physical distance between us. A voice has also been given to disenfranchised communities and people thanks to the rise of social media, which has provided a forum for these groups and individuals to express their thoughts and share their experiences [4].

Another area where the effects of social media may be felt is in politics. Political parties and leaders at all levels have recently come to understand the usefulness of

social media as a tool for communication and election campaigns. Because of the widespread use of social media, politicians may now reach a wider audience, particularly younger voters who are more engaged on social media. Politicians now run their campaigns on social media platforms, and they employ a number of strategies to reach voters, including targeted advertising, influencer marketing, and hashtag campaigns. Another new trend is for companies to use social media as a way to sell to customers and talk to them. Platforms for social media give companies the chance to communicate with consumers, market their goods and services, and reach a large audience. The rise of social media has had a big impact on how businesses are run as well. These days, many firms use social media platforms in order to make it easier for employees to work remotely and collaborate online.

Additionally, the entertainment business has been revolutionized by social networking. Traditional celebrity culture has been shaken up as a result of the growth of social media influencers, who have now become important people in their own right. Artists, musicians, and filmmakers now have an additional venue to exhibit their work and communicate with a larger audience thanks to the proliferation of social media. In addition to this, it has made it possible for new types of entertainment to emerge, such as live streaming, online gaming, and experiences that use virtual reality.

On the other hand, social media doesn't always have a positive effect. Using social media has been linked to a number of bad things, such as harassment, mental health problems, and the spread of fake news and wrong information. The algorithms that run social media platforms give precedence to information that is more likely to keep users engaged, which may lead to the proliferation of viewpoints that are divisive and extreme. Many users share personal information on social media platforms without fully comprehending the repercussions of doing so, which is another developing issue over the effects of social media on users' privacy.

The significance of the function that social media plays in society as well as the impact it has on many elements of our life are both readily apparent. It has completely altered the ways in which we communicate with one another, take in information, and engage with one another. The use of social media has not only altered the commercial

world but also the entertainment industry, bringing with it both new possibilities and new obstacles. It is abundantly obvious that social media will continue to play an important part in molding the future of society, despite the fact that the effects of social media are not uniformly favorable. As a result, it is very necessary to make sure that social media is used properly and that suitable protections are put into place to limit the harmful impacts of its usage.

1.2.1 History and Rise of Social Media

Social media refers to many online platforms that enable users to generate and distribute information, make connections with other people, and take part in online communities. It is possible to trace the origins of social media all the way back to the early days of the internet, when the most common forms of online communication consisted of online message boards and chat rooms [5].

Six Degrees, which began in 1997, is generally acknowledged as being the first social networking site. Users were able to build profiles, communicate with friends and connect with new people, and send messages. However, owing in large part to the restricted access to the internet that existed at the time, Six Degrees was never able to achieve global recognition.

At the beginning of the 2000s, social media platforms such as Friendster and MySpace saw a surge in popularity, leading to the recruitment of millions of new users who created accounts, connected with friends, and shared material on the sites. In instance, MySpace was very popular, reaching its zenith with more than 100 million active members.

Facebook was first made available to college students shortly after its introduction in 2004. It became quite well-known very rapidly, and by 2008, it had over one hundred million active members. The success of Facebook may be attributed, in large part, to the company's emphasis on the user experience, privacy, and security. In addition to this, it was the first platform to provide new features like the news feed and the like button, both of which went on to become fundamental components of social media platforms.

In the years that followed, more social media platforms such as Twitter, LinkedIn, and Instagram were introduced; each of these platforms offers a distinctive collection of features and caters to a distinct user demographic. For instance, Twitter was first conceived as a platform for microblogging, but LinkedIn was conceived as a platform for professionals and corporations.

During the latter part of the 2000s and the beginning of the 2010s, the proliferation of smartphones and mobile internet connections further accelerated the expansion of social media. Users were now able to access social media sites at any time and from any location, and they could post material in real time. The advent of mobile-based social media platforms like Snapchat and TikTok has proved the value of user-generated content and short-form video content in particular.

In the world we live in now, using social media has become an important part of our daily lives. People use social media sites for an average of 2.5 hours a day, and there are more than 4 billion regular social media users around the world. The rise of social media has changed the way we talk to each other, get information, and interact with each other in basic ways. Social media also continues to shape the future of human society.

The development and expansion of social media have been marked by ongoing innovation, the introduction of new features, and shifting patterns of user activity. Social media platforms have become an important part of our everyday lives, as billions of people around the world use them to stay in touch with friends and family, share information, and take part in online communities. Emerging technologies such as virtual and augmented reality, artificial intelligence, and blockchain will allow new kinds of participation, creativity, and social interaction, and it is anticipated that these technologies will have a significant impact on the future of social media.

1.2.2 Problems Caused by Social Media

Unquestionably, social media has completely changed the manner in which individuals interact with one another, connect with others, and take in information[6]. However, it has also given birth to a variety of issues that may have major negative consequences on people, communities, and society as a whole. These difficulties can

have serious negative repercussions on individuals, communities, and society as a whole. In the following paragraphs, we will talk about some of the issues that are brought about by social media.

1. **Cyberbullying**: One of the most significant issues generated by social media is cyberbullying. Harassment, hate speech, and other types of abusive conduct may have catastrophic impacts on victims, leading to despair, anxiety, and even suicide in extreme cases. Online harassment is only one form of cyberbullying.
2. **Addiction:** Users of social media platforms spend hours each day browsing through their feeds, checking alerts, and replying to messages, which is evidence that social media platforms may be addicting. This addiction may cause problems at work, school, and in personal relationships, as well as feelings of worry and sadness. It can even lead to physical health problems.
3. **Misinformation and disinformation**: The proliferation of social media has made it much simpler for erroneous information to circulate rapidly and widely. Especially when it comes to situations involving health and safety concerns, the dissemination of false information may result in confusion, fear, and even physical injury.
4. **Concerns about users' privacy**: The platforms for social media platforms gather huge quantities of user data, which may be used for a variety of reasons, including targeted advertising. This gives rise to significant issues over the users' right to privacy and the safety of their data, in particular in situations when sensitive information is involved.
5. **Radicalization through the Internet**: Extremist organizations have used social media platforms to disseminate their ideology and attract new members. This may result in online radicalization, which is when susceptible people are led into violent views and actions via the use of the internet.
6. **A culture of comparison**: The usage of social media platforms may lead to a culture of comparison, in which users compare themselves to others based on their looks, accomplishments, and lifestyle choices. This may result in a lack of confidence, feelings of inadequacy, and even sadness in some people.

7. **Division**: The use of social media may worsen political and social division, since users often gravitate toward groups and people who share their views. This might result in people being only exposed to information and ideas that validate their preexisting beliefs and prejudices, which can lead to the establishment of echo chambers.

The proliferation of social media has resulted in the emergence of a number of significant issues that have the potential to have a harmful impact not only on individuals but also on communities and society as a whole. Although there are numerous advantages to using social media, it is essential to be aware of the potential drawbacks and to take measures to address them[7]. This may mean putting restrictions on the amount of time spent on social media, educating oneself on issues related to privacy and security, and actively seeking out a variety of ideas and points of view.

1.2.3 Social media platforms

1. **Facebook**: As of March 2021, Facebook has over 2.8 billion members that were actively using the network on a monthly basis. This makes it the biggest social media platform in the world. Users are granted the ability to create profiles, establish connections with friends and family members, participate in groups, and exchange material. More than sixty percent of Facebook's users are at least 35 years old, making this demographic one of the platform's most active. Facebook's popularity with younger audiences has been declining, with many of these audiences choosing more recent platforms such as Instagram and TikTok.
2. **YouTube**: YouTube is a platform for sharing videos that enables users to both post and view videos shared by other users. As of the month of May 2021, it has over 2 billion monthly active users. Over 80% of YouTube viewers are between the ages of 15 and 25, making the platform especially popular with younger audiences. Additionally, it is the second most widely used search engine, behind only Google.
3. **WhatsApp**: WhatsApp is a messaging program that enables users to communicate text messages, voice messages, and make phone calls to one

another. As of the month of February 2021, it has over 2 billion monthly active users. The countries of India, Brazil, and Mexico are leading the pack in terms of WhatsApp use.

4. **Instagram**: Instagram is a platform for sharing photos and videos that gives users the ability to create and share visual material with one another. As of April 2021, it has more than 1.2 billion monthly active users worldwide. Over 70 percent of Instagram's user base is under the age of 35, making it one of the most popular social media platforms among younger demographics. Additionally, many companies and influencers use it as a platform to share their content.
5. **TikTok**: TikTok is a platform for sharing short videos that enables users to make videos set to music and share them with other users. As of the month of February 2021, it has over 1 billion monthly active users. TikTok is especially well-liked among younger audiences, with more than 60 percent of its users falling in the 16-24 age range. Additionally, it is gaining popularity among people of older generations.
6. **Twitter** is a platform for microblogging that enables users to publish and exchange brief messages known as tweets. Twitter is also known as "tweets." As of the month of April 2021, it has over 330 million monthly active users. Twitter is especially popular among journalists, politicians, and celebrities, and it is often used as a venue for public discussion and the dissemination of breaking news.
7. **LinkedIn**: LinkedIn is a platform for professional networking that enables users to establish professional profiles, interact with colleagues and peers, and look for work in addition to searching for employment opportunities. As of April 2021, it has over 740 million subscribers worldwide. LinkedIn is especially well-liked among working professionals and companies, and it is often put to use for the purposes of recruiting and professional advancement.
8. **Snapchat** is a messaging software that enables users to transmit photographs and brief movies to one another that vanish after a few seconds. Snapchat was developed by the company Snap Inc. As of the month of March 2021, it has more than 280 million daily active users. Snapchat has over 75% of its users

between the ages of 13 and 34, making it especially popular with younger audiences. Additionally, many companies and advertising make use of this platform.

9. **Pinterest**: Pinterest is a platform for visual discovery and bookmarking that enables users to store photos and ideas and share them with other people. As of the month of April 2021, it has more than 450 million monthly active users. Over 70% of Pinterest's users are female, indicating that the platform's primary appeal lies with women. Additionally, it is a well-liked platform for doing e-commerce and purchasing online.

Social media platforms like snapchat, YouTube, Twitter etc.[8] continue to develop new features and gain more users, while new platforms are always being introduced into the market. The platforms that were just mentioned are some of the most well-known and prominent in the modern environment. Each of these platforms has its own set of characteristics, audiences, and chances for interaction.

1.3 Twitter

Twitter is a platform for social media that enables users to submit what are known as tweets, which are brief communications. Tweets may be as long as 280 characters and can contain text, images, videos, and links in addition to the standard 140 characters. Twitter was first launched in 2006 and has since grown to become one of the most widely used social media platforms throughout the globe [9][10][11].

Users on Twitter are able to "follow" other users and view the tweets that they post in their own feed. Hashtags are another way for users to organize their tweets and increase the likelihood that other users will find and read them. Twitter has rapidly grown in popularity as a venue for the dissemination of breaking news, updates in real time, and public dialogues on a broad range of subjects.

Twitter offers a variety of extra tools and services in addition to its main features. These include Twitter Ads, which enables companies to advertise on the network, and Twitter Analytics, which gives statistics and insights about the success of tweets and audience interaction. Both of these tools are available to users. Twitter has also been used as a tool for social and political activism, with hashtags and tweets being used to

raise awareness about social problems and to organize demonstrations and rallies. Activists have found this to be a useful way to further their causes.

1.3.1 Popularity of Twitter

There are several factors that have contributed to Twitter's rise to prominence as a significant social media network, including the following:

1. **Real-time updates**: Because Twitter is updated in real time, users are able to exchange information and updates on current events, news, and other subjects of interest in a way that is both fast and simple. Because of this, Twitter is an excellent medium for the dissemination of breaking news and live events, as well as for conversations and debates over topical themes.
2. **Coverage of a diverse variety of subjects** Because Twitter is such an open platform, users are free to debate and exchange information on nearly any subject they want. Twitter has become a popular venue for a broad variety of interests, ranging from sports and entertainment to politics and social concerns, in large part as a result of this.
3. **Ease of use**: Twitter is a platform that is incredibly simple to use, with an easy-to-understand UI and features that are straightforward to use. Because of this, it is usable by people of varying ages and levels of technological expertise, which has contributed to its growing popularity all over the globe.
4. **Features that promote user involvement**: Twitter provides users with a variety of features that encourage user interaction, such as the ability to retweet, like, and utilize hashtags. Twitter has become a very engaging and social platform thanks in large part to the existence of a number of tools that assist to amplify messages and foster discussions.
5. **Access to prominent people**: Twitter has become a popular venue for public figures, such as celebrities, politicians, and business leaders, to express their ideas and communicate with their fans and followers. This includes the ability to send direct messages (DMs). Because of this, Twitter has become a very visible platform, which has led to the network's rise in popularity.

The real-time updates, broad variety of subjects, simplicity of use, interesting features, and access to prominent figures that are all available on Twitter have contributed to the platform's rise to the position of one of the most popular social networking sites [12][13]. Twitter is a vital tool for people, companies, and organizations alike because of its capacity to ease communication, disseminate information, and drive involvement.

1.3.2 Twitter's Functionality

Twitter is functional because it enables users to publish brief messages to the platform in the form of "tweets." These tweets may be as long as 280 characters and can contain text, photographs, videos, and links in addition to the aforementioned media types. Hashtags are another way for users to organize their tweets and increase the likelihood that other users will find and read them.

When a user publishes a tweet, the tweet is published to the person's profile and is seen in the feeds of people who follow that user. Users have a variety of options available to them for interacting with tweets, such as retweeting, like, and responding to tweets. When one person retweets another user's tweet, the tweet is then shared with the user's own followers, which may serve to enhance the reach of the initial tweet [14]. When a person likes a tweet, it is added to the list of tweets that they have liked, and when a user responds to a tweet, their answer is shown underneath the tweet that they were responding to.

Twitter provides its users with a variety of additional tools and services, in addition to its main features, that enable them to personalize the experience they have while using the network. For instance, users may establish lists of accounts that they follow to help them keep track of various subjects or interests, or they can use sophisticated search capabilities to identify tweets that are connected to certain keywords or phrases. Both of these options are available to users. Businesses are able to target particular audiences on Twitter because of the variety of advertising options that the site provides, such as Promoted Tweets and Promoted Accounts, which Twitter gives to its users.

Because of its open structure and emphasis on real-time updates, Twitter has become a popular medium for breaking news, public dialogues on a broad variety of issues, and live events. It is a very social platform that fosters engagement and discussion among its users, and it is easy to use and has interesting features, both of which have contributed to creating it this way. Overall, the one-of-a-kind features and services offered by Twitter have contributed to the development of an interactive and interesting platform that has evolved into a useful resource for people, corporations, and other organizations.

1.3.3 How people use Twitter.

1.3.3.1 Twitter as a Marketing Tool

Twitter has emerged as a popular tool for marketing among companies and organizations that are interested in connecting with their clients and promoting their goods or services. Because of its real-time nature and open communication channels, the platform is an excellent choice for platforms that are used for interacting with consumers and increasing brand recognition. Building consumer awareness of a company's brand is one of the ways in which companies may utilize Twitter as a marketing strategy. Establishing a robust online presence and elevating a company's profile in the eyes of prospective clients may be facilitated for companies by tweeting on a regular basis with content that is both educational and interesting. Tweets might contain updates about new goods or services, news about the firm, or insights about the industry that highlight a brand's competence and provide value [15][16][17].

Engaging with clients in real time is another way that Twitter may be used effectively as a marketing tool. Businesses have the ability to reply to questions, comments, and concerns raised by consumers; doing so may help businesses create connections with their customers and demonstrate that the businesses are sensitive to and attentive to the requirements of their customers. This has the potential to result in greater levels of satisfaction and loyalty among customers. Sharing material, such as blog entries, films, or infographics, may also be accomplished via the usage of Twitter. Businesses may position themselves as thought leaders in their area by publishing material that is both relevant and instructive. This will also boost the number of followers and interaction the company receives. Additionally, companies may use Twitter to

promote promotions, discounts, and special offers, all of which can help boost sales and conversions. Twitter is a great tool for this.

Twitter also provides a variety of advertising solutions, which enable companies to boost their presence on the network by targeting certain demographics and expanding their customer base. While Twitter Ads enables companies to develop targeted campaigns based on the demographics, interests, and behaviors of their target audiences, Promoted Tweets and Promoted Accounts enable businesses to attract new consumers and boost the number of followers they have. In general, Twitter is an adaptable and efficient marketing tool that may assist companies in developing their brand, increasing consumer engagement and content sharing, and driving sales. Businesses have the opportunity to broaden their reach to new audiences, improve their exposure, and foster closer connections with the clients they already have by effectively using Twitter.

1.3.3.2 Twitter as a Social Messaging Tool

Twitter is a popular social messaging site where people can send and read short messages, called "tweets," from people who follow them. Because of its real-time nature and open communication channels, the platform is an excellent resource for establishing connections with other people, exchanging information, and taking part in public dialogues. Connecting with friends and family members is one of the ways that Twitter may be used as a social messaging platform. Users have the ability to follow other users and get their tweets in their feeds, which enables them to keep tabs on what is happening in the lives of their friends and family members. In addition, users have the ability to utilize Twitter to send direct messages to other users, which enables them to engage with one another in a manner that is both more private and direct.

Participating in public discussions is another manner in which Twitter may be utilized as a social communications tool[18]. Hashtags allow users to participate in conversations on a broad variety of subjects, ranging from current events and politics to sports and entertainment. This has the potential to boost users' level of engagement and exposure, as well as their ability to interact with people who have similar interests and points of view. Sharing information and ideas with others through Twitter may

also be done in the capacity of a social messaging tool. Users are able to submit links to articles, movies, and other stuff that they think other users will find interesting or helpful. This gives users the opportunity to share their knowledge and skills with one another. Twitter users also have the ability to interact with one another by asking questions or seeking advice from other users, which may help users learn and develop.

1.4 Spam in Twitter

The term "spam" on Twitter refers to any material that is either undesired or unsolicited that is sent out to users of Twitter [19][20][21]. It may manifest itself in a variety of ways, such as direct messages, mentions, and responses, and its effects can vary from vexatious to destructive. The following is a list of some of the most prevalent forms of spam on Twitter:

1. **Unsolicited mentions** Unsolicited mentions are mentions or tags in tweets that are not connected to the user or the user's interests[22]. These may be obnoxious and distracting, as well as fill up a user's Twitter feed, making it difficult to follow conversations.
2. **Direct message spam**: Unwanted direct messages from other Twitter users, often including links to phishing sites or advertisements for goods or services; also known as "direct message spam." If the links in these messages go to malicious websites or include malware, then the messages themselves might be hazardous.
3. **Phishing scam**: Phishing scams are tweets or direct messages that include links to websites that are meant to steal a user's personal information, like usernames, passwords, or credit card information. These websites are designed to acquire this information in order to commit identity theft.
4. **Hashtag spam**: The use of unrelated hashtags in tweets in an effort to get more visibility or followers is an example of the practice known as "hashtag spam." These tweets, which often have nothing to do with the hashtag in question, may be confusing and frustrating to readers.
5. **Tweetbot spam**: TweetBot spam is the use of automated Twitter bots to send out spam messages, often marketing goods or services. This kind of spam is

known as "tweetbotting." These bots may be quite unpleasant, and it might be difficult to stop them.

6. **Fake followers**: Fake followers are Twitter accounts that have been established with the sole intention of fraudulently increasing the number of people who follow a certain person. These accounts are often generated by bots, and you can tell they are fake by looking for a lack of activity and followers on the account.

Users have the ability to report spam accounts or material to Twitter, block or mute other users who are sending spam, and utilize third-party applications to filter undesired information from their Twitter feeds in order to fight spam on Twitter. In order to avoid falling victim to phishing schemes, it is essential to exercise extreme caution before opening direct messages from unknown people or clicking on links sent by them, as well as to choose passwords that are both robust and unique.

1.4.1 Spam Detection in Twitter

On Twitter, spam may manifest itself in a variety of ways, such as via direct messages, mentions, and replies. There are a number of different approaches that may be used to find and report spam accounts and material on Twitter in order to prevent spam [23][24][25]. The following is a list of some of the most frequent ways for detecting spam on Twitter:

1. **Pattern recognition using machine learning algorithms** Pattern recognition using machine learning algorithms may be used to identify patterns in tweets that are often associated with spam. These algorithms are able to be trained using enormous datasets of labeled data, which may contain both instances of spam tweets and examples of tweets that are not spam. These algorithms are able to detect accounts and material that are likely to be spam or bots by assessing characteristics like the content of tweets, the behavior of users, and the structure of the network. After a spam account or piece of content has been located, machine learning techniques may be used to automatically flag and delete the offending material.

2. **Content analysis**: Content analysis is a process that entails evaluating the content of tweets and searching for keywords and phrases that are often connected with spam. For instance, the use of promotional language, links that seem suspect, or information that is repeated may all be signs of spam. Content analysis can determine whether tweets and accounts include spam by examining not just the text but also any additional material that may be there.
3. **Network analysis**: Network analysis is a process that includes evaluating the connections between Twitter accounts and searching for patterns that are often connected with spam. Indicators of spam might include, for instance, groups of accounts that follow one other or engage in questionable conduct. Network analysis may be used to discover spam accounts and material by doing a study of Twitter's social graph as well as the patterns of interaction that occur between accounts.
4. **Analysis of user behavior:** In order to detect spam accounts and material, user behavior analysis is a technique that includes examining the activity of individual Twitter users. Users that engage in suspicious activities, such as following a high number of accounts in a short period of time or sending out numerous tweets that are identical to one another, for instance, may be symptoms of spam on Twitter. The identification of spam accounts and material is made possible through user behavior analysis, which works by examining patterns of user activity.
5. **Crowdsourcing**: Utilizing the Collective Intelligence of a Large Group of Users to detect Spam Accounts and Content Crowdsourcing refers to the practice of utilizing the collective intelligence of a large group of users to detect spam accounts and content. This strategy, which may be used to enhance automatic detection methods, can entail the use of user reports or the human assessment of material and accounts that are suspected of being spam. Crowdsourcing is a method that can help detect and eliminate spam from Twitter by drawing on the knowledge and expertise of the platform's user base as a whole.

The identification of spam on Twitter is a difficult and time-consuming activity that calls for a hybrid approach consisting of both automatic and human processes. Twitter is able to detect and delete spam accounts and material by using machine learning algorithms, content analysis, network analysis, user behavior analysis, and crowdsourcing. This allows Twitter to give users a safer and more pleasurable experience.

1.4.2 Need for Twitter Spam Detection

When considering the effects that spam has on Twitter and the people that use the network, it is possible to have a better understanding of the need and significance of Twitter spam detection. Detecting spam on Twitter is very necessary for a number of important reasons, including the following:

1. **Preserving the confidence of users**: Spam has the potential to damage user faith in a platform by inundating users with material that is either irrelevant or deceptive. Because of this, consumers may get frustrated, disengaged, and finally stop using the product. A reliable spam detection system may assist in the preservation of user trust by ensuring that users are only presented with material that is relevant to their needs and beneficial to them.
2. **Safeguarding the privacy and safety of users**: Spam may be used to distribute links that lead to dangerous software, phishing scams, and other types of harmful information. Because of this, consumers run the danger of having their personally identifiable information stolen, of having their devices infected, or of having their accounts hacked. Twitter is able to assist in the protection of its users' privacy and security by identifying and deleting spam.
3. **Encouraging honest competition**: Spam may be used to inflate artificially the number of followers and engagement metrics, giving some users an unfair edge over other users. Twitter is able to create healthy competition among its users and guarantee that its engagement metrics are accurate and dependable by eradicating spam from the platform.
4. **Preserving the integrity of the platform**: The credibility and integrity of the platform might be put at risk by spam since it lowers the overall quality of the material that is provided by users. By ensuring that users only encounter

material of the highest possible quality and authenticity, efficient spam detection contributes to the upkeep and preservation of the platform's integrity.

In general, the identification of spam on Twitter is vital for maintaining a high-quality user experience, protecting user privacy and security, promoting fair competition, and keeping the platform's integrity intact. Twitter is able to guarantee that users may interact with relevant and important material and that the platform continues to be a trusted and valued resource for its users by efficiently identifying and eliminating spam from its users' feeds. This is accomplished by deleting spam from the site.

1.4.3 Machine Learning and Deep learning for Twitter Spam Detection

Due to their massive user bases and potential for viral spread, social media sites like Twitter have emerged as a top target for spammers. Machine learning and deep learning methods have been used to detect and delete spam accounts and material in order to fight spam on Twitter. In this article, we'll talk about how machine learning and deep learning are used to identify Twitter spam.

Algorithms are used in machine learning to examine and spot patterns in data. Machine learning algorithms may be taught to recognise spam accounts and material on Twitter based on characteristics including tweet content, user behaviour, and network structure. For instance, machine learning algorithms may examine tweets' vocabulary to find terms and phrases that are often used in spam, including sales-oriented jargon or dubious links. Similarly, machine learning algorithms can analyze user behavior patterns, such as the frequency of tweets, retweets, and likes, to identify accounts that are likely to be spam.

Contrarily, deep learning is a branch of machine learning that uses artificial neural networks for data analysis and learning. Deep learning algorithms may be used to Twitter spam detection to find spam accounts and material by examining the format and content of tweets. Deep learning algorithms, for instance, may be used to analyse the language and graphics included in tweets and spot trends that are often linked to spam. Similar to this, deep learning algorithms may be used to Twitter's social network and patterns of account activity to identify accounts that are likely to be spam.

The ability to train machine learning and deep learning algorithms on big data sets of labelled data is one benefit of adopting these methods for Twitter spam identification. To train machine learning and deep learning algorithms to recognize and eliminate spam, for instance, datasets of well-known spam accounts and material might be employed. Additionally, when new forms of spam appear and spammers' strategies change over time, machine learning and deep learning methods may be utilised to constantly enhance the accuracy of spam detection.

The ability to detect spam accounts and material in real-time is another benefit of employing machine learning and deep learning for Twitter spam detection. This is significant because spammers often attempt to disseminate their communications widely before they are found and eliminated. Twitter can swiftly delete spam accounts and material by utilising machine learning and deep learning algorithms to analyse and detect spam in real-time. This prevents spam from spreading.

1.5 Motivation

Twitter has emerged as a prominent medium for communication, the dissemination of information, and business promotion. However, an increase in the number of people using the platform has also contributed to an increase in the amount of spam that is posted on Twitter. This may negatively impact the user experience, put user privacy and security at risk, and hurt the network's trustworthiness. As a direct consequence of this, there is an ever-increasing need for efficient spam detection strategies in order to preserve the honesty and reliability of Twitter.

In recent years, approaches including machine learning and deep learning have shown a great deal of promise in identifying and eliminating spam on Twitter. These methods are able to evaluate massive amounts of data, recognize patterns and trends, and learn from previous data in order to increase their accuracy over time. However, despite the significant amount of research that has been conducted in this field, there is still a great deal of work to be done in order to build efficient AI-based spam detection algorithms that are able to keep up with the ever-evolving strategies that spammers utilize.

Therefore, research on AI-based Twitter spam detection is necessary. It can help safeguard users from the harmful impacts of spam, encourage fair competition among users, and guarantee that the platform continues to be a trusted and important resource for communication and information exchange if we create and refine these strategies. In addition, the development of efficient AI-based spam detection technologies has crucial ramifications that extend beyond Twitter. This is because similar approaches may be used to other social media platforms and online communication channels in order to fight spam and safeguard users.

1.6 Problem Formulation

Spam on Twitter has become a big problem, which may negatively impact the user experience, put user privacy and security at risk, and undermine the trust of the network. Learning approaches such as machine learning and deep learning have shown some promise in identifying and eliminating spam on Twitter, but there is still a lot of work to be done to construct models that are both successful and efficient. The study will concentrate on identifying spam on Twitter via the use of machine learning and deep learning techniques. The investigation will focus on a variety of spam, such as account spam, content spam, and network spam. This research will contribute to the development of effective and efficient machine learning and deep learning models for Twitter spam detection. These models have the potential to help protect users from the negative effects of spam, promote fair competition among users, and ensure the integrity and trustworthiness of the platform.

1.7 Objectives

- To study and analyse various existing spam detection models and techniques for twitter datasets.
- To collect a dataset from twitter for spam detection.
- To preprocess the dataset using selective features which will reduce High dimensionality, Class Imbalances and Twitter spam drift.
- To design and implement the proposed framework for spam detection in twitter dataset using metaheuristic approaches.
- To validate and evaluate the proposed framework using standard metrics.

The selective features which reduce the high dimensionality, Class imbalances can be achieved through:

1. **Tweet Feature Extraction**: Tweet features and user account features are extracted from the input tweet. These features include metadata about the tweet, such as timestamps, user information, and other relevant attributes.

2. **Feature Selection with Whale Optimization Algorithm**: The extracted tweet and user account features are subjected to feature selection using the Whale Optimization Algorithm. This step aims to identify the most relevant and informative features for further analysis, improving the efficiency of the model.

   Spam detection in twitter can be achieved through the proposed model which involves the following:

3. **Text Embedding with GloVe**: The tweet text is processed using GloVe (Global Vectors for Word Representation) to convert the text into numerical vectors. This vectorization step captures the semantic meaning of words in the tweets.

4. **LSTM Deep Learning Model**: An LSTM (Long Short-Term Memory) deep learning model is employed to train the extracted features. LSTM, known for its effectiveness with sequence data, is used to analyse text data like tweets. The model is trained to learn patterns and relationships within the tweet features and GloVe representations.

5. **Integration of Modules**: Finally, the results from the feature selection module (Whale Optimization Algorithm) and the text analysis module (LSTM model) are combined. These combined modules work together to detect spam tweets effectively. Feature selection helps identify relevant features, and the LSTM model processes the tweet text to make predictions or classifications.

1.8 Thesis Organization

Chapter 1 - Introduction: In this chapter, the research topic is introduced, and an overview of the thesis is provided. The history and rise of social media, along with the

problems associated with it, are covered, with a specific focus on Twitter as the chosen social media platform. The prevalence of spam on Twitter and the necessity for spam detection employing machine learning and deep learning techniques are also discussed.

Chapter 2 - Literature Survey: The existing literature concerning automatic spam detection in social media is reviewed in this chapter, with particular emphasis on Twitter. The utilization of machine learning and deep learning for spam detection is explored, and various feature selection techniques are discussed. The foundation for the proposed models in the thesis is established within this chapter.

Chapter 3 - Tweet Spam Detection using Metaheuristic Features and Swarm Optimization Techniques: This chapter introduces the proposed model for Twitter spam detection, incorporating metaheuristic features and swarm optimization techniques. The model's components, including the Whale Optimization Algorithm, Stochastic Gradient Descent, and Adaboost Classifier, are discussed. Experimental results and conclusions drawn from this model are also presented.

Chapter 4 - GLoVe Language Model for Twitter Spam Detection using Bidirectional LSTM: In this chapter, another proposed model that utilizes GLoVe word embeddings and Bidirectional LSTM for Twitter spam detection is introduced. The model's architecture, experimental results, and a comparison with other methods, including CNN and feature-based approaches, are covered.

Chapter 5 - Conclusion: The final chapter of the thesis summarizes the key findings and contributions of the research. A conclusion to the study is provided, and the implications of the work are discussed. Additionally, future research directions in the field of Twitter spam detection are suggested.

Chapter 2

Literature Survey

The amount of material published on social media platforms has grown exponentially because of their rising popularity. With this growth, social media platforms have also become a fertile ground for spam and malicious activities, which can harm users and the credibility of the platforms themselves. Therefore, detecting and blocking spam on social media platforms has emerged as a crucial study field, garnering considerable interest from both academics and industry professionals.

In this literature review chapter, recent papers are explored that investigate different approaches for automatic spam detection in social media, with a focus on Twitter. Specifically, we will examine papers that use machine learning (ML) and deep learning (DL) techniques for spam detection, as well as papers that explore the use of feature selection to improve the accuracy of spam detection models. We will also examine papers that investigate the problem of spam drift, where spammers adapt their tactics over time to evade detection by spam filters.

Overall, the literature suggests that ML and DL techniques can be highly effective for detecting spam on social media platforms, with some approaches achieving high levels of accuracy. However, there remain challenges in dealing with the constantly evolving tactics of spammers, and more research is needed to improve the robustness of spam detection models over time. The literature also highlights the importance of feature selection in improving the accuracy of spam detection models, as well as the need for real-time detection and response mechanisms to combat spam drift.

2.1 Spam detection in social media using traditional techniques.

Sanjeev Rao et al [26] offered an informative guide to social spam, the technique of spamming, and the many classifications of social spam. The extensive study discusses several dimensionality reduction approaches that are used for feature selection and extraction, features that are utilised, as well as several machine learning and deep learning techniques that are utilised for the detection of social spam and spammers, along with the benefits and demerits of each methodology. Deepfake is a kind of text,

picture, and video spam that was made possible by artificial intelligence and deep learning; the defences against it are being investigated.

In combination with the self-attention mechanism, Rao et al. [27] used a variety of methodologies such as dataset balancing, sophisticated word embedding methods, machine learning, and deep learning approaches to improve the effectiveness of the social spam detection system. Improving the functioning of the system was the goal that they set for themselves. The datasets are standardized in the proposed framework by applying the Near Miss and Smote Tomek approaches to produce input for a variety of machine learning models. This input can then be used by the models. Because of this, the predicted accuracy of the models will be increased to its full potential. Following this step, the baseline machine learning models, as well as the ensemble models based on voting that are suggested by the research, are evaluated using both the unbalanced and balanced datasets. This is done in the second phase of the process.

Aljabri et al. [28] carried out a study with the purpose of compiling and analysing the most current developments in Machine Learning-based algorithms. This research was published in the journal Computers in Human Behaviour. These platforms are representative of a wide range of well-known social media websites. The authors provide a clear and simple summary of supervised, semi-supervised, and unsupervised methods, as well as in-depth information on the datasets that were made accessible to researchers. In addition to this, they carry out a comprehensive investigation into the many feature categories that are produced from the database.

Zineb Ellaky et al [29] focused to identify the most effective methods for the recognition of SMBs. The research that was published between 2008 and 2022 is covered by this SLR. Because of the findings of this investigation, the authors were able to categorise OSN profiles as either actual, verified, or bogus accounts. SMBs, spam bots, Sybil, and cyborgs, stegobots, political bots, and gaming bots are all examples of different forms of malevolent SMBs.

Goksu et al. [30] did research to determine the most recent publications on the systematic literature review approach for identifying false news in social networks.

This was accomplished by searching electronic resources that were regarded as being both thorough and reputable. In order to acquire a better understanding of how well these tools' function, the purpose of the research was to evaluate how successful they are in a variety of settings.

Verma et al. [31] presented a method that they named UCred (User Credibility) with the intention of determining whether user accounts are legitimate. To accomplish profile classification, the proposed model makes use of a mixture of three separate machine learning techniques. By using this strategy, the number of votes allotted to each categorization will be increased, which will result in an improvement to the system's accuracy.

Deep learning is a branch of artificial intelligence that is based on multi-layered artificial neural networks. Macas et al. [32] did research to investigate the possible uses of deep learning, which is a subfield of AI, in a variety of security-related activities. The examination carried out by the researchers yielded some really encouraging results. The first thing they do is talk about the underlying properties of certain typical deep learning architectures that are used in cybersecurity applications. In addition, they discuss the implications of these trends. They highlight the limits of the works that have been examined, and they provide a picture of the current issues that are being faced in the domain. In doing so, they offer helpful insights and best practises for academics and developers who are working on problems that are related.

Trivikram Muralidharan et al [33] introduced the first completely automated system for detecting fraudulent emails utilising deep ensemble learning to examine all segments of an email (the content, the header, and any attachments). As a result, there is no longer a requirement for human expert assistance in the feature engineering process. They show how this can be done by comparing the performance of the ensemble framework to the performance of individual deep learning classifiers.

Mohammed Ayub et al [34] conducted a comprehensive analysis of the published works on machine learning strategies is carried out to defend against DDoS assaults. Five search engines are utilised to locate research that are pertinent, the results are filtered based on certain selection criteria, and a total of 48 papers are ultimately

chosen for further examination. There are more than 20 different datasets that are utilised for training machine learning models, and the research shows that there are significant differences across the datasets that are employed. Most of the research have used the accuracy metric to carry out performance assessments. More than 30 different modelling methods were used throughout the construction of the ML models.

Exhaustive research was carried out by Hangloo et al. [35] that largely focused on deep learning (DL) approaches as the leading option for solving the problem of false news in online media. The investigation also took into consideration the relevance of multimodality in relation to this setting. In this study, we investigate a variety of DL frameworks, pre-trained model techniques, and transfer learning strategies, and then provide an in-depth analysis of each. However, because to the limited availability of multimodal datasets at the time this research was written, the emphasis of the study is placed on numerous data gathering approaches that may be used. The study throws light on various problems that have still to be addressed as well as obstacles that are related with this technique, with the goal of addressing and overcoming them.

Table 2. 1: Spam detection in social media using traditional techniques.

Author name	Methods Used in the paper	Merits	Demerits
Kornraphop Kawintiranon et al., [2022][36]	Pre-trained BERT model	This paper demonstrates that there is Twitter context-specific spam. Context-specific spam is included under a comprehensive taxonomy of conversation pollution.	Focused only on the content-based features
Peng et al., [2021][37]	Bi-LSTM with ALBERT	The self-attention Bi-LSTM neural network model in conjunction with ALBERT, a lightweight word vector model of BERT, powers the spam detection technique.	The model requires more computational time and resources due to the addition of the self-attention mechanism.
Al-Zoubi et al.,[2021][38]	Naïve Bayes, Decision Trees, MLP,KNN,RF	This paper focused on constructing an effective spam detection algorithm by extracting a huge range of public information from	No proper information on the selected features for classification

		Twitter profiles.	
Gadiraju et al., [2018][39]	Combination of multiple machine learning algorithms Like Random Forest, SVM, Decision tree, Naïve Bayes	The detection of spam accounts on Twitter is done using an innovative approach based on deep learning technology. One advantage of these techniques is that, in contrast to typical machine learning algorithms.	Focused only on the Graph based features

2.2 Spam detection in twitter using ML based approaches.

María Novo-Lourés et al [40] explained how different features can be used to supplement synset-based and bag-of-words models of texts when using traditional ML methods to filter spam. Even though there are a lot of traits that go together, to make this study more useful, the authors chose only those that can be calculated no matter what communication method is used to send information.

Rahul A. Patil et al [41]suggested the methods for identifying Twitter spammers. Additionally, the methods used by Twitter to separate spam are ranked according on how well they can identify fake data, a URL, and spam patterns.

Saud Alshammari et al [42] conducted experimental research with the use of machine learning algorithms to determine whether the tweet in question is spam. The Bayes theorem, which is a probabilistic theory that was presented by Naive Bayes, may be used to accomplish this goal. The information is taken from the KAGGLE website, which has both spam and authentic tweets inside its database. The information that has been pre-processed by standard articulations with the purpose of excluding information that is unwanted. Applying each of the many techniques for arranging things to the information will result in the element being transformed into a vector. When converting text into vector form, a tool known as a Term Frequency Inverse Document Frequency (TF-IDF) vectorizer will be used.

Kübra Nur Güngör et al [43] provided an approach to the identification of spam. It was determined to use the Naive Bayes, J48, along with Logistic machine learning algorithms.

Using a variety of machine learning and deep learning strategies, Dalia Alsaffar, and colleagues [44] undertook a research study with the objective of identifying whether or not a tweet may be categorized as spam. Seven different machine learning algorithms as well as one deep learning approach known as Recurrent Neural Network (RNN), were put to the test in this research to see which performed the best. The assessment consisted of carrying out a variety of experiments, including cross-validation and percentage split tests.

Marouane Kihal et al [45] introduced a new deep multimodal decision-level fusion system that has the potential to successfully identify spam in multimedia formats. Feature extraction and selection are handled by CNN, which are used by the method that the authors have suggested. To get an accurate representation of the information, the recovered characteristics are sorted and organised into three separate vectors known as visual, textual, and audio (VTA) vectors.

Hamdy Mubarak et al [46] presented a big collection of Arabic tweets that have been carefully tagged with information about advertisements (Spam). The authors do an analysis on the properties of these tweets that set them apart from other tweets, and they determine the subjects and targets of these tweets. In addition to this, they do research on the characteristics of spam accounts.

K. R. Vidya Kumari et al [47] utilizing machine learning, divide the tweets into spam as well as non-spam categories and determine which categories provide the best results.

Research was carried out by Nour El-Mawass and colleagues [48] to investigate whether it would be possible to make use of previously suggested supervised classification methods in order to detect spammers. These algorithms have a notable capacity to identify spammers on a constant basis, even though their memories are not flawless. This assumption serves as the foundation for the key argument that will be presented throughout the study. To accomplish this goal, the researchers devised an analysis tool that is known as a Markov Random Field. This tool focuses on a network of users that have a few characteristics in common. They established their previous beliefs by applying a wide variety of creative classifiers from a variety of sources.

They used a method known as Loopy Belief Propagation to produce posterior predictions about the users.

Research that demonstrated the presence of context-specific spam as well as its detectability was carried out by KornraphopKawintiranon and colleagues [49]using a variety of datasets obtained from Twitter. The purpose of this study was to investigate which model is more successful than others at recognizing spam that has both generic and context-specific components.

Somya Ranjan Sahoo et al [50] discussed the technology, which is built on an extension for the Chrome web browser and can identify bogus accounts in the Twitter environment by analysing several features. The goal is to identify the phoney account by doing research into the many features that are responsible for the dissemination of dangerous information in a real-time setting. The creation of a fake profile involves stealing the identity of a real user to steal their profile information and then recreating the profile using their credentials. In a subsequent step, the profile is corrupted to cast aspersions on the real owner of the profile while also sending a friend request to the user's buddy.

The problem of identifying spam on Twitter was addressed by Abdullah M. Alkadri and colleagues [51], who offered an integrated approach to solve the problem. Their strategy overcomes the unique obstacles presented by the identification of Arabic spam. They use word embedding methods and include pre-trained word embedding vectors in order to improve the data. Several different types of machine learning techniques are applied in the process of identifying spam. In order to illustrate the usefulness and practicability of their suggested technique, the researchers compiled and annotated a real-world dataset consisting of Arabic tweets.

Saksham Gupta et al [52] provided an in-depth analysis of the various approaches to spam filtering that make use of machine learning techniques. On a dataset consisting of tweets from Twitter, postings from Facebook, and comments made on YouTube, the spam filtering techniques. In addition, a comprehensive analysis of each approach has been offered in this study's accompanying discussion.

M. Ghiassi et al [53] proposes an unsupervised method for text categorization that is very easy to use across different problem domains and offers accuracy on par with or better than existing options.

Alok Kumar et al [54]provided a distributed, decentralised, and unsupervised method for locating and removing spam from social networks. The authors describe a novel approach that can identify spams from a single message stream and is based on fuzzy logic. They use the method to operate on the MapReduce platform to manage massive amounts of data in networks.

In their work, E. Elakkiya and colleagues [55] employed a feed-forward neural network to detect spam in complex data. To enhance the accuracy of the model and its training process, they focused on fine-tuning various hyperparameters, including the learning rate, momentum term, neural network architecture, activation function, training technique, weight initialization ranges, and initial weight tuning. The study introduces reinforcement learning and k-Norm factor-based shuffling frog leaping algorithms as potential approaches for determining the optimal parameter combination for the neural network. These methods were explored due to the lack of a comprehensive and specialized solution tailored specifically for this task.

Minyoung Lee et al [56] focused for quickly identifying vishing. Due to a lack of research on spam detection using low-resource languages, the authors use simple machine-learning models to identify phishing in the Korean language. In order to identify spam using natural language processing methods, they transformed the audio recordings from real vishing damage data into text. Instead of developing models, the main goal is to see whether vishing can be quickly recognised.

Deepali Dhaka et al [57] examined approaches for detecting spam across several domains, including email and online spam, social spam, opinion spam, and comparisons of these types of spam. This is an effort to present a variety of different difficulties in this field. This is the first comprehensive literature analysis that has been conducted in the topic of cross-domain spam detection, as far as the knowledge and understanding go.

Tabassum Gull Jan et al [58] offered a quick method for creating labelled datasets so that time may be saved, and mistakes made by humans can be avoided. For improved efficiency and broader application, the suggested strategy depends on instantly accessible features. With the use of Twitter streaming, this effort intends to compile a user's most recent tweets and create a recent Twitter dataset.

The research conducted by Vimala Balakrishnan and colleagues [59] focused on the development of an automated cyberbullying detection system that makes use of psychological features of Twitter users. These qualities include personalities, moods, and emotions. Big Five and Dark Triad models were used to do personality analysis on users. The #Gamergate hashtag was used to collect 5453 tweets, which were then carefully annotated by professionals by hand in the Twitter dataset. The baseline algorithm employed a selection of Twitter-based characteristics, including text, user, and network information.

Sarra Ouni and colleagues [60] introduced a novel framework that aims to combine contextual BERT embeddings with subject-based features. The final feature vector is generated, and it is subsequently utilized as input for the supervised classifier to perform classification.

Radwa M.K. Saeed and colleagues [61] proposed four distinct methods for detecting Arabic spam reviews, with a particular emphasis on developing and evaluating an ensemble approach. These methods are designed for the identification of Arabic spam reviews. This approach also incorporates content-based aspects to enhance the detection process.

C. Vanmathi et al [62] employed the Naive Bayes algorithm, a supervised learning approach, to ban the users. It is also possible to analyse the users who have been barred each month, which will aid in the research of users and rumour information.

Somya Ranjan Sahoo and colleagues [63] devised a real-time system for detecting the content of spam messages based on behavioural analysis by combining various machine learning strategies with the genetic algorithm. The research aims to provide distinct features based on profiles and content of spam messages to facilitate spam identification. The process begins by structuring the task around social networking

sites' anti-spam regulations. Next, data is collected from multiple social networks like Facebook, Twitter, and Instagram to create a dataset containing both spam and non-spam accounts. Genetic algorithm and other classifiers are employed to generate appropriate feature selections.

Oğuzhan Çıtlak et al. [64] investigated notable spam detection algorithms, analyzing their strengths and weaknesses and how they differentiate genuine users from fraudulent ones.

Woo Hyun Park et al. [65] proposed a spam detection technique based on natural language processing. This approach utilizes a least-squares model for topic modification and a gradient-descent with altering-least-squares (AMALS) model to address missing data using TF-IDF and uniform distribution.

Rasheed G. Jimoh et al [66] recommended the use of unigrams and bigrams to detect spam in brief communications. The effectiveness of these suggested features was evaluated using four categorization approaches.

Ramesh Paudel et al. [67] introduced a graph-based strategy to identify potential instances of spam. This method leverages the connections between mentioned entities in tweets and the documents addressed by the URLs in those tweets. By combining multiple data types into a single graph, the authors aim to detect distinctive patterns that reflect fraudulent activities, patterns that are difficult for spammers to replicate.

Table 2. 2: Spam detection using Machine Learning based approaches

Author name	**Methods Used in the paper**	**Merits**	**Demerits**
Lipas Das et al[2022][68]	Various methodologies in Machine Learning SVM, DT, Logistic regression etc.,	The effectiveness and characteristics of Twitter spam detection are reviewed in this article along with a summary of the advantages and disadvantages of each approach.	This paper provided a review on various ML based techniques but failed to ensure the demerits associated with each methodology
Anisha P Rodrigues et al[2022][69]	Stochastic gradient descent, support vector machine, logistic regression	This paper demonstrated that the features taken from the tweets may be used to reliably determine if a given tweet is spam or not, as well	Discrimination amid spam accounts exploiting other interactions functions should be investigated

		as to build a learning model that can correlate tweets with specific sentiments.	further
Sundararaja and Palanisamy, [2020][70]	Random Forest, Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting, AdaBoost, Logistic Regression, and Decision Tree.	Prediction on user's mood influencing the sarcasm and vice versa is done. Tweets before and after specific sarcastic kinds are attained. Thereby modelling the user emotion change through past tweet histories collection.	Users affected with their mood levels on the basis of sarcasm
Chen et al [2018][71]	Semi-Supervised Clue Fusion (SSCF)-SSCF acquires a linear weighted function	Identify spammers with increased detection rate via multiple aspects, such as content, behavior, relationship, and interaction	Small size of primarily labeled instances
Alsaffar et al [2019][72]	Random Forest (RF), Naive Bayes (NB), Bayesian Network (BN), SVM, KNN, and Multi-Layer Perceptron (MLP) and Recurrent Neural Network (RNN)	Improved outcomes with minimal error rate and highest classification accuracy rate by means of RF	Higher computation on Twitter spam detection

Muhammad Adeel Abid and his team [73] developed a system that employs supervised machine learning methods to distinguish between spam and ham SMS messages. Techniques such as TF-IDF and bag-of-words are utilized to extract

features from the data. To address the dataset's class imbalance, over- and under-sampling techniques are applied. The performance of the models is evaluated using accuracy, precision, recall, and F1 score metrics on the SMS dataset.

Haoyu Wang et al [74] conducted tests using Bayesian linear regression and decision forest regression methods on a dataset obtained from the UCI Machine Learning Repository. The authors assess the quantitative data to select a better prediction technique and employ the trained models to determine if a letter is spam.

2.3 Spam detection with feature selection

Rozita Talaei Pashiri et al [75] utilized the sine-cosine algorithm (SCA) to develop a feature selection-based strategy in this study that lowers the spam detection error. In the suggested approach, the SCA updates the feature vectors to choose the best features for instructing the ANN.

Aliaksandr Barushka et al [76] proposed an innovative plan for screening spam on social networking sites while keeping costs in mind. The strategy that has been suggested can be broken down into two steps. This is accomplished by reducing the number of characteristics that are required for spam filtering. After then, the strategy makes use of cost-sensitive algorithms for ensemble learning, with regularised deep neural networks serving as the basis learners.

FaezeAsdaghi et al [77] presented a novel method known as backward elimination, for the purpose of feature selection. This approach is similar to the sequential backward selection in that its primary objective is to evaluate the effect of removing a group of features rather than a single feature in order to determine how it affects the overall performance of a classifier. This approach searches for the biggest feature subset possible, with the goal of excluding those features from the whole set of features in such a way that it not only lowers the classification accuracy but also raises it.

Poria Pirozmand et al. [78] introduced an innovative approach for identifying spam across various social networks. Their method involved enhancing a Support Vector Machine (SVM) using a combination of the Genetic Algorithm (GA) and the Gravitational Emulation Local Search Algorithm (GELS) to select the most relevant spam features.

E. Elakkiya et al. [79] proposed a novel multi-evaluation method that combines feature group selection with the evolutionary algorithm called GAMEFEST. The effectiveness of this approach was evaluated by utilizing data from Twitter, Apontador, and YouTube to assess its performance.

Aakanksha Sharaff et al [80] provided a classification algorithm and feature selection methods. This method improves accuracy and allows us to choose better

characteristics. Using feature selection approaches, the redundant and unnecessary features that do not improve the model's accuracy are eliminated. As fewer features are sought, the model's complexity is decreased, and the easier-to-understand simplified model is more intuitive.

An alternative method was presented by M. Salih Karakaşl et al [81] to categorize spam users on Twitter. This method does not depend on a predetermined set of attributes and instead makes use of methods that include machine learning. They proposed grouping users who have similarities and using a dynamic feature selection process that considers several characteristics that are unique to each user group rather than using a static feature set. This method would integrate many characteristics that are unique to each user group.

An automated system that was provided by Saleh Beyt Sheikh Ahmad et al [82] was mainly created for the purpose of recognizing spam tweets. This strategy places an emphasis on the extraction of features and the preprocessing of data, considering the one-of-a-kind quality of tweets. To do an accurate analysis of the issue, the preprocessing phase is absolutely necessary. Following preprocessing, just the text content of each tweet is stored, which makes it much simpler to determine whether or not a tweet should be considered spam.

V. Sri Vinitha et al [83] addressed the issue of email spam detection by exploring a range of feature selection techniques. By performing feature selection before classification, the authors aimed to enhance the effectiveness of spam filtering and improve efficiency.

Hossam Faris et al [84] suggested an intelligent system for detecting email spam that is based on the Genetic Algorithm (GA) and the Random Weight Network (RWN). The suggested system also has an automatic recognition feature that helps find the most important traits during the discovery process. Three large collections of emails are used to test the proposed method in several detailed studies.

Table 2.3: Spam detection based on features selection.

Author name	**Methods Used in the paper**	**Merits**	**Demerits**
Poria Pirozmand et al., [2023][85]	SVM, Genetic algorithms, Gravitational Emulation Local Search Algorithm	Exploits complex features existing in high-dimensional data on social network spam	poor capacity to deal with high-dimensional datasets
Zhao et al [2020][86]	heterogeneous stacking-based ensemble learning framework	meta classifier with the individual errors of classifiers from the previous stage for any biased behaviour detection mitigated imbalanced class distributions influence on classification performances	Increased time complexity
Chiew et al [2019][87]	Hybrid Ensemble Feature Selection (HEFS). In the first phase of HEFS, a novel Cumulative Distribution Functio n gradient (CDF-g) algorithm, perturbation ensemble is utilized	It is exploited to produce primary feature subsets	Not adaptable to different datasets with significant performance gain
Faeze Asdaghi et al.,[88]	Naïve Bayes Classifier	Evaluates the effect of removing a group of features rather than a single feature, which is comparable to sequential backward selection, on the classifier's performance.	Poor capacity to deal with the real time applications rather than Webspam

2.4 Spam detection in twitter using DL based approaches.

Zulfikar Alom et al [89] demonstrated a fresh method that makes use of deep learning (DL) methodologies. The method for identifying spammers makes use of both the content of tweets and the meta-data associated with users (such as the age of an account, the number of people it follows and people it is followed by, and so on).

Akhil Pratap Singh et al [90] suggested an email spam detection method that uses the idea of deep learning to find junk emails. Compared to other ML methods, it has a higher chance of correct identification.

A hybrid modulated approach was developed by Chanchal Kumar and colleagues [91] for the purpose of identifying spam on Twitter. Their method consisted of using the SMOTE-ENN sampling algorithm in order to identify whether or not a given tweet constitutes spam. They were able to create balanced data for input into a variety of deep learning classification methods thanks to the combination of SMOTE and Edited Nearest Neighbours (ENN).

Deep learning as well as more conventional approaches to machine learning were used by Sanaa Kaddoura et al. [92] to categorize Arabic tweets according to a variety of criteria. To produce a trustworthy dataset, they used hand tagging on a tweet corpus that was obtained from the Twitter API. The dataset underwent feature extraction, and N-gram models in the form of uni-grams, bi-grams, and char-grams were applied in accordance with the various feature extraction strategies. The dataset was extended using a method called synthetic minority oversampling to solve the class imbalance that was found in the data.

Jenifer Darling A Multi-Objective Genetic Algorithm and a CNN-based Deep Learning Architectural Scheme were suggested by Rosita P et al. [93] as the major approach for identifying spam on Twitter. This method is referred to as MOGA–CNN–DLAS.

An innovative deep learning architecture for the identification of spam was developed by Gauri Jain and her colleagues [94] and is based on CNN and LSTM. WordNet and ConceptNet, both of which are types of knowledge bases, were deployed to improve the representation of individual words inside the database. These knowledge bases provided more accurate semantic vector representations for the test words, which allowed the model's performance to be enhanced as a result. To build a structural context representation, CNN and BiLSTM were used. This representation included both global semantic dependence traits and local semantic features.

TextSpamDetector is the name given to the approach that was suggested by E. Elakkiya and colleagues [95] to identify spam at the text level using deep learning. To combat attention drift, it made use of a conjoint attention mechanism, in addition to the standard attention mechanism and the context preserving attention mechanism. The attention processes focused on the representations of the context and the words that provided relevant information within the input text. CNN and BiLSTM were used to build a structural context representation that included global semantic dependence features. This was accomplished by including them.

Loukas Ilias and his colleagues [96] created two innovative ways to discriminate between authentic users and bots that are based on Natural Language Processing (NLP). The first approach, which employed feature extraction, was used to identify accounts that were publishing automated messages, and the second method, which used machine learning algorithms, followed it. The second approach consisted of using a deep learning architecture that was coupled with an attention mechanism to differentiate between tweets that were published by real people and those that were produced by bots.

For identifying spam on social networks, Razan Ghanem, and colleagues [97] suggested a deep learning architecture that they referred to as CBLSTM (Contextualised Bi-directional Long Short Term Memory neural network). This design made use of language model embedding and was constructed using bidirectional long short-term neural networks.

For spam categorization, Gauri Jain, and her colleagues [98] used deep learning, more especially the LSTM component of the Recursive Neural Network (RNN). This technique learns abstract feature representations as opposed to manually generating features.

Zhiwei Guo and his colleagues [99] built a model for the identification of spammers using Deep Graph neural networks and gave it the name DeG-Spam. To build a framework for graph neural networks, the model considered both occasional relations and intrinsic links in a distinct manner. This resulted in the production of feature expressions for the social graph. When compared to more conventional methods, the

accuracy of spammer identification has been significantly enhanced because to the mining of new feature components.

[100] Vanyashree Mardi and her coworkers suggested a technique for recognizing tweets containing text that were classified as spam using Naive Bayes Classification in conjunction with an artificial neural network. According to the findings of a performance investigation, Artificial Neural Networks are superior to the Naive Bayes Classification method in terms of accuracy.

For determining the value of photographs, Aaisha Makkar, and her colleagues [101] developed a framework that they named PROTECTOR. This framework merged textual information, connecting information, and metadata information that was linked with the photos. By comparing this information with other data pertaining to the picture, a rank score was produced for it.

Md. Rafiqul Islam and his colleagues [102] carried out an exhaustive study of automated misinformation detection (MID), which considered incorrect information, rumors, spam, fake news, and disinformation. Deep learning (DL) was used to provide improved outcomes and scalability in real-world MID applications by automatically analyzing and extracting global information. They also brought attention to the difficulties as well as the opportunities for future advancement in the sector.

Deep learning was used by Akrivi Krouska and colleagues [103] to categorize the degree to which tweets are positive or negative. The classification challenge made use of a total of four pre-trained word vectors, namely Word2Vec, Crawl GloVe, Twitter GloVe, and FastText.

For doing high-dimensional data analysis in a platform environment that is representative of the actual world, Merly Thomas and his colleagues [104] suggested using a Deep Neuro Fuzzy Network (DNFN) that was built on Chimp Sailfish Optimization (ChSO). The approach that was suggested exhibited excellent dependability, resulted in better results, and significantly decreased the complexity of the computing process.

S. Sumathi et al [105] presented Random Forest and connected it with Deep Neural network in order to find out how accurate the categorization is. When building its decision trees, the Random Forest method applies a probability that has been determined in advance to each characteristic. When ranking the key characteristics, the Gini measure comes into play. This needs all of the training data to be learned at the same time. A dynamic adjustment was made to the detector process in order to accommodate the newly discovered data patterns up until it approaches the spam coverage.

Sunita Dhavale et al [106] suggested C-ASFT (CNN-based Anti-Spam Filtering Technique) is a server-side CNN-based solution. For efficient text-based spam identification and filtering, C-ASFT utilises a three-tiered one-dimensional CNN layer model. The spatial structure or invariant properties contained in the word order in the input mail text data are learned using one-dimensional CNN layers. The email may be classified as spam or not spam by the email server, giving the client the option of reading or deleting it.

Insaf Kraidia et al [107] suggested using a Deep Learning (DL) system to categorise various types of harmful tweets. This allows us to assure the effective filtering of spam that may be buried in either text or images. After that, a fusion model is applied to the data to determine whether the tweet contains harmful content.

The research that was carried out by Aditya Anil and colleagues [108] included a comprehensive analysis of the performance of many different machine learning and deep learning models when integrated with natural language processing strategies. The study article investigates a variety of ways that may properly identify spam and finds the approaches that have shown to be the most effective in reaching the intended goal. In this research, a wide variety of datasets, such as emails, SMS messages, and tweets, are analyzed using a few different algorithms. According to the data, random forest had the most accuracy in recognizing spam included within tweets, while deep learning models got the highest accuracy in recognizing spam contained inside SMS messages and emails.

Atheer S. Alhassun et al [109] gathered an Arabic dataset that could be used for spam identification and used it to solve the problem of identifying spam accounts using Arabic on Twitter. The dataset included information from Twitter's premium features and was compiled with the help of Twitter's premium API. The labelling of the data was done by marking suspended accounts as active. A combination framework that is based on deep-learning approaches has been presented. This framework has various benefits, including the ability to provide more accurate and quicker results while using less computer resources. The authors made use of two distinct kinds of data: text-based data, which was modelled using convolution neural networks (CNN), and metadata, which was modelled using basic neural networks. Accounts were either categorised as spam or not spam based on the combined output of the two algorithms.

Zhiming Xu et al [110] conducted research on the unique topic of modelling and integrating human knowledge of various forms of network anomalies in order to discover attributed network anomalies. To be more specific, the authors begin by modelling past human knowledge by using an innovative data augmentation technique. After that, they make use of a carefully crafted contrastive loss to incorporate the modelled information into the encoder of a Siamese graph neural network. In the end, they train a decoder to rebuild the original networks based on the node representations that were learnt by the encoder. As the anomalous metric, they score nodes according to the reconstruction error that the decoder generates.

Researchers Sepideh Bazzaz Abkenar et al. [111] performed research with the aim of increasing the percentage of spam that could be detected in actual Twitter datasets. They suggested an approach that is a combination of two different techniques, namely the Synthetic Minority Over-sampling Technique (SMOTE) and the Differential Evolution (DE) methodologies. DE is used to optimize the hyperparameters of Random Forest (RF), while SMOTE is used to address the uneven class distribution in the datasets, which eventually improves the accuracy of classification. Together, these two methods are referred to as data engineering.

Kangyang Chen and colleagues [112] created a sophisticated model for the detection of spam that they named deep cascade forest. This deep learning model, as opposed to approaches that use backpropagation, simplifies the management of training costs

because to the lower amount of hyperparameters that it utilizes. The simplified nature of the model provides benefits to the effective operation of the training process.

An intelligent algorithm that can identify between phishing messages and authentic communications was developed by Aakanksha Sharaff and her colleagues [113]. They used regular expression (Regex), machine learning (ML), and deep learning (DL) models in tandem with one another. The spam messages included inside the dataset were used to develop Regex rules, which ultimately resulted in the improvement of the dataset.

Ankit Kumar Jain et al [114] provided a method for the detection of communications that are spam. The authors have discovered an efficient collection of features for text messages that can accurately categorise messages as spam or ham. These features may be applied to text messages. To produce a feature vector for each normalised text message, the technique for selecting features is applied to the normalised text messages first. To evaluate the performance of the various machine learning methods, the produced feature vector is put through its paces. In addition to that, this work offers a comparative examination of several methods, all of which are used to implement the characteristics. In addition to that, it discusses the roles played by a variety of characteristics in the identification of spam. The Artificial Neural Network Algorithm that makes use of the Back Propagation approach operates in the most efficient way once it has been implemented and according to the set of characteristics that have been chosen.

Ashish Singh et al [115] presented LSTM, a kind of deep learning method, to detect the subject matter of bogus reviews. The combination of these two techniques results in a detection rate that is more accurate for opinion spam in comparison to other models that are currently in use. The dataset known as the "Deceptive Opinion Spam Corpus v1.4" is utilised for the benchmark.

Carlos Lago et al [116] focused on conducting research on the potential applications of deep learning methods to three distinct challenges faced by the field of cybersecurity: filtering SPAM, detecting malware, and identifying adult material to demonstrate the advantages of using such methods. The authors put a broad range of

methods to the test, including image augmentation techniques, LSTMs for spam filtering, DNNs for malware identification, and lastly CNNs in conjunction with Transfer Learning for adult content detection. In addition, they employed picture augmentation methods to enrich the dataset.

Geetanjali Sharma and colleagues [117] investigated on the numerous machine learning. This investigation was accomplished using a process known as a systematic review. This comprehensive evaluation will improve the planning and execution of a fresh, effective method for automatically identifying and removing objectionable or abusive content from user messages and posts. To prevent the spread of hatred and harassment via social media, this in-depth examination of the available strategies will also be of great service to individuals, society, the government, and social platforms.

Donia Gamal et al. [118] presented a novel deep learning architecture with the intention of identifying the strength of emotions in four independent binary balanced Arabic datasets of varying sizes. This was done to achieve their goal. The proposed framework incorporates five distinct types of deep neural networks to address the challenges that are associated with Arabic sentiment analysis. This all-encompassing methodology tries to address some of the shortcomings of the Arabic sentiment analysis approaches that are already in use.

Dipalee Borse et al [119] performed a poll, and its findings have been split into three parts, such as detecting spam, detecting spam in real time, and detecting spammers. The authors also talked about how different Twitter functions are used to find trash, how well they work, and what problems they pose for current study.

To tackle the evolving nature of spam content and emerging spamming patterns, Mahdi Washha and his colleagues [120] devised a framework called "spam drift." This framework utilizes unsupervised machine learning to update a real-time supervised spam detection model at the tweet level in batch mode. By learning from unlabeled tweets, the system adapts to changing spam characteristics and patterns.

Recently, Darshika Koggalahewa and colleagues [121] introduced a hierarchical test-based strategy for detecting spam drift over time. The system autonomously acquires features without explicit instructions and uses the difference in feature similarity, KL

divergence, and Peer Acceptability to identify and confirm changes in spam user behavior in real time. Through drift detection, the system continuously updates its learning model and provides users with up-to-date categorization.

Reem Alharthi and her colleagues [122] proposed a fine-grained real-time categorization technique specifically designed to identify various types of low-quality Arabic tweets, including promotional tweets, phishing tweets, and spam tweets. Instead of relying on manually engineered features, the system leverages deep learning algorithms to automatically extract textual features, eliminating the need for time-consuming and specialized characteristics. Additionally, the researchers proposed a straightforward method for real-time identification of spamming Twitter accounts using a selected set of textual qualities.

Monal R. Torney et al [123] examined the performance of the outcomes produced by employing different datasets. The authors attempt to analyse the appropriate domain datasets that would provide the best results after using different approaches, strategies, and algorithms by comparing the results and performance.

Table 2.4: Spam detection using Deep learning techniques.

Liang and Yan [2019][124]	Deep Bidirectional LSTM model	Classified malicious domains based on lexical features for comparison.	Higher computation on malicious URLs detection
Le et al [2018][125]	Convolutional Neural Networks to both characters along with URL String words for URL embedding learning in a joint optimized framework.	Numerous rare words exist here. Various URLNet. components need to be examined. No expert features is necessitated	Very fast approach (necessitating a basic database lookup), low False Positive rates
Madisetty and Desarkar [2018][126]	CNNs, feature-based model uses content-based, user-based,	consistent spam detection techniques for	In the future, trained word embeddings such as glove, fast

	and n-gram features	spam detection at tweet level	Text, and deep learning models like LSTM, RNN, GRU can be evaluated for classifying praises and complaints.
Kudugunta, and Ferrara [2018][127]	Long-Term Short-Term Memory (LSTM) architecture	Model can achieve an extremely high accuracy exceeding 96% AUC. a near perfect user-level detection accuracy (> 99% AUC)	Complexity lies in massive amounts of posted data labelling which may be erroneous. It is not suitable for practical applications
Abdi and Wenjuan [2017][128]	Convolutional Neural Network (CNN)	Gives improved detection rate for malicious URL detection	This doesn't depend on the features it becomes difficult to apply them to current social networks

2.5 Research Gaps

Spam detection in social networks is highly necessary due to several concerns, including user privacy security, public opinion research, network environment security, etc. Prior studies have utilized blacklists and crowdsourced data to identify anomalous accounts in addition to filtering for the purpose of maintaining social network security via spam identification.

More than 90% of users click on malicious links before blocking via blacklisting takes place. However, because active information identification requires personal participation, these methods are seen as time-consuming. Numerous researchers use graph analysis-based techniques to extract features from social graph structures based on follower and follower interactions through node similarity, which significantly improves detection performances. However, a lot of spammers employ artificial intelligence technology to mimic the social interactions of regular users and fabricate

their link relationships. Effectively detecting illegal accounts so becomes a very difficult process.

A strategy based on Deep Learning (DL) is presented for Twitter spammer detection. Nevertheless, because of the large dimensionality issue, the current DL-based approach is less accurate. Further research is necessary to identify features, which have a bigger impact on the accuracy of spam detection. Even though maintaining an acceptable feature set size for the purpose of verifying prediction efficiency is regarded as a non-trivial activity, it is imperative that feature selection and decision-making be done to ensure prediction proficiency for model training. Prior studies have shown that for any given task, DL-based algorithms do not perform much better than ensemble learning techniques. Improved classification performances can be obtained by training several classifiers. Ensemble learning, which includes both homogeneous and heterogeneous methods, performs better than single classifiers.

To achieve improved performance, heterogeneous ensemble learning makes use of many basis classifier kinds, while homogeneous ensemble learning relies on one type numerous classifier instances. Thus, heterogeneous ensemble learning is chosen as the common base for spam identification.

Chapter-3

Tweet spam detection using metaheuristic features and swarm optimization techniques.

3.1 Introduction

Spam is a phenomenon that emerged after the creation of internet. It is also no secret that it continues to be a substantial cause of disruption and a barrier to productivity. Spam mail may negatively affect practically everything in addition to making it more difficult to explore your inbox. In a similar vein, it can have an impact on a variety of variables, including the financial performance of an industry and the development and uptake of technical or scientific concepts. Regardless of all the other things that may be done online, spammers only care about utilising the internet as a special means of making other people's life more difficult.

In its most basic form, spam refers to unsolicited commercial emails that have jammed your inbox for no apparent reason. To have a better grasp on what it is, however, you need realise that it refers to online material that has not been requested and is often sent in large quantities from anonymous or unknown sources for the aim of advertising, phishing, spreading viruses, and other similar activities. They often arrive in the form of junk mail. However, it is nothing out of the usual to come across spam communications sent via instant messaging (IM) services, text messages (SMS), recorded phone calls, or social networking websites. Additionally, not only can spam messages waste your time, but they also run the risk of infecting your device with a virus and, in many cases, use a significant amount of internet bandwidth.

In any case, it's an interesting titbit to know that the original form of spam was the dish known as "SPAM – the conserved meat product made from gammon," which was very well-liked back in the day. The tins of SPAM might occupy significant amounts of space at practically every shop that you went to. This is the origin of the term "spam," which refers to unsolicited messages sent through the internet. Let me clarify. When web-based mass messaging first started to become popular, someone, somewhere in an online forum, coined the term "spam." The canned meal known as

SPAM has become so famous that the word is now often used to refer to unwanted mass internet advertising material, also known as junk mail. This is because SPAM was originally a brand name for the canned food. The moniker has endured, and it has, up to this point, become even more well-known than the culinary product SPAM itself.

Most of the spam is annoying and wastes a lot of time, but some types of spam may really be rather hazardous to deal with. Email scams often include an attempt to trick you into divulging your banking information so that the con artists may either steal your identity or take money from your account.

Phishing scams and advanced fee fraud are examples of these mails. Keep an eye out for:

- something that provides you a benefit without cost.
- anything that seems to be asking you for money information.
- anything pertaining to your accounts that have embedded links to follow.
- Anything that requests your secrecy.

The letter is obviously a fake, but other forgeries like this one may be difficult to spot without careful examination of the mail headers. Most people have a hard time understanding mail headers, and many email clients make it much more difficult to see them.

If there is an issue with your account, and you need to confirm your information to prevent the account from expiring.

- due to the discovery of suspicious activities, your account has been frozen.
- They want you to join up for a new service they are providing.
- a transaction on your account has been refused because it has to be verified.
- there has been an issue with your shipment, and you must log in to examine the specifics; and so on.

The link in the message may be clicked to access the website. However, it should be noted that the link will direct the user to a fraudulent website operated by con artists rather than the official website of the institution. It can be difficult to determine the

URL that the link leads to, and some email clients may not display this information properly. Consequently, the user may be directed to a different site than the one claimed by the client. Some of these fake websites may appear very convincing. Alternatively, the email may request the user to provide their login information in the reply, which will be sent to an external email account that the thief can access.

3.1.1 Twitter Spam

Twitter is just one example of a platform that has benefited immensely from the recent explosion in popularity of microblogging. As a result of this development, businesses and media outlets are looking for more methods to make use of Twitter to gather information on how users perceive the products and services they provide. This is happening as a direct consequence of the growth that has occurred. Due to the shorter character constraints of microblogging and informal language, there has been far less study conducted on the ways in which feelings are conveyed. In recent years, many companies have exploited data from Twitter and have achieved tremendous upside potential for firms going into a variety of areas. On the other hand, spambots and fraudsters have been actively flooding Twitter with dangerous links and fraudulent material, which has resulted in legitimate users being misled because of this activity.

The number of people who utilise various social networking platforms has been steadily rising over the last several years. The functionalities of journals, bulletin boards, and email are only some of the ways in which social networking services digitise contacts with other people. Users are increasingly finding that Online Social Networks (OSNs) are becoming essential communication tools for their day-to-day lives. Users sign up for accounts on social media platforms to communicate with their friends, family, and other people who are important to them by posting messages, sharing photographs and videos, expressing their opinions, and spreading the news. Users of social media platforms can engage in discussion with one another, exchange information with one another, and create material that may be published on the internet. Other forms of social media include instant messaging, video-sharing sites, podcasts, and widgets. These are only a few instances of each kind of social media. Twitter, Facebook, Instagram, and YouTube are just some of the most prominent

examples of social media platforms. them have access to a useful resource in the shape of Twitter Analytics, which enables them to dive more deeply into the success of the Twitter campaigns they have ran by using the Analytics dashboard. This gives them a competitive advantage in the social media marketing space. The dashboard not only makes it easy for you to track the progress and outcomes of your Twitter advertising campaigns, but it also helps you build a better knowledge of the demographic that you are trying to reach.

Data analysts make use of the information that is posted on social media platforms such as Twitter to determine the mindsets of users and the views of customers, reveal trends in the market, discover business insights, evaluate the public's reaction to new items, and keep track of complaints. Twitter, which is a well-known social network, has more than 229 million active members as of the year 2022, and the number of tweets that are sent out each day has reached 500 million. Although tweets on Twitter are effective at disseminating information and have the benefit of being able to extensively broadcast their own information, they also have the disadvantage of being misused by spam.

Twitter is a massively popular social networking tool that has millions of active users. Because of this, multiple spammers have been prompted to send tweets containing dangerous information to several different persons. Because of this, both Twitter and researchers utilise a variety of detecting techniques to combat spammers. On Twitter, all that is required to possess an account is to set up a Twitter ID and a password. To put it another way, anybody who has access to a Twitter account and knows the associated password may make a tweet using the identity of another user. Spammers take use of this characteristic to take control of another user's Twitter account, publish spam messages, and propagate those messages farther. 84% of the accounts that tweet spam are general accounts that are run by spammers, whereas only 16% of the accounts that tweet spam are spam accounts that are automatically propagated by bots.

The malicious spamming activities have created a huge risk to the normal users' information security as well as their personal privacy. Spammers use a broad array of tactics to avoid detection by security devices so that they may continue to send

unwanted messages. These messages are often unwanted ads that the victim does not like to receive, or they are intended to lure victims into clicking on harmful URLs that are incorporated in spam tweets. In any case, the victim should delete these messages immediately. There are some users of Twitter who will only tweet links to their own websites, blogs, or products. You could get spam messages from other people, or other people might spam you by tweeting rubbish themselves. Twitter often has issues with overloading and crashing as a direct consequence of the vast number of users it serves. One further approach for marketers to breach the users' right to data privacy is via the deployment of tweets that are regarded as spam.

There is research being done to identify spam text, as well as research being done to identify spam in e-mail; however, since this study utilises mail-specific information such as headers, it cannot equate to spamming on Twitter. They can identify accounts that distribute spam even though Twitter's spam detection services and research are available. It is thus hard to distinguish between spam that has been uploaded by a spammer and spam that has been submitted by a general account that has been hacked. Researchers have made use of machine learning methods in their efforts to identify spammers operating online.

One option for gaining access to the data that Twitter has is to seek for datasets that have already been compiled and made public by other academics to accomplish their research objectives. Preprocessing and standardisation of the obtained data should be done to get rid of duplicate and missing values, and resampling should be done if the datasets were biassed. Then, to differentiate spam from non-spam, feature engineering is used to extract the features of tweets that are the most useful, and a model that fulfils the researcher's goals is selected.

In addition, supervised machine learning techniques identify a portion of a dataset as spam or non-spam. Then, the chosen model is educated using this labelled dataset. On the other hand, unsupervised machine learning techniques train the model via the use of an unlabeled dataset. After the training phase is complete, the model is evaluated using what is known as testing data, which is a new dataset that has not been used before. This evaluation determines how well the model can recognise new inputs. At long last, responses may be provided to the inquiries about the predictions.

This chapter presents a model for detecting spam on Twitter by using a swarm optimisation approach. The goal of this method is to identify spam on a tweet-by-tweet basis. To identify spam in tweets, the machine learning model must first be trained using a dataset. The swam optimisation[132]process is then used to choose the important characteristics that will be used in the classification process. The detection of spam tweets is the goal of the machine learning model, which is developed with the assistance of a dataset. The input characteristics taken from the dataset are what serve as the foundation for the development of the metaheuristic features. The Whale swam Optimisation Algorithm[133] is used before conducting classification to determine the pertinent qualities to concentrate on. This is done before the classification process. The classical objective function of WOA is converted into the stochastic gradient descent (SGD) [134]algorithm so that the process of feature selection may be carried out. A certain set of characteristics was chosen to instruct the Adaboost classifier on how to recognise spam in tweets. This was done for the goal of educating the classifier. The Adaboost classifier[135], when used in combination with WOA and SGD, produced conclusions of the greatest possible quality. The literature review is presented in the second part, the suggested model is discussed in the third section, and the experimental findings are discussed in the fourth section, which is followed by a conclusion and a list of references.

3.2 Proposed model

A swarm optimisation technique for spam detection is presented, and it would be used on a tweet-by-tweet basis. In order to identify spam in tweets, the machine learning model must first be trained using a dataset. The metaheuristic features are generated as a result of the input features included within the dataset. Using the WOA, the relevant features are selected first, then the categorization process begins. To select features, this technique presents a modification of the conventional objective function of WOA that makes use of SGD. The suggested model is shown in the form of a block diagram in figure 3.1.

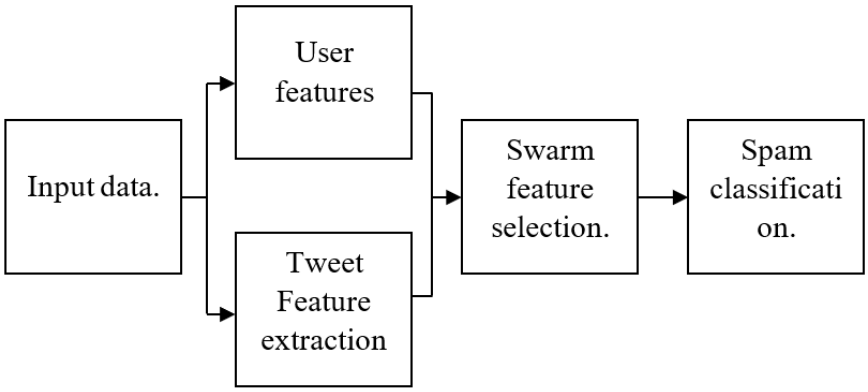


Figure 3. 1: block diagram of the proposed model

Figure 3.1 demonstrates that in order to recognize spam in tweets, the machine learning model must first be trained using a dataset. This is necessary in order to accomplish the task. In order to produce metaheuristic features, it is necessary to utilize the characteristics of the dataset that are being entered as a basis. When using the WOA, the relevant characteristics are chosen initially, and then the process of categorization may commence. This method uses SGD to provide an alternative to the traditional objective function of WOA, which can then be used for the task of selecting features. A certain set of characteristics was chosen in order to instruct the Adaboost classifier on how to recognize spam in tweets. This was done for the goal of educating the classifier. The Adaboost classifier, when used in combination with WOA and SGD, produced conclusions of the greatest possible quality.

3.2.1 Swarm Optimization Techniques

Swarm optimization algorithms are a category of optimization methods that are based on the collective behavior of a number of people or agents. These persons or agents are referred to as "swarms." The behavior of social swarms such as flocks of birds, schools of fish, and colonies of ants served as an inspiration for the development of

these algorithms. For the purpose of resolving difficult optimization issues, swarm optimization algorithms are increasingly finding widespread use across several disciplines, including engineering, finance, and robotics.

The goal of swarm optimization algorithms is to accomplish efficient and effective problem-solving by imitating the collective intelligence of social swarms. This is the major reason for the development of these algorithms. These algorithms are meant to search for the best possible solutions by interacting with a population of agents in order to uncover hidden patterns in the search space. Typically, the agents who make up the population are depicted as points in a space that has a high dimension, and each point in this space represents a different potential solution to the optimization issue. The repeated process of repositioning the agents in the search space is what allows the swarm optimization algorithm to steadily make improvements to the solutions.

The following are the stages that are involved in the functioning of swarm optimization methods. To begin, a population of agents is first seeded in the search space using a randomization method. Each agent illustrates a different approach that may be used to solve the optimization issue. Second, the agents communicate with one another by exchanging information about their locations and the speeds at which they are moving at the moment. This interaction is represented using a set of rules that describe how the agents move and update their locations in the search space. These rules define how the agents move and update their positions in the search space. Thirdly, the fitness of each agent is assessed using a fitness function, which assesses how well the agent fits the optimization requirements. This step helps determine which agents are the most likely to succeed. The agents will then modify their locations and velocities such that they are consistent with the fitness function and the interaction rules. This stage is repeatedly continued until either an optimum solution is identified, or a stopping criterion is satisfied, whichever comes first.

In comparison to more conventional optimization strategies, swarm optimization algorithm provide a number of distinct benefits. To begin, these algorithms are highly parallelizable, which indicates that they are able to make effective use of the processing capacity provided by contemporary parallel computing architectures. Secondly, swarm optimization algorithms are appropriate for addressing complicated

and noisy optimization issues because they are very resilient to noise and uncertainty in the optimization problem. This makes WOA a good choice. Thirdly, these algorithms are scalable, meaning that they may be used to solve optimization issues on a much larger scale. Fourthly, swarm optimization algorithms are adaptive, meaning that they are able to dynamically alter their search methods depending on the features of the optimization issue. This is a significant advantage over traditional search-based optimization algorithms.

Swarm optimization algorithms are a strong family of optimization methods that are inspired by the collective behavior of social swarms. These techniques were first developed by Google and were named after the term "swarm." These algorithms attempt to solve problems in an efficient and effective manner by modeling their actions after those of social swarming. Swarm optimization methods are becoming more popular for usage in a broad variety of contexts to tackle difficult optimization challenges. These algorithms offer various benefits over more conventional methods of optimization, including the capacity to run in parallel, resistance to noise and uncertainty, scalability, and adaptivity. Since swarm optimization algorithms continue to show promise in resolving a diverse variety of optimization issues, it is expected that their already substantial popularity will continue to grow in the years to come.

3.2.2 Whale Optimization Algorithm (WOA)

A programme known as the Whale Optimisation programme (WOA) was developed by concentrating on the behaviours of whales that are associated with predation. Whales hunt their prey by engaging in a swarm activity known as bubble nets.

As seen in figure 3.2, the bubble net seems to be an activity of tracking down and devouring one's prey while simultaneously drawing a circle.

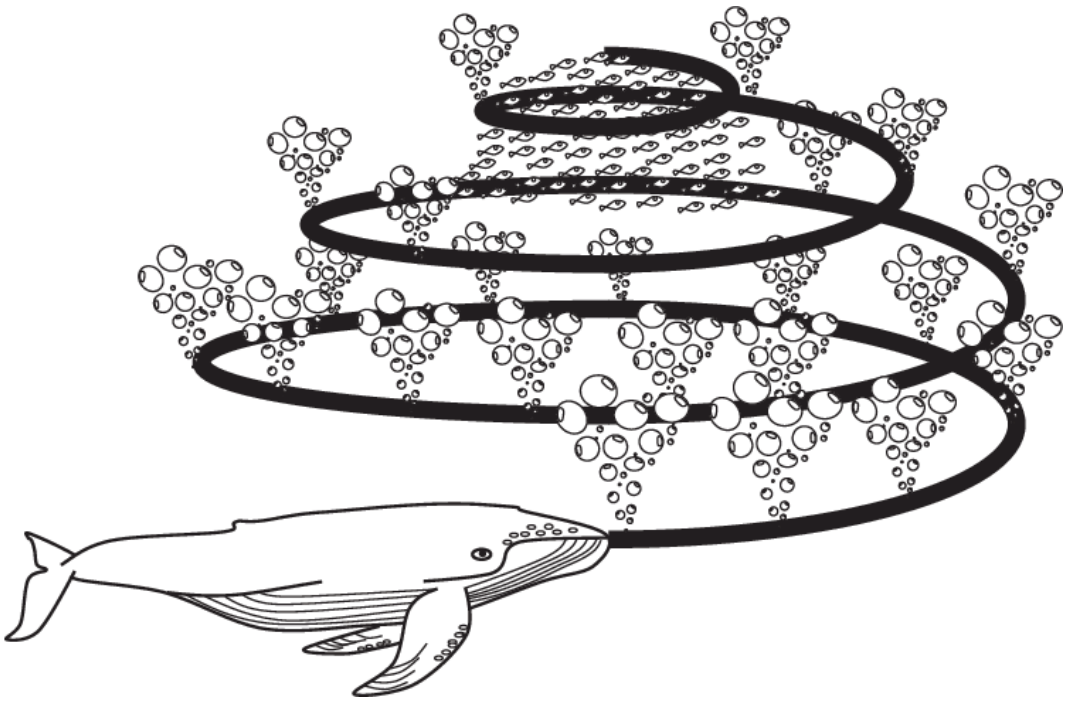


Figure 3. 2: Whale hunting behaviour

The whales find the prey by choosing one of the three actions of:

- Approaching the prey: The exploration phase where the whale searches for the prey.
- Encircling the prey: The whale rounds up the fish in this phase.
- Attacking the prey: In this phase, the spiral bubble bets are used by the whale to catch the prey.

Approaching the prey (Exploration phase): In this phase, the search agents look for the best solution randomly. The update equations in this phase are as follows:

$$\vec{D} = |\vec{C} * \vec{X}_{rand} - \vec{X}| \quad (1)$$

$$\vec{X}(t+1) = |\vec{X}_{rand} - \vec{A} * \vec{D}| \quad (2)$$

Where, $\vec{X}_{rand}$ = is a random position vector, $\vec{D}$ is the distance vector, and $\{\vec{A}, \vec{C}\}$ = coefficient vectors, which are calculated by:

$$\vec{A} = 2 * \vec{a} * \vec{r} - \vec{a} \quad (3)$$

$$\vec{C} = 2 * \vec{r} \tag{4}$$

Where, $\vec{r}$ = random vector between 0 and 1, $\vec{a}$ decreases from 2 to 0 linearly and updated based on the following equation:

$$a = 2 - t * \frac{2}{\max\_iter} \tag{5}$$

Where, t is the current iteration and $\max\_iter$ is the maximum iteration count assigned during the beginning.

Encircling the prey: In this phase, the search agents encircle the prey and the best solution is updated in such a way that the agents move close to the optimal solution. The update equations in this phase are given by:

$$D = \left|\vec{C} * \overrightarrow{X'}(t) - \vec{X}(t)\right| \tag{6}$$

$$\vec{X}(t+1) = \left| \overrightarrow{X'}(t) - \vec{A} * \vec{D}\right| \tag{7}$$

Where, $\overrightarrow{X'}(t)$ = position of the best solution and $\vec{X}$ = position vector of a solution.

Attacking the prey: In this phase, the search agents move in spirals around the prey by creating bubble nets as a trap. While forming the bubble nets, the search agents move closer to the prey, shrinking the spiral after each iteration. The update equations are given as follows:

$$\vec{X}(t+1) = \overrightarrow{D''} * e^{bl} * \cos(2\pi l) + \overrightarrow{X'} \tag{8}$$

Where, (X, Y) denote the position of the search agent and the (X', Y') denotes the position of the prey. The distance between the search agent and the prey is denoted as $\overrightarrow{D''}$ which is given by:

$$\overrightarrow{D''} = |\overrightarrow{X'}(t) - \vec{X}(t)| \tag{9}$$

l = random number between [-1,1] and b is a constant.

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X'}(t) - \vec{A} * \vec{D} \ \ if\ p < 0.5 \\ \overrightarrow{D''} * e^{bl} * \cos(2\pi l) + \overrightarrow{X'} \ \ if\ p \geq 0.5 \end{cases} \tag{10}$$

Where, p denotes the probability.

Whale Optimization Algorithm

Step 1: Initialize random search agents (X_i) and the number of iterations (t) needed.

Step 2: Begin the exploration phase with the help of equations (3), (4) and (5).

Step 3: For every search agent, evaluate the fitness function using Stochastic Gradient Descent (SGD) approach.

Step 4: Update the position vectors of the search agents as follows:

If p is greater than or equal to 0.5, the positions are updated using equations (9) and (10)

If p is less than 0.5,

If $\vec{A}$ is greater than 1, update the position using equations (6) and (7)

If $\vec{A}$ is less than 1, update the position using equations (1) and (2)

Step 5: While t is less than max_$iter$, repeat steps 3 and 4.

Step 6: When t is equal to max_$iter$, obtain the best solution.

Identifying spam tweets is a classification problem based on a set of features. In the proposed model, WOA is used to select the features and reduce the dimension of the input data. Here, the search agents are randomly selected subset of features from the input dataset. Conventional WOA uses Euclidian distance as the fitness function. The distance between the cluster members is calculated and minimized over the iterations. In the proposed model, after each iteration, the fitness function, SGD, calculates the classification accuracy of the best selected subset of features. If the new fitness is better than, previous best, the best fitness is updated along with the optimal subset of features. During the updating phase, new subset of features is calculated and the steps in the algorithm are followed to obtain the final best solution, that is the subset of features which produce the best classification accuracy.

3.2.3 Stochastic Gradient Descent

The Gradient Descent algorithm is a general-purpose optimisation method that may discover the best answers to a broad variety of challenging situations. The overarching goal is to reduce the cost function by any means necessary, which will be accomplished by systematically adjusting various parameters. Since it is responsible

for determining the size of the steps that are performed, the learning rate hyperparameter is an extremely important component of the Gradient Descent (GD) method. Finding a happy medium is of the utmost importance. If the learning rate is set too low, the method will need many iterations before it can converge, which will cause the computation time to be significantly increased. If, on the other hand, the learning rate is set too high, there is a possibility of going beyond what would be the value that is ideal.

Three types of Gradient Descent:

1. Batch Gradient Descent
2. Stochastic Gradient Descent
3. Mini-batch Gradient Descent

A procedure or operation is stochastic if it is tied to a random probability in some manner, shape, or form. Therefore, in the procedure that is known as stochastic gradient descent, rather of picking all the samples from the data set for each iteration, just a few are picked at random at each stage of the process. In Gradient Descent, the term "batch" refers to the total number of samples from a dataset that are utilised for calculating the gradient for each iteration. The word "batch" is used in the context of the Gradient Descent algorithm. The quantity denoted by this number is referred to as the "batch size." When doing a conventional optimisation using Gradient Descent, such as Batch Gradient Descent, the batch is believed to represent the whole dataset. This is because Gradient Descent uses a standard algorithm. Even while using the whole dataset is a very useful tool for discovering the minimum in a manner that is less noisy and less haphazard, when our dataset is too huge, we run into a problem, even though using the entire dataset is a very helpful tool.

To find a solution to this problem, the method known as the Stochastic Gradient Descent must be executed. Only one sample is taken into consideration while performing an iteration using SGD. A batch size of one is another way of referring to this situation. The sample is selected for usage in the iteration after first being combined in a haphazard sequence prior to said selection.

Stochastic Gradient Descent (SGD), a version of the Gradient Descent approach, is used to improve machine learning models. In this variant, the gradient is generated by making use of only one random training sample, and the parameters are changed after each iteration by making use of that one example. The use of SGD is associated with several advantages and disadvantages, which are outlined in the following list:

In the gradient method, the concept of moving parameters is an important concept. This concept is explained better with the help of linear regression as an example. Here, a linear regression of one variable is described. A one-variable linear regression is a given number of points.

$$(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$$

A straight line that minimizes the sum of errors $y = f(x)$ is used to approximate the points. For example, in the figure below, the red straight line is the solution.

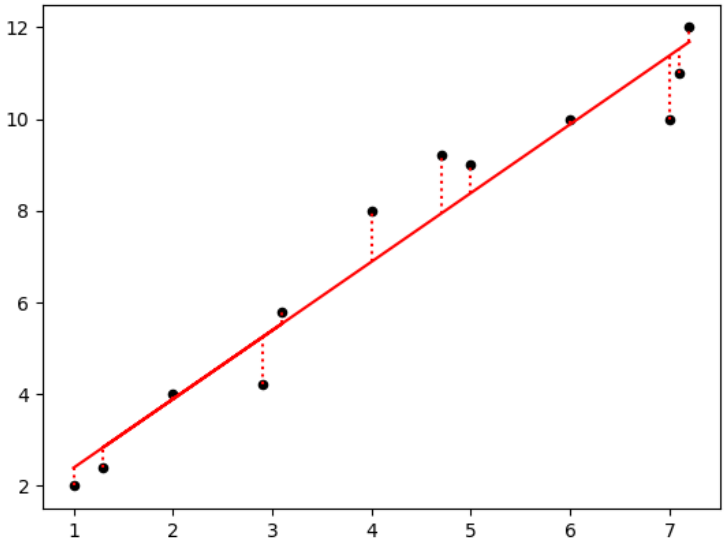


Figure 3. 3: error plot

The sum of the errors of a straight line and multiple points is generally calculated as the sum of squares of residuals. Specifically, it is given by the following formula.

$$sum\ of\ squares\ of\ residuals = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Now, $y = f(x)$ is a is a straight line, which can be represented as $f(x) = ax + b$, where a is the slope of a line, and b is an intercept. Parameters a and b make the line move. The value of the objective function (error) moves when the parameter is moved. Gradient descent method is a method of minimizing the error by using the relationship between the parameter and the objective function. The gradient descent method suffers from drawback like computation complexity and excess time of execution. These drawbacks can be overcome by using SGD. The basic mechanism of SGD is that only one randomly selected data is used for each parameter update. Because the technique only calculates the gradient for a single observation at a time rather than for the full dataset, the result is just an estimate of the true gradient. In other words, the amount of calculation is greatly reduced by using only one data instead of using all the data for each parameter update. SGD updates the parameters by considering only the distance between the straight line corresponding to the current parameters (red straight line in the figure below) and one randomly selected point.

The SGD algorithm is implemented using the following steps:

SGD algorithm

Step 1: The slope/gradient of the input data is calculated with respect to each parameter.

Step 2: Select a random set of input parameters and calculate the partial derivative of the output with respect to each input parameter.

Step 3: Update the gradient function by setting the set size

$$step\ size\ =\ gradient\ *\ learning\ rate$$

Step 4: find the new parameters:

$$new\ params\ =\ old\ params\ -step\ size$$

Step 5: Repeat steps 2 to 4 until gradient is almost 0.

3.2.4 Adaboost Classifier

In the realm of machine learning, the AdaBoost algorithm is a form of boosting strategy utilized within Ensemble Methods. It is referred to as "Adaptive Boosting" because it reallocates weights to each instance, assigning greater weights to

incorrectly classified examples. This adaptive approach aims to enhance the overall classification accuracy. Boosting, a technique in supervised learning, aims to reduce bias and variance simultaneously. The fundamental principle behind boosting is that learners progressively acquire new knowledge at higher levels. Each subsequent learner, except the initial one, is built upon the learners developed in previous iterations. In other words, learners with lower capabilities can be transformed into learners with higher capabilities. The AdaBoost algorithm operates on the same basic idea as the boosting method, but with a few key distinctions. Let's go into more depth about this distinction.

To begin, let's talk about how boosting really works. During the time that it spends "training" on the data, it generates "n" different decision trees. In the process of creating the first decision tree or model, the record that was initially misclassified in the first model is given precedence. The only records that are used as input for the second model are these ones. The procedure will continue until we decide on the total number of foundational learners that we want to produce. Keep in mind that the playing of the same record more than once is permitted with all the boosting approaches.

This image explains how the first model is constructed, and it also demonstrates how the algorithm accounts for any flaws that may have been introduced by the first model. The record that has been erroneously categorised is considered as an input for the next model. This procedure is carried out many times till the required condition is satisfied. As can be seen in the image, a 'n' number of models are produced when the mistakes from one model are included into the production of subsequent models. The process of boosting works like this. The models 1, 2, 3,..., N are all separate models that together make up what are called decision trees. The fundamental idea behind every single sort of booster model is the same.

Since we now know the boosting principle, it should not be too difficult for us to comprehend the AdaBoost algorithm. Let's go into the inner workings of AdaBoost. The programme creates a 'n' number of trees whenever the random forest data structure is utilised. It creates correct trees that have a root node and several leaf nodes in each branch. In a random forest, some trees will likely be larger than others,

but the overall depth will not be consistent. However, when using AdaBoost, the method will only produce a node with two leaves, which is referred to as a Stump.

The figurative representation of the stump may be seen here. It is obvious that there is only one node between the two leaves that it possesses. These stumps are poor students, yet boosting strategies favour them because of their little potential for growth. In AdaBoost, the sequence in which the stumps are placed is of the utmost importance. The lesson learned from the mistake made with the first stump is carried over to the subsequent stumps. Let's look at an example of this to better comprehend it.

The following is an example dataset that consists of just three characteristics and produces output in the categorical form. The data set is shown here in its real form, as seen in the picture. As a result of the output being in binary or categorical form, we now have a difficulty with categorization. In the actual world, the dataset may include records and characteristics in any quantity that the user desires. Let us examine 5 datasets for explaining reasons. The results are presented in a categorical format, which in this case takes the shape of a yes or no. A sample weight will be applied to each one of these records. "W=1/N" is the formula that is utilised for this, where "N" refers to the total number of records. Due to the small size of this dataset, which only contains 5 records, the sample weight will initially be 1/5. Each record is given the same amount of weight. In this instance, the answer is 1/5.

A learner generated by using a classification algorithm once is called a weak learner. An example of a weak learner is a decision tree. In addition, the final classifier that can be used is called a strong learner. Creating a strong learner based on a weak learner is called ensemble learning. Boosting is one method of ensemble learning.

Booting is one method of ensemble learning. The general flow of boosting is as follows.

Boosting Algorithm
1. Make a weak learner $f_1(x)$

2. Consider the result of $f_1(x)$ and make the following weak learner $f_2(x)$

3. Create the following weak learner $f_t(x)$ in order, considering the result of $f_{t-1}(x)$.

4. After making f_k, finally make a strong learner $f(x)$ by collecting $f_1(x)$to $f_k(x)$.

An overview of AdaBoost for binary classification problems

Consider creating a binary classifier $f(x)$ based on the training data (x_i, y_i) (where i = 1, ..., n). Since it is a binary classification problem y is -1 or 1. Also, the output of each classifier should be 1 or -1.

1. Make a weak learner $f_1(x)$.

First, create a weak learner $f_1(x)$ that minimizes the training error.

$$E_1 = \frac{1}{n}\sum_{i=1}^{n}[y_i - f_1(x_i)]$$

Where y_i is the actual output.

2. Consider the result of $f_1(x)$) and make the following weak learner $f_1(x)$. Next, the training error, E2 is given by

$$E_2 = w_i^{(2)}\sum_{i=1}^{n}[y_i - f_2(x_i)]$$

$w_i^{(2)}$represents the importance of each sample when f_2 was created (calculated so that $f(x)$ is more important for misclassified samples). Samples of high importance have a large penalty if a mistake is made, so f_2 is created with an emphasis on avoiding mistakes as much as possible.

3. In turn, create the following weak learner $f_t(x)$, taking into account the results of $f_{t-1}(x)$.

Similarly, create a weak learner $f_t(x)$ such that the training error is minimum.

$$E_t = w_i^{(t)} \sum_{i=1}^{n} [y_i - f_t(x_i)]$$

$w_i^{(t)}$ represents the "importance" of each sample when creating f_t (calculated so that $f_{t-1}(x)$is more important for misclassified samples).

4. Finally, make a strong learner $f(x)$ by collecting $f_1(x)$ to $f_t(x)$.

$$f(x) = sign\{\sum_{t=1}^{k} \alpha_t f_t(x)\}$$

α_t is the weight to be applied to the t-th learner, and sign is a sign function (a function that returns 1 for positive inputs and -1 for negative inputs).

$$\alpha_t = \frac{1}{2}\ln(\frac{1-E_t}{E_t})$$

$E_t$is the error of the t-th learner.

3.3 Experimental Results

This section presents the experimental results carried out in order to evaluate the proposed model. The proposed model is compared with other swarm optimization techniques namely Particle Swarm Optimization (PSO), Moth-Flame Optimization (MFO) and Mean-variance optimization (MVO). The machine learning algorithms under study are Stochastic Gradient Descent (SGD), Support Vector Machine (SVM) and Decision Tree (DT). The dataset contains 11,968 entries for training and 630 entries for testing. Each entry has the following attributes.

- Tweeted text
- Number of followers of the tweet and the user
- The actions performed on the tweet.
- Location of the user.
- Type: Either Quality or Spam

The tweets are classified as spam based on their motive. These include:

- Politically Motivated
- Automatically generated content
- Meaningless content
- Click Bait

The Metaheuristic features crated from the tweet:

1. Tweet length
2. Number of # tags
3. Number of @ tags
4. Number of weblinks (URLs)
5. Number of capitalized words
6. Number of Exclamation symbols
7. Number of Question marks

The data that was entered are initially passed to a module called the feature selection module, which uses a method called swarm optimisation. Particle Swarm Optimisation (PSO), Moth Flame Optimisation (MFO), Mean-variance Optimisation (MVO), and Weighted Overall Average (WOA) are the four swarm optimisation approaches that were used in this study. Classifiers such as Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Decision Tree (DT), and Adaboost (AB)**Error! Reference source not found.** have been used to classify the dataset u sing the reduced set of features determined by each method. This was accomplished by first using the algorithms to choose the optimal feature subset and then using the classifiers to classify the dataset using the reduced set of features. The graphics that follow demonstrate how the algorithms function when combined.

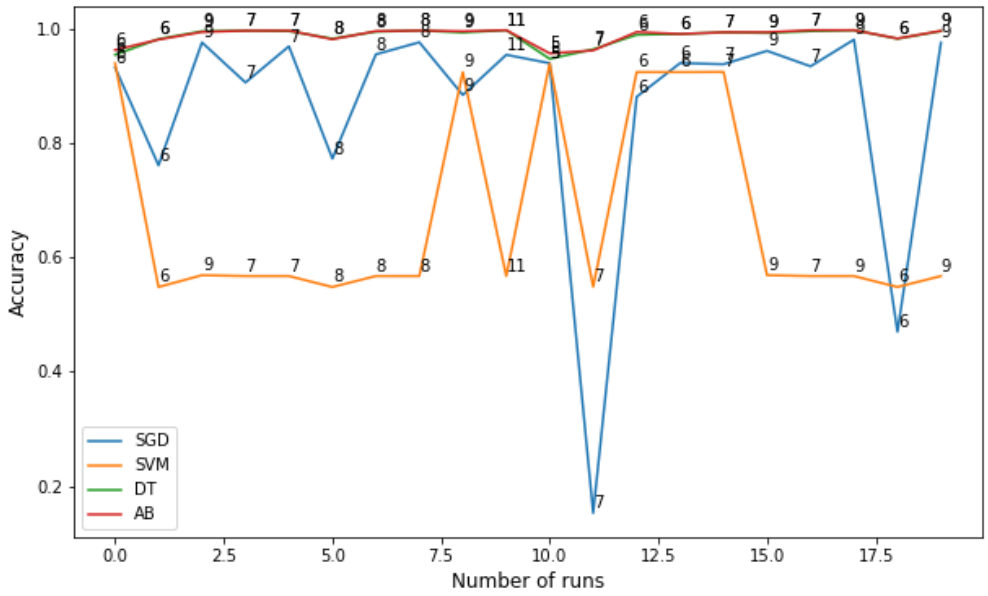


Figure 3. 4: Accuracy of PSO based feature reduction with machine learning techniques

Figure 3.4 shows the accuracy of PSO algorithm with different classifiers. In each iteration, PSO selects a subset of features which are then classified using SGD, SVM, DT and AB. In the graph, a number denotes the count of selected subset of features by SVM. From the graphs, it can be seen that SVM has the least average accuracy while classifying the data. SGD is the next best classifier after SVM but not the overall best. DT and AB have almost equal accuracy at each iteration, all close to 98%. The minimum subset of features selected by the algorithm is 6 when producing high accuracy.

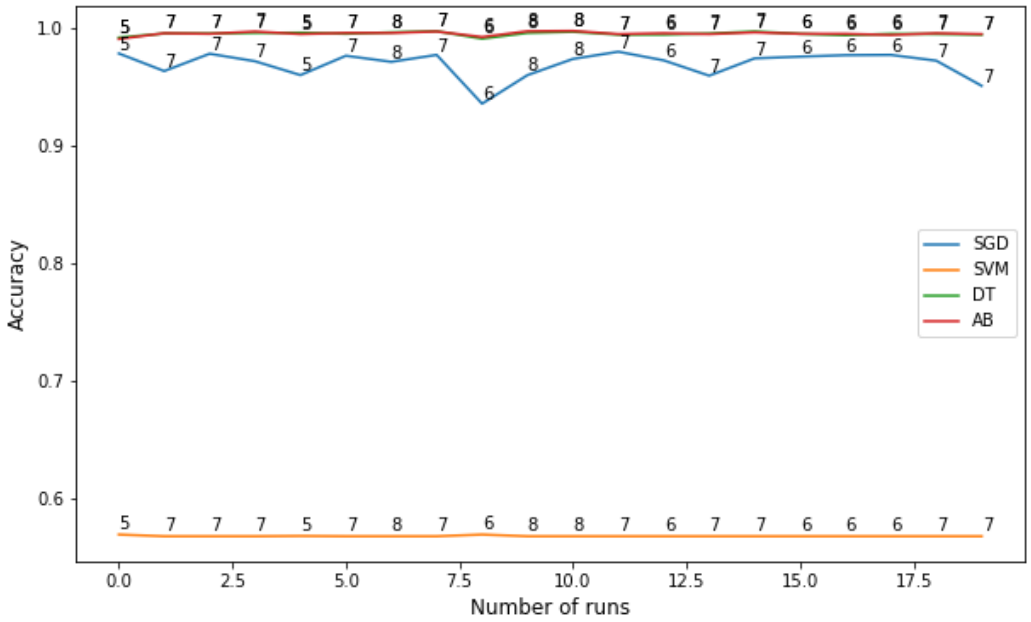


Figure 3. 5: Accuracy of MFO based feature reduction with machine learning techniques

The accuracy of the MFO method using various classifiers is shown in Figure 3.5. MFO chooses a subset of features for each iteration, which are subsequently categorized using SGD, SVM, DT, and AB. The graph shows the selected subset of features at each iteration. The graphs show that while categorizing the data, SVM has the lowest average accuracy, never more than 60%. The next best is the SGD classifier, and the best results are obtained by both DT and AB. The minimum subset of features selected by the algorithm is 6 when producing high accuracy.

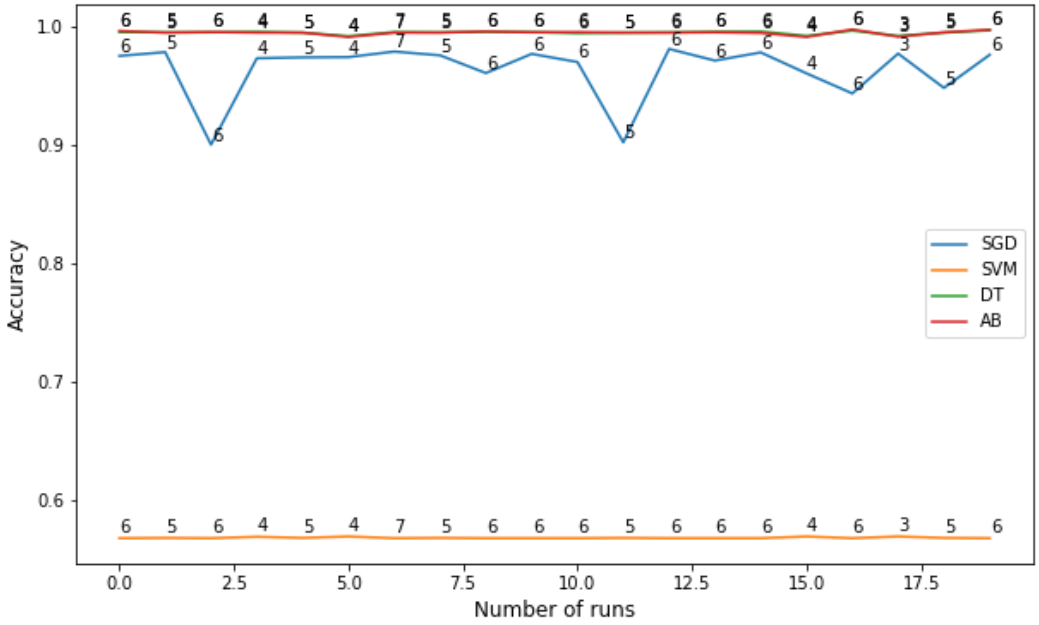


Figure 3. 6: Accuracy of MVO based feature reduction with machine learning techniques

Figure 3.6 shows the result of MVO algorithm. Like the other techniques, with MVO, DT and AB have produced the best accuracies at each iteration. The minimum subset of features selected by the algorithm is 3.

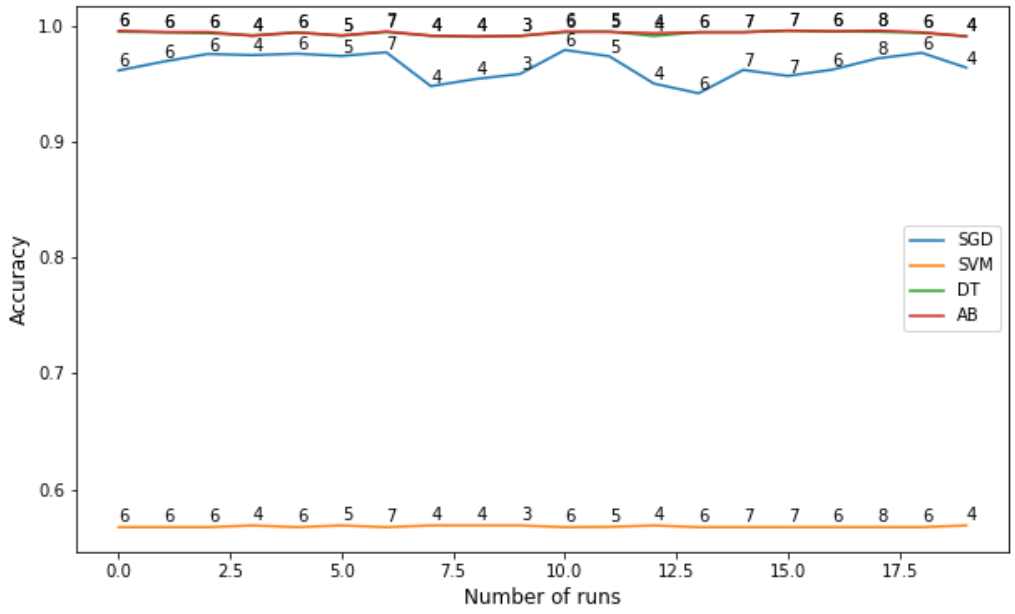


Figure 3. 7: Accuracy of WOA based feature reduction with machine learning techniques

WOA is the selected optimization model. As discussed in section 3, the modified WOA has selected the best subset of features while producing the best accuracy. The highest accuracy was produced by adaboost algorithm. The combination of WOA and Adaboost produced the highest accuracy with minimum features and in the least amount of time. Table 3.1 shows the numerical parameter analysis of the proposed model.

Table 3. 2: Parameter analysis with other optimization techniques

Optimizer	Classifier	Accuracy	Execution time	Selected features
PSO	**SGD**	0.980344	25.72834754	[1 1 1 1 1 1 0 1 1 0 0 1]
PSO	**SVM**	0.938821	24.52714443	[0 0 1 1 1 0 0 1 0 1 1 0]
PSO	**DT**	0.996806	27.24336624	[1 1 1 1 1 1 1 1 1 0 1 1]
PSO	**AB**	0.996806	27.24336624	[1 1 1 1 1 1 1 1 1 0 1 1]
MVO	**SGD**	0.98059	22.41409969	[1 1 1 1 1 0 0 0 0 1 0 0]
MVO	**SVM**	0.568796	22.90994859	[0 1 1 1 0 1 0 0 0 0 0 0]
MVO	**DT**	0.996314	20.33446574	[1 1 1 1 0 0 0 1 0 0 1 0]
MVO	**AB**	0.997052	20.88643289	[1 1 1 1 1 0 0 1 0 0 0 0]
MFO	**SGD**	0.979607	30.5992732	[1 1 1 1 0 0 0 0 0 1 1 1]
MFO	**SVM**	0.568796	25.80968833	[0 1 1 1 0 0 1 0 0 0 0 1]
MFO	**DT**	0.997052	30.73957872	[1 1 1 1 0 1 0 1 0 0 1 0]
MFO	**AB**	0.997297	24.77219057	[1 1 1 1 1 1 0 1 0 0 1 0]
WOA	**SGD**	0.97887	18.89253712	[1 1 1 1 0 0 1 0 0 0 0 1]
WOA	**SVM**	0.568796	19.98783708	[0 1 1 1 0 0 0 0 0 0 1 0]
WOA	**DT**	0.995577	17.92719698	[1 1 1 1 0 0 0 1 0 0 1 1]
WOA	**AB**	0.998577	17.92719698	[1 1 1 1 0 0 0 1 0 0 1 1]

Table 3.1 shows the comparative analysis of the algorithms with the proposed model. The table shows the combinations of optimization techniques and classifiers that are used in the experimental analysis. For each classifier and optimization technique, the accuracy, the execution time, and the selected subset of features are indicated. PSO algorithm took 25.73 seconds and produced an accuracy of 98% with SGD. With

SVM, PSO took 24.5 seconds while producing an accuracy of 93%. The Decision tree took 27 seconds and produced an accuracy of 99.6% which is same as AB. The best accuracy of 99.68% was obtained with a combination of PSO+AB and PSO+DT. With MVO, SGD produced 98%, SVM produced 56%, DT produced 99.6% and AB produced 99.7% in around 20 to 22 seconds. The best accuracy obtained with the combination of MVO and AB is 99.7%. MFO with DT and AB produced 99.7% accuracy in 30 seconds and 24 seconds respectively. Coming to the WOA, DT produced an accuracy pf 99.55% in 17.92 seconds while AB produced an accuracy of 99.85% in 17.92 seconds. The selected count of features is 7 out of 12. The modified Whale optimization algorithm was the fastest with AB and produced a highest accuracy of 99.85% with 7 features only. The selected features are:

1. Following
2. Followers
3. Actions
4. Tweet_length
5. Weblinks
6. Question_marks
7. Fullstops.

Table 3. 3: Comparative analysis

Algorithm	**Accuracy**
MLP [129]	92%
SVM [130]	93%
PSO + DT [131]	99.6%
Proposed model	99.85%

The proposed model obtained the highest accuracy of 99.85% when compared with existing techniques like MLP which obtained an accuracy of 92%, the SVM obtained an accuracy of 93% and PSO + DT which obtained an accuracy of 99.6%.

3.4 Conclusion

On a tweet-by-tweet basis, a swarm optimization strategy for spam detection is suggested. The machine learning model is trained using a dataset for the identification

of spam tweets. Based on the input features in the dataset, metaheuristic features are produced. Before the classification process begins, the WOA technique is used to choose the necessary properties. When selecting features, the SGD algorithm, which is a variation of the conventional objective function of WOA, is used. The Adaboost classifier is educated to identify spam in tweets by making use of the selected subset of attributes during training. The Adaboost classifier produced the greatest results when used in conjunction with WOA and SGD. In testing using the smallest possible subset of seven features and in the least amount of time (17.9 seconds), an accuracy of 99.85% was achieved.

Chapter-4

GLoVe Language Model for Twitter Spam Detection using Bidirectional LSTM

4.1 Introduction

The term "Twitter spam" refers to information or messages that are broadcast on the Twitter network that have not been requested and are not desired. It includes promotional or unrelated information with the intention of misleading or manipulating consumers for the sake of personal benefit. A few examples of this include automated tweets, links that lead to phishing sites, bogus accounts, content that is repetitive, and excessive advertising. Twitter employs automated tools and relies on user reports to identify and eliminate spam to achieve its goal of improving the user experience, ensuring users' safety, and upholding its terms of service.

The proliferation of social media platforms, such as Twitter, in recent years has made it possible to participate in more productive kinds of communication and has increased chances for such activities. However, in addition to the advantages, there has been a rise in the appearance of spam, which presents a variety of issues for users as well as the administrators of the platform. The use of models based on artificial intelligence (AI) has shown itself to be a useful answer to this problem, which must be addressed. This article investigates how artificial intelligence models can efficiently identify spam in tweet content, which contributes to the development of Twitter's security and user experience.

By using a wide variety of strategies and procedures, artificial intelligence models have proved that they are capable of effectively identifying spam. Natural Language Processing, often known as NLP, is an extremely important component in both comprehending and processing text data. The linguistic patterns, mood, context, and semantic meaning of tweet content are analyzed by AI models using natural language processing (NLP) methods. NLP can differentiate between valid material and communications that include spam by recognizing patterns, which are linked with spam, and extracting attributes associated with spam. To extract useful information

from twitter text, AI models make use of several feature engineering approaches. When trying to detect spam trends, certain characteristics, such as the number of links, hashtags, mentions, and repeated material, are taken into consideration. In addition, an examination of the speaker's attitude as well as the peculiarities of their language could aid to the detecting procedure.

By doing an analysis of the textual content as well as the linguistic properties of the tweets, vocabulary models are an extremely useful tool for detecting spam tweets. The following is a list of the many ways that vocabulary models contribute to the identification of spam:

1. **Lexical Analysis**: Specific terms and phrases that are regularly connected with spam material may be identified with the use of vocabulary models, which are used in lexical analysis. These algorithms have been trained on massive datasets, and as a result, they are able to identify patterns that point to suspicious activity. The process of obtaining significant information from lexical analysis includes looking for things like suspicious URLs, excessive usage of specific words, and recognized spam keywords. The program is able to identify material that may include spam by comparing the vocabulary of a tweet to a list of phrases that are often associated with spam.
2. **An Understanding of Context**: Vocabulary models can grasp the semantic meaning as well as the context of twitter content. They do an analysis of the connections between words and sentences, gaining a grasp of how these elements are used within a certain setting. This helps the model to differentiate between legal and spammy text, which is useful given that spam tweets often make use of language that is either odd or incomprehensible. The model can recognize patterns that are indicative of spam or material that is misleading since it makes use of contextual information.
3. **The Detection of Attempts at Phishing**: Phishing is a prevalent method used by spammers to deceive users into divulging critical information. Phishing attempts may be identified using vocabulary models, which achieve this by identifying URLs or domain names that are related with known phishing sites. The program can determine whether or not a tweet contains potentially

hazardous links by comparing the URLs included inside the tweet to a database that contains the URLs of known malicious websites.

4. **Spam Repetition and Redundancy**: Vocabulary models can identify tweets that are spam because of their repetitive or redundant character. Spammers often make use of automated systems to manufacture and disseminate enormous numbers of tweets that are either identical to one another or are very close to one another. The program can recognize these recurrent spam messages and mark them appropriately by doing an analysis of the language and linguistic patterns involved.
5. **Analysis of Sentiment**: Vocabulary models can evaluate the tone that is communicated in tweet text. Tweets that are part of a spam campaign may often utilize wording that seems suspiciously favourable or too promotional to get readers to interact with the material. The program can discover abnormalities and possibly flag them as spam by analysing the sentiment of the tweet and comparing it to the usual sentiment distribution seen in valid tweets. This process is called sentiment analysis.
6. **Linguistic Features**: Vocabulary models may be used to assess the linguistic features of tweets, such as problems in grammar, punctuation, or spelling. By purposefully misspelling words and using strange grammatical structures, spammers may obscure the meaning of the messages they send and avoid being discovered. Through the examination of these linguistic features, the model can recognize potentially malicious patterns and label tweets as candidates for being spam.

Vocabulary models provide the groundwork for recognizing spam tweets by doing an analysis of the textual content and the linguistic characteristics of the tweets. They successfully recognize tweets as spam or valid material by using lexical analysis, contextual understanding, detection of phishing efforts, identification of repetition and redundancy, sentiment analysis, and linguistic characteristics. These models may adapt to newly discovered spamming strategies and increase their accuracy over time if they are continually trained and updated with fresh data.

4.2 Proposed Model

The proposed model processes the input data consisting of tweets and additional information, such as follower counts and user behaviors. It comprises two independent components. The first component focuses on the textual content of the tweets. It utilizes the GLoVe language model to extract relevant features related to the vocabulary used in the tweets. These features are then used by an LSTM deep learning model to identify spam messages. The second component utilizes the information associated with the tweets, along with additional meta-heuristic aspects, including tweet length and the presence of question marks. A CNN model is employed to classify the tweets based on these attributes. The final decision is reached by combining the data obtained from both the LSTM and CNN models. This integration allows for a comprehensive and accurate assessment of the tweets, specifically determining whether they are spam or not. Figure 4.1 shows the proposed model architecture.

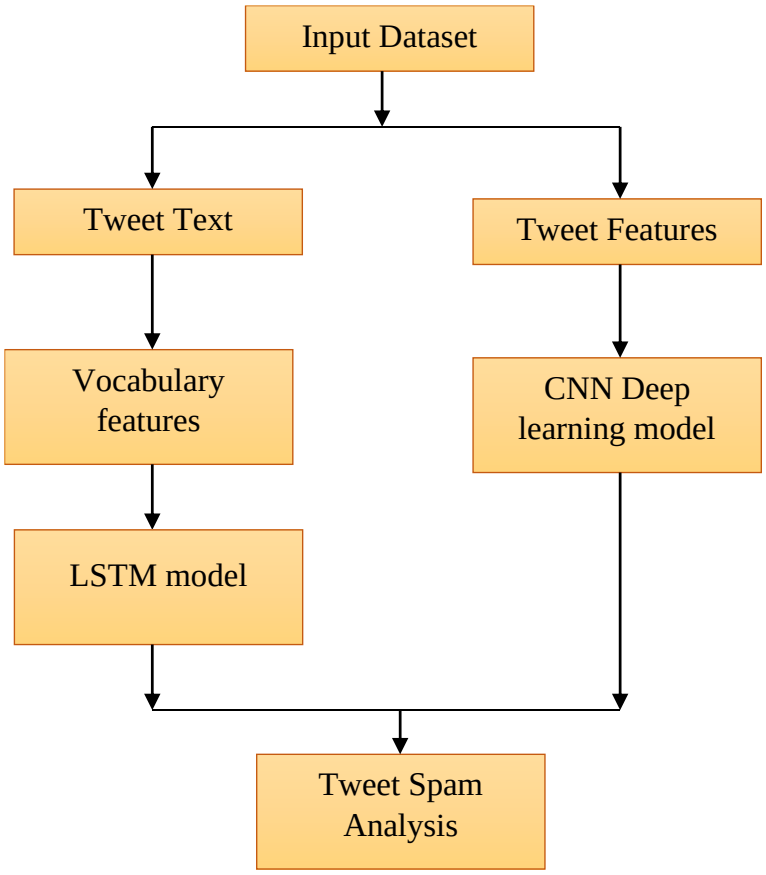


Figure 4. 1: proposed model

Proposed Algorithm

1. **Input Data Processing**: Input data consists of tweets and additional information, including follower counts and user behaviours.
2. **Two-Component Model**: The proposed model comprises two independent components for spam detection.
3. **Textual Content Analysis**: Utilize the GLoVe language model to process the textual content of the tweets. Extract relevant features related to the vocabulary used in the tweets.
4. **LSTM Model for Textual Content**: Use an LSTM (Long Short-Term Memory) deep learning model to analyze and classify tweets based on the extracted features. The LSTM model is employed to identify spam messages in the tweets.
5. **Information-Based Analysis**: Utilize information associated with the tweets, including meta-heuristic aspects such as tweet length and the presence of question marks.
6. **CNN Model for Information-Based Analysis**: Employ a CNN (Convolutional Neural Network) model to analyze and classify tweets based on attributes such as tweet length and question mark presence. The CNN model is used for classifying tweets as spam or non-spam based on these attributes.
7. **Combination of Results**: Combine the outputs and decisions obtained from both the LSTM and CNN models. This integration allows for a comprehensive and accurate assessment of the tweets.
8. **Final Spam Detection Decision**: Based on the combined data and decisions from the LSTM and CNN models, determine whether the tweets are classified as spam or not.

4.2.1 GLoVe word embeddings

Global Vectors for Word Representation (GLoVe) is an unsupervised learning algorithm used for creating word embeddings. GLoVe was developed with the express purpose of accomplishing the acquisition of these embeddings via the use of global statistics generated from large-scale text corpora. The distributional characteristics of words are the major focus of this project's goals. In order to achieve this goal, the model does an analysis of the co-occurrence statistics of terms inside a corpus. Specifically, it looks at the frequency with which word pairs are found together in context. The fundamental presumption is that words that have a similar meaning or are often used in comparable contexts tend to display greater co-occurrence rates.

4.2.1.1 Word Embedding Algorithm

The word embedding[136] algorithm based on matrix decomposition is a method that utilizes global statistical information. First, a word co-occurrence matrix or a document-word matrix needs to be constructed in the corpus. The following is a simple example to illustrate, assuming that the corpus contains the following three documents, the corresponding word co-occurrence matrix or document-word matrix can be constructed:

Document 1: I have a cat

Document 2: cat eat fish

Document 3：I have an apple

	I	have	a	cat	eat	fish	apple
I	0	2	2	1	0	0	1
have	2	0	2	1	0	0	1
a	2	2	0	1	0	0	1
cat	1	1	1	0	1	1	0
eat	0	0	0	1	0	1	0
fish	0	0	0	1	1	0	0
apple	1	1	1	0	0	0	0

Figure 4. 2: Word co-occurrence matrix

	I	have	a	cat	eat	fish	apple
Document 1	1	1	1	1	0	0	0
Document 2	0	0	0	1	1	1	0
Document 3	1	1	1	0	0	0	1

Figure 4. 3: Document word matrix

Figure 4.2 shows Word co-occurrence matrix and Figure 4.3 shows Document word matrix

In the word co-occurrence matrix, the word "I" and the word "have" co-occur in two documents, so their connection weight is 2, in the document- word matrix, document 1 contains a word "I", so it is 1. When constructing the document-word matrix, TF-IDF can be used as weights. After obtaining the word co-occurrence matrix or document-word matrix, the LSA algorithm can be used to learn the word vector. The LSA algorithm (latent semantic analysis) is mainly used for text topic analysis. By decomposing the document-word matrix, documents and topics, Links between words and topics. The matrix X (M×N) represents the document-word matrix, which contains M documents and N words. LSA uses SVD to decompose the matrix X to obtain two low-dimensional matrices U (M×k) and V (N×k), and each row of V is a word vector of a word.

$$X_{M*N} = U_{M*k}\Sigma_{k*k}V_{N*k}^{T} \quad (1)$$

The advantage of the method based on matrix decomposition is that it can effectively utilize the global statistical information. The disadvantages are: 1. The time complexity of the SVD algorithm is too large, and it is not suitable for large data sets; 2. It is mainly used to obtain the similarity of vocabulary, and the performance of the vocabulary analogy task is not as good as the method based on shallow window prediction.

Shallow window-based methods are also called prediction-based methods, and representative algorithms include NNLM, Word2Vec, etc. Shallow window-based methods usually use the local information of the corpus to generate a local context window during training. By using the context word to predict the the Skip-Gram model of Word2Vec, the central word is mainly used to predict the context word, maximizing P (context word | central word); while the CBOW model in Word2Vec mainly predicts the central word through the context word, maximizing P (central word | context word).

The previous article introduced Word2Vec, so I won't go into details. The advantages of the shallow window-based method are: 1. The prediction method is used in the training process, and the performance in the vocabulary analogy task is better; 2. The training is faster and can adapt to large data sets; 3. It can learn between words Complex patterns beyond similarity. The disadvantages are: 1. It cannot use global statistics well; 2. It requires many data sets. Both the matrix decomposition and the shallow window-based method have some limitations, and the logic of the Glove algorithm is to combine the advantages of the two types of algorithms, and then focus on understanding the Glove algorithm.

4.2.1.2 Glove word co-occurrence matrix and co-occurrence probability matrix

The GloVe model combines the advantages of LSA and Word2Vec, using both the global statistical information of the corpus and the local contextual features (sliding window). Glove initiates the process by generating a matrix that records the frequency with which words appear together. It presents the idea of a co-occurrence probability

matrix. Based on the word "co-occurrence matrix," one may determine how to construct the co-occurrence probability matrix.

A. Glove word co-occurrence matrix

There are some differences between Glove and LSA when constructing the word co-occurrence matrix. A context window needs to be limited. The construction process is as follows:

1. Construct a Nonempty matrix whose value is 0.
2. Define a sliding crisp mouth, the size is c.
3. Start from the first word in the corpus as the central co- moving seat, and the central word is in the centre of the window.
4. There are c-1 monotones on the left and right sides of the centre tone, which is the context monotone.
5. Count the number of occurrences of the left and right context words in the statistics centre and add them to the matrix.
6. Screw sliding and refreshing.

For example, given the sentence "I have a cat" and a context window size of 3, the following windows can be constructed. When traversing to the third window "have a cat", the central word is "a", and statistical information should be added to the word co-occurrence matrix X at this time. X (a, have) += 1, X (a, cat) += 1. Note that the word co-occurrence matrix X constructed by this method is a symmetric near word is large.

Table 4. 1: Centre word and Window

Centre word	Window
I	I have
have	I have a

a	have a cat
cat	a cat

Table 4.1 shows Centre word and Window.

B. Glove co-occurrence probability matrix

After the co-occurrence matrix X is counted, X ij can be used to indicate the number of co-occurrences of word i and j, and X i is the sum of all X ij , P ij = P (j|i) means that word j appears in the context of word i probability.

$$X_i = X_{ij}$$

$$P_{ij} = P(j|i)| = \frac{X_{ij}}{X_i} \quad (2)$$

Glove proposed the concept of co-occurrence based on the above, and the co-occurrence probability can be understood as the ratio of the above conditional probability. The following is an example in the original paper. Given the central words ice (ice) and steam (water vapor), they can be judged by the ratio of different context words k to the conditional probabilities of the central words ice and steam Ratio (ice, steam, k).

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

$$Ratio = P(k|ice)/P(k|stean)$$

When the correlation between word k and ice is relatively large, such as k = solid (solid), Ratio (ice, steam, k) will be relatively large; When the word k is highly correlated with steam, such as k = gas (gas), Ratio (ice, steam, k) will be relatively small; When k is related to both ice and steam, such as k = water (water), the value of Ratio (ice, steam, k) will be close to 1; When k is not related to ice and steam, such as k = fashione (fashion), the value of Ratio (ice, steam, k) will be close to 1; Through

this ratio Ratio (ice, steam, k) can well distinguish words related to ice (solid), words related to steam (gas) and some words that are not very important to ice and steam (water, fashion). Therefore, good word vector can encode Information about Ratio (i, j, k).

4.2.1.3 Derivation of Glove algorithm

Use w(x) to represent the word vector of word x, and w'(x) to represent the word vector when word x is used as the context. Then given the central word i, j and the context word k, Glove hopes that the word vector can encode the information of Ratio (i, j, k), and there will be a function F that makes the following formula hold.

$$F\big(w(i), w(j), w'(k)\big) = Ratio(i,j,k) = \frac{P_{ik}}{P_{jk}} \tag{3}$$

The right part of the above formula is calculated through the word co-occurrence matrix, and then the formula needs to be simplified. The author of Glove believes that the word word vector space is a linear structure, for example, the difference of "man" - "women" is very similar to the difference of "king" - "queen". Therefore, an intuitive method is to simplify the formula by the difference of word vectors.

$$F(w(i) - w(j), w'(k) = \frac{P_{ik}}{P_{jk}} \tag{4}$$

The right side of the above formula is a scalar, while the left side of the F function is a vector. To avoid the function F (F can be very complicated, such as using a neural network to learn) to learn some useless things and confuse the linear structure that Glove hopes to obtain, so the formula is further simplified, and the function in the formula F is also changed to a scalar.

$$F(\big(w(i) - w(j)\big)^T w'(k)) = \frac{P_{ik}}{P_{jk}} \tag{5}$$

unchanged. However, the above formula does not satisfy this condition, so the above formula must be changed to satisfy homomorphism.

$$F\left(\big(w(i) - w(j)\big)^T w'(k)\right) = \frac{F(w(i)^T w'(k))}{F(w(j)^T w'(k))}$$

$$F\big(w(i)^T w'(k)\big) = P_{ik} = \frac{X_{ik}}{X_i} \tag{6}$$

It can be seen from the formula that F is an exponential function, that is, F = exp, so the above formula can be transformed to get the following formula.

$$exp\left(w(i)^T w'(k)\right) = \frac{X_{ik}}{X_i}$$

$$w(i)^T w'(k) = \log\left(\frac{X_{ik}}{X_i}\right) = \log(X_{ik}) - \log(X_i)$$

Note that exchanging the positions of i and k on the right side of the above formula will change the symmetry of the formula. To ensure the symmetry, the author made the following transformation, adding two bias terms b(i) and b'(k).

$$w(i)^T w'(k) + \log(X_i) = \log(X_{ik})$$

$$w(i)^T w'(k) + b(i) + b'(k) = \log(X_{ik}) \quad (7)$$

Therefore, it is ultimately necessary to minimize the following objective function:

$$J = \sum_{i,j=1}^{V} f(X_{ij})\, w(i)^T w'(k) + b(i) + b'(k) - \log(X_{ik})^2$$

$$f\left(X_{ij}\right) = \begin{cases} \left(\frac{X_{ij}}{X_{max}}\right)^{\alpha} & if \quad X_{ij} < X_{max} \\ 1 & otherwise \end{cases}$$

The objective function is the square error, where f($X_{ij)}$ represents the weight of the loss function. The author uses the above formula to calculate f(X ij), which guarantees: 1. The more times the two words co-occur, the greater the weight of the loss function; 2. When the number of co-occurrences of two words exceeds a threshold, the weight does not continue to increase, and the maximum weight is 1; 3. The number of co-occurrences of two words is 0, then the weight is 0. The graph of f (X_{ij}) is as shown in Figure 4.4.

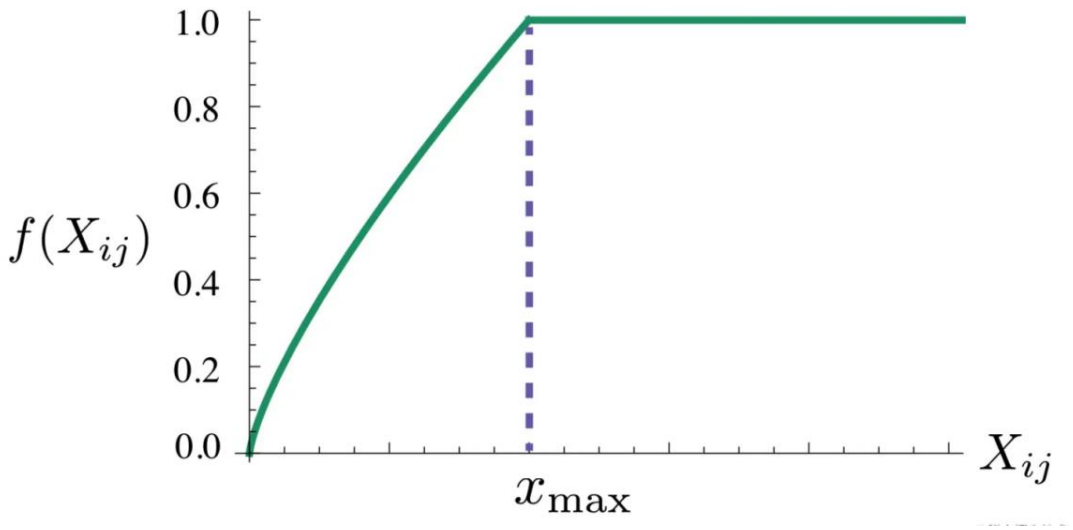


Figure 4. 4: Weight of the loss function

The objective function is optimized by the stochastic gradient descent method, and the non-zero items in the word co-occurrence matrix X are randomly selected for optimization. X is a sparse matrix, Glove usually optimizes faster than Word2Vec, because each pair (center word, context word) of the corpus in Word2Vec is a training sample, and the number of samples is large.

4.2.2 Long Short-Term Memory model

Recurrent Neural Network (RNN) is a special kind of artificial neural network that was developed for the exclusive purpose of processing sequential data, such that found in text or time series. RNNs feature a recurrent connection that enables them to keep a hidden state, sometimes known as a memory, of past inputs. This contrasts with standard feedforward neural networks, which analyze input data in a single forward pass. RNNs can recognize relationships and patterns concealed within sequential data because to this hidden state. The capability of RNN to receive inputs of varying duration and effectively manage sequential data is the RNN's defining characteristic. The RNN will take an input at each stage of the process, combine it with the newly learned information about the hidden state, and then create an output while simultaneously updating the information about the hidden state. Since RNNs are recurrent, they can store knowledge gleaned from previous inputs and use this information to guide or direct subsequent predictions or outputs.

RNNs have found widespread use in a variety of tasks relating to natural language processing, including language modeling, machine translation, sentiment analysis, and voice recognition. RNNs, on the other hand, are plagued by an issue known as "vanishing gradient," which hinders their capacity to accurately capture long-term relationships. Variations of RNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), have been developed as a solution to this problem. These RNNs contain gating mechanisms to better regulate the flow of information across the network and alleviate the vanishing gradient problem. LSTM stands for "Long Short-Term Memory," while GRU stands for "Gated Recurrent Unit."

The term "vanishing gradient" refers to the computation and backpropagation process that takes place during training. At each step, the gradient either becomes flatter or steeper. Over time, the gradient may converge to zero, leading to vanishing gradients, or diverge to infinity, resulting in exploding gradients. In other words, the challenge with long-term dependency arises because, as the time interval increases, the RNN loses its ability to connect to information that is further away.

As depicted in figure 4.5, the memory ht at time t may lose the ability to capture information related to time 0 as the time point t expands. This occurs because the time gap between time t and time 0 becomes relatively large when the time interval between them is significant. Consider the example where the input at X0 is "I live in Hyderabad," and additional words are subsequently added, leading to the input at Xt as "I work in the municipal government." Since X0 is located far away from Xt, when the RNN processes Xt, the memory ht at that moment has lost the information stored at X0. Consequently, the neural network at time Xt is unable to comprehend in which city's municipal government the individual is functioning.

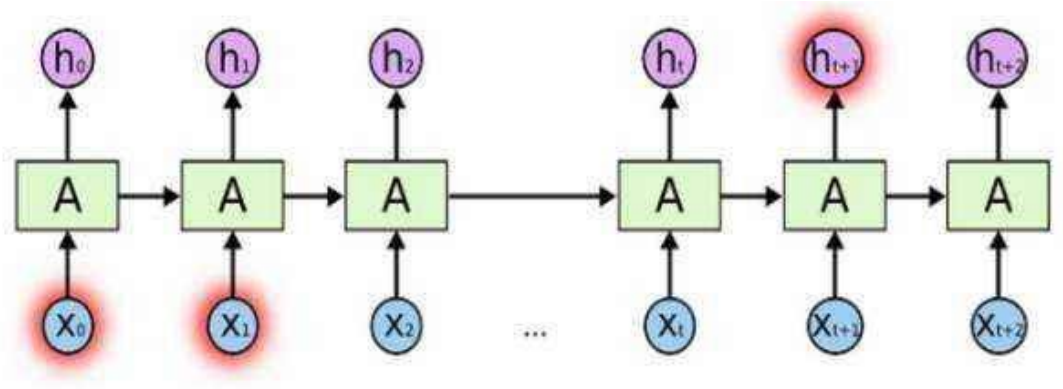


Figure 4. 5: Architecture of RNN

In principle, RNNs should be able to deal with dependencies that are held over a longer period of time. It is feasible to address the essential aspects of these problems if attention is used while picking the relevant parameters. In actual operations, however, RNNs have a difficult time properly capturing and using this information. The Long Short-Term Memory, sometimes known as the LSTM, was developed as a solution to the problem of long-term reliance. The LSTM was developed specifically to address this issue by purposely structuring its architecture in a way that sidesteps the problem.

LSTMs, as opposed to more conventional RNNs, are innately endowed with the capacity to recall information for protracted periods of time. This capability is inherent to their architecture, therefore acquiring it does not call for a significant amount of work on the user's part. Each RNN may be seen as a collection of individual neural network modules that are linked together. The recurrent module of a regular RNN is often quite straightforward and straightforward, frequently consisting of a simple structure like as a tanh layer.

LSTMs, on the other hand, use a different strategy. They have more complex processes, including as memory cells and gating mechanisms, which enable them to remember or forget information in a selective manner[137]. These mechanisms allow them to assimilate knowledge. Because of their purposefully designed architecture, LSTMs can successfully capture and remember long-term dependencies, which makes them a good choice for jobs that require sequential data processing. Figure 4.6 shows LSTM Cells.

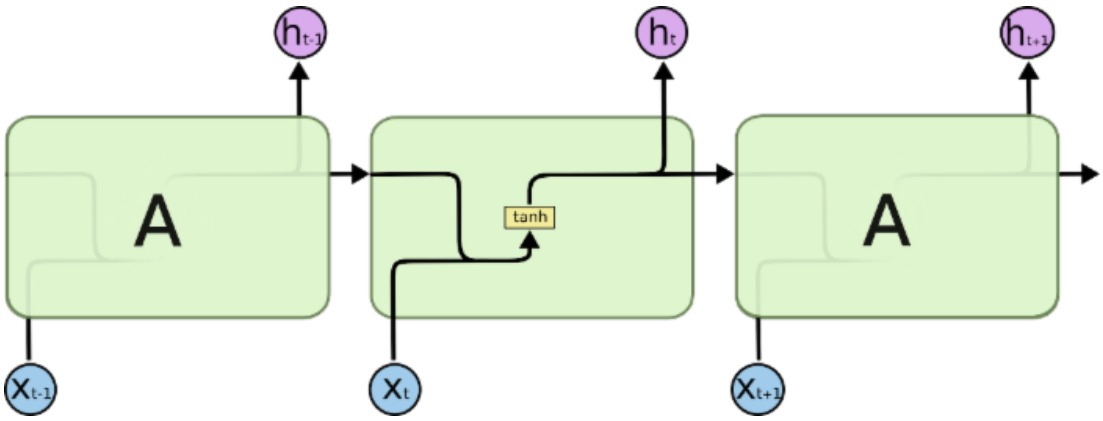


Figure 4. 6: LSTM Cells

The LSTM always has the same structure, however the modules that are repeated always have a different structure. In this example, there are four layers of the neural network, each of which interacts with the others in a very particular manner. Figure 4.7 shows LSTM cell structure.

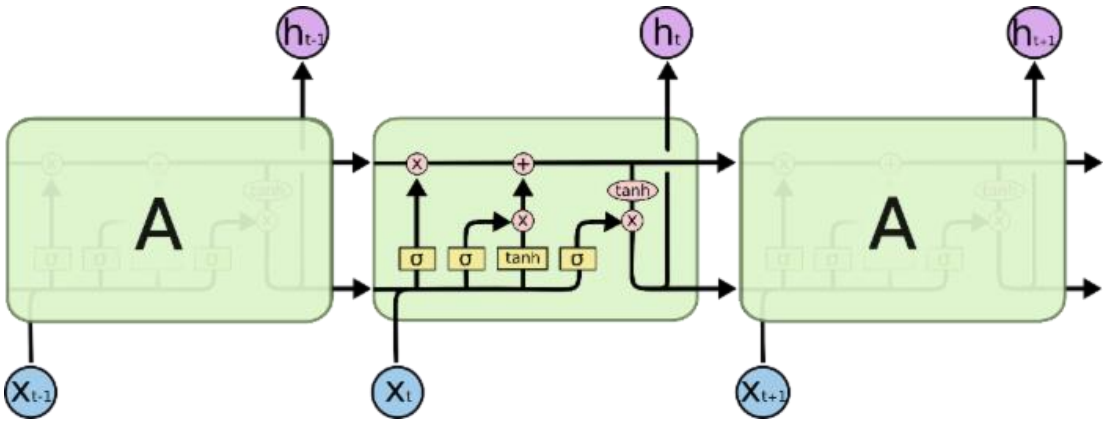


Figure 4. 7: LSTM cell structure

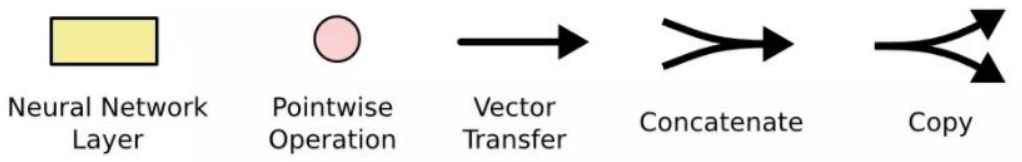


Figure 4. 8: LSTM cell components

Figure 4.8 shows LSTM cell components. The information that travels through a LSTM network is shown in graphical form above in the figure. Each individual black line denotes the movement of a complete vector from the output of one node in the

network to the input of the next node in the network. This movement is shown as arrows moving from left to right throughout the diagram. While the yellow matrix shows the layer of the neural network that has been taught, the pink circle indicates pointwise operations such as vector summation. The presence of lines that intersect one another denotes a connection between vectors, but the presence of lines that branch out from one another denotes the duplication of information that is subsequently dispersed to various locations.

There are three key components that make up each memory cell, which are indicated by the letter "A" in the figure 4.7. These components are the forget gate, the input gate, and the output gate. In addition, there is a state of the cell that is designated by the letter "Ct." These gate structures are responsible for either deleting information from or adding information to the state of the cell, and they make it possible for selective transmission of information.

When a cell transitions to a new state, the forget gate decides whether information from the previous state should be ignored or forgotten. It does so by analyzing the current input and determining which aspects of the prior cell state are not relevant to the calculation that is currently being performed. The quantity of fresh information that is added to the cell state is under the control of the input gate, which controls the gate. It analyzes the new data coming in and decides which of the new pieces of information should be saved while also selectively updating the cell state. Finally, the output gate is responsible for controlling how information travels from the cell state to the LSTM's output. It establishes which aspects of the cell state need to be used in order to create the output at the present time step.

Within an LSTM network, these gate structures and the cell state provide the network the ability to selectively preserve vital information over lengthy durations while rejecting irrelevant or stale information. Because of this technique, LSTMs can successfully collect and make use of long-term dependencies even while performing sequential data processing tasks.

1. Cell state (Ct)

The information stored in memory at time t is accessed to preserve crucial data. The information points that we have learnt in the past are saved in it, just as they are in our notebook. The horizontal line in the picture below passes through the top of the figure and runs straight on the whole chain. Because of this, it is simple for information to flow throughout the chain while being unmodified. Figure 4.9 shows the cell state

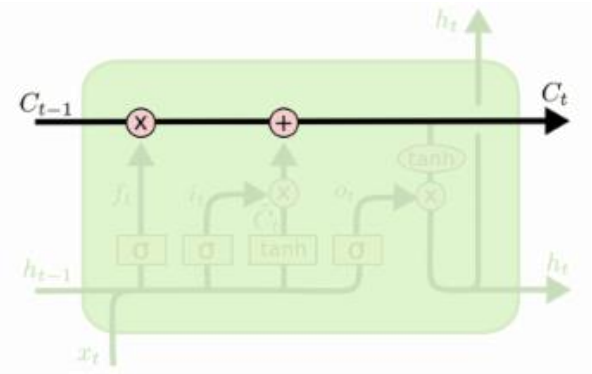


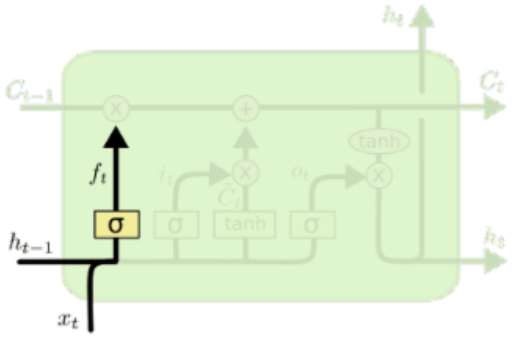
Figure 4. 9: Cell state

2. Forgotten Gate

The forgetting of the content in the cell state of the previous layer is controlled by considering the previous hidden state (ht-1) and the current input (Xt) as inputs. The purpose is to determine which contents from the previous cell state should be forgotten and which should be retained. To achieve this, a forget gate is employed, which utilizes the sigmoid activation function. The sigmoid function is chosen because it can produce values close to 0 or 1, which correspond to complete forgetting or complete remembering of each value in the vector input. Unlike other activation functions, such as the step function which has a gradient of 0 everywhere, the sigmoid function allows for the computation of gradients during the training process.

It is important to note that the input to the forget gate is in vector form. Each element of the vector corresponds to a specific value that needs to be evaluated for forgetting or retention. By using the sigmoid activation function, the forget gate can selectively determine the importance of each element and control the forgetting process

accordingly. It is worth mentioning that while other neural networks may allow for modification of the activation function, it is not recommended to change the activation function of the LSTM. The sigmoid function serves a crucial role in the functionality of the forget gate and altering it may negatively impact the network's performance. Figure 4.10 shows the Forget gate.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figure 4. 10: Forget gate

In the context of a language model, the cell state in an LSTM network may capture essential information relating to the subject being addressed, such as whether it contains a single or plural form. For example, "discussion" is a singular version of "discussion." For instance, if the current subject is "Manasa" and the input is "students," the forget gate of the LSTM network will be activated, which will cause it to forget information related with "Manasa" and the singleton subject. This will occur if the input is "students." The reason for this is because the input of "students" contrasts with the topic that is now being discussed, which is "Manasa," suggesting a transition from singular to plural.

3. Input gate

There are two separate stages involved in the process of updating the cell state in an LSTM. First, the information that represents the current position in the sequence, which is represented by the input, is examined to locate the pertinent data that has to be updated. After this information has been figured out, it is then converted into a format that is appropriate so that it may be put to the cell state. This transition is made possible using the input gate, which consists of a sigmoid layer that assists in determining which incoming information should be assimilated into the current state

of the cell. Applying the tanh function to produce a new candidate vector is the next step that has to be taken. The outputs of the first phase are efficiently used by LSTM to decide the precise information that should be added to the cell state. This is accomplished by translating all the information into a form that is compatible with its addition to the cell state. Figure 4.11 shows Input gate.

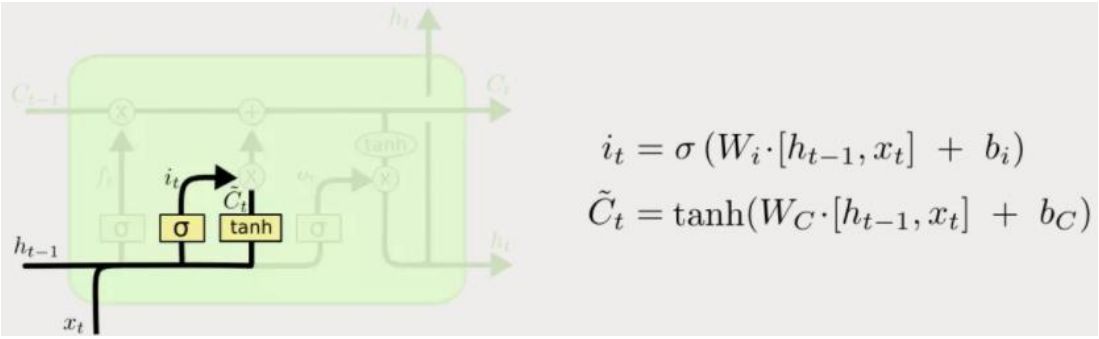


Figure 4. 11: Input gate

Because of the addition of the forget gate and the input gate, it is now possible to change the state of the cell from Ct1 to Ct. This capacity was not before had. The following graphic provides a visual representation of the procedure that is being described. The information that has to be discarded is denoted by the symbol ft x Ct1, and the information that has been most recently introduced is denoted by the symbol it x Ct. Through the incorporation of these gates, it is possible for the cell state to selectively preserve critical information while discarding irrelevant or out-of-date information. Figure 4.12 shows Change of state.

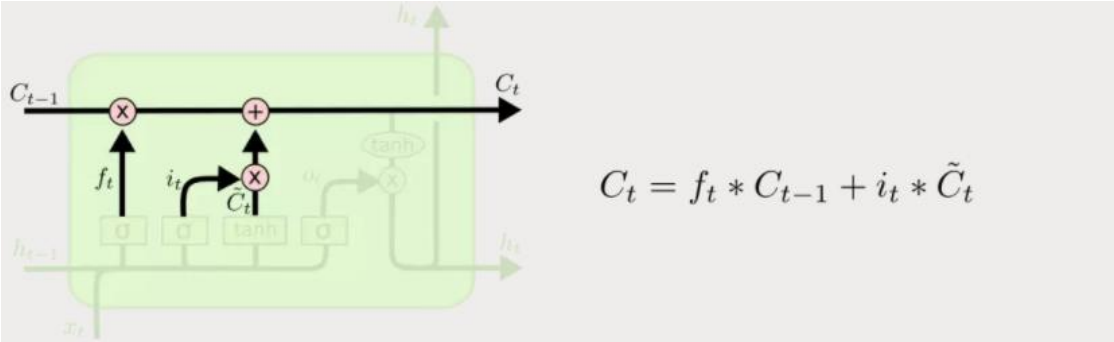


Figure 4. 12: Change of state

4. Output gate

After considering all factors, the responsibility of deciding what should be output, based on the current state of the cell, lies with the user. In other words, the content of the cell state can be selectively outputted. The output gate uses the sigmoid activation function to determine which portion of the information needs to be output, just like the updating process of the two components of the input gate. Additionally, it processes the contents of the cell state using the tanh activation function. It is worth noting that each value of Ct obtained from the calculation corresponds to a different activation function. If any value falls outside the range of -1~1 to be processed by the tanh function, modification is necessary. By multiplying these two components, the desired output component is obtained. Figure 4.13 shows Output gate.

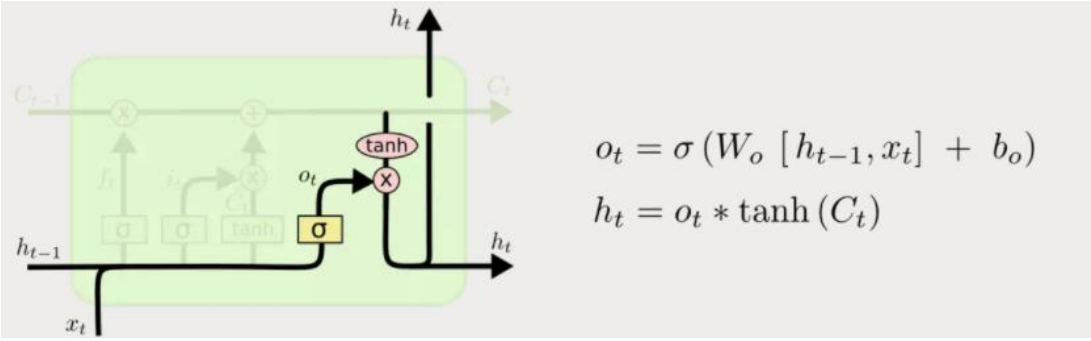


Figure 4. 13: Output gate

In the language model, the cell state is where a lot of relevant information is stored. This information covers a variety of elements, including the identification of a solitary subject, an indication of the past tense, an attribution of male gender, and more. When the input is about a subject, it is reasonable to assume that information about verbs will be required for the output. To be more specific, the goal at this level is to produce just the singular form and tense, even if the gender of the subject is not specifically mentioned in the output. Because of this, the model is able to recognize the shift in the part of speech of the verb even if it does not explicitly output the gender of the subject.

4.2.3 Bidirectional LSTM

Bidirectional Long Short-Term Memory (Bi-LSTM) is AI model that processes sequences of data in both the forward and backward directions. It is often employed in

language modelling because of its ability to do these tasks in both ways. A sequence may be processed by a regular LSTM network in either the left-to-right or the right-to-left direction, but a Bi-LSTM network can process a sequence in both directions at the same time. since a consequence of this, the network will be able to extract more contextual information from the sequence, since information coming from both directions will be able to be utilised in the prediction process.

The input sequence is split up and fed into two different LSTM layers in a BiLSTM. One of these LSTM layers processes the sequence from left to right, while the other processes it from right to left. Each layer of the LSTM computes separately until it reaches the output layer, which is where all the layers' weights and biases are concatenated into a single value. The output of the BiLSTM is a mixture of the outputs from both LSTM layers, forward and backward.

BiLSTMs have found widespread use in language modeling due to its capacity to accurately represent the intricate connections that exist between the words in a phrase. A BiLSTMis able to take into consideration the context both before and after a given word since it processes a phrase in both directions. This may be especially helpful for tasks such as named entity identification and sentiment analysis. In addition, BiLSTMs may be used in combination with other methods, such as attention mechanisms, to enhance the degree of precision that language models possess. Figure 4.14 shows the proposed deep learning model.

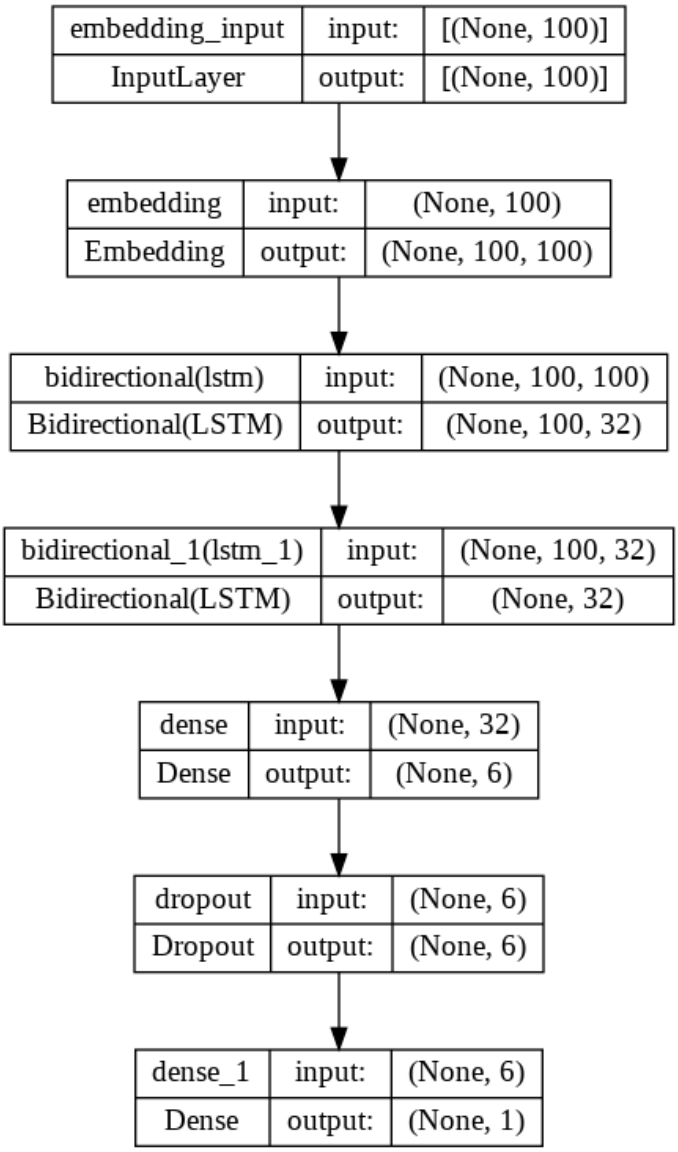


Figure 4. 14: Proposed deep learning model

4.2.4 Convolutional Neural Network

A one-dimensional convolutional neural network (1D CNN) is a specialized neural network architecture designed specifically for processing one-dimensional input data. It is commonly used for analyzing sequential data like time series, audio signals, and text. Unlike standard CNNs used for image processing, 1D CNNs operate on a single dimension, making them well-suited for sequential data analysis.

A 1D CNN consists of convolutional layers, activation functions, pooling layers, and fully connected layers[138]. Convolutional layers employ filters to scan and identify

local patterns and features in the input data. Each filter performs a convolution operation, calculating a dot product with a local portion of the input and generating a feature map. Activation functions, such as ReLU, introduce non-linearities to capture complex relationships in the data.

Pooling layers down sample the feature maps, reducing their dimensionality while preserving crucial information. Max pooling selects the maximum value within a local region, while average pooling computes the average value. Pooling helps abstract and summarize the features obtained from convolutional layers. After convolution and pooling, the resulting feature maps are typically flattened into a one-dimensional vector and passed through fully connected layers. These layers combine the extracted features and make predictions based on the specific task, such as classification or regression.

1D CNNs have demonstrated success in various applications, including voice recognition, emotion analysis, time series forecasting, and biological signal analysis. They excel at identifying local patterns and relationships in sequential data, making them valuable for evaluating and extracting relevant information from one-dimensional sequences.

Convolutional Layer

This is the initial layer that is used to extract the various features that are present in the input photographs. A mathematical operation known as convolution is performed at this layer, which is located in between the input image and a filter with a certain size denoted by the notation MxM. By sliding the filter over the input image, one may acquire the dot product (MxM) between the filter and the sections of the input picture with reference to the size of the filter. This can be done by saying that one moves the filter across the input image. The resulting document is called the Feature map, and it contains information about the image, such as the location of the picture's borders and corners. After that, this feature map is passed on to subsequent layers so that they may learn more features from the input image.

Pooling Layer

A Pooling Layer is typically added after a Convolutional Layer. This layer's main goal is to scale down the convolved feature map in order to conserve the required computer resources. This is accomplished one step at a time on each feature map, as well as by cutting down on the number of linkages between the layers. The technique that is used might result in a variety of different kinds of pooling activities being carried out. In its most basic form, it is a condensed version of the features that are generated by a convolution layer. The feature map provides the most contribution to Max Pooling and may be found there. The method known as "average pooling" is used to calculate the component averages contained inside an image segment of a certain size. Sum The entire sum of all of the sections' individual components is determined via the process of pooling. In most cases, the Pooling Layer will perform the function of a connection between the FC Layer and the Convolutional Layer.

By generalizing the information collected by the convolution layer, this CNN approach enables the networks to identify the features on their own. This helps reduce the number of computations that take place inside a network.

Dense Layer

The dense layer of a neural network is intimately linked to the other levels of the network. This is due to the fact that all of the neurons in the layer below transfer information to the neurons in the dense layer. This suggests that the thick layer can access information from all of the neurons that are located below it in the network. It has been found that the thick layer is used rather often in model construction owing to the increased accuracy that it provides. As a consequence of this, the size of the matrices and vectors that make up the background is increased by a factor of two because of the existence of the thick layer. Backpropagation is used to train and update the parameters necessary to create the values that are used in the matrix, and this training and updating may take place at any time. Backpropagation is also used to train and update the values that are used in the matrix. The most important impact that the thick layer has is a change in the dimensions of the vector, which ultimately

results in the production of a vector that has m dimensions. In addition, thick layers carry out other operations, such as translation, scaling, and rotation, on the vector. Figure 4.15 shows the proposed CNN model.

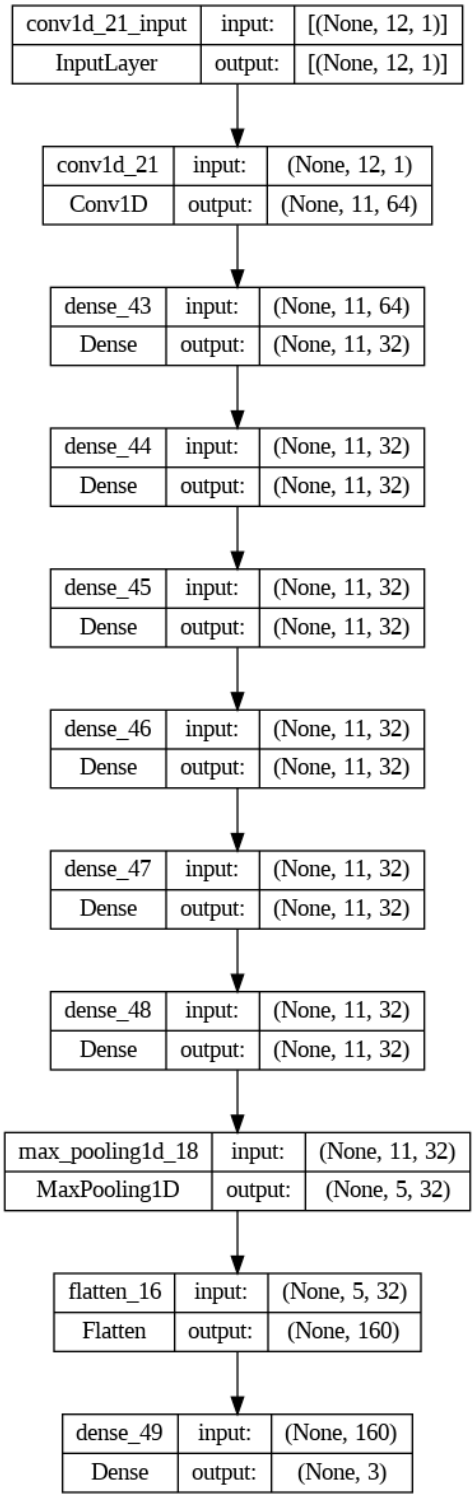


Figure 4. 15: Proposed CNN model

4.2.5 Twitter Spam Drift:

Twitter spam drift is the process through which the features and patterns of spam material on Twitter shift over time[139]. This is a word that is used to characterize the ever-changing patterns of behavior shown by spam on the site. Spammers are always innovating and refining their methods in order to circumvent spam detection systems and communicate with a larger audience. They may make use of a variety of ways, including sending out new kinds of spam messages, changing the way they fool people, or taking use of weaknesses in the Twitter network.

The unpredictable nature of spam on Twitter presents a problem for the algorithms and systems used to identify spam. It is possible that spam detection algorithms that have been trained on past spam data may become less successful if spammers add new patterns or approaches that depart from the typical behavior of spam. It is essential to regularly monitor and update spam detection systems in order to adjust to the changing features of spam on Twitter in order to effectively battle the spread of spam.

Twitter wants to deliver a cleaner and more trustworthy platform for its users by monitoring spam drift and keeping spam detection algorithms up to date. This will minimize the effect of spam and preserve the quality of content on the site.

4.2.6 Hate Speech Detection

Social media sites like Twitter have significantly impacted our daily lives in recent years. These platforms create relationships and make it possible for people all over the world to share and discuss their thoughts. On the other hand, due to the open nature of these platforms, the problem of hate speech has become more urgent in recent years. Hate speech is a huge problem since it not only endangers the health and safety of individuals but also deteriorates the cohesiveness of the social fabric that holds together online communities. Finding instances of hate speech on social media sites like Twitter and taking steps to remove it are very necessary steps in preserving a positive atmosphere online. This case study explores the use of deep learning methods

for the identification of hate speech on Twitter. It outlines the challenge, the motivation, the significance, and the future applications of these approaches.

Because of the rising prevalence of hate speech on Twitter, there is an immediate and pressing need for reliable detection techniques. The term "hate speech" refers to a variety of hurtful utterances, such as racial slurs, threats, discriminatory remarks, and insulting remarks directed at certain groups or persons. When there is such a large volume of user-generated material, it might be difficult to differentiate between speech that is motivated by hatred and expressions of lawful free speech. The constantly shifting nature of hate speech and the myriad of contextual subtleties render rule-based techniques, which have been used for decades, often ineffective[140].

The goal to make public internet areas more secure is at the heart of what prompted the creation of this case study. In addition to serving as venues for social contact, the various social media platforms also function as conduits for the distribution of information and the conduct of public dialogue. Hate speech has the potential to impede productive conversations and discourage people from participating on online forums. The goal of building strong models for detecting hate speech is to provide platform administrators and users alike the ability to quickly identify and eliminate harmful information, which will ultimately lead to the creation of an online environment that is courteous and welcoming to all users.

Finding instances of hate speech on Twitter is of the utmost significance for a number of different reasons. To begin, it protects the health and safety of users by guaranteeing that people are able to freely express themselves without the risk of being harassed or harmed. Second, it maintains the reputation of the platform and the confidence of its users, since addressing hate speech displays a commitment to the safety of the company's customers. Third, the identification of hate speech contributes to conformity with legal and ethical norms, which in turn reduces the likelihood of incurring legal obligations. Lastly, it helps contribute to a larger social purpose of combating online hatred and encouraging digital diversity. This is an important aspect of the digital age.

Deep learning has a wide range of potential applications, one of which is the identification of hate speech on Twitter. It is a tool that may be used by academics, content moderators, and social media companies to proactively detect and combat hate speech. These models may also be used as instructional tools, guiding users toward a better understanding of the parameters that define courteous online conversation. In addition, the technology may be used to assist in the production of awareness campaigns and the establishment of policies by offering data-driven insights about the frequency of hate speech on the platform as well as the patterns associated with it. The application's scope is multifaceted and comprehensive, including a wide range of parties involved in the effort to combat hatred expressed online.

4.3 Experimental Results

This section focuses on the identification of spam tweets within the Twitter network. We will present the outcomes of our experiments, which involved evaluating different methods using the Twitter dataset as a benchmark. Additionally, we will provide a concise overview of the evaluation metrics employed in this study. Subsequently, we will conduct a detailed analysis of the data collected from each methodology and present our findings.

During the investigation, our primary objective was to develop effective techniques for detecting and classifying spam tweets on Twitter. We performed rigorous experiments, comparing various approaches and assessing their performance against the Twitter dataset. The dataset served as a reliable reference point for evaluating the effectiveness of each method.

We used an assessment criterion to evaluate the accuracy, precision, and recall of the various spam detection strategies. This allowed us to guarantee that the review was as complete as possible. The use of these indicators allowed for a systematic and quantitative analysis of the effectiveness of the strategies. After the assessment, we moved on to the next step, which was to conduct an analysis of the data acquired using each approach. In order to do this, we had to investigate the characteristics and

patterns that were retrieved from the tweets. For example, we looked for tweets that had questionable links, an excessive number of hashtags, or text that was repetitious. We conducted in-depth research on the performance of each method, considering important aspects such as total detection accuracy as well as false positives and false negatives.

Our results give information on the efficacy of the approaches that were examined in terms of spotting spam tweets and differentiating them from real tweets. We were able to acquire useful insights into the strengths and shortcomings of each technique by combining the study of the assessment criteria with the extensive investigation of the data that was gathered. These insights will add to the continuing work to prevent spam and enhance the general quality and dependability of material on the Twitter platform. Those efforts will be improved thanks to these insights.

4.3.1 Datasets

Dataset 1: UtkML's Twitter Spam Detection dataset

This dataset contains a comprehensive collection of labeled tweets that may be used to train and evaluate spam detection models. It was maintained by the Utkal University Machine Learning (UtkML) research group. The dataset used for Twitter Spam Detection contains a wide variety of tweets, including tweets that are considered spam as well as tweets that are considered to be valid. The text and attributes of each tweet in the dataset are evaluated to determine whether or not it should be classified as "spam" or "ham" (non-spam). For optimal training of machine learning models, it is essential to have a dataset that is intended to be balanced. This ensures that the dataset has an equal number of tweets that are spam and tweets that are not spam.

Dataset 2: Social Honeypot Dataset

The Social Honeypot Dataset is a comprehensive collection of data that may be exploited for the purpose of researching and evaluating a wide variety of illegal acts, such as spamming, phishing, fraud, and other dishonest behaviors. This dataset was gathered via the use of social honeypots, which allow researchers to draw the

attention of harmful actors and capture their interactions, messages, and other important data by establishing accounts that imitate legitimate users. This was accomplished by using accounts that mimicked real users.

The collection contains a broad variety of information, including user profiles, messages, URLs, photos, and the metadata connected with the interactions between users. Researchers now have the ability, thanks to the wealth of data that is already accessible, to investigate the features, patterns, and techniques that spammers and other harmful actors adopt in their efforts to mislead or exploit users of social media platforms.

A total of 2,353,473 tweets are connected to the 22,223 accounts that are included in the dataset. Additionally, the number of followers that each of these accounts has fluctuated throughout the course of a certain period of time. In addition to this, it includes 19,276 real users who are included in the dataset, as well as the total number of tweets that these users have made and the number of followers that they have accrued over the course of time.

4.3.2 Performance Metrics

Model evaluation in machine learning usually includes evaluating performance using metrics like accuracy and loss.

Accuracy Plot: Accuracy plot demonstrates how the model's accuracy varies across training epochs.

Loss Plot: A loss plot shows how the model's error, or loss, varies over training epochs. A declining loss suggests that the model is improving its ability to predict outcomes.

4.3.3 Spam Detection using Long Short Term Memory Model

In the first phase of the process, linguistic characteristics are extracted from the text of tweets using the GLoVe language model. Following that, the LSTM deep learning model will utilize these features to determine whether or not messages are spam. GLoVe is able to simplify the process of learning word embeddings, which is accomplished by using a co-occurrence matrix of words found within a corpus. This

matrix maintains a record of the frequency with which certain word combinations occur together in the same setting, such as inside a sentence or paragraph. The GLoVe model takes this matrix as input for further processing.

GLoVe is a model that produces word embeddings in several NLP applications. Because of their valuable capacity to capture both semantic and syntactic information about words, these word embeddings are often utilized as input in neural networks for the aforementioned tasks. Word embeddings are prized for their ability to capture both semantic and syntactic information about words. When compared to other widely used language models, such as word2vec, the performance of GLoVe has been shown to be better in a number of different contexts.

In addition, GLoVe has been used to build pre-trained word embeddings, which customers may then download and use in a variety of natural language processing applications of their choosing. Within the framework of the LSTM model, certain lexical features are taken into consideration as sources of data input. Figure 4.16 is an illustration of the training results that were achieved using the LSTM model.

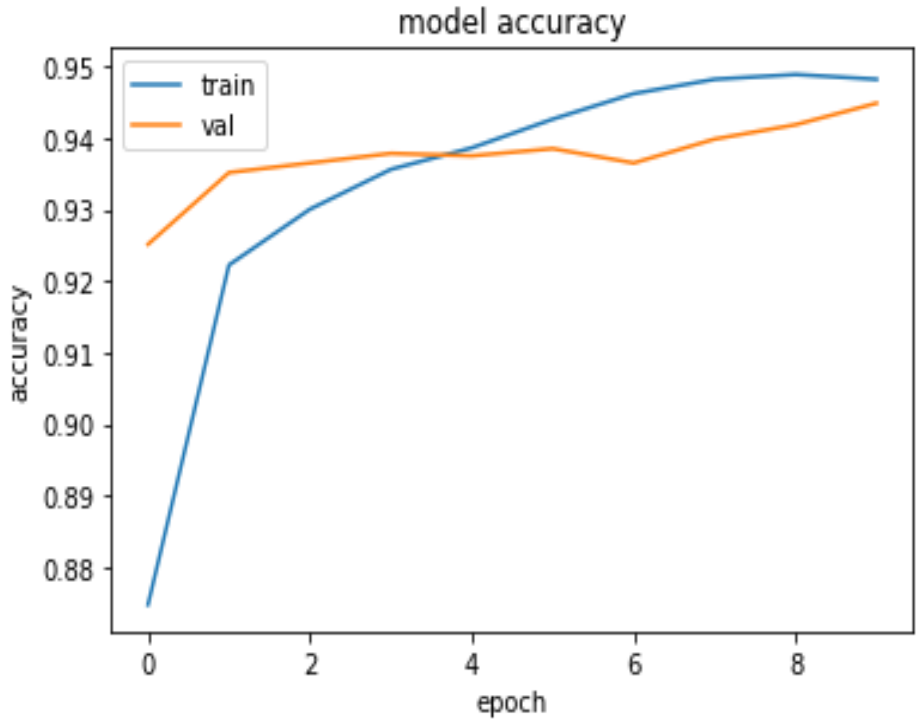


Figure 4. 16: LSTM Model Performance Plots

Figure 4.16 presents a visual depiction of the accuracy plot, which offers a visual picture of the performance of the LSTM model during the process of training and validating the model. This demonstrates the high degree of precision that the model can reach. In most cases, the y-axis indicates the level of accuracy achieved by the model, while the x-axis indicates the total number of training epochs or iterations. This visual representation provides an illustration of the model's performance with regard to its level of accuracy.

The graphic presents two separate curves: one for the accuracy of the training data, and another for the accuracy of the validation data. The accuracy of the model with respect to the training data is reflected along the training accuracy curve for each epoch or iteration. This curve tends to begin at a relatively low value and gradually grow as the model learns and improves its fit to the training data. Initially, this curve tends to start at a relatively low value.

On the other hand, the validation accuracy curve illustrates how accurate the model is using a distinct validation set for each epoch or iteration. The validation set is used in order to evaluate how well the model performs on data that it has not previously seen and to avoid overfitting. It is very uncommon for the validation accuracy curve to follow a pattern that is similar to the training accuracy curve, initially growing along with it. This is because both curves are intended to reflect the same level of precision. However, if the model starts to overfit the training data, there may come a time when the validation accuracy curve starts to fall. This might happen at a certain point.

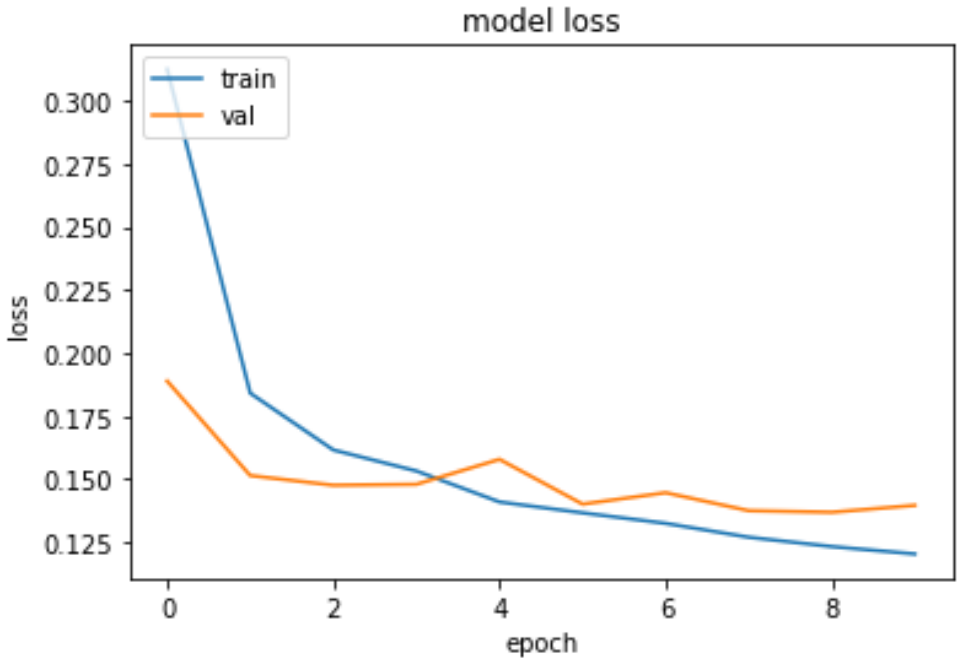


Figure 4. 17: LSTM Model Loss Plot

The y-axis in Figure 4.17 indicates the loss of the model, while the x-axis relates to the number of training epochs or iterations that have been successfully completed. As seen on the y-axis of the graph, the amount of information that is either ignored by the model or not captured by it is shown to be rather extensive. Loss is a statistic that indicates how well a model is able to make predictions by using the information that it has learned from its training data. It provides a numerical measure of the difference between the values that were expected and the values that were actually obtained from the experiment. The primary goal of the training process is to improve the model's prediction skills, therefore seeing a reduction in loss is an indication that the model is making progress in this regard. The training has to be continued until the loss reaches the level that is considered successful.

The LSTM (Long Short-Term Memory) architecture is highly recommended since it was developed to manage sequential data, which is typical in Natural Language Processing (NLP) jobs. Because this model is able to recognize relationships and patterns across time, it is well-suited for applications such as language modeling, voice recognition, and machine translation. A number of studies have shown that

LSTM models are a useful tool for modeling the complicated sequences of data seen on Twitter.

4.3.4 Spam Detection using Convolutional Neural Network

When categorizing tweets, the CNN model considers a variety of different meta-heuristic criteria, such as the length of the tweet, the amount of question marks, and the existence of tags. By training the model using backpropagation and stochastic gradient descent, the performance of the model may be increased while simultaneously minimizing the loss function. Testing the model with data that is distinct from the data it was trained on is necessary in order to assess how well it generalizes. The CNN model successfully categorizes Twitter data, capturing essential characteristics and trends by using word embeddings and detecting the hierarchical structure of the text. This is accomplished by exploiting word embeddings. Figures 4.18 and 4.19 respectively exhibit the accuracy and loss graphs of the CNN model.

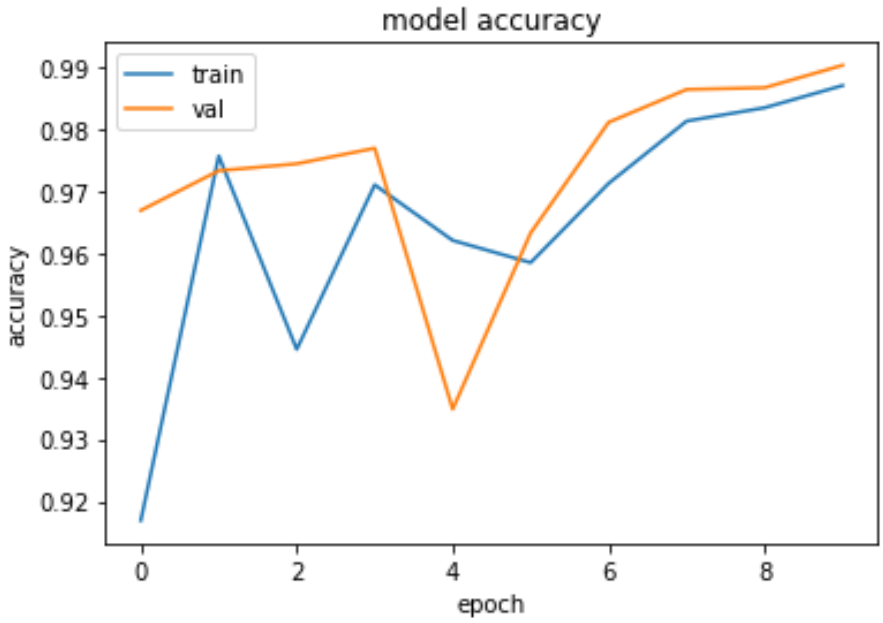


Figure 4. 18: CNN Modell Accuracy Plot

Figure 4.18 displays the outcomes of the performance evaluations that were conducted during the model's training and validation phases. The degree to which the model accurately depicts the real environment is shown on the y-axis. Accuracy refers

to a model's ability to correctly classify test data. In order to improve the model's ability to make more accurate predictions, training is done to obtain the highest level of prospective accuracy. The number of epochs or iterations that the model has completed throughout the training process is shown along the x-axis of the plot. This information may be gleaned from the history of the model. Iterations are single adjustments that are made to the model's parameters based on a collection of training samples. These adjustments are made in order to build upon the findings of the iteration that came before them. Epochs, on the other hand, are representations of iterations that span the whole of the training dataset, beginning with its inception and ending with its completion.

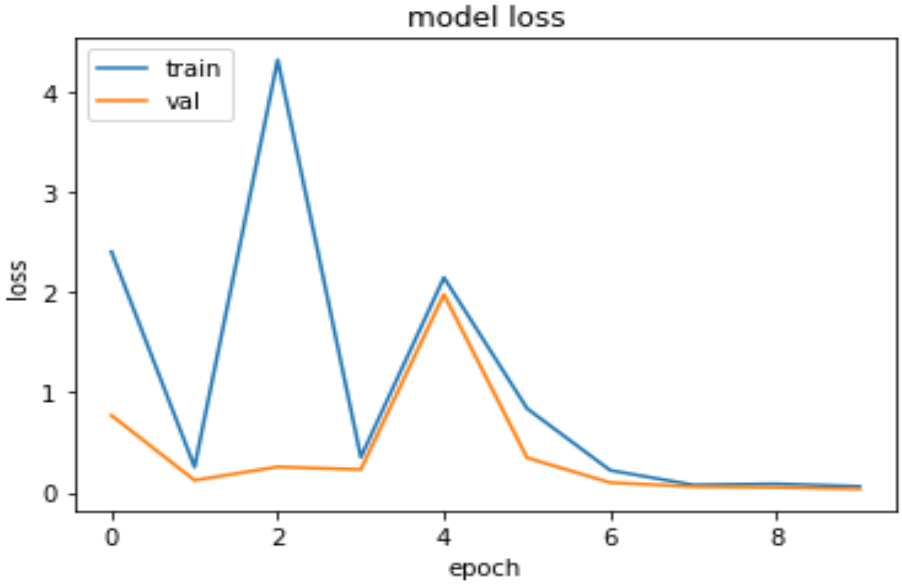


Figure 4. 19: CNN Model Loss Plot

Figure 4.19 displays the results of the performance evaluations that were carried out during the model's training and validation phases.

4.3.5 Spam detection using Tweet and Vocabulary features

The combination of CNN and LSTM models is a successful strategy that has the potential to increase text categorization accuracy dramatically. It is feasible to reduce the limits of each individual model and increase the overall performance of the models when they are combined using the strengths that both the LSTM and CNN models bring to the table. Every model comes with both positives and negatives

specific to it. While LSTM models are excellent at discovering temporal correlations in text data, CNN models perform well when it comes to locating local features and patterns. We are able to obtain a more complete comprehension of the text data and create more accurate predictions if we use the findings of many models. Table 4.2 illustrates how accurate this integrated model is when taken as a whole.

Table 4. 2: Proposed model performance evaluation

	Precision	Recall	F1-Score
Spam	0.99	0.99	0.99
Non-Spam	0.99	0.99	0.99
Accuracy	99%		

In this method, a multi-input model is used, and the LSTM model and the CNN model each conduct their own analysis of the text data. The results of these analyses are combined and then used. The results obtained from each of these models are then combined before being sent into the ultimate classifier. The frameworks for deep learning include a variety of different methods for combining or concatenating these layers. Table 4.3 contains the findings that were obtained by analyzing the suggested model in light of the most recent findings from related studies.

Table 4. 3: Comparison results

Spam Detection models	**Accuracy**
Bag of words + LSTM	92.7%
TF-IDF + LSTM	95%
BERT	97.1%
Proposed model	99%

An accuracy of 92.7% was achieved as a consequence of the employment of LSTM technology in conjunction with Bag of Words feature extraction. The TF-IDF

algorithm and the LSTM were able to reach an accuracy of 95%, respectively. The accuracy achieved by the BERT language model was 97.1% overall. An accuracy of 99% was reached by using the proposed model.

4.3.6 Hate Speech Detection in Twitter

The proposed model identifies hate speech accurately.

Figure 4. 20: Twitter hate speech example 1

Tweet: It's a good ass fight when you gotta run out with ps4 controller

Result: The tweet is categorized as "hate and abusive" with a probability of 0.9023 because of the words and phrases "good ass fight gotta run".

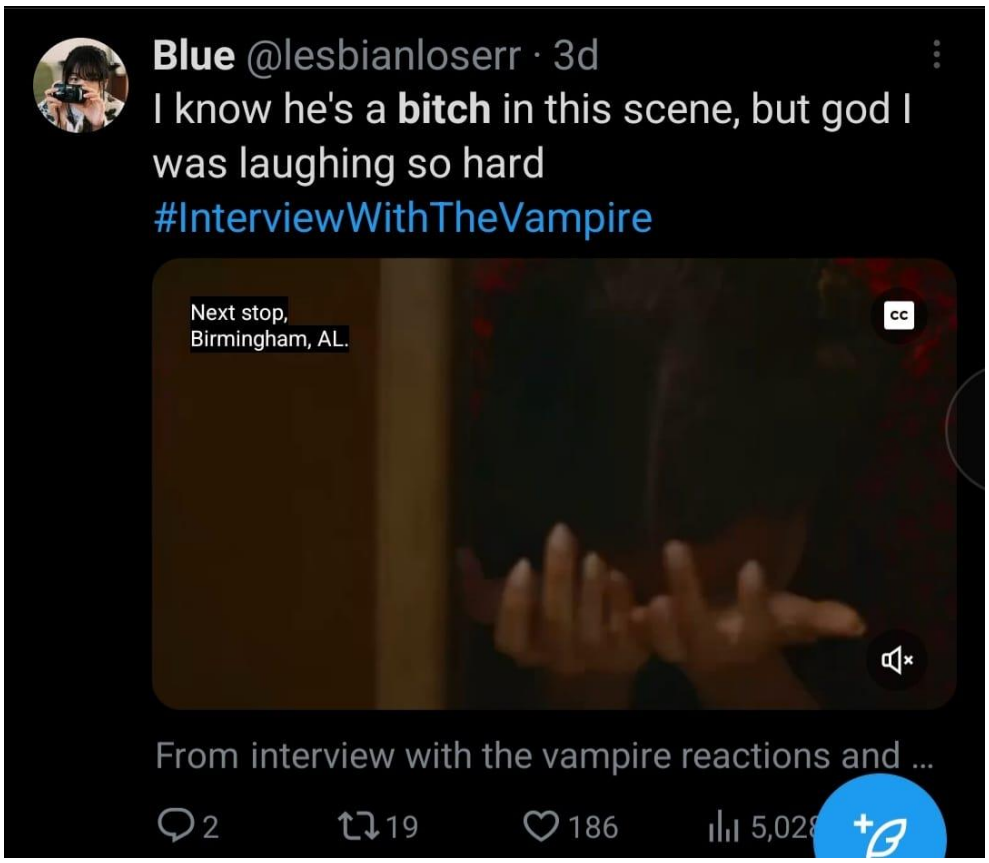


Figure 4. 21: Twitter hate speech example 2

Tweet: I know he is a bitch in this scene, but God I was laughing so hard

Result: The tweet is categorized as “hate and abusive” with a probability of 0.9833 because of the words and phrases “know bitch scene”.

Figure 4. 22: Twitter hate speech example 3

Tweet: Saying fuck the nazis isn’t same thing as saying fuck the jews and never let Elon Musk convince you otherwise

Result: The tweet is categorized as “hate and abusive” with a probability of 0.9994 because of the words and phrases “fuck nazis isnt thing sying fuck jews”.

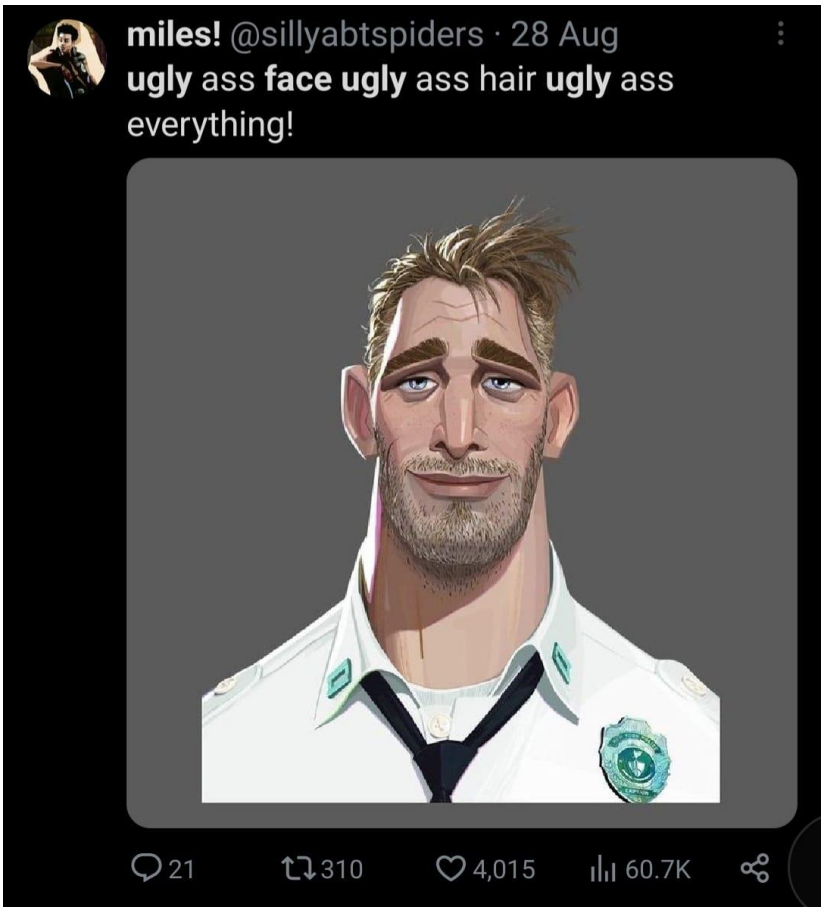


Figure 4. 23: Twitter hate speech example 4

Tweet: ugly ass face ugly ass hair ugly ass everything

Result: The tweet is categorized as “hate and abusive” with a probability of 0.9981 because of the words and phrases “ugly ass face ugly ass hair ugly ass everything”.

4.4 Conclusion

Deep learning techniques have gained more attention in recent years as a way to combat spam on Twitter, and this tactic has proven to be effective. The deployment of a deep learning-powered spam detection system by Twitter has the potential to significantly enhance user experience, conserve time and resources, boost security and confidence, and guarantee adherence to applicable laws and regulations. A model for

identifying spam on Twitter that is based on deep learning is presented in this research. As part of its input dataset, the model takes into consideration not just individual tweets but also other information, such as the number of followers and activities. The dataset is required for both of the model's portions in order to ensure that the model is able to work appropriately.

The content of the tweets is the primary focus of the first stage, which involves the use of a GLoVe language model for the purpose of extracting lexical features from the tweets. The features are then retrieved and fed into an LSTM deep learning model to facilitate spam identification. In the second stage, a CNN model is used to the task of classifying tweets. This model considers a number of extra meta-heuristic variables in addition to the embedded information in the tweets. A tweet's total character count and whether or not it contains asterisks are two examples of these supplementary characteristics. When the data from the LSTM model and those from the CNN model are combined, a more comprehensive set of findings may be obtained, which can then be used to derive the ultimate conclusion. The tweet data was successfully categorized with an accuracy of 99% by using the proposed model.

Chapter-5

Conclusion

The issues caused by spam on Twitter can be solved with the help of AI-based tweet spam identification. Through the utilization of advanced deep learning algorithms, this research has demonstrated the potential for effectively identifying and filtering out spammy content from the vast number of tweets generated on the platform. The findings of this thesis emphasize the importance of leveraging AI techniques, such as machine learning and deep learning, in addressing the challenges posed by tweet spam. Traditional rule-based approaches and keyword filtering methods often fall short in keeping up with the constantly evolving nature of spam, making AI-based solutions a necessity.

The implementation of deep learning models, such as RNNs and CNNs, has showcased their capability to capture complex patterns, semantic cues, and contextual information from tweet content. The integration of additional metadata, including user information and engagement metrics, has further enhanced the accuracy and robustness of the spam detection system.

The proposed approach for detecting tweet spam involves utilizing a swarm optimization strategy on an individual tweet basis. Using a dataset created specifically for recognising spam tweets, a machine learning model is built. Metaheuristic features are produced from the input features in the dataset. The WOA approach is used before the classification procedure to identify the relevant properties for classification. The SGD algorithm, a modified version of the standard WOA objective function, is used during the feature selection process. By using the chosen subset of features, the Adaboost classifier is trained to identify spam in tweets. The best results are produced when WOA, SGD, and the Adaboost classifier are combined. During testing, an excellent accuracy of 99.85% was achieved utilising a small subset of only seven features and in a prompt period of 17.9 seconds.

Deep learning-based spam detection on Twitter might enhance user experience, conserve time and resources, improve security and trust, and guarantee compliance to

pertinent laws and regulations. This work presented a deep learning-based strategy for identifying tweet spam. Tweets and other meta data, such as the number of followers and activities, are included in the input dataset that the model processes. Two parts of the model, which use this dataset as input, are divided. In the initial step, the focus is on the content of the tweets. A GLoVe language model is utilized to extract lexical features from the tweet content. In the second phase, the tweets are categorised using a CNN model that contains the embedded meta data and extra meta-heuristic properties, and these features are then input into an LSTM deep learning model to identify spam. These characteristics encompass factors like tweet length and the presence of question marks. The combination of these features aids in the classification process.

The ultimate final result is obtained by consolidating the findings from both the LSTM and CNN models into a unified set of results. The proposed model demonstrated remarkable accuracy, achieving a classification accuracy of 99% for the tweet data. This approach not only improves spam detection but also contributes to a more secure and trustworthy Twitter environment, ensuring compliance with regulations and providing a better overall user experience.

Here are some of the research's major contributions:

1. **Whale Optimization Algorithm for Feature Selection**:

One of the novel aspects is the utilization of the Whale Optimization Algorithm for feature selection. While many existing works in this field employ traditional feature selection methods or rely solely on deep learning models, this research introduces a metaheuristic approach to optimize the selection of tweet and user account features. The use of Whale Optimization Algorithm can enhance the efficiency and effectiveness of feature selection, potentially leading to improved model performance.

2. **Integration of Textual and User Account Features**:

The research integrates both tweet features and user account features. While some existing works focus exclusively on the content of the tweets, this approach recognizes the significance of user-related information, such as follower counts and

user behaviors. This comprehensive consideration of features provides a more holistic view of the data, which can lead to better hate speech detection accuracy.

3. **GloVe-Based Text Feature Extraction**:

The extraction of GloVe (Global Vectors for Word Representation) features from tweet text is another unique aspect. Many previous works may use simpler text representation techniques, but GloVe embeddings capture the semantic meaning of words in tweets, allowing for a more nuanced analysis of textual content. This can result in a deeper understanding of the language used in tweets and, consequently, more accurate hate speech detection.

4. **Combination of LSTM and CNN Models**:

The research combines two different deep learning models, namely the LSTM model for textual content analysis and a CNN model for information-based analysis. This hybrid approach leverages the strengths of both models to detect hate speech. Existing works often focus on a single model type, whereas this research harnesses the advantages to improve overall performance.

5. **Comprehensive Spam Detection**:

The final step involves the comprehensive integration of results from both the LSTM and CNN models. This dual-model fusion approach aims to achieve more accurate and robust hate speech detection. Existing works may not incorporate such a comprehensive combination of deep learning models for spam detection.

Future Scope

The future of AI-based Twitter spam detection holds great promise in tackling the persistent challenge of spam on the platform. As technology continues to advance, there are several areas where AI can further contribute to the detection and prevention of spam tweets, ensuring a safer and more enjoyable user experience. One aspect lies in the refinement of AI algorithms to achieve even higher accuracy rates in detecting spam. CNNs and RNNs are two examples of more sophisticated deep learning techniques that researchers can investigate to enhance the models' capacity to identify spam tweets more accurately. By continually refining the algorithms and training

them on large, diverse datasets, the accuracy of AI-based spam detection can be significantly enhanced.

Real-time detection is another crucial area for future advancements. The ability to identify and flag spam tweets in real time is essential in minimizing the spread of malicious content and protecting users from potential scams or harmful links. By employing faster processing techniques and optimizing the model's architecture, AI-based spam detection can be made more efficient and capable of swift action in response to emerging spam incidents on Twitter. A potential area for future development lies in the integration of contextual understanding into spam detection algorithms. AI models can be trained to consider various contextual factors, such as user behavior, relationship with followers, and the overall sentiment of the conversation, to better differentiate between genuine tweets and spam. AI-based systems that incorporate contextual information can better recognise spam tweets in different situations and react to the dynamic nature of spam.

Multilingual spam detection is another area with significant potential. Twitter is a global platform used by individuals from diverse linguistic backgrounds. Developing AI models that can effectively detect spam in multiple languages can help address the issue of spam across different regions and language communities. By training the models on multilingual datasets and incorporating language-specific features, AI-based spam detection can become more inclusive and effective in combating spam on a global scale. User feedback integration is a valuable aspect of future development. Users' comments on spam detection's precision and the ability to report spam can be used to improve AI models. By incorporating mechanisms for users to contribute to the training process and incorporating user feedback into the learning algorithms, AI-based spam detection systems can continuously learn and adapt to the evolving tactics used by spammers.

References

[1] Pasquini, Cecilia, Irene Amerini, and Giulia Boato. "Media forensics on social media platforms: a survey." EURASIP Journal on Information Security 2021, no. 1 (2021): 1-19.

[2] Yang, C., Harkreader, R., & Gu, G. (2013). Empirical evaluation and new design for fighting evolving Twitter spammers. IEEE Transactions on Information Forensics and Security, 8(8), 1280–1293

[3] Fazil M., M. Abulaish , A hybrid approach for detecting automated spammers in twitter, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018), pp. 2707–2719.

[4] Voorveld, Hilde AM, Guda Van Noort, Daniël G. Muntinga, and Fred Bronner. "Engagement with social media and social media advertising: The differentiating role of platform type." Journal of advertising 47, no. 1 (2018): 38-54.

[5] Ortiz-Ospina, Esteban, and Max Roser. "The rise of social media." Our world in data (2023).

[6] Ali, Imran, Maria Balta, and Thanos Papadopoulos. "Social media platforms and social enterprise: Bibliometric analysis and systematic review." International Journal of Information Management 69 (2023): 102510.

[7] Salminen, Joni, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. "Developing an online hate classifier for multiple social media platforms." Human-centric Computing and Information Sciences 10 (2020): 1-34.

[8] Hruska, J., & Maresova, P. (2020). Use of Social Media Platforms among Adults in the United States—Behavior on Social Media. Societies, 10, 27. https://doi.org/10.3390/soc10010027.

[9] Logghe, Heather J., Marissa A. Boeck, and Sam B. Atallah. "Decoding Twitter: understanding the history, instruments, and techniques for success." Annals of surgery 264, no. 6 (2016): 904-908.

[10] Jensen, Marion, Tom Caswell, Justin Ball, Joel Duffin, and Rob Barton. "TwHistory: Sharing history using twitter." (2010).

[11] Giles, David. "Context, History, and Twitter Data: Some Methodological Reflections." In Analysing Digital Interaction, pp. 41-63. Cham: Springer International Publishing, 2021.

[12] Wu, Bo, and Haiying Shen. "Analyzing and predicting news popularity on Twitter." International Journal of Information Management 35, no. 6 (2015): 702-711.

[13] Zhang, Liwei, and Jue Wang. "What affects publications' popularity on Twitter?." Scientometrics 126, no. 11 (2021): 9185-9198.

[14] Bruns, Axel, and Jean Burgess. "Researching news discussion on Twitter: New methodologies." Journalism studies 13, no. 5-6 (2012): 801-814.

[15] Jiang, Lan, and Mehmet Erdem. "Twitter-marketing in multi-unit restaurants: is it a viable marketing tool?." Journal of foodservice business research 20, no. 5 (2017): 568-578.

[16] Kinney, Lance, and Jennifer Ireland. "Brand spokes-characters as Twitter marketing tools." Journal of Interactive Advertising 15, no. 2 (2015): 135-150.

[17] Leonowicz-Bukała, Iwona, Andrzej Adamski, and Anna Jupowicz-Ginalska. "Twitter in Marketing Practice of the Religious Media. An Empirical Study on Catholic Weeklies in Poland." Religions 12, no. 6 (2021): 421.

[18] X. Jin, C. Lin, J. Luo, and J. Han. A data mining-based spam detection system for social media networks. Proceedings of the VLDB Endowment, 4(12):1458–1461, 2011.

[19] Alom, Zulfikar, Barbara Carminati, and Elena Ferrari. "A deep learning model for Twitter spam detection." Online Social Networks and Media 18 (2020): 100079.

[20] Madisetty, Sreekanth, and Maunendra Sankar Desarkar. "A neural network-based ensemble approach for spam detection in Twitter." IEEE Transactions on Computational Social Systems 5, no. 4 (2018): 973-984.

[21] Grier C., K. Thomas, V. Paxson , M. Zhang, @ spam: the underground on 140 characters or less, in: Proceedings of the 17th ACM conference on Computer and communications security, ACM, 2010, pp. 27–37 .

[22] Verma, Monika, and Sanjeev Sofat. "Techniques to detect spammers in twitter-a survey." International Journal of Computer Applications 85, no. 10 (2014).

[23] Fazil, Mohd, and Muhammad Abulaish. "A hybrid approach for detecting automated spammers in twitter." IEEE Transactions on Information Forensics and Security 13, no. 11 (2018): 2707-2719.

[24] Çıtlak, Oğuzhan, Murat Dörterler, and İbrahim Alper Doğru. "A survey on detecting spam accounts on Twitter network." Social Network Analysis and Mining 9 (2019): 1-13.

[25] Aswani, Reema, Arpan Kumar Kar, and P. Vigneswara Ilavarasan. "Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspired computing." Information Systems Frontiers 20 (2018): 515-530.

[26] Rao, Sanjeev, Anil Kumar Verma, and Tarunpreet Bhatia. "A review on social spam detection: Challenges, open issues, and future directions." *Expert Systems with Applications* 186 (2021): 115742.

[27] Rao, Sanjeev, Anil Kumar Verma, and Tarunpreet Bhatia. "Hybrid ensemble framework with self-attention mechanism for social spam detection on imbalanced data." *Expert Systems with Applications* 217 (2023): 119594.

[28] Aljabri, Malak, Rachid Zagrouba, Afrah Shaahid, Fatima Alnasser, Asalah Saleh, and Dorieh M. Alomari. "Machine learning-based social media bot detection: a comprehensive literature review." *Social Network Analysis and Mining* 13, no. 1 (2023): 20.

[29] Ellaky, Zineb, FaouziaBenabbou, and Sara Ouahabi. "Systematic Literature Review of Social Media Bots Detection Systems." *Journal of King Saud University-Computer and Information Sciences* (2023).

[30] Goksu, Murat, and Nadire Cavus. "Fake news detection on social networks with artificial intelligence tools: systematic literature review." In *10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions-ICSCCW-2019*, pp. 47-53. Springer International Publishing, 2020.

[31] Verma, Pawan Kumar, Prateek Agrawal, Vishu Madaan, and Charu Gupta. "UCred: fusion of machine learning and deep learning methods for user credibility on social media." *Social Network Analysis and Mining* 12, no. 1 (2022): 54.

[32] Macas, Mayra, Chunming Wu, and Walter Fuertes. "A survey on deep learning for cybersecurity: Progress, challenges, and opportunities." *Computer Networks* 212 (2022): 109032.

[33] Muralidharan, Trivikram, and Nir Nissim. "Improving malicious email detection through novel designated deep-learning architectures utilizing entire email." *Neural Networks* 157 (2023): 257-279.

[34] Ayub, Mohammed, Omar Lajam, Abdullatif Alnajim, and Mahmood Niazi. "Use of Machine Learning for Web Denial-of-Service Attacks: A Multivocal Literature Review." *Arabian Journal for Science and Engineering* (2022): 1-16.

[35] Hangloo, Sakshini, and Bhavna Arora. "Combating multimodal fake news on social media: methods, datasets, and future perspective." *Multimedia Systems* 28, no. 6 (2022): 2391-2422.

[36] Kawintiranon, K., Singh, L., & Budak, C. (2022). Traditional and context-specific spam detection in low resource settings. *Machine Learning*, *111*(7), 2515-2536.

[37] Xu, G., Zhou, D., & Liu, J. (2021). Social network spam detection based on ALBERT and combination of Bi-LSTM with self-attention. *Security and Communication Networks*, *2021*, 1-11.

[38] Al-Zoubi, A. M., Alqatawna, J. F., Faris, H., & Hassonah, M. A. (2021). Spam profiles detection on social networks using computational intelligence methods: the effect of the lingual context. Journal of Information Science, 47(1), 58-81.

[39] Gadiraju, N., Raju, G., & Amruta, G. (2018). Spam Detection on Online Social Media Networks. International Journal of Engineering & Technology, 7, 631.

[40] Novo-Lourés, María, David Ruano-Ordás, Reyes Pavón, Rosalía Laza, Silvana Gomez-Meire, and Jose R. Mendez. "Enhancing representation in the context of multiple-channel spam filtering." *Information Processing & Management* 59, no. 2 (2022): 102812.

[41] Patil, Rahul A., and Chetana C. Chaudhari. "Use of a Recurrent Neural Network to Identify Spammers on Twitter." In *Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021*, pp. 949-957. Springer Singapore, 2022.

[42] Alshammari, Saud, Eman Aljabarti, and Yusliza Yusoff. "Protection of Users Kids on Twitter Platform Using Naïve Bayes." In *Kids Cybersecurity Using Computational Intelligence Techniques*, pp. 109-120. Cham: Springer International Publishing, 2023.

[43] Güngör, Kübra Nur, O. Ayhan Erdem, and İbrahim Alper Doğru. "Tweet and account based spam detection on twitter." In *Artificial Intelligence and Applied Mathematics in Engineering Problems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2019)*, pp. 898-905. Springer International Publishing, 2020.

[44] Alsaffar, Dalia, Amjad Alfahhad, BashaierAlqhtani, Lama Alamri, Shahad Alansari, Nada Alqahtani, and Dabiah A. Alboaneen. "Machine and deep learning algorithms for Twitter spam detection." In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019*, pp. 483-491. Springer International Publishing, 2020.

[45] Kihal, Marouane, and Lamia Hamza. "Robust multimedia spam filtering based on visual, textual, and audio deep features and random forest." *Multimedia Tools and Applications* (2023): 1-19.

[46] Mubarak, Hamdy, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish. "Spam detection on arabic twitter." In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pp. 237-251. Springer International Publishing, 2020.

[47] Vidya Kumari, K. R., and C. R. Kavitha. "Spam detection using machine learning in R." In *International Conference on Computer Networks and Communication Technologies: ICCNCT 2018*, pp. 55-64. Springer Singapore, 2019.

[48] El-Mawass, Nour, Paul Honeine, and Laurent Vercouter. "SimilCatch: Enhanced social spammers detection on twitter using Markov random fields." *Information processing & management* 57, no. 6 (2020): 102317.

[49] Kawintiranon, Kornraphop, Lisa Singh, and Ceren Budak. "Traditional and context-specific spam detection in low resource settings." *Machine Learning* 111, no. 7 (2022): 2515-2536.

[50] Sahoo, Somya Ranjan, and B. B. Gupta. "Real-time detection of fake account in twitter using machine-learning approach." In *Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2019*, pp. 149-159. Springer Singapore, 2021.

[51] Alkadri, Abdullah M., Abeer Elkorany, and Cherry Ahmed. "Enhancing Detection of Arabic Social Spam Using Data Augmentation and Machine Learning." *Applied Sciences* 12, no. 22 (2022): 11388.

[52] Gupta, Saksham, Amit Chhabra, Satvik Agrawal, and Sunil K. Singh. "A Comprehensive Comparative Study of Machine Learning Classifiers for Spam Filtering." In *International Conference on Cyber Security, Privacy and Networking (ICSPN 2022)*, pp. 257-268. Cham: Springer International Publishing, 2023.

[53] Ghiassi, Manoochehr, Sean Lee, and Swati Ramesh Gaikwad. "Sentiment analysis and spam filtering using the YAC2 clustering algorithm with transferability." *Computers & Industrial Engineering* 165 (2022): 107959.

[54] Kumar, Alok, Maninder Singh, and Alwyn Roshan Pais. "Fuzzy string matching algorithm for spam detection in Twitter." In *Security and Privacy: Second ISEA International Conference, ISEA-ISAP 2018, Jaipur, India, January, 9–11, 2019, Revised Selected Papers 2*, pp. 289-301. Springer Singapore, 2019.

[55] Elakkiya, E., and S. Selvakumar. "Stratified hyperparameters optimization of feed-forward neural network for social network spam detection (SON2S)." *Soft Computing* 26, no. 21 (2022): 11915-11934.

[56] Lee, Minyoung, and Eunil Park. "Real-time Korean voice phishing detection based on machine learning approaches." *Journal of Ambient Intelligence and Humanized Computing* (2021): 1-12.

[57] Dhaka, Deepali, and Monica Mehrotra. "Cross-Domain Spam Detection in Social Media: A Survey." In *Emerging Technologies in Computer Engineering: Microservices in Big Data Analytics: Second International Conference, ICETCE 2019, Jaipur, India, February 1–2, 2019, Revised Selected Papers 2*, pp. 98-112. Springer Singapore, 2019.

[58] Jan, Tabassum Gull, Surinder Singh Khurana, and Munish Kumar. "Semi-supervised labeling: a proposed methodology for labeling the twitter datasets." *Multimedia Tools and Applications* 81, no. 6 (2022): 7669-7683.

[59] Balakrishnan, Vimala, Shahzaib Khan, and Hamid R. Arabnia. "Improving cyberbullying detection using Twitter users' psychological features and machine learning." *Computers & Security* 90 (2020): 101710.

[60] Ouni, Sarra, Fethi Fkih, and Mohamed Nazih Omri. "BERT-and CNN-based TOBEAT approach for unwelcome tweets detection." *Social Network Analysis and Mining* 12, no. 1 (2022): 144.

[61] Saeed, Radwa MK, Sherine Rady, and Tarek F. Gharib. "An ensemble approach for spam detection in Arabic opinion texts." *Journal of King Saud University-Computer and Information Sciences* 34, no. 1 (2022): 1407-1416.

[62] Vanmathi, C., and R. Mangayarkarasi. "An Analysis of Machine Learning Approach for Detecting Automated Spammer in Twitter." In *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2020*, pp. 443-451. Springer Singapore, 2021.

[63] Sahoo, Somya Ranjan, Brij B. Gupta, Chang Choi, Ching-Hsien Hsu, and Kwok Tai Chui. "Behavioral analysis to detect social spammer in online social networks (OSNs)." In *Computational Data and Social Networks: 9th International Conference, CSoNet 2020, Dallas, TX, USA, December 11–13, 2020, Proceedings 9*, pp. 321-332. Springer International Publishing, 2020.

[64] Çıtlak, Oğuzhan, Murat Dörterler, and İbrahim Alper Doğru. "A survey on detecting spam accounts on Twitter network." *Social Network Analysis and Mining* 9 (2019): 1-13.

[65] Park, Woo Hyun, Isma Farah Siddiqui, Chinmay Chakraborty, Nawab Muhammad Faseeh Qureshi, and Dong Ryeol Shin. "Scarcity-aware spam detection technique for big data ecosystem." *Pattern Recognition Letters* 157 (2022): 67-75.

[66] Jimoh, Rasheed G., Kayode S. Adewole, Tunbosun E. Aderemi, and Abdullateef O. Balogun. "Investigative Study of Unigram and Bigram Features for Short Message Spam Detection." In *International Conference on Emerging Applications and Technologies for Industry 4.0 (EATI'2020)*

Emerging Applications and Technologies for Industry 4.0, pp. 70-81. Springer International Publishing, 2021.

[67] Paudel, Ramesh, Prajjwal Kandel, and William Eberle. "Detecting spam tweets in trending topics using graph-based approach." In *Proceedings of the Future Technologies Conference (FTC) 2019: Volume 1*, pp. 526-546. Springer International Publishing, 2020.

[68] Das, L., Ahuja, L., & Pandey, A. (2022, April). Analysis of Twitter Spam Detection Using Machine Learning Approach. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 764-769). IEEE.

[69] Rodrigues, A. P., Fernandes, R., Shetty, A., Lakshmanna, K., & Shafi, R. M. (2022). Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. *Computational Intelligence and Neuroscience*, *2022*.

[70] Sundararajan, K., & Palanisamy, A. (2020). Multi-rule based ensemble feature selection model for sarcasm type detection in twitter. *Computational intelligence and neuroscience*, *2020*.

[71] Chen, H., Liu, J., Lv, Y., Li, M. H., Liu, M., & Zheng, Q. (2018). Semi-supervised clue fusion for spammer detection in Sina Weibo. *Information Fusion*, *44*, 22-32.

[72] Alsaffar, D., Alfahhad, A., Alqhtani, B., Alamri, L., Alansari, S., Alqahtani, N., & Alboaneen, D. A. (2019, October). Machine and deep learning algorithms for Twitter spam detection. In *International conference on advanced intelligent systems and informatics* (pp. 483-491). Cham: Springer International Publishing.

[73] Abid, Muhammad Adeel, Saleem Ullah, Muhammad Abubakar Siddique, Muhammad Faheem Mushtaq, Wajdi Aljedaani, and Furqan Rustam. "Spam SMS filtering based on text features and supervised machine

learning techniques." *Multimedia Tools and Applications* 81, no. 28 (2022): 39853-39871.

[74] Wang, Haoyu, Bingze Dai, and Dequan Yang. "A Comparative Study of Two Different Spam Detection Methods." In *Dependability in Sensor, Cloud, and Big Data Systems and Applications: 5th International Conference, DependSys 2019, Guangzhou, China, November 12–15, 2019, Proceedings 5*, pp. 95-105. Springer Singapore, 2019.

[75] Talaei Pashiri, Rozita, Yaser Rostami, and Mohsen Mahrami. "Spam detection through feature selection using artificial neural network and sine–cosine algorithm." *Mathematical Sciences* 14 (2020): 193-199.

[76] Barushka, Aliaksandr, and Petr Hajek. "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks." *Neural Computing and Applications* 32 (2020): 4239-4257.

[77] Asdaghi, Faeze, and Ali Soleimani. "An effective feature selection method for web spam detection." *Knowledge-Based Systems* 166 (2019): 198-206.

[78] Pirozmand, Poria, Mehdi Sadeghilalimi, Ali Asghar Rahmani Hosseinabadi, Fatemeh Sadeghilalimi, SeyedsaeidMirkamali, and Adam Slowik. "A feature selection approach for spam detection in social networks using gravitational force-based heuristic algorithm." *Journal of Ambient Intelligence and Humanized Computing* (2021): 1-14.

[79] Elakkiya, E., and S. Selvakumar. "GAMEFEST: Genetic Algorithmic Multi Evaluation measure basedFEature Selection Technique for social network spam detection." *Multimedia tools and applications* 79 (2020): 7193-7225.

[80] Sharaff, Aakanksha. "Spam detection in SMS based on feature selection Techniques." In *Emerging Technologies in Data Mining and*

Information Security: Proceedings of IEMIS 2018, Volume 2, pp. 555-563. Springer Singapore, 2019.

[81] Karakaşlı, M. Salih, Muhammed Ali Aydin, Serhan Yarkan, and Ali Boyaci. "Dynamic feature selection for spam detection in Twitter." In *International Telecommunications Conference: Proceedings of the ITelCon 2017, Istanbul*, pp. 239-250. Springer Singapore, 2019.

[82] Ahmad, Saleh Beyt Sheikh, Mahnaz Rafie, and Seyed Mojtaba Ghorabie. "Spam detection on Twitter using a support vector machine and users' features by identifying their interactions." *Multimedia Tools and Applications* 80, no. 8 (2021): 11583-11605.

[83] Vinitha, V. Sri, and D. Karthika Renuka. "Feature selection techniques for email spam classification: a survey." In *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019*, pp. 925-935. Springer International Publishing, 2020.

[84] Faris, Hossam, Al-ZoubiAla'M, Ali Asghar Heidari, Ibrahim Aljarah, Majdi Mafarja, Mohammad A. Hassonah, and Hamido Fujita. "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks." *Information Fusion* 48 (2019): 67-83.

[85] Pirozmand, P., Sadeghilalimi, M., Hosseinabadi, A. A. R., Sadeghilalimi, F., Mirkamali, S., & Slowik, A. (2021). A feature selection approach for spam detection in social networks using gravitational force-based heuristic algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 1-14.

[86] Zhao, C., Xin, Y., Li, X., Yang, Y., & Chen, Y. (2020). A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data. *Applied Sciences*, *10*(3), 936.

[87] Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S., & Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, *484*, 153-166.

[88] Asdaghi, F., & Soleimani, A. (2019). An effective feature selection method for web spam detection. *Knowledge-Based Systems*, *166*, 198-206.

[89] Alom, Zulfikar, Barbara Carminati, and Elena Ferrari. "A deep learning model for Twitter spam detection." *Online Social Networks and Media* 18 (2020): 100079.

[90] Singh, Akhil Pratap, Ashish Singh, and Kakali Chatterjee. "A Comparative Approach for Email Spam Detection Using Deep Learning." *Intelligent Computing and Communication Systems* (2021): 187-200.

[91] Kumar, Chanchal, Taran Singh Bharti, and Shiv Prakash. "A hybrid Data-Driven framework for Spam detection in Online Social Network." *Procedia Computer Science* 218 (2023): 124-132.

[92] Kaddoura, Sanaa, Suja A. Alex, Maher Itani, Safaa Henno, Asma AlNashash, and D. Jude Hemanth. "Arabic spam tweets classification using deep learning." *Neural Computing and Applications* (2023): 1-14.

[93] Jacob, W. Stalin. "Multi-objective genetic algorithm and CNN-based deep learning architectural scheme for effective spam detection." *International Journal of Intelligent Networks* 3 (2022): 9-15.

[94] Jain, Gauri, Manisha Sharma, and Basant Agarwal. "Spam detection in social media using convolutional and long short term memory neural network." *Annals of Mathematics and Artificial Intelligence* 85, no. 1 (2019): 21-44.

[95] Elakkiya, E., S. Selvakumar, and R. Leela Velusamy. "TextSpamDetector: textual content based deep learning framework for social spam detection using conjoint attention mechanism." *Journal of Ambient Intelligence and Humanized Computing* 12 (2021): 9287-9302.

[96] Ilias, Loukas, and Ioanna Roussaki. "Detecting malicious activity in Twitter using deep learning techniques." *Applied Soft Computing* 107 (2021): 107360.

[97] Ghanem, Razan, and Hasan Erbay. "Spam detection on social networks using deep contextualized word representation." *Multimedia Tools and Applications* 82, no. 3 (2023): 3697-3712.

[98] Jain, Gauri, Manisha Sharma, and Basant Agarwal. "Optimizing semantic LSTM for spam detection." *International Journal of Information Technology* 11 (2019): 239-250.

[99] Guo, Zhiwei, Lianggui Tang, Tan Guo, Keping Yu, Mamoun Alazab, and Andrii Shalaginov. "Deep graph neural network-based spammer detection under the perspective of heterogeneous cyberspace." *Future generation computer systems* 117 (2021): 205-218.

[100] Mardi, Vanyashree, Anvaya Kini, V. M. Sukanya, and S. Rachana. "Text-Based Spam Tweets Detection Using Neural Networks." In *Advances in Computing and Intelligent Systems: Proceedings of ICACM 2019*, pp. 401-408. Springer Singapore, 2020.

[101] Makkar, Aaisha, and Neeraj Kumar. "PROTECTOR: An optimized deep learning-based framework for image spam detection and prevention." *Future Generation Computer Systems* 125 (2021): 41-58.

[102] Islam, Md Rafiqul, Shaowu Liu, Xianzhi Wang, and Guandong Xu. "Deep learning for misinformation detection on online social networks: a survey and new perspectives." *Social Network Analysis and Mining* 10 (2020): 1-20.

[103] Krouska, Akrivi, Christos Troussas, and Maria Virvou. "Deep learning for twitter sentiment analysis: the effect of pre-trained word embedding." *Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications* (2020): 111-124.

[104] Thomas, Merly, and B. B. Meshram. "ChSO-DNFNet: Spam detection in Twitter using feature fusion and optimized Deep Neuro Fuzzy Network." *Advances in Engineering Software* 175 (2023): 103333.

[105] Sumathi, S., and Ganesh Kumar Pugalendhi. "Cognition based spam mail text analysis using combined approach of deep neural network classifier and random forest." *Journal of Ambient Intelligence and Humanized Computing* 12 (2021): 5721-5731.

[106] Dhavale, Sunita. "C-ASFT: convolutional neural networks-based anti-spam filtering technique." In *Proceeding of International Conference on Computational Science and Applications: ICCSA 2019*, pp. 49-55. Springer Singapore, 2020.

[107] Kraidia, Insaf, Afifa Ghenai, and Nadia Zeghib. "HST-Detector: A Multimodal Deep Learning System for Twitter Spam Detection." In *Computational Intelligence, Data Analytics and Applications: Selected papers from the International Conference on Computing, Intelligence and Data Analytics (ICCIDA)*, pp. 91-103. Cham: Springer International Publishing, 2023.

[108] Anil, Aditya, Ananya Sajwan, Lalitha Ramchandar, and N. Subhashini. "Advanced Spam Detection Using NLP and Deep Learning." In *Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 1*, pp. 319-332. Singapore: Springer Nature Singapore, 2022.

[109] Alhassun, Atheer S., and Murad A. Rassam. "A combined text-based and metadata-based deep-learning framework for the detection of spam accounts on the social media platform twitter." *Processes* 10, no. 3 (2022): 439.

[110] Xu, Zhiming, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. "Contrastive attributed network anomaly detection with data augmentation." In *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia*

Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II, pp. 444-457. Cham: Springer International Publishing, 2022.

[111] Bazzaz Abkenar, Sepideh, Ebrahim Mahdipour, Seyed Mahdi Jameii, and Mostafa Haghi Kashani. "A hybrid classification method for Twitter spam detection based on differential evolution and random forest." *Concurrency and Computation: Practice and Experience* 33, no. 21 (2021): e6381.

[112] Chen, Kangyang, Xinyi Zou, Xingguo Chen, and Huihui Wang. "An automated online spam detector based on deep cascade forest." In *Science of Cyber Security: Second International Conference, SciSec 2019, Nanjing, China, August 9–11, 2019, Revised Selected Papers 2*, pp. 33-46. Springer International Publishing, 2019.

[113] Sharaff, Aakanksha, Vrihas Pathak, and Siddhartha Shankar Paul. "Deep learning-based smishing message identification using regular expression feature generation." *Expert Systems* (2022): e13153.

[114] Jain, Ankit Kumar, Diksha Goel, Sanjli Agarwal, Yukta Singh, and Gaurav Bajaj. "Predicting spam messages using back propagation neural network." *Wireless Personal Communications* 110 (2020): 403-422.

[115] Singh, Ashish, and Kakali Chatterjee. "A Comparative Approach for Opinion Spam Detection Using Sentiment Analysis." In *Proceedings of First International Conference on Computational Electronics for Wireless Communications: ICCWC 2021*, pp. 511-522. Springer Singapore, 2022.

[116] Lago, Carlos, Rafael Romón, Iker Pastor López, Borja Sanz Urquijo, Alberto Tellaeche, and Pablo García Bringas. "Deep Learning Applications on Cybersecurity." In *Hybrid Artificial Intelligent Systems: 16th International Conference, HAIS 2021, Bilbao, Spain, September 22–24, 2021, Proceedings 16*, pp. 611-621. Springer International Publishing, 2021.

[117] Sharma, Geetanjali, Gursimran Singh Brar, Pahuldeep Singh, Nitish Gupta, Nidhi Kalra, and Anshu Parashar. "An Exploration of Machine Learning and Deep Learning Techniques for Offensive Text Detection in

Social Media—A Systematic Review." In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3*, pp. 541-559. Singapore: Springer Nature Singapore, 2022.

[118] Gamal, Donia, Marco Alfonse, Salud María Jiménez-Zafra, and Mostafa Aref. "Arabic Sentiment Classification on Twitter Using Deep Learning Techniques." In *Advances in Intelligent Systems, Computer Science and Digital Economics IV*, pp. 236-251. Cham: Springer Nature Switzerland, 2023.

[119] Borse, Dipalee, and Swati Borse. "State of the art on Twitter spam detection." *Applied Computational Technologies: Proceedings of ICCET 2022* (2022): 486-496.

[120] Washha, Mahdi, Aziz Qaroush, Manel Mezghani, and Florence Sedes. "Unsupervised collective-based framework for dynamic retraining of supervised real-time spam tweets detection model." *Expert systems with Applications* 135 (2019): 129-152.

[121] Koggalahewa, Darshika, Yue Xu, and Ernest Foo. "A Drift Aware Hierarchical Test Based Approach for Combating Social Spammers in Online Social Networks." In *Data Mining: 19th Australasian Conference on Data Mining, AusDM 2021, Brisbane, QLD, Australia, December 14-15, 2021, Proceedings*, pp. 47-61. Singapore: Springer Singapore, 2021.

[122] Alharthi, Reem, Areej Alhothali, and Kawthar Moria. "A real-time deep-learning approach for filtering Arabic low-quality content and accounts on Twitter." *Information Systems* 99 (2021): 101740.

[123] Torney, Monal R., Kishor H. Walse, and Vilas M. Thakare. "A Comprehensive Survey of Datasets Used for Spam and Genuineness Views Detection in Twitter." *Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022* (2022): 223-237.

[124] Liang, Y., & Yan, X. (2019, May). Using deep learning to detect malicious urls. In *2019 IEEE International Conference on Energy Internet (ICEI)* (pp. 487-492). IEEE.

[125] Le, H., Pham, Q., Sahoo, D., & Hoi, S. C. (2018). URLNet: Learning a URL representation with deep learning for malicious URL detection. *arXiv preprint arXiv:1802.03162*.

[126] Madisetty, S., & Desarkar, M. S. (2018). A neural network-based ensemble approach for spam detection in Twitter. *IEEE Transactions on Computational Social Systems*, *5*(4), 973-984.

[127] Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, *467*, 312-322.

[128] Abdi, F. D., & Wenjuan, L. (2017). Malicious URL detection using convolutional neural network. *Journal International Journal of Computer Science, Engineering and Information Technology*, *7*(6), 1-8.

[129] Inuwa-Dutse, Isa, Mark Liptrott, and Ioannis Korkontzelos. "Detection of spam-posting accounts on Twitter." Neurocomputing 315 (2018): 496-511.

[130] Ahmad, Saleh Beyt Sheikh, Mahnaz Rafie, and Seyed Mojtaba Ghorabie. "Spam detection on Twitter using a support vector machine and users' features by identifying their interactions." Multimedia Tools and Applications 80, no. 8 (2021): 11583-11605.

[131] Senthil Murugan, N., and G. Usha Devi. "Detecting streaming of Twitter spam using hybrid method." Wireless Personal Communications 103, no. 2 (2018): 1353-1374.

[132] Bharti, K. K., & Pandey, S. (2021). Fake account detection in twitter using logistic regression with particle swarm optimization. *Soft Computing*, *25*(16), 11333-11345.

[133] Ala'M, A. Z., Faris, H., Alqatawna, J. F., & Hassonah, M. A. (2018). Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts. *Knowledge-Based Systems*, *153*, 91-104.

[134] Vidyashree, K. P., & Rajendra, A. B. (2023). An Improvised Sentiment Analysis Model on Twitter Data Using Stochastic Gradient Descent (SGD) Optimization Algorithm in Stochastic Gate Neural Network (SGNN). *SN Computer Science*, *4*(2), 190.

[135] Bhardwaj, U., & Sharma, P. (2023). Email spam detection using bagging and boosting of machine learning classifiers. *International Journal of Advanced Intelligence Paradigms*, *24*(1-2), 229-253.
[136] Yi, H.; Liu, J.; Xu, W.; Li, X.; Qian, H. A Graph Neural Network Social Recommendation Algorithm Integrating the Multi-Head Attention Mechanism. Electronics 2023, 12, 1477.
[137] Jain, G., Sharma, M., & Agarwal, B. (2019). Optimizing semantic LSTM for spam detection. International Journal of Information Technology, 11, 239-250.
[138] Jain, G., Sharma, M., & Agarwal, B. (2019). Spam detection in social media using convolutional and long short term memory neural network. Annals of Mathematics and Artificial Intelligence, 85(1), 21-44.
[139] Imam, N., Issac, B., & Jacob, S. M. (2019). A semi-supervised learning approach for tackling Twitter spam drift. International journal of computational intelligence and applications, 18(02), 1950010.
[140] Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. Sensors, 19(21), 4654.

LIST OF PUBLICATIONS

Sno.	**Title of paper with author names**	**Name of journal / conference**	**Published date**	**Issn no/ vol no, issue no**	**Indexing in Scopus/ Web of Science/UGC-CARE list (*please mention*)**
1.	Conference	Algorithms, Computing and Mathematics Conference	2021	1613-0073	Scopus
2.	Journal Paper	IEEE Transactions on Computational Social Systems	29/12/2022	2329-924X	SCI

3.	Conference	International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)	6/4/2023	978-1-6654-5499-5	Scopus
4.	Journal	SN Computer Science	01/2024	2662-955X	Scopus
5.	Journal	Webology	2022	ISSN: 1735-188X	UGC