NOVEL METHODOLOGIES FOR PREDICTIVE ANALYSIS IN CRIME DATA OVER ONLINE SOCIAL MEDIA

A Thesis Submitted
In Partial Fulfillment of The Requirements
For The Degree Of

DOCTOR OF PHILOSOPHY

in
Computer Science and Engineering

by

Monika
(2K18/PHD/CO/14)

Under the Supervision of

Prof. Aruna Bhat
Department of Computer Science and Engineering
Delhi Technological University

Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

July, 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I Monika, hereby certify that the work which is being presented in the thesis entitled "Novel Methodologies for Predictive Analysis in Crime Data over online Social Media" in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from 2018 to 2024 under the supervision of Prof. Aruna Bhat.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Monika** (2K18/PHD/CO/14) has carried out their search work presented in this thesis entitled **"Novel Methodologies for Predictive Analysis in Crime Data over online Social Media"** for the award of **Doctor of Philosophy** from Department of Computer Science and Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Prof. Aruna Bhat
Supervisor
Department of CSE
Delhi Technological University

Place: New Delhi
Date: 25th July 2024

ABSTRACT

Social media platforms have become integral for communication and information exchange, yet they also pose challenges such as cybercrime and hacking. The escalating incidents of crimes on platforms like Twitter necessitate proactive measures, including crime prediction. This research employs a comprehensive approach, utilizing Twitter data for crime prediction through data preprocessing and feature extraction techniques. Techniques such as Bag of Words, Glove, TF-IDF (Term Frequency-Improved Document Frequency), and feature hashing are employed, with feature selection using a Modified Tree Growth Algorithm (MTGA) and clustering via Fuzzy Manta Ray Foraging (FMRF). The crime detection is performed using a hybrid Wavelet Convolutional Neural Network with World Cup Optimization (WCNN-WCO). The proposed method outperforms existing ones in terms of precision, accuracy, F1 measure, and recall, addressing the rising social issue of social media crimes.

Furthermore, the study introduces DAC-BiNet, a robust Deep Attention Convolutional Bi-directional Aquila Optimal Network, specifically tailored for crime detection on the Twitter platform. The model undergoes a multi-stage process involving pre-processing, feature extraction, and clustering through Possibilistic Fuzzy LDA (Latent Dirichlet Allocation). Experimental results demonstrate the effectiveness of DAC-BiNet (Deep Attention Convolutional Bi-directional Aquila Optimal Network), achieving increased accuracy, precision, recall, specificity, and F1 score.

Additionally, the paper explores the application of Apache Pig with Hadoop in large-scale crime data analysis. Utilizing incident-level crime data, the study showcases the efficacy of Apache Pig in analyzing vast datasets, aiding decision-makers, policymakers, and governments in minimizing crime.

In conclusion, these research works collectively highlight the significance of technology-driven approaches in addressing and mitigating the complex issues surrounding crime, be it on social media platforms or in large-scale datasets. The findings provide valuable insights for law enforcement, policymakers, and governments to make informed decisions and formulate effective strategies in the realm of crime prevention and control.

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

ACKNOWLEDGEMENTS

I owe a debt of my gratitude to my supervisor, Prof. Aruna Bhat, Professor, Department of Computer Science and Engineering, Delhi Technological University, for their valuable guidance, motivation, constant support and encouragement, which has helped me in formulating an urge for this research work. Their exemplary hard work and insightful guidance have always helped me overcome difficulties.

I would also like to extend my heartfelt thanks to the DRC and SRC committees for their valuable insights, suggestions, and critical evaluation of my research work.

I would like to take this opportunity to thank all the faculty members of the Computer Science & Engineering Department, Delhi Technological University, for their encouragementand support.

I am also thankful to all the staff members and research fellows for their untiring support.

Moreover, I am profoundly grateful to my parents (Mr. Bharat Singh and Mrs. Urmila Devi), my husband (Mr. Rahul), my daughter (Ms. Ivanka Chandra) and my siblings (Mr. Vikas and Mr. Ritin) whose unconditional support and blessings have been a cornerstone of my achievements.

I recognize that this accomplishment would not have been possible without the collective efforts and contributions of all those mentioned above. Their presence in my life has been truly invaluable, and I am forever indebted to them for their unwavering support and belief in my abilities.

Monika

2K18/PHDCO/14

Department of CSE

Delhi Technological University

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

φ	Power reduction rate.
γ	regulates the influence of the closest tree
δ	Angle between 0 and 1
δ_s	The hybrid WCNN with WCO classifier classification error rate.
$\|\|$	The relationship between the measured data and the cluster center.
ℓ	the weight coefficient
β	mean value
ω_i	feature weight of the tweet data
γ	feature length parameter
λ	volume many instances
ϕ	Dirichlet parameter
ζ	document level topic variables
η	PFCM's objective function
α	variation in the sample
$\oplus$	addition operation

LIST OF ABBREVIATIONS

ROC	Receiver Operating Characteristic
ML	Machine Learning
SL	Supervised Learning
UL	Unsupervised Learning
RL	Reinforcement Learning
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
WCNN-WCO	Wavelet CNN With Weighted Overlap Combination
BiLSTM	Bi-Directional Long Short-Term Memory
MNB	Multinomial Naive Bayes
DRNN	Deep Recurrent Neural Network
LR	Logistic Regression
GIS	Geographic Information Systems
TM	Text Mining
DM	Data Mining
CV	Computer Vision
DAC-BiNet	Deep Attention Convolutional Bi-Directional Aquila Optimal Network

ITF-IDF	Improved Term Frequency-Improved Document Frequency
NLP	Natural Language Processing
NLG	Natural Language Generation
ARIMA	Autoregressive Integrated Moving Average
BDA	Big Data Analytics
WCNN	Wavelet Convolutional Neural Network
WCO	World Cup Organization
MTGA	Modified Tree Growth Algorithm
FMRF	Fuzzy Manta Ray Foraging
TS	Twitter Streaming
API	Application Programming Interface
BoW	Bag-Of-Words
MLlib	Machine Learning Library
ANN	Artificial Neural Network
KNN	K-Nearest Neighbor
COPS	Community Oriented Policing Services
LSTM	Long Short-Term Memory
SLR	Systematic Literature Review

BDA	Big Data Analytics
CSV	Comma-Separated Values
IPC	Indian Panel Code
SLL	Special And Local Laws
TFIDF	Frequency-Inverse Document Frequency

CHAPTER 1

INTRODUCTION

This chapter presents about the offences in social network such as cybercrimes and social media crimes, criminal offences committed on social media like coordination, manipulation, and abuse, then types of crime rate prediction, and various crime predictive analytics. Also, provides with the problem statement, motivation and research objectives of this research.

1.1 Overview

Crime is a well-known social issue that has an impact on a society's economic growth and overall quality of life in all of its forms. Research indicates that there is a correlation among reduced economic growth and crime rates in both national and local contexts, including cities and metropolitan areas [1]. Criminal law and sociology experts have always been interested in information about crimes. Studies on the behavioral development of criminals and its relationships to certain features of the communities in which they were raised, lived, and behaved date back to the beginning of the twentieth century [2].

Many insights have been gained from both individual and group perspectives, when learning the effects on behavioral improvement of variables such as revelation to particular peer networks, neighborhood features (e.g., absence/presence of educational/recreational facilities), as well as poverty indexes. The majority of research in the domains of criminology, psychology, sociology, as well as economics focuses on the connections between socioeconomic factors including unemployment, income level, ethnicity, and education [3].

Social media is becoming a more popular platform for people to share news, opinions, and ideas related to events happening in the regions they enjoy. So, one well-known topic of research for enhancing public safety is crime prediction. Research

demonstrates that crime analytics and forecasting can be applied to social media data [4]. Technology is advancing quickly, which has made big data analysis possible. However, as cities get more congested, crime rates rise. Any country should prioritize maintaining its national security. Nations have made investments in criminology research to learn about the unique traits of criminals and the use of data mining in this application. Law enforcement organizations conduct crime analysis to examine trends and patterns in criminal activity and conduct a systematic evaluation [5].

So, in order to solve criminal cases, crime analysis is crucial. Technology has advanced quickly, allowing for quicker criminal investigations. Uncontrolled migration and population increase are major contributors to urban violence. Large volumes of crime data are gathered by law enforcement and intelligence agencies in order to forecast future incidents [6]. Due of the large amount of data, manual data analysis techniques are ineffective. Consequently, one of the most important issues that law enforcement organizations still need to resolve is crime analysis. There are several approaches to crime analysis. Numerous crime data mining procedures are available, containing clustering systems, classification, association rule mining, sequential pattern mining, and string evaluation. Data from SM (social media) platforms can be utilized to analyze and forecast crime.

1.2 Offences in Social Network

Social media has also led to a rise in worries about crime in general. People frequently fall victim to scams on social media sites. Not all of the news is bad, though. Among other things, criminal justice organizations now have more ways to solve crimes because of social media. As a result, similar to many other technical innovations, social media offers advantages and disadvantages when it comes to how it interacts with the legal system and criminal justice system.

1.2.1 Cyber crime

Cybercrimes are offenses committed on or through the usage of the Internet. These encompass a wide range of illicit activities. Due to the anonymous aspect of the internet, there are a lot of unsettling actions taking place in cyberspace that could allow the offenders to engage in a variety of cybercrimes [7]. Since technology is the weapon

used in cybercrimes, the majority of those who commit them are technically proficient individuals who have a deep understanding of computers and the internet. Cyberstalking, cyberterrorism, email spoofing, bombing, cyberpornography, cyberdefamation, polymorphic viruses, worms, and other recently developed cybercrimes are some of the most common ones [8]. If certain traditional crimes are done online or via the internet, they may also be classified as cybercrimes. The Indian Penal Code lists several types of offenses that are punishable, including theft, mischief, fraud, deception, pornography, intimidation, and threats [9].

Cybercrime is currently a global issue that is growing not just in India but also worldwide. The frequency of this crime is directly correlated with a nation's advancements in computer technology. More than 50% of websites in the US, Canada, and Europe have suffered security breaches and cyberterrorism threats, according to the United Nations International Review of Criminal Policy on Prevention and Control of Computer Crime19 report. This presents a significant challenge to law enforcement agencies. Terrorist training is a new concept that has developed in recent years among militants. For militants, the internet has evolved into a vital teaching tool that they utilize to instruct new recruits in cyberterrorist training facilities.

1.2.2 Social Media Crimes

Individuals of all ages and genders are increasingly creating profiles on online social networks so they can communicate with one another in this virtual environment. Some people have thousands or even hundreds of friends and followers split over several accounts. However, the expansion of phony profiles is also occurring at the same time [10]. Oftentimes, fraudulent profiles bombard authentic people with unsuitable or unlawful stuff. Additionally, false profiles are made, portraying well-known individuals in order to harass them. The following are the most popular targeted websites/apps used to create "Fake Profiles": 1. Facebook, 2. Twitter, 3. Instagram, 4. LinkedIn, and 5. WhatsApp, etc.

An online site that focuses on forming social networks or associations between people who might want to exchange images, videos, backdrops, sports, or real-life connections, as well as talk and scaring, is known as a social media service. A social community service provides a user's profile, which is an overview of the user's social

relationships, as well as a variety of other services. The majority of social community services are focused on the internet and provide ways for users to instantaneously message one another or converse online. Online social offerings are group-oriented, whereas social community services typically explain as individual-oriented services. Despite this, online network services are classified as social network services. Social networking websites give users the ability to share thoughts, interests, sports, and pastimes with those in their private networks. The details are no longer private once they are posted on a social networking platform. The more information a customer publishes, the more opportunities they may have. Even with the highest security settings enabled, friends or websites may unintentionally leak people's information.

1.3 Criminal offences committed on social media

This section examines the usage of social media (SM) for criminal activities, as well as the various offenses that are perpetrated through SM platforms. There are three distinct categories for the utilization of SM in the command of crimes: manipulation, coordination, and abuse. The specific crimes associated with each of these categories are discussed accordingly.

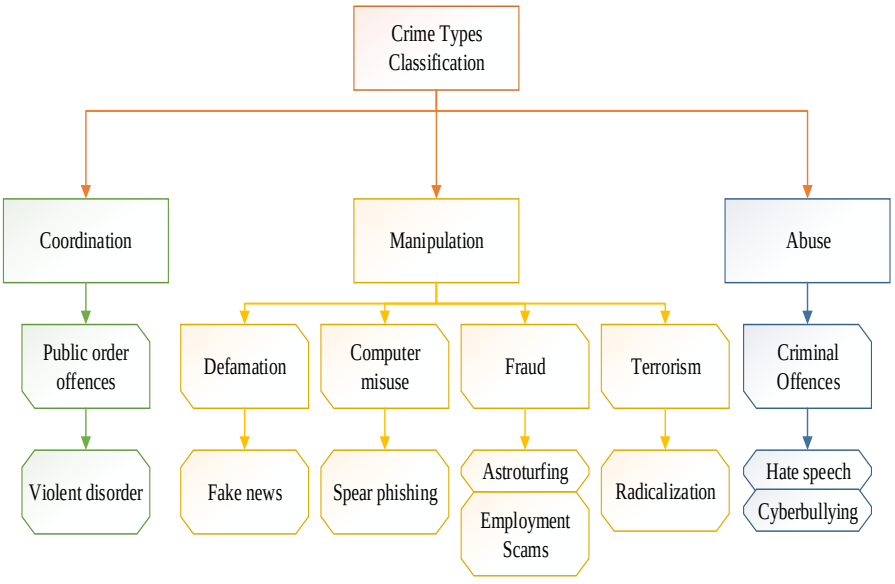

Figure 1.1: Crime type classification based on exploitation

1.3.1 Coordination

Coordination involves the utilization of SM platforms to orchestrate and facilitate criminal activities by organizing one or more individuals to engage in unlawful acts [11]. Social media is employed as a means to incite or communicate with individuals, encouraging them to partake in criminal actions.

1.3.1.1 Public Order offences

Public order offenses encompass a range of unlawful activities that disrupt or compromise societal harmony and safety. Among these offenses are actions such as affray, violent disorder, and unlawful assembly. Affray refers to the engagement in violent and tumultuous behavior that causes fear among the public, often occurring in public spaces [12]. Violent disorder involves multiple individuals engaging in violent conduct, creating a risk of harm to others. Unlawful assembly pertains to the gathering of individuals with the intent to carry out actions that breach public peace or safety. These public order offenses are critical concerns for law enforcement, as they involve disturbances that can lead to broader social unrest and jeopardize the well-being of the community. Addressing and preventing public order offenses are essential components of maintaining a secure and orderly society.

Violent disorder

Violent disorder is a criminal offense characterized by a group of individuals engaging in tumultuous and violent behavior that poses a threat to public order. This disorderly conduct typically involves multiple participants acting in a way that causes fear, alarm, or distress to the public. Examples may include rioting, brawls, or other forms of collective violence where the actions of the group escalate beyond what is deemed acceptable within the norms of a peaceful society [13]. Law enforcement agencies often intervene to control and disperse such disorderly groups to restore public safety and prevent further escalation of violence. Prosecution for violent disorder aims to deter and address collective violent behavior that undermines public peace and security.

1.3.2 Manipulation

A manipulation exploit is usually when a criminal attacker tries to trick a victim or target into doing something they wouldn't ordinarily do. An assailant uses language to control their victim. Language, whether used in posts or direct messages, aims to influence a target's behavior in the virtual or real world. The manipulation exploit type is associated with the following five offenses: 1. Employment scams; 2. Spear phishing; 3. Radicalization; 4. Astroturfing; and 5. False news. Spear-phishing attacks are targeted at tricking people into disclosing private information, like bank account details [14]. Frauds known as employment scams trick its victim into committing crimes by posing as job offers. The process of "radicalization" involves manipulating someone to become more violently inclined. Techniques that present an inaccurate story about a topic or person include astroturfing and fake news.

1.3.2.1 Computer misuse

Computer misuse, refers to unauthorized activities and offenses related to computer systems, networks, and data [15]. This can include various malicious actions carried out with the intent to compromise the integrity, confidentiality, or availability of information stored on computers. While legal definitions and specific offenses may vary by jurisdiction, common aspects of computer misuse often include unauthorized access, data breaches, and disruptive actions. Perpetrators of computer misuse may employ various techniques to compromise computer systems, ranging from exploiting software vulnerabilities to social engineering tactics. The motivations behind such activities can vary widely, including financial gain, information theft, political motives, or simply causing disruption. Computer misuse poses a significant threat to individuals, organizations, and even governments, as the reliance on digital systems continues to grow. Activities falling under the umbrella of computer misuse can include hacking, malware distribution, denial-of-service attacks, and unauthorized access to sensitive information.

Efforts to combat computer misuse typically involve legal frameworks and legislation designed to deter and prosecute individuals engaged in these activities. International cooperation is often crucial, as computer misuse can transcend national borders. Treaties, agreements, and collaborative efforts between countries aim to address the

global nature of cybercrime and facilitate the prosecution of offenders regardless of their physical location. It's important for legal systems and law enforcement agencies to stay updated on emerging threats and adapt their approaches to effectively address the evolving landscape of computer misuse. Public awareness and education also play key roles in mitigating the risks associated with computer misuse, as individuals and organizations need to implement robust cybersecurity measures to protect against potential threats. This broad offense encompasses one specific crime, known as spear phishing. Because spear phishing aims to obtain unauthorized admittance to a computer system, it is prohibited by the Computer Misuse Act.

Spear Phishing

Neural networks and other machine learning techniques can be utilized to automate spear phishing assaults directed at specific people with social media presences [16]. The '5C' phases of automated spear phishing are: Collect, Construct, Contact, Compromise, and Contagion [14].

- Phase of Collection: The attacker obtains details about the targets of their attack during this phase. In order to find appropriate profiles and retrieve personal information from tweets and other publicly accessible content, data collecting techniques include the use of APIs and keywords.
- Construct Phase: Using the data gathered in the preceding phase, the attacker creates messages and content. Creating messages that are convincing and realistic for use in the ensuing contact phase is the focus of this phase.
- Phase of Contact: Using an account that the target has followed or friended, the attacker gets in touch with the victim. The personal material that the target account generates is frequently used to automatically produce the contact messages. There has been a reported click-through rate of 66% for the communications, which usually have a malicious URL payload.
- Phase of Compromise: During this stage, malware is loaded into the target's device, giving the attacker access to the social media network without authorization [17].
- Phase of Contagion: The attack is then propagated by using the compromised account to contaminate additional accounts connected to the target account.

When using data that has been compromised from social media, automated spear phishing is especially effective. This attack is based on creating believable emails that targets are likely to believe and clicking on a link that either takes them to a compromised website where credentials are stolen or to a payload that compromises the target's device. Natural Language Processing (NLP) improvements like BERT [17], GPT-2, GPT-3, and XLNET have improved Natural Language Generation (NLG) as well as Natural Language Understanding. These advancements help exploit tools produce more realistic, artificial exploit emails, which in turn raises the click-through rates of harmful emails. This demonstrates the increasing complexity of automated Spear Phishing attempts and emphasizes the need for constant vigilance and the use of cutting-edge security tools to thwart such threats.

1.3.2.2 Fraud

In the realm of computer misuse on a global scale, fraudulent activities are a prevalent concern. While specific legal frameworks may vary between jurisdictions, a common thread in addressing computer-related fraud can be found. Instances of computer fraud typically involve unauthorized access, deceptive practices, or the use of false representations to gain advantages or cause harm. Such fraudulent activities transcend geographical boundaries, necessitating a collective and international effort to combat them effectively. One notable international framework addressing computer-related fraud is the Council of Europe's Convention on Cybercrime, commonly recognized as the Budapest Convention. In order to combat fraud and other types of cybercrime, this convention seeks to promote international collaboration, align national laws, and improve investigative techniques. Additionally, initiatives by international organizations and collaborations between countries contribute to the development of strategies to prevent and prosecute fraudulent activities in the digital realm

Astroturfing

Astroturfing is characterized as a deceptive online practice described as "a fake grassroots activity on the Internet". Its primary objective is to exert influence over lawmakers, elections, as well as campaigns of election. This deceptive tactic can be executed through the collaboration of nefarious groups, the use of automated methods such as social bots, or a combination of both [18]. Astroturfing campaigns typically

involve multiple sources disseminating and promoting the same message, a strategy believed to enhance the persuasive impact on the intended audience. Detection of Astroturfing campaigns on social media has become a subject of interest, with several techniques emerging in the existing researches. In the review of existing literature, two prominent methods for identifying Astroturfing campaigns were identified: author attribution as well as text classification. Both methods contribute to the development of automated systems aimed at recognizing and mitigating the impact of Astroturfing across various online platforms.

Employment Scams

The usage of social media in employment scams is common. Offers of work that appear genuine but really trick the recipient into committing crimes unintentionally are known as employment scams. Social networking is a technology used in recruitment that helps identify and target vulnerable people with fictitious job offers. Money mules and reshipping were two examples of representative employment scams that were commonly seen. Those that transfer money obtained through illicit means are known as money mules [19]. Mules receive very less compensation and frequently aren't aware that the money transfers are unlawful. Mules frequently think they are employed legally. It is necessary to locate and recruit mules, and SM can be a great resource for recruiters to do just that.

Reshipping scams enlist people to work from home, sending products purchased using credit cards that have been stolen to them, who then forward the goods to the original thieves. Because the goods are being distributed to several valid locations, this fraud makes it more difficult for the merchants to identify illegal activity. Employment Recruitment Job offers on social media platforms can be easily manipulated. There are a number of telltale signs that a job offer shared on social media is fake. Interviews with "scammers" have revealed that: "Posts that promise large amounts of money for very little work" and "Employers that use the candidates own bank account to transfer their money" are likely to be scams, notwithstanding the lack of formal academic research on the subject. Recruiters target users that fit a particular profile, in this case, those who are struggling financially, as is the case with some illegal acts carried out on social media.

1.3.2.3 Terrorism

Terrorism, on a global scale, encompasses a range of acts intended to instill fear, coerce, or intimidate populations or governments. It often involves violent actions carried out by individuals or groups with ideological, political, religious, or social motivations. The nature of terrorist activities varies widely, including bombings, hijackings, kidnappings, and cyber-attacks. Counterterrorism efforts worldwide are aimed at preventing, mitigating, and responding to these threats, often involving collaboration between nations and international organizations. Legal frameworks, such as the United Nations' Global Counter-Terrorism Strategy, provide a basis for international cooperation to address the challenges posed by terrorism. Additionally, many countries have enacted legislation to define and combat terrorism within their jurisdictions, with a focus on preventing radicalization, disrupting terrorist networks, and prosecuting those involved in planning or executing acts of terror. International cooperation and information sharing play crucial roles in the global effort to combat terrorism, as threats often transcend national borders.

Radicalization

Radicalization, a complex and multifaceted phenomenon, can be broadly understood as the procedure by which individuals accept extreme beliefs, ideologies, or viewpoints that deviate significantly from the mainstream societal norms. For the purposes of this, two key dimensions of radicalization, as delineated by a particular source [20], are considered: violent and non-violent radicalization. Violent radicalization is characterized as a progression in which an individual increasingly embraces the usage of undemocratic or violent means, which may include terrorism and nationalism, in pursuit of specific political or ideological objectives. This form of radicalization involves a departure from democratic principles and a willingness to resort to extreme measures. On the other hand, non-violent radicalization is defined as the active pursuit of or support for substantial societal changes that pose a potential threat to the democratic legal order. This pursuit may involve the endorsement of undemocratic approaches that, while not necessarily violent, can still jeopardize the functioning of the democratic legal order. Non-violent radicalization, therefore,

encompasses activities that seek far-reaching societal transformations, potentially undermining the democratic foundations of a legal order.

1.3.2.4 Defamation

The act of making untrue statements about a person or organization that damage their reputation is referred to as defamation. Such statements, often termed defamatory, can take the form of spoken words (slander) or written words (libel) [21]. Defamation typically involves communicating false information that harms the subject's character, integrity, or standing in the community. The legal implications of defamation aim to protect individuals and organizations from unjust harm to their reputation. To establish a defamation claim, certain elements are commonly considered, such as the false nature of the statement, its publication to a third party, the resultant harm to the subject's reputation, and, in some jurisdictions, a lack of privilege or a recognized defense. Laws regarding defamation vary across jurisdictions, but they generally seek to balance the right to freedom of expression with the need to protect individuals from false and damaging statements. In many legal systems, the onus is on the claimant to demonstrate that the defamatory statement meets certain criteria, such as being false and causing actual harm to their reputation. Defamation laws may also provide for various defenses, such as truth, privilege, and in some cases, public interest.

Fake News

Certain methods of detecting fake news on SM use specific traits to identify and flag the content for deletion. “1. The false knowledge it carries, 2. Its writing style, 3. Its propagation patterns, and 4. The credibility of its creators and spreaders” are the major traits of fake news that can be exploited by automated systems.

On social media, fake news can be disseminated automatically by social bots as well as manually by troll farms and retweet networks. Similar to how they are explained in the section on spear phishing, social bots automatically post content that appeals to people's preconceptions, which is subsequently shared by those in their network. Given the results of [22], it is unlikely that these strategies have been particularly effective given the small amount of people who have actually come into touch with fake news.

1.3.3 Abuse

This technique involves offenders leveraging SM platforms to engage in the abuse of specific individuals or groups. Abuse tactics may encompass various forms, including the issuance of threats, the use of racial epithets, or the hurling of insults dependent on a protected characteristic of the targeted group or individual [23]. The anonymity and accessibility provided by social media platforms can amplify the impact of such abuse, enabling offenders to reach and harm their targets on a broader scale. This form of online abuse raises concerns about the potential psychological and emotional impact on victims, highlighting the need for effective strategies to address and prevent such behavior in the digital realm. Legal actions, platform rules, and educational programs to encourage a safer and more polite online environment are frequently combined in attempts to prevent online harassment.

1.3.3.1 Criminal Offences

Abusive behavior on SM is addressed by various legislations globally. Legal frameworks typically outline the offenses related to abusive messages as well as behavior on social media. Two offenses identified in the existing studies fall under the abusive categorization: Hate Speech as well as Cyberbullying. Hate speech involves the use of derogatory and insulting language targeting minority groups depend on their protected characteristics. Cyberbullying, on the other hand, entails the use of bullying language, often containing threats and insults directed at the target. These offenses highlight the diverse nature of criminal behavior on social media platforms, emphasizing the need for legal frameworks to address and mitigate such actions globally.

Hate Speech

SM messages that are deemed to be "offensive" may be illegal in some jurisdictions, and its authors may be subject to fines, jail time, or public censure. It should be mentioned that hate speech is not illegal in a number of countries, including the USA, thanks to free speech rights. Hate speech derives from mature issues including angry message, cyberbullying, and "flaming". Because hate speech is well-defined by the victim group, it can be difficult for automated systems to identify it [24]. For example,

a group that would normally view criticism as acceptable may interpret it as hate speech against another group. The ambiguous character of hate speech is demonstrated by the recent problem of the Gender Critical community's criticisms of the Trans community.

Cyberbullying

Cyberbullying, the practice of utilizing social media as well as other Internet-enabled tools to abuse vulnerable persons, is undoubtedly the precursor to hate speech. Cyberbullying can be characterized as "an intentional, aggressive act or behavior against a recipient who is unable to defend themselves, carried out by a group or an individual, using electronic forms of contact, repeatedly and over time". Cyberbullying can result in "depression, low self-esteem, behavioral problems, and substance abuse" for the victim, despite the fact that it may not involve violent acts. Social media businesses have a responsibility to identify and remove posts that engage in cyberbullying due to the tangible effects of such behavior.

1.4 Crime Rate Prediction

The substantial volume of big data generated from urban areas is characterized by noise, extensive scale, heterogeneity, and dynamism. Consequently, addressing this data requires effective and efficient computational solutions. Various proposed computational approaches aim to improve crime research in urban areas. In this section, discussed about key computational tasks and appropriate algorithms designed to predict criminal activity in urban settings. Crime rate prediction forecasts the future crime rate in a certain area. This category classifies crime prediction models depend on the data sources they utilize, such as social media, environmental context, and crime statistics.

1.4.1 Prediction Based on Crime Data

In the realm of crime rate prediction, one approach involves leveraging crime data as a primary source of information. This method involves analyzing historical crime data to identify trends, patterns, and correlations that can contribute to forecasting future criminal activities. Various algorithms and analytical techniques are employed to extract meaningful insights from the available crime data, aiding in the development

of predictive models for anticipating crime rates in specific areas. This data-centric approach shows a crucial role in enhancing the accuracy and effectiveness of crime rate predictions.

In smaller locations, such police precincts, several research studies have suggested crime prediction up to thirty days in advance. The predicted accuracy of the police methods was compared with that of univariate time series models. A fixed effect regression model with a 100% prediction error states that in order to obtain a prediction error of less than 20%, the average number of offenses must be higher than thirty. Holt exponential smoothing is also known to be the most accurate model for predicting crime at the precinct level [25]. An autoregressive integrated moving average (ARIMA) is a tool for forecasting potential future property crimes. In order to forecast crimes one week in advance, the ARIMA model uses data from 50 weeks of property crimes. When considering forecast accuracy and fit, the ARIMA model outperforms exponential smoothing.

1.4.2 Prediction Based on Environmental Context Data

Forecasting crime based on environmental context data involves leveraging information about the surroundings and external factors that may influence criminal activity. This category of prediction models analyzes data related to the physical environment, weather conditions, and other contextual elements to infer potential patterns in criminal behavior. By examining how environmental factors intersect with crime data, these models contribute valuable insights into understanding and predicting urban criminal activity. Algorithms in this category aim to decipher the nuanced relationships between environmental features and crime occurrences, providing a holistic approach to crime rate prediction.

1.4.3 Prediction Based on Social Media Data

Prediction dependent on SM data is a dynamic and evolving field that harnesses the vast quantity of information created on social media platforms for anticipating various outcomes, including potential criminal activities. Social media data prediction involves the analysis of user-generated content, interactions, and patterns to forecast trends and behaviors. One key aspect of social media data prediction is the utilization

of ML algorithms to analyze and interpret user activities. These algorithms can identify patterns, anomalies, and correlations within social media data, enabling the prediction of events or trends, including those related to criminal activities.

NLP methods play a vital role in comprehending the textual content shared on social media. Sentiment analysis, entity recognition, and topic modeling are some of the NLP methods employed to extract meaningful insights from user posts, comments, and messages. This information contributes to predicting potential criminal behavior or events based on the sentiments and discussions within the social media sphere. Furthermore, social network analysis is instrumental in examining the relationships and connections between users. By studying the structure of social networks, authorities can identify influential individuals, potential collaborators in criminal activities, and emerging clusters of interest. This network-centric approach enhances the accuracy of predictions related to criminal collaboration and coordination. The real-time nature of social media data provides a valuable advantage in prediction, allowing law enforcement agencies to stay abreast of evolving situations. By integrating social media data into predictive models, authorities can proactively address and prevent criminal activities, leveraging the wealth of information generated by users on these platforms.

1.5 Crime predictive analytics

Several methods based on DL and ML are utilized in crime predictive analysis. A pre-processing procedure receives the prepared data set under certain parameters. Following the completion of the pre-processing method, text format is used to extract the features. Next, a cluster is created based on certain demands using categorization [26]. Prediction is the word for the classification method, which uses training and test models in supervised learning mode to identify information about unknown crimes. In social networks, this architecture can be applied to crime prediction analysis. The success and peace among the nationals are ensured by this system. Several of the approaches shown in Fig. 5 are used to carry out the crime predictive analysis. It has multiple states, and machine learning is used in each state to predict crime clusters based on crime documentation. The information gathered from the traits and actions of the offender is used to generate the data set.

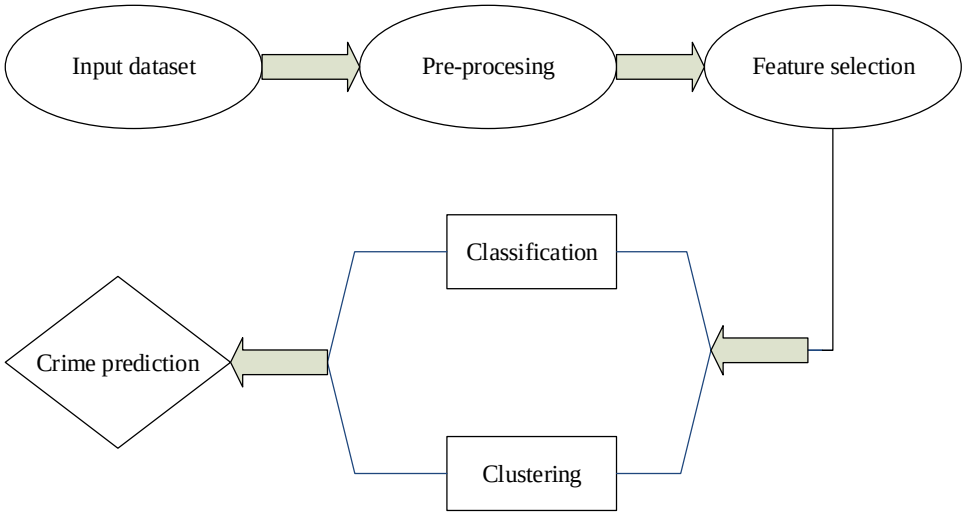

Fig 1.2: Crime prediction technique

Today, identifying criminal activity in social media postings, comments, likes, and tweets is one of the most important methods. ML has been used extensively in research and development. The user is under pressure from the compromised accounts; in some cases, the criminals have left the victim with no other option except to commit suicide. Thus, based on their posts, each social media user's behavioral profile is generated by the COMPA algorithm. Every new action the user takes is analyzed against the datasets that were previously provided. In the event that behaviors deviate from the dataset, the algorithm can identify an irregularity. It can therefore distinguish between regular and hacked social media accounts, even those with large followings. The method of identifying spam is promising in 83% of spam accounts. However, almost all of them are impartial, meaning they don't have any previous ideas regarding the information they are given. To detect cyberbullying and abuse, the platform uses machine learning, data mining, and language evaluation techniques. The results of the writers' analyses never offer any legally-sound justification for harassment and cyberstalking. Giving them evidence-based policies to deal with harassment improves their performance. It is unrealistic for the author to expect every client to be lovely and trustworthy. Some clients are misinformed, and their behavior could undermine other clients' presumptions. Research on criminal analysis is divided into three categories: criminal analysis based on sentiment analysis, criminal analysis based on data mining techniques, and criminal analysis based on different prediction techniques. As seen in Fig. 6, the criminal analysis based on different prediction techniques is further divided

into the following categories: text content as well as NLP depend technique; crime patterns and evidence-oriented technique; spatial and geolocation-oriented technique; prisoner-based technique; and communication-oriented technique.

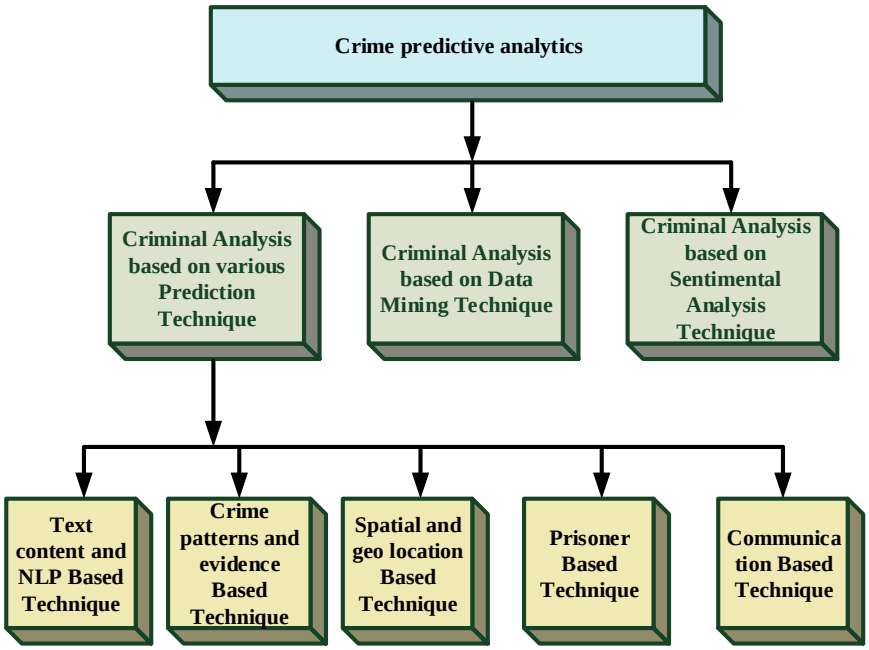

Fig 1.3: Types of Criminal Analysis Techniques

1.5.1 Criminal analysis methods for crime data prediction

In the realm of predicting social media crimes, various criminal analysis methods have been developed to improve the accuracy as well as efficiency of crime data forecasts. Notably, text content and NLP based methods play a decisive role in deciphering patterns within textual data, enabling a deeper understanding of potential criminal activities. Crime pattern and evidence-based methods leverage historical data and patterns to identify trends and provide insights into the modus operandi of criminals. Special and geo-location-based methods focus on the spatial aspects of crimes, utilizing geographical information to predict potential hotspots and crime-prone areas. Predictive analysis of crime involving prisoners delves into the analysis of individual criminal histories and behaviors to anticipate future offenses. Additionally, communication-based predictive analysis explores the relationships and interactions

between individuals on social media platforms, offering valuable insights into potential criminal collaborations and activities. The amalgamation of these diverse methods contributes to a comprehensive approach in predicting and preventing social media crimes.

1.5.2 Crime Data Mining Techniques for Predictive Analysis of Crime Data

Crime data mining methods are instrumental in extracting valuable patterns, trends, and insights from vast datasets, facilitating predictive analysis for crime prevention. One key technique involves association rule mining, which identifies relationships and connections within crime data, revealing potential patterns of criminal behavior. Clustering algorithms group similar crime incidents together, aiding in the identification of crime hotspots and enabling proactive law enforcement strategies. Another effective crime data mining technique is anomaly detection, which focuses on identifying unusual or suspicious patterns in the data that may indicate criminal activities. This method is particularly useful in recognizing outliers and anomalies that deviate from expected norms. Classification algorithms play a pivotal role in predictive analysis by categorizing crime data into different classes, helping law enforcement agencies anticipate and prioritize potential criminal activities. Additionally, decision trees provide a structured framework for analyzing crime data, enabling authorities to make informed decisions dependent on identified patterns and criteria. Furthermore, temporal data mining techniques consider the temporal aspects of crime data, such as time trends and seasonality, to improve the accuracy of predictions. This ensures that law enforcement is equipped to address fluctuations and patterns that may vary over time. The integration of these crime data mining techniques offers a comprehensive and sophisticated approach to predictive analysis, empowering law enforcement agencies to proactively address and prevent criminal activities based on insights derived from data-driven methodologies.

1.5.3 Crime Prediction Using Sentiment Analysis

Sentiment analysis is a valuable tool in crime prediction, leveraging the examination of emotions, opinions, and attitudes expressed in text data to discern potential criminal activities. This technique involves the analysis of social media content, news articles, and other textual information to gauge the sentiments associated with specific topics

or events. In the context of crime prediction, sentiment analysis can be applied to assess public sentiment towards certain issues, events, or locations. By monitoring online discussions and reactions, law enforcement agencies can gain insights into the public's perceptions and concerns, potentially identifying areas of heightened risk or emerging threats. Additionally, sentiment analysis can be employed to analyze the emotional tone of specific messages or posts, helping to identify potential indicators of criminal intent or activity. Unusual patterns in sentiment, such as a sudden surge in negative emotions or alarming language, may signal the need for further investigation. Social media platforms, being a rich source of real-time data, are particularly conducive to sentiment analysis for crime prediction. Monitoring user sentiments on these platforms provides a dynamic and timely perspective on evolving situations, enabling law enforcement to respond promptly to potential threats. By integrating sentiment analysis into crime prediction models, authorities can enhance their ability to anticipate and mitigate criminal activities, utilizing the power of public sentiment as an additional layer of information for proactive decision-making.

1.6 Problem statement

The increasing crime rates, exacerbated by population growth, underscore the critical need for efficient crime analysis as a legislative function. Crime analysis plays a pivotal role in identifying disorder, patterns, and trends in criminal activities. The surge in crimes, including dowry deaths, attempted rapes, thefts, extortions, robberies, and dacoities, emphasizes the urgency for comprehensive crime surveying. To address this, the government must make informed decisions to maintain law and order and ensure the safety of citizens. The application of Big Data Analytics (BDA) to raw, disorderly data from numerous sources such as social media, manufacturing, and education sectors emerges as a promising solution. However, challenges persist in the realm of online social networking, where criminal activities pose a threat to users. Existing methods for Twitter-based crime detection exhibit limitations, including issues related to security, privacy, feature selection algorithms, and the complexity of event/data collection from the Twitter stream. Moreover, the prevalence of irregular language in tweets, the use of code for illegal communications, and the manual recognition of crime tweets present significant obstacles. The proposed study aims to address these challenges by developing an advanced crime detection model in the

Twitter environment, offering a more effective and efficient solution to the complexities associated with crime detection in social media datasets.

1.7 Motivation

The pervasive use of social media as a platform for expressing views, sharing personal events, and connecting with others highlights its integral role in contemporary communication. However, the dual nature of social media, where individuals exploit it for criminal activities, underscores the importance of leveraging technology for defense, control, deterrence, and investigation of crimes. The prediction of crimes through advanced technology not only safeguards the national critical information infrastructure but also ensures the security and privacy of users. Twitter, as a widely used online platform, provides a substantial dataset for crime detection, presenting an opportunity to enhance user privacy and security. The motivation behind this study lies in recognizing the transformative potential of crime detection on Twitter in establishing a more secure and private communication process for millions of users. The imperative to address the escalating growth of illegal activities on SM, which directly impacts users' lives, is underscored by instances such as suicide attempts resulting from privacy breaches. The complexity of crime detection, attributed to the vast array of features, necessitates the application of DL techniques to enhance accuracy and system performance, making this research essential for the advancement of technology-driven crime prevention on social media platforms like Twitter.

1.8 Objectives of the Research

Some of the major objectives of the research works are given below:

To Perform review of existing literature on Predictive Analytics in Crime Data over online social media.

- To analyze extensive crime data using Apache Pig and a variety of grunt shell commands in conjunction with the Hadoop distributed file system using a big data analytic approach.

To propose a framework for Crime Prediction Modeling.

- To predict different crimes on the online Twitter platform with less time complexity, the hybrid wavelet convolutional neural network (WCNN) with world cup optimization (WCO) is provided.
- To improve CNN performance, it is integrated with the wavelet approach, and the function of loss is then adjusted utilizing WCO.
- To use the modified tree growth algorithm (MTGA), an efficient feature selection approach, to lower the dimension of the features and enhance the overall performance of crime detection.
- The fuzzy manta ray foraging (FMRF) method is applied to improve the clustering process. MRF method applied to clustering based on fuzzy logic.

To propose algorithm and method for Crime Detection on Social Media Content.

- In order to provide strong Twitter crime detection, DAC-BiNet, an efficient Deep Attention Convolutional Bi-directional Aquila Optimal Network model, is presented.
- To create a cluster-based model by minimizing the computational complexity through the use of Possibilistic Fuzzy LDA, which reduces the size of the original feature set.

To do a comparative Analysis of Existing and Proposed Technique for Predictive Analytics in Crime Data over online SM.

1.9 Thesis Organization

Chapter 1: Introduction

This chapter delves into social network offenses, encompassing cybercrime and various social media crimes, with a focus on categories such as coordination, manipulation, and abuse. It further explores crime data prediction, classifying data into crime data, environmental context data, and SM data. The discussion extends to crime predictive analytics, categorizing techniques into text content/NLP, crime patterns/evidence, spatial/geolocation, prisoner-based, and communication-based methods. The chapter addresses the problem statement, motivation, objectives, and

provides an organizational overview of the thesis. In conclusion, a summary encapsulates the key insights gleaned from the comprehensive exploration of criminal offenses and predictive analysis in the realm of social media.

Chapter 2: Literature Survey

This chapter reviewed crime data prediction models, presenting a comparative analysis of their strengths and limitations in a tabular format. Each model's advantages and disadvantages were carefully examined to provide a comprehensive overview of their respective efficacy. The chapter completes with a concise summary, encapsulating the key findings and insights derived from the comparative assessment of crime data prediction models.

The following paper has been communicated from this work:

- A Research Paper entitled "Predictive Analytics of Crime Data in Social media: A Systematic review, incorporating framework, and future investigation schedule" in **SN Computer Science, Springer Publisher, Scopus Indexed**

Chapter 3: Automatic Twitter Crime Prediction Using Hybrid Wavelet Convolutional Neural Network with World Cup Optimization

This chapter discussed about automatic twitter crime prediction utilizing hybrid wavelet CNN with WCO. In this discussed about Data cleansing, extraction of feature, Feature selection using MTGA, Clustering utilizing FMRF, and Crime detection using hybrid WCNN with WCO. Additionally result and discussion for the proposed model also provided and finally concluded with summary.

The following paper has been published from this work:

- Monika, Aruna Bhat, "Automatic Twitter Crime Prediction Using Hybrid Wavelet Convolution Neural Network with World Cup Optimization", World Scientific Publisher, International Journal of Pattern Recognition and Artificial Intelligence (2022), Volume 36, No.5, DOI: 10.1142/S0218001422590054. (**SCIE, IF:1.6**)

Chapter 4: DAC-BiNet: Twitter crime detection using deep attention convolutional bi-directional aquila optimal network

This chapter discussed about the proposed DAC-BiNet with Aquila optimal network. It includes about the pre-processing, feature extraction using improved TF-IDF, feature hashing and glove modelling, Possibilistic Fuzzy LDA based clustering for feature reduction and DAC-BiNet depend twitter crime detection and classification. Additionally result and discussion of the proposed model are provided. Finally concluded with summary.

The following paper has been published from this work:

- A Research Paper entitled "DAC-BiNet: Twitter Crime Detection using Deep Attention Convolution Bi-Directional Aquila Optimal Network" in Multimedia Tools and Applications (2023), Springer Publisher, DOI: 10.1007/s11042-023-17250-4. (**SCIE, IF:3.6**)

Chapter 5: ANALYZING EXTENSIVE CRIME DATA WITH APACHE PIG AND HADOOP

This chapter discussed into the analysis of crime data utilizing Apache Pig on big data. It encompasses the design framework and implementation, featuring an examination of the frequency of crimes against girls and women over four years, the frequency of crimes in India categorized by types, and the frequency of crime accusations in specific states along with the corresponding results. Finally, this chapter concluded with summary.

The following paper has been published from this work:

- **Monika, Aruna Bhat**. "An analysis of Crime Data under Apache Pig on Big Data", 2019 3rd IEEE International Conference on I-SMAC (IOT in Social, Mobile, Analytics and Cloud), Palladam, Tamil Nadu, India, 11-13 December, 2019 (**IEEE Xplore Digital Library**) **Scopus Indexed**

Chapter 6: PREDICTIVE ANALYTIC IN CRIME DATA ON SOCIAL MEDIA: COMPARATIVE ANALYSIS

This chapter discussed about the comparative analysis of different methods for predictive analytics in crime data over online SM. Also provided with results and discussion for the different techniques. Finally concluded with summary.

The following paper has been published from this work:

- A Research Paper entitled "Comparative Review of Different Techniques for Predictive Analytics in Crime Data Over Online Social Media", 5th IEEE International Conference on Data & Information Sciences (ICDIS-2023), Bichpuri, Agra, India, 16-17 June, 2023. (**Scopus Indexed)**

Chapter 7: CONCLUSION AND FUTURE SCOPE

This chapter discuss about the conclusion of the entire thesis along with the future enhancements.

List of Publications: This section lists published/accepted/communicated papers relating to this research work in International/National Journals/Conferences of repute.

References: This section is the list of references cited in this research work.

1.10 Summary

In this chapter, extensively examined social network offenses, covering cybercrime and various types of social media crimes. The discussion focused on categories like coordination, manipulation, and abuse within social media contexts. Moving beyond, the exploration delved into crime data prediction, where data was systematically classified into crime data, environmental context data, and social media data. A detailed analysis of crime predictive analytics ensued, categorizing techniques into text content/NLP, crime patterns/evidence, spatial/geolocation, prisoner-based, and communication-based methods. The chapter also addressed crucial elements including the problem statement, motivation, objectives, and provided a comprehensive organizational overview of the thesis. The concluding section summarized key insights derived from the thorough examination of criminal offenses and predictive analysis within the domain of social media, underscoring the significance of understanding and addressing emerging challenges in this evolving landscape.

CHAPTER 2

LITERATURE SURVEY

In this chapter, a meticulous exploration of prediction models applied to crime data was undertaken, featuring a comprehensive tabular analysis that intricately weighed the merits and demerits of each model. The primary aim was to conduct a nuanced assessment of the strengths and weaknesses inherent in these models, providing a well-rounded perspective on their effectiveness. By diligently evaluating the advantages and disadvantages, the chapter sought to offer a balanced viewpoint, enriching the understanding of these models' practical utility. The subsequent section distills the main findings and noteworthy revelations derived from the comparative study of crime data prediction models, encapsulating them in a concise yet insightful overview.

2.1 Overview

Social media is a computer-based platform that allows ideas, opinions, and information to be shared instantly. A growing number of people were making fake internet personas in an attempt to threaten, harass, intimidate, and stalk other people [27]. In many countries, this had resulted in a marked rise in crime rates. Security had become increasingly important as a result of technological improvements, which have forced thieves to adopt cunning tactics to commit their crimes. "Crimes" were defined as actions that cause physical or psychological injury to another individual or piece of property and were subject to criminal prosecution. SM platforms such as Twitter, Facebook, and Instagram have grown to be important sources of sentiment-rich data [28]. With 300 million users, Twitter was well-known for its messages, or tweets, which let users engage with one another in a virtual environment. However, since tweets reflect the

goals of the users, obtaining information from Twitter takes a lot of work. Care should be taken in the analysis of each tweet in order to increase the efficacy of criminal detection.

Internet crimes have increased recently, which had a big impact on social and economic activity. Crime episodes immediately affect the nation's social structures, economy, and international reputation. Stronger security measures need to be put in place in order to recognize and lessen online fraud and criminal behavior [29]. Recent technological advancements have made it necessary to replace traditional approaches for forecasting Twitter crime events with a variety model of DL-based. A wide range of crimes, such as cyberbullying, cybercrimes, election outcomes, crime against women, box office projections, crime against children, and other applications, are analyzed by the automated detection system [30]. Lawmakers and law enforcement can profit from the DL-dependent effective crime detection in order to quickly and wisely construct models that can stop online crime and support national growth. Large-scale criminal data was evaluated utilizing big data analysis based on Apache Pig and Hadoop distributed file system. Hardware assets were the planned use for the Hadoop file system. It was inexpensive and quickly deployable to fault in crimes. Large data sets can be processed quickly and easily using Hadoop's versatility. It was made up of several parts, such as the Map Reduce and Hadoop Distributed File System, which were utilized to store, process, and analyze data more effectively [31]. The data set was processed as a node in a map, providing reduced data in an excellent format.

Hadoop constituents include sqoop, oozie, hbase, apache pig, apache spark, and apache kafka. In addition to being a channel for illegal communication, SM also acts as a venue for community conversation. Law enforcement agencies were continuously keeping an eye on online communications activity in order to identify criminal activity. But also confront obstacles comprising cloud computing, unstructured text monitoring, tracking website troubles, huge data, privacy concerns, difficulty with real-time monitoring, anonymous identifications monitoring of multilingual text, and more. Tracking images, sounds, and video, among other things, was another challenge. A machine learning-based crime detection system that makes use of facial and text recognition algorithms was presented in the study [32]. Criminals use social media to plan, carry out, and complete crimes, and law enforcement uses it to keep an eye on,

prevent, defend against, and look into criminal activity. With 300 million users, Twitter offers a special means of communication for followers as well as a virtual online platform for users to interact with each other and the outside world. It takes a lot of effort to extract information from Twitter because tweets were thought of as people' intentions and can be transmitted as icons or in any format. Cyberbullying, online stalking, cyberhacking, cyberscam, and cyberharassment were examples of serious crimes committed via social media. Due to the danger and detrimental effects these crimes have on people's safety and privacy, it was critical to identify crimes as soon as feasible.

As more people create accounts to enter the virtual world, social media was a computer-based platform for exchanging ideas and information. On the other hand, people were also making fictitious profiles in order to harass, threaten, and stalk others. As technology develops, criminals adapt cunning ways to commit crimes, leading to an rise in the number of crimes committed. On the streets and in the realm of associations, like the Internet, crime was best noticed. It was essential to develop and adapt strategies to support intelligent creation and implement safety measures against these offenses [33]. For the purpose of detecting patterns in the future, predictive crime analysis methods like data mining, ensemble, machine learning, deep learning, and clustering were crucial for making fast and accurate predictions about criminal designs. A systematic review concentrating on methodologies and methods for predictive analysis of crime data in social media was part of the research being done to establish an intuition of predictive crime data mining tools.

This chapter reviews the relevant literature and identifies the key benefits and drawbacks of using the suggested techniques. These papers were displayed below, along with a comparison table that goes with them.

2.2 Literature Review

Social media crime predictions were based on a variety of predictive criminal analyses. Examine the best methods for analyzing and predicting crime statistics in this instance. The following was a description of some current research studies that have been conducted on crime detection utilizing various methodologies.

Lal, S. et al. [34] had to classify and examine tweets on cybercrime. This investigation focuses on the content and messages that were exchanged on Twitter about crimes. concentrating on the subsequent hashtags: The tweet details include things like #rape, #crime, #violence, #murder, #fraud, #cybercrime, #harassment, and #robbery. TF-IDF, hashtag expansion, sentence segmentation, stop word removal, punctuation removal, and extending slang and acronyms were among the several methods used to initially clean the data. The manually collected data contained tweets about crimes as well as tweets about non-crimes. The tool for the experimentation was WEKA. Accurate classification was carried out and analyzed using a number of classifiers, comprising J48 (97.0%), NB (97.83%), ZeroR (61.5%), and RF (98.1%). The primary limitation was the RF classifier's high temporal complexity, which took 7.24 seconds to complete.

AlGhamdi, M. A., & Khan, M. A. [35] developed a clever model. Using the TS, API, data from Arabic tweets was collected. Tweets that were deemed suspicious or not were manually assigned the labels 0 and 1. To make the suggested model better, it was filtered. A variety of supervised classifiers were used to model the tweet classification process. BoW based feature extraction was used to process the classifiers KNN, LDA, SVM, and DT, as opposed to the LSTM and ANN classifiers, which use word embedding. The accuracy of the SVM (86.72%) classifier was greater. The weaknesses were binary tweet classification and small data sets.

Santhiya, K. et al. [36] had focuses at the automated classification of tweets, which seems to be a particularly challenging task due to the nature, heterogeneity, and volume of data involved. Internet live statistics show that users give their opinions on almost 500 million tweets every day regarding various subjects. A decision support system that was automated was created to examine the tweets pertaining to crimes against women and children. Because of the nature of the data, the issue was seen through the lens of big data. The two technologies that will be developed as part of this project were the Hadoop MapReduce and the Apache Spark framework for big data programming. A distributed and parallel classification of several criminal categories is achieved by the hierarchical domain lexicon-based method. Moreover, the criminal classification tool is built on a hybridized technique combining machine learning and processing of natural language. The location of Twitter users was

predicted using a multinomial Naive Bayes classifier trained on Location Indicative terms and other significant criteria (such as city/country names, #hash tags, and @mentions). The technique performs better in terms of classification accuracy, mean, and median error distance among algorithms based on criteria such Location Indicative words, #hash tags, and city/country names.

Boukabous, M., & Azizi, M. [37] had described a method for classifying and detecting crimes that combines the BERT deep learning algorithm with a lexicon-based approach. Using the Twitter dataset, the approach presented in this paper achieved 94.91% classification accuracy, 94.94% precision, 16.26% loss, 94.91% recall, and 94.92 F1-score. The experimental outcomes show that the approach taken in this paper had potential efficacy in the detection of criminal activity. Additionally, it was evident that the hybrid approach depend on BERT performs better in crime detection than previous works.

Hissah, A. S., & Al-Dossari, H. [38] had presented methods for classifying and identifying crimes using text mining algorithms based on tweets written in Arabic. The Twitter network was used to collect data, and human labeling was done. Among the techniques used to show the pre-processing of the tweets were tokenization, which is filtering, normalization, and stemming. TF-IDF based technology was utilized to extract tweets in Arabic. Four distinct classifiers were used to analyze the tweet classification: DT (82.46%), KNN (78.06%), SVM (91.55%), and CNB (88.17%). KNN had the lowest value, however the largest classification was estimated using SVM. The CNB (Complement NB) classifier had the fastest processing time and speed. The study's failure to account for temporal and spatial information was one of its weaknesses.

AlJanabi, K. B. [39] provides an overview of the most popular studies carried out in this area and discusses the experiments, algorithms, data kinds, data quantities, and findings produced. One of the fundamental ideas in the arena of data mining was the prediction of the class label of labelled objects, or ambiguous class label. among its various methodologies, algorithms, and applications. The classification techniques (DT, Bayes, SVM, KNN, and others) that embody supervised learning were the most widely used methods in this field. Nevertheless, the novel clustering-classification

strategy was employed since there were frequently no target class labels and boundaries accessible to carry out the prediction. The study discovered that prior to classification, clustering approaches increased classification accuracy and shortened experiment execution times. With a summarization rate of more than 50%, the Cluster Classifier proved to be effective at reducing the size of the test dataset. Preprocessing the data, such as selecting and extracting features, increased classifier accuracy and performance while lowering classification error.

Krishnendu, S. G. et al. [40] had to anticipate the crime statistics, the sites were divided into many clusters. To cluster the input unlabeled data, an iterative unsupervised algorithm called K-means was used. The number of clusters that must be denoted by a constant K was determined in this initial stage. Then choose the centroids' K point at random. It was possible that these centroids were not derived from the data. After that, each data point was linked to the closest centroid, creating K clusters. Choose a location for the new centroid for each of the freshly formed clusters. Lastly, allocate each data point to the most recent closest centroid once more. The area was divided into five groups using this strategy. Every cluster had its own peculiarity, role, weapon, technique, and place of operation. The accuracy that the K-means algorithm attained was of 78%, precision of 60%, f measure of 44%, Recall of 45%, respectively.

Pepsi, M. B. B., & Kumar, S. N. [41] had discussed engineering students' learning process and the challenges encountered while studying through their tweets. The massive volume of data collected necessitates processing using the Apache Hadoop Map Reduce environment. The system pre-processes tweets, computes the F1 measure, identifies prominent categories, determines the likelihood of words and categories, and then assigns tweets to the appropriate categories. Heavy study loads, low social engagement, and sleep issues were detected using supervised learning approaches including logistic regression, multiclass SVM based Platt Scaling, and Naïve Bayes. When comparing the obtained findings, SVM yields an accuracy score of 84%, which was 5–10% higher than that of the Naïve Bayesian approach and Logistic Regression

Al-Saqqa, S. et al. [42] had provide fresh sentiment analysis evaluation tests using the Apache Spark data processing system on a sizable dataset of online customer reviews.

Three classification approaches were applied from the scalable MLlib of Apache Spark: LR, SVM, and NB. The accuracy metric was used to assess the outcomes. A SVM classifier outperforms the LR model and naïve bayes classifiers, according to the findings of the trial.

Savaş, S., & TOPALOĞLU, N. [43] had conducted in Turkey using Twitter data gathered over nearly a month, the most talked-about subjects throughout the study period were bomb attacks and terror incidents. The analysis discovered that there was no word root from the keywords of demonstration, smuggling, and prostitution, and that the keywords "bomb" and "terror" and "rape" have a strong association. Future research, according to the report, might employ dynamic keyword modifications to focus on intelligence and reach a larger audience. Government representatives might exercise prudence by adhering to the relevant associations. The methods and instruments employed in this study can be extended to commercial research with the goal of finding word connections to enhance goods and services. Compared to sentiment analysis research conducted for commercial purposes, the outcomes of this in-depth investigation will be more thorough. Commercial enterprises can benefit greatly from the application of these strategies in the areas of customer satisfaction research, association regulations, assistance, and feedback activities.

Garima, A., & Alaiad, A. [44] had presents the Crime Data Information System design. The Crime Database performs two procedures for crime analysis in addition to data pre-treatment. The outcomes of these two methods were compared and validated against ground truth. Identified two methods for locating the criminal hotspots. To identify crime hotspots and conduct in-depth investigations into crime data, have mined historical crime data sets using the standard K-means and spatial clustering algorithms. Also have also set up visualizations on Google Earth. Also discovered some fascinating information about high and low crime rate areas by analyzing the results with some real-world data. This project should assist Police in tracking crime episodes in real time throughout the city of Chicago, retrieving historical crime incidents for specific wards, and optimizing resource utilization in high-crime areas. Furthermore, like to suggest increasing the number of security personnel stationed in high-crime areas, such as wards #5, #6, #7, #8, #9, #10, #24, and #28.

Panja, B. et al. [45] had discovered Crime mapping analysis, which was made simpler by the use of ANN and KNN algorithms. The Office of Community Oriented Policing Services (COPS) was responsible for both doing and funding crime mapping. Research with an evidence base aid in the analysis of crimes. Utilizing data mining techniques, ascertain the crime rate by examining past data. Crime analysis uses analytical techniques along with qualitative and quantitative data to solve crimes. One area of study that was vital to public safety was the mapping of criminal activities. Using data mining tools, identify the places with the highest frequency of crimes.

Baby, A. et al. [46] had suggested model was a strategy derived from the findings of integrating the sentiment polarity and the polarity values of target speech, hate speech, and foul language in order to enhance the measure of cybercrime detection. The model's persistence lies in its capability to evaluate a sentence or consider cybercrime by considering the parameters of the crime and the emotions surrounding it. This research broadsheet offers an improvement over the existing method of identifying criminality on social media through prima facie detection. By combining sentiment analysis with three different crime detection parameters, able to measure and qualify targeted abuse detection and open up a lot of new possibilities. The combination of four detection parameters had created a path for improvement over the current system, which relied on either one for analysis.

Sharrab, Y et al. [47] had enhanced the use of deep neural networks to identify questionable content on social networking platforms. Any text that was strange or out of the norm, or that was likely to be connected to illegal conduct, was considered suspicious text. Through its assistance in enhancing the detection of such text, this study had the potential to significantly impact the field of text data forensics. For text data forensics research, made use of the "CIC Truth Seeker Dataset 2023", which was regarded as a thorough and representative dataset. Over 180,000 tweets that have been expertly categorized as genuine or fake news were included in the collection. In this work, improve text data forensics in SM by utilizing deep neural networks' potent analytical powers. More particularly, look into how well Long Short-Term Memory (LSTM) can identify questionable text. Initial evaluations yielded a 96% accuracy rate, which was rather encouraging. Future research on the model's possible uses, such as

the detection of fraudulent activity, the prevention of online harassment, and the identification of criminal conduct, was something intend to investigate.

Alduailaj, A. M., & Belghith, A [48] had implemented a sizable Arabic dataset with about 30,000 comments to train the SVM model. Because Twitter was a popular service for gathering text data for the purpose of classifying comments related to cyberbullying, evaluated the SVM model on a different dataset after that. Demonstrated that the best classification of cyberbullying was obtained by combining the TF-IDF vectorizer with SVM utilizing Farasa NLTK. After that, the obtained results were contrasted with the output of the NB classifier using various ngram range parameters and supplementary feature extraction, like BoW. To create fixed-length vectors, BoW uses CountVectorizer to count the number of times each word appears in the text. With a percentage of 95.742%, the results demonstrated that SVM continued to perform better than NB in identifying cyberbullying content. Because of model's great accuracy, users will be better shielded from the behaviors of bullies on social media.

Siddiqui, T et al. [49] had attempted to locate Twitter content that engages in cyberbullying. Additionally, carried out a comparison analysis of a number of classification methods, including LR, NB, DT, and SVM. Dataset was created from Twitter data that had been manually tagged and verified by linguists. In this study, 1065 data with a label distribution were employed, of which 638 had a label indicating that the data were not bullying and 427 had a label indicating that the data were bullying. The Bag Of Word (BOW) approach, which employs three weighing features, was used to weight each word. Three-word vector weighting features—bigram, trigram, and unigram were employed. Two scenarios were tested in the experiment, one of which involved determining the optimal accuracy value using the three features. The performance of the algorithm was compared overall across all features in the following scenario. The experimental results show that, when it came to measuring accuracy weighting depend on features as well as algorithms, the SVM classification method performed better than other algorithms, with a percentage of 76%. Next, for the weighting depend on average recall, the DT categorization method outperformed the other algorithms by an average of 76%. The SVM classification algorithm, which uses an F-measure of 82% to evaluate overall performance (F-measure) depend on

accuracy as well as precision, performed better than other algorithms. The SVM classification system can identify terms that include material related to cyberbullying on social media, according to multiple studies done.

Jenga, K. et al. [50] had investigated the most advanced crime prediction methods that have been developed in the last ten years, talk about potential difficulties, and have a conversation about potential future research in the field of crime prediction. Despite the fact that many studies attempt to forecast crimes, there was a wide range of datasets and approaches utilized. In order to assist scientists and law enforcement agencies in mitigating and preventing future crime incidents, seek to gather and synthesize the necessary knowledge about machine learning-based crime prediction through the use of a Systematic Literature Review (SLR) methodology. The main focus was on 68 machine learning studies that have been chosen for their crime prediction. Develop eight research objectives and note that most articles take a supervised machine learning method, assuming the availability of previously labeled data; nevertheless, in certain real-world settings, there may not be any labeled data. Also talked about the major difficulties the researchers encountered when working on some of their investigations. Believe that this research opens the door for other studies to assist nations and governments in combating crime and reducing it for improved safety and security.

Saikia, S. et al. [51] had demonstrated a real-time system that can identify items connected to interior spaces, such bedrooms. Employed the Faster R-CNN algorithm's cutting-edge VGG16 network architecture, which can compute area proposal inside the network. This feature makes the technique extensively applicable to the creation of real-time object detection systems. Developed a test-set called "ImageNet-Room Objects" made up of photos that were frequently seen indoors in order to assess the system, and were able to reach state-of-the-art accuracy. The Karina dataset had also been used to test the system; however, the poor quality of the photos had resulted in low accuracy. We will use image super-resolution methods to train a new model with several categories depend on other object types that the police may have found all through their crime scene investigation for addressing this issue in the future. Finally, the technique can be utilized as a surveillance tool to quickly recognize subjects in images and videos so that various crime scenes can be investigated.

Jain, R. et al. [52] had suggested a useful criminal detection system based on machine learning that makes use of facial and text recognition methods. Parking lots, toll booths, airports, border crossings, and other locations will find usage for these systems. The suggested system's text recognition function entails removing the characters from Indian license plates, and the anticipated result was then cross-referenced with the database of registered vehicles. Concurrently, the face recognition feature entails mapping the appropriate coordinates with the criminal database and detecting criminal faces based on specific face areas. With more than 85% of successful recognitions under normal working settings, the method presented in this research study strives to provide flexible results while accounting for accuracy and time restrictions. The ultimate objective was to provide a valuable real-time detection by effectively detecting crimes utilizing machine learning methods such as KNN, SVM, and facial identification classifiers.

Pande, V et al. [53] had suggests to take information out of crime record databases so that can do data mining on it. In order to forecast and predict, data classification and regression algorithms were suggested to be used. This was done by training a set first, then using the learnt rules to the test set to ascertain the expected output. With this information, law enforcement organizations can gain a better understanding of the pattern of crime in a provided area or over a specific period of time. These organizations can then use the data to take preventive measures to stop the growth of specific crimes in specific areas or at specific periods. Would save a ton of money, time, and effort by doing this. The system suggests mining this data in order to apply the proper algorithms to it. Using a data visualization tool such as the K-means clustering algorithm, this projected output might likewise be shown to the user as clusters. The ultimate result may therefore be a system that uses training crime data sets to make certain predictions about the future. The output could then be visually appealing to make it easy for the user to understand.

Saiba nazah et al. [54] had employed the Systematic Literature Review (SLR) technique in an effort to give researchers and experts in the field of cyber security guidance and information on new crime dangers in the Dark Web. In order to address predetermined research topics, the 65 most pertinent papers from top electronic databases were chosen for extraction of data and synthesis for this SLR. The outcome

of this systematic review of the literature offers (i) in-depth information about the rise in crimes committed through the Dark Web; (ii) an assessment of the moral, legal, and financial consequences of the cybercrimes committed there; and (iii) an examination of the challenges, tried-and-true techniques, and tactics for apprehending the criminals as well as their flaws. The study demonstrates that more comprehensive research was required to identify criminals on the Dark Web in a more noticeable way, that forensic investigations primarily rely on the examination of Dark Web discussion forums and crypto currency markets, that the anonymity provided by these services can be exploited as a tool for criminal capture, and that digital evidence ought to be examined and handled in a way that facilitates law enforcement's efforts to apprehend offenders and shut down illicit websites on the Dark Web..

Table 2.1: The Comparison Table of Merits and De-Merits of Reviewed Papers

Sl. No.	**Author**	**Methods**	**Objectives**	**Merits**	**De-Merits**
1.	Lal, S. et al. [34]	RF, ZeroR, NB, and J48	The goal of the study is to develop a model that can differentiate between tweets about crimes and those about non-crimes.	This could expedite the investigation into the incident, benefiting both the victim and the police.	The feature extraction process required enhancement through the use of a proficient technique.
2.	AlGhamdi, M. A., & Khan, M. A. [35]	KNN, DT, LDA and SVM	The objective was to categorize tweets into suspicious and non-suspicious categories, with a specific dataset containing tweets falling into the suspicious category.	The system was smart enough to read tweets written in Arabic and recognize any	The experiment had a limitation due to the lack of additional tweet classification.

				offensive material.	
3.	Santhiya, K. et al. [36]	NLP, Naive Bayes	Focuses on developing the big data programming framework Apache Spark and the Hadoop MapReduce system.	The classification system, built using Apache Spark's machine learning framework and Map Reduce, has been proven reliable, scalable, and effective in numerous implementations.	To enhance classification efficiency, additional features were added to the feature vector to broaden and improve the framework.
4.	Boukabous, M., & Azizi, M. [37]	DL, BERT	The DL model was a hybrid approach, combining deep learning and vocabulary-based learning, with BERT serving as the primary model.	Social media platforms like Twitter are utilized to promptly identify potential security risks.	The approach's drawback is its expensive hardware specification and its application for identifying illegal conduct in pictures, videos, and audio is not effective.

5.	Hissah, A. S., & Al-Dossari, H. [38]	CNB, SVM, DT, and KNN	The study aims to identify and categorize offenses in tweets published in Arabic, a widely spoken language globally.	Out of all the classifiers, SVM had the highest accuracy.	The study was unable to conduct a comprehensive geographical and temporal analysis to identify historical crime outbreak sites, dates, and potential future locations.
6.	AlJanabi, K. B. [39]	DT, Bayes, SVM, KNN	This synopsis covers the most renowned studies in this field, detailing their experiments, algorithms, data types, amounts, and conclusions.	Decreased the classification error and improved the classifier's accuracy and performance.	Hardware was so expensive
7.	Krishnendu, S. G. et al. [40]	K means algorithm	Concentrated on forecasting the area with highest rates of crime and age groups that have a tendency toward crime.	to reduce the time complexity and boost output efficiency	low robustness
8.	Pepsi, M. B. B., & Kumar, S. N. [41]	SVM, Bayes and logistic regression	The aim of this undertaking was to classify the tweets according to the issues that the students were having.	Performance was better than other devices	The computational complexity was high and the heavy load was present.

10.	Al-Saqqa, S. et al. [42]	NB, SVM, and logistic regression	The paper presents innovative experiments utilizing logistic regression, SVM, and NB to categorize sentiment in large-scale data using Spark's MLlib.	Increase the classification's accuracy.	Need to improve its computational performance
11.	Savaş, S., & TOPALOĞLU, N. [43]	SCI, Zemberek-NLP	The case study examines Twitter intelligence, examining collective wisdom and whispers on the platform for cyber intelligence purposes related to security.	Considered the real time dataset	Reconnection is crucial for maintaining data stream continuity in case of disruptions due to Twitter permissions and available bandwidth.
12.	Garima, A., & Alaiad, A. [44]	K-means	Objective was to use the Poisson model to identify spatial crime clusters.	Track crime episodes in real time throughout the city of Chicago, access past crime incidents for specific wards, and optimize resource usage in high-	The average cost of renting is extremely expensive.

				crime areas.	
13.	Panja, B. et al. [45]	KNN, ANN, COPS	Rejoining is essential to preserving the continuation of the data stream in the event that disruptions arise from Twitter permissions and available bandwidth.	It is appropriate for identifying concealed and disregarded information at any given time.	The unwanted data were collected automatically with the required data.
14.	Baby, A. et al. [46]	MLP, ADAM, TF–IDF, CSV, SVM	Provide a model that uses sentiment analysis, hate speech analysis, foul language detection, and targeted speech analysis to detect crimes on Twitter.	to identify illegal activity on Twitter by employing MLP sentiment analysis in conjunction with classification algorithms to identify hate speech, targeted speech, and filthy language.	Need to enhance the system's execution speed and accuracy.

15.	Sharrab, Y et al. [47]	DNN, FNN, LSTM	The identification of illegal activity, including fraud, child sex abuse, and terrorism; the detection of false information and fake news.	This involves recognizing illegal activities like fraud, child sex abuse, and terrorism, identifying false information, and understanding how ideas and information spread on social media.	The dataset used was very small and expensive, and real-time systems were not utilized.
16.	Alduailaj, A. M., & Belghith, A [48]	NB, SVM	Intends to detect cyberbullying automatically in Arabic with machine learning.	Obtain high accuracy	Small dataset
17.	Siddiqui, T et al. [49]	NB, DT, LR, and SVM	Seeks to locate Twitter content that engages in cyberbullying.	The language's structure was improved, leading to improved accuracy.	The model's poor dataset quality will increase its susceptibility to bias and over fitting during the dataset transfer process, though the effect is not as significant as anticipated.

18.	Jenga, K. et al. [50]	Systematic Literature Review (SLR)	to gather and combine the necessary information about machine learning-based crime prediction in order to assist scientists and law enforcement in reducing and preventing crime in the future.	assist nations and governments in combating crime and reducing it to improve safety and security	real world data were not considered
19.	Saikia, S. et al. [51]	R-CNN	Present an object detection system built on the Faster R-CNN algorithm.	suitable to be utilized as a real-time application	time needed to detect
20.	Jain, R. et al. [52]	KNN, SVM	The study unveiled a face- and text-recognition technology-driven machine learning crime detection system.	The proposed solution enhances monitoring, tracking of illegal activities, and license plate identification, potentially preventing human reliance and stopping criminal activities in case of	The proposed system has limitations on facial expressions, light illumination, disguised identity, and facial emotions, affecting prediction accuracy.

				carelessness.	
21.	Pande, V et al. [53]	ARIMA, Bayesian Network Algorithm, ANN	This initiative generates a dataset and preprocesses it in order to mine the massive amounts of raw data.	The system aims for maximum efficiency, security, and improved accuracy, ensuring data integrity for reliable output and efficient data processing.	Real-time data was not utilized, and users were unable to select from various forecasting, prediction, and categorization models by comparing their outcomes side by side.
22	SAIBA NAZAH et al. [54]	SLR	The Systematic Literature Review (SLR) technique was utilized to provide guidance and information on new cyber security threats in the Dark Web to researchers and experts.	More efficient	High complex system

In the past, ML models have been utilized to forecast crime data; however, because deep learning models perform better, were developing quickly. ML models require more data and time, but DL models require less accuracy, time, and data. Due to the fact that many investigations fail to produce unique effects, a comparison study on crime data prediction models is now necessary. Many datasets were analyzed to determine their efficacy, as were essential for forecasting crime statistics on social media.

2.3 Summary

Credit card fraud, cyber infiltration, data breaches, and disaster fraud were examples of crimes connected to social media. Criminals obtain user personal information by processing lost credit cards or gaining control of bank accounts. After a tragedy, criminals establish private groups to gather information for their own benefit. This is known as disaster fraud. Even fabricate groups in order to make money off of the distress of others. Social media was used to commit a wide range of crimes, such as robberies, violent crimes, murders, property crimes, crimes using guns, sexual offenses, and organization-based crimes. These crimes pose a serious risk to the security and safety of users and can be committed by one or more people.

The criminal detection system provides high protection for user data and helps stop rumors from spreading on social media. Due to its excellent microblogging service, Twitter had become more and more well-known in the last several years. The Twitter data also includes a lot of crimes that have significant negative effects on society. Many methods were currently employed to find criminal activity in Twitter data. But the preceding algorithms become irrelevant when illicit tweets were manually identified. The earlier methods had trouble detecting crime data because could not handle greater inputs, which led to increased processing complexity problems. Social media collections typically contain a sizable number of user reviews that contain both actual user and crime data. Thus, the analysis of crimes in vast databases was becoming a challenging task. Other challenges that have been found with the detection of illicit tweets include large Twitter handles, noisy tweets, and a high volume of tweets. In order to solve the problems raised by the existing approaches, the suggested study creates a novel crime detection model for the Twitter environment.

CHAPTER 3

AUTOMATIC TWITTER CRIME PREDICTION USING HYBRID WCNN WITH WCO

This chapter intricately explores the domain of automatic Twitter crime prediction by employing a sophisticated hybrid approach, integrating wavelet CNN with World Cup Optimization (WCO). The discussion unfolds with a thorough examination of crucial stages, including data cleansing, feature extraction, and feature selection utilizing a metaheuristic technique, namely Modified Tree Graph Algorithm (MTGA). Furthermore, the chapter delves into the intricacies of clustering through Fuzzy Manta Ray Foraging (FMRF) for improved data organization. The culmination of these preparatory steps sets the stage for crime detection, ingeniously executed through the application of the hybrid Wavelet CNN model with WCO.

In tandem with the implementation of the planned model, the chapter delivers a complete presentation of results and engages in a comprehensive discussion to unravel the nuances, strengths, and limitations of the hybrid approach. This analytical journey not only sheds light on the efficacy of the novel methodology but also enriches our understanding of its practical implications in the realm of Twitter crime prediction.

Finally, the chapter achieves with a succinct summary, encapsulating the key answers, methodological contributions, and potential avenues for upcoming investigation. This holistic approach ensures that the reader gains a comprehensive grasp of the innovative processes and outcomes presented in the study.

3.1 Overview

Social media are digital environments that enable users to generate and share ideas, content, and other types of expression with others through online communities [55]. Individuals typically download applications offering social media features onto their

devices or utilize web-based platforms on their desktop computers and laptops to engage with social media platforms [56]. Groups, companies, and people can publish, collaborate on, debate, interact with, and update user-generated or self-curated content as they connect with these platforms [57]. Early identification of crime rates provides information regarding various criminal activity types. Some social media platforms offer some of the most effective resources for crime prevention [58]. Social networking technologies will therefore significantly lower crime rates [59]. This information tool gives users the ability to access their location and find global crime statistics in different places. It is imperative to take immediate action to decrease the increasing amount of criminal activity on social media environments [60]. Platforms like Facebook, Twitter, Instagram, and others help with communication, but they also have drawbacks like hacking and cybercrime [61]. Predicting the crimes is necessary in order to reduce the increasing number of crimes committed through these environments. The information that is gathered from Twitter is pre-processed for the purpose of information purification in order to forecast crimes. Later, the features are retrieved using a variety of methods, including feature hashing, Glove, TF-IDF, and BoW. FMRF is used for clustering and MTGA for feature selection. Ultimately, the WCNN-WCO hybrid wavelet CNN is used for crime detection. The analysis is done using the Twitter user dataset, and the implementation is performed with the help of Python tool. The findings demonstrated that the suggested strategy beats the current approach in terms of memory, F1 measure, accuracy, and precision.

3.2 Proposed Methodology

Several researchers have employed diverse methodologies to forecast criminal activity on Twitter. There are certain drawbacks to the current crime prediction techniques. These restrictions are taken into account to make the system better. By using the suggested method, the data is categorized as either non-crime or criminal. In addition to giving users security, this also helps to increase user security. Figure 3.1 displays the block diagram for the suggested approach.

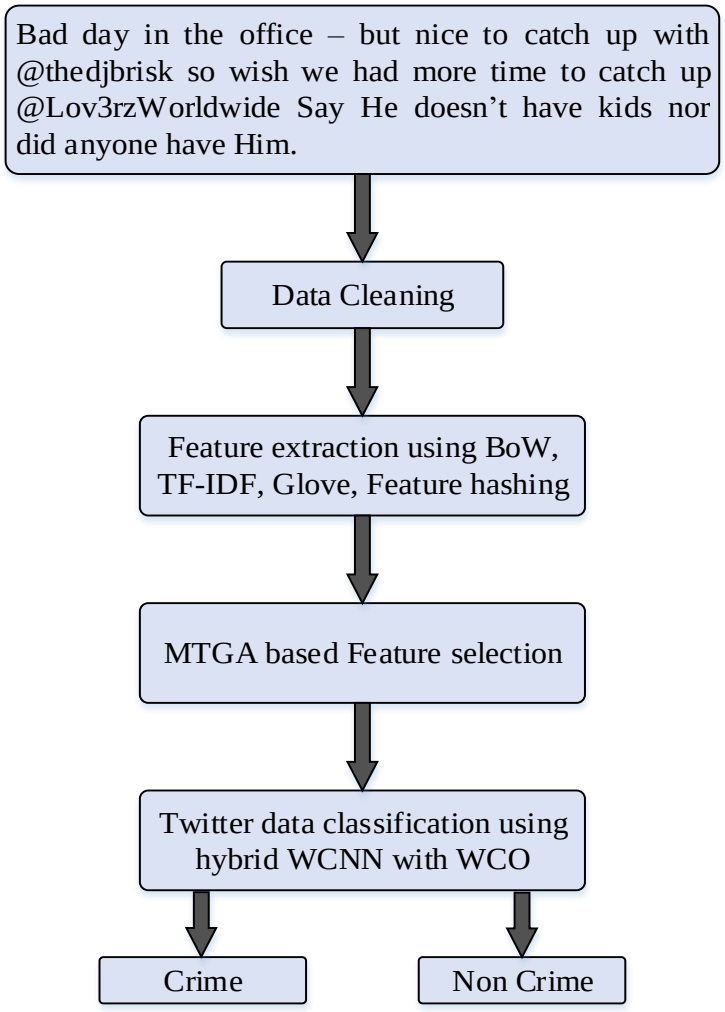

Figure 3.1: Block structure of proposed model

The proposed method consists of five processes such as, data purification, feature extraction, feature selection, clustering, and classification. The pre-processing stage cleans and gets the data ready for the next stage. A technique for removing particular attributes from a dataset is called feature extraction. For feature extraction, four methods are used, Glove, BoW, TF-IDF, and feature hashing. The features from the extracted feature set are chosen by the use of MTGA in the feature selection procedure. The FMRF method then completes the clustering process. Hybrid WCNN with a WCO procedure makes the prediction.

3.2.1 Data cleansing

Twitter provides disorganized and disorderly information. Data cleansing procedures are used to decrease noise in data. The first step in predicting Twitter crime is data

purification. It comprises tweet extraction and filtering, noise reduction, vocabulary cleansing (removing unnecessary words), and tweet modification. The process of extracting tweets involves stemming, tokenization, and stop word removal.

3.2.1.1 Tokenization

The initial stage of data cleaning is to segment or tokenize tweets. In text analysis, tokenizing words is a basic fundamental component.

3.2.1.2 Stop word removal

The terms that are utilized in the text the most are stop words. The number of terms in the text indicates a relatively low term value.

3.2.1.3 Stemming

Another crucial step in cleaning up data is termed as stemming. Stemming reduces the text to its most basic form by removing affixes that save time and space.

3.2.2 Feature extraction

Selecting or combining variables into features is referred to as feature extraction. Feature extraction is necessary in order to achieve optimal text analysis quality and streamlines the procedure. Accurately extracting data from a large dataset without sacrificing crucial information is helpful. The feature extraction stage speeds up the processing method. Various features can be extracted using various strategies for Twitter crime detection. BoW, TF-IDF, Glove, and feature hashing are the methods employed for the extraction.

3.2.2.1 BoW

A text document's features can be extracted using the BoW [62]. It provides information about the word's appearance in a document. Since BoW pulls consecutive "n" words out of the data, it is often referred to as "n-gram." In this strategy, the word's frequency of occurrences is the primary factor taken into account. Weights are allocated to each word that is taken out of the text in order to indicate how important it is throughout the text.

BoW shows the words that are most frequently associated with two concepts. They are listed as follows:

a) Vocabulary of known words
b) Measure of presence of known words

BoW is a simple to operate and adaptable method for feature extraction. These two basic text documents, D1 and D2, are attached.

D1: “The Sun is a star. Sun is beautiful.”

D2: “The Moon is a satellite.”

A dictionary is created using these two text documents as a foundation.

{“The”: 1 “Sun”: 2 “is”: 3 “a”: 4 “star”: 5 “beautiful”: 6 “Moon”: 7 “satellite”: 8}

Documents include eight unique words. Every document is signified as an 8-element vector with each item referring to the count of the associated entry in the dictionary: [1, 2, 2, 1, 1, 0, 0] [1, 0, 1, 1, 0, 0, 1, 1].

3.2.2.2 TF-IDF

The social crime dataset's feature words are chosen using the TF-IDF technique [63]. A word loses significance and weight in an article when it appears frequently. The formula for TF-IDF is as follows:

$$TF - IDF(s,r) = TF(s,r) \cdot IDF(s) \tag{3.1}$$

Where, $TF(s,r)$denotes the frequency of a feature, $t \in F$, and TF is computed as:

$$TF(s,r) = \frac{n(s,r)}{|r|} \tag{3.2}$$

Where, $n(s,r)$ is the frequency occurrence of feature s in the document r , normalized by $|r|$, with the length of r .

$$IDF(s,) = \log\left(\frac{|A|}{DF(s)}\right) \tag{3.3}$$

Where, the document frequency of the feature t is indicated by $DF(s)$. It determines how many papers in A have s in them. The total number of documents in the corpus is denoted by $|A|$. The weight of words that look in a significant portion of the corpus A is decreased by the $IDF(s,)$ parameter.

3.2.2.3 Glove

The words are represented in vector form by glove. The vector representation of each word is found by mapping it into a space. This method links the semantic similarity of words to their distances from one another. By taking into account a matrix of word-word co-occurrence counts (B), the Glove [64] can be understood. The matrix's entries (B_{ji}) list the frequency with which a word appeared in relation to word j. The word that appears in the context of word j the number of times is provided by,

$$B_j = \sum_l B_{jl} \tag{3.4}$$

The formula that follows calculates the probability that word i is to appear in relation to word j is given by,

$$D_{ji} = D(i/j) = B_{ji}/B_j \tag{3.5}$$

Words with greater relevance are distinguished from those with less relevance by the ratio. Furthermore, it may distinguish between two pertinent words.

3.2.2.4 Feature hashing

The hashing approach [65], also known as feature hashing, is used to vectorizing the features from the cleaned Twitter data. In this method, the arbitrary features are transformed into their vector representation. A hash function is then functional to these features, and the technique operates based on the resulting hash values. Hashing involves converting an input into a numerical value and subsequently recovering the desired output a fundamental aspect of hashing techniques. Generally, information is converted into a numeric hash during this process. Although the hash function plays a major role in hash computation, all hash functions seem to have similar properties.

3.2.3 Feature Selection using MTGA

In the feature selection stage, the retrieved features from the feature extraction stage are utilized. This step aims to enhance the performance of the planned method while reducing computational complexity by reducing the input variables. Through this process, the most influential features contributing to the output are selected. In this study, Modified Tree Graph Algorithm (MTGA) is employed for feature selection. MTGA, an evolved version of Tree Graph Algorithm (TGA), allows for the optimization of power reduction rates. Inspired by nature, MTGA operates akin to a tree-based algorithm. The tree population is separated into four groups: the first group comprises the best-performing trees, while the second group involves moving trees closer to their two nearest neighbors. In the third group, unhealthy trees are removed and replaced with new ones. First, the population of trees is generated at random. The trees are ordered in ascending order based on their fitness values. The finest trees belong to the first group. Using the provided equation, a new tree is generated in this group.

$$B_i^{s+1} = \frac{B_i^s}{\varphi} + nB_i^s \tag{3.6}$$

Where, n is a random number between [0,1], B_i is the solution at i, n is the number of iterations, and φ is the power reduction rate. The current tree will be replaced if the created tree has a higher fitness score. An essential component of the tree growth algorithm is the parameter φ. The value shouldn't alter while the processing is being done. In the algorithm, determining the correct φ is both a crucial and time-consuming phase. Thus, the linear rising method is used to modify the parameter. Equation (3.7) can be used to enhance the value of φ.

$$\varphi = 0.5 \times \left(1 + \frac{3u}{n_{iter}}\right) \tag{3.7}$$

Where, n_{iter} denotes the total number of iterations and u denotes the number of iterations that are currently occurring. The trees are then shifted to a location that is closer to two adjacent trees. This distance can be computed with the following formula.

$$Q_i = \left(\sum_{i}^{N_1+N_2} \left(B_{N_2}^{s} - B_i^{s} \right)^2 \right)^{\frac{1}{2}} \tag{3.8}$$

Where, B_i is used to denote the i^{th} tree and B_{N_2} is used to denote the current tree. The Euclidean distance approaches infinity when B_{N_2} equals B_i . The tree then goes to the trees that are closest to it in an attempt to compete for light. The formula for the linear combination of the two closest trees is (3.9).

$$A = \gamma V_1 + (1-\gamma) V_2 \tag{3.9}$$

Where, V_1 represents the closest tree, and V_2 represents the second-closest tree, γ stands for the parameter that regulates the influence of the closest tree. The tree's position in the second group has been improved by.

$$B_{N_2}^{s+1} = B_{N_2}^{s} + \delta A \tag{3.10}$$

Where, δ represents the angle that ranges between 0 and 1. In the third group, new trees are planted in place of the poorest trees (N_3). The N_3 computation takes place through,

$$N_3 = N - N_1 - N_2 \tag{3.11}$$

Where, N represents the size of the population, N_1 represents the trees in the first group, and N_2 represents the trees in the second group. Fitness determines how the population is ordered. The best N trees are then selected for the following iteration. The optimal tree is ultimately determined to be the most effective solution. The (minimum) number of features in the feature subset and the (maximal) classification accuracy that can be attained with the chosen features are the two criteria that can be used to evaluate a solution's goodness of fit because the solution represents the feature subset in the MTGA for FS. To integrate these two parameters, an objective function is used as.

$$Fitness = \min\left[v\delta_s + \vartheta \frac{|A|}{|P|} \right] \tag{3.12}$$

Where, $|A|$ is the number of selected features, $|P|$ is the number of original features in the dataset, and δ_s is the hybrid WCNN with WCO classifier classification error rate. The weighting variables for the subset's cardinality and the classification error rate are u and v.

3.2.4 Clustering using FMRF

FMRF technology is used for the Twitter data grouping. It's a hybrid approach that combines the MRF algorithm with fuzzy-based clustering. The first step in the fuzzy based clustering of cybercrime offenses is to make the objective function easier.

$$K_n = \sum_{i=1}^{C}\sum_{j=1}^{M} U_{ij} \left| b_i - m_j \right|^2, 1 \le n \le \infty \tag{3.13}$$

Where, n is a real number larger than 1, b_i is the i^{th}-dimensional data, U_{ij} is the membership degree of b_i in a cluster j, and m_j is the representation of the cluster center. The notation $\left| \right|$represents the relationship between the measured data and the cluster center. Regular updates are made to the membership U_{ij} and cluster center by,

$$U_{ij} = \frac{1}{\sum_{K=1}^{m} \left[\frac{\left| b_i - m_j \right|}{\left| b_i - m_K \right|} \right]^{\frac{2}{n-1}}} \tag{3.14}$$

$$m_j = \frac{\sum_{i=1}^{m} U_{ij}^{n} b_1}{\sum_{i=1}^{N} U_{ij}^{n}} \tag{3.15}$$

Where, K stands for the step of iteration. The iteration finishes when $\max_{ij} \left\{ \left| U_{ij}^{K+1} - U_{ij}^{K} \right| \right\}$ is less than the fixed value ∂. The MRF optimization algorithm can be used to optimize the cluster center m_j. It quickens the fuzzy-based grouping of

offenses related to cybercrime. Marine animals with a flat body and pectoral fins are called manta rays. They feed on planktons and lack teeth. There are three steps in the MRF mathematical model. These three types of foraging are somersault, cyclone, and chain. During the chain foraging stage, manta rays follow the planktons in front of them and migrate in the direction of the area where they are concentrated. Every iteration updates the solution for each individual and represents the best answer so far. The chain foraging mathematical formula is represented like follows:

$$a_i^p(s+1)=\begin{cases} a_i^p(s)+q\cdot\left(a_{best}^p(s)-a_i^p(s)\right)+w\cdot\left(a_{best}^p(s)-a_i^p(s)\right) & i=1 \\ a_i^p(s)+q\cdot\left(a_{i-1}^p(s)-a_i^p(s)\right)+w\cdot\left(a_{best}^p(s)-a_i^p(s)\right) & i=2,....,N \end{cases} \tag{3.16}$$

$$w=2\cdot q\cdot\sqrt{|\log(q)|} \tag{3.17}$$

Where, w is the weight coefficient, q is the random variable with a range of [0, 1], and $a_i^p(s)$ is the location of i^{th} person at time s in the p^{th} dimension. The plankton with the highest concentration is the $a_{best}^p(s)$. During MRF's cyclone foraging stage, manta rays follow the lead ray and use a spiral path to aim for the food. Both exploration and exploitation are improved at this stage. Instead of trailing the person in front of it, an individual travels in a spiral pattern in the direction of the meal. The following mathematical formula can be used to simulate the manta rays' spiral motion in two-dimensional (2D) space:

$$a_i^p(s+1)=\begin{cases} a_i^p(s)+q\cdot\left(a_{best}^p(s)-a_i^p(s)\right)+\ell\cdot\left(a_{best}^p(s)-a_i^p(s)\right) & i=1 \\ a_i^p(s)+q\cdot\left(a_{i-1}^p(s)-a_i^p(s)\right)+\ell\cdot\left(a_{best}^p(s)-a_i^p(s)\right) & i=2,....,N \end{cases} \tag{3.18}$$

$$\ell=2\cdot e^{q_1\frac{S-s+1}{S}}\cdot\sin(2\pi q_1) \tag{3.19}$$

Where, S is the maximum number of iterations, q_1 is a random number inside [0, 1], and ℓ stands for the weight coefficient. The position is regarded as a hinge in the behavior of somersault foraging. This is a method to generate the algebraic model:

$$a_i^p(l+1) = a_i^p(l) + F \cdot \left(q_2 \cdot a_{best}^p - q_3 \cdot a_i^p(l)\right), \quad i = 1,\ldots..,N \tag{3.20}$$

Where, $F = 2$, q_2 and q_3 are random values inside [0, 1], and F is the somersault factor that establishes the somersault range of manta rays. Following the clustering phase, the clustered features advance to the detection stage. Hybrid WCNN with WCO is used to do the classification process at this phase.

3.2.5 Crime detection using hybrid WCNN with WCO

Following data clustering, hybrid WCNN with WCO is used to identify the cybercrime. In this approach, the CNN [66] wavelet principle is integrated to reduce the features by compressing the number of layers in the CNN, and the loss function value is optimized using the WCO optimization algorithm.

The CNN algorithm can be summarized as follows:

(i) The neurons' bias and the weights between layers are determined.

(ii) Propagation forward.

(iii) Find each sample's MSE by using the loss function.

(iv) Deriving the results of the chain rule derivation, which are the back propagation errors for each layer.

(v) The weights and bias are adjusted using the optimization algorithm (WCO) in accordance with the backpropagated errors.

(vi) Steps (ii) to (v) until the MSE is within acceptable bounds, then evaluate accuracy, precision, and efficiency.

Wavelet transform can be used to change the size and translation to learn various features. CNN may learn richer characteristics by including the wavelet transformation. The first part of the suggested WCN is the WCPNN, and the second part is the FCNN. The training algorithm of WCNN also consists of three steps:

(i) Forward propagation of WCNN

(ii) Reverse propagation of WCNN

(iii) Modification of WCNN weight and bias

Forward propagation of WCNN: WCNN employs the same forward propagation method as CNN. The WCNN receives the feature of training samples. From the first layer of WCNN (the input of WCNN), the pooling layer, convolutional layer, and fully connected layer decide the output of WCNN. Forward propagation of a convolution layer: If layer p is a convolutional layer, its input (net_{ab}^{p}) can be ascertained using equation (3.24).

$$\begin{aligned} net_{ab}^{p} &= convolution\left(\mathrm{E}^{p-1}, \alpha^{p}\right) + h^{p} \\ &= \sum_{c=0}^{\text{size}^{p}-1} \sum_{n=0}^{\text{size}^{p}-1} \left(\mathrm{E}_{x+c,\, y+n}^{p-1}, \alpha_{c,n}^{p}\right) \end{aligned} \tag{3.21}$$

The outcome of the layer $\left(\mathrm{E}_{x,n}^{p-1}\right)$ is represented as,

$$\mathrm{E}_{ab}^{p} = \chi\left(\frac{net_{ab}^{p} - ac^{p}}{bc^{p}}\right) \tag{3.22}$$

Where, ac^{p} and bc^{p} stand for the activation function's scale transformation. In the WCNN convolutional layer, the activation function $\chi_{wc}(a)$ is expressed as Equation (3.23)

$$\chi_{wc}(a) = \cos(1.75a) \cdot e^{-\frac{a^2}{2}} \tag{3.23}$$

The forward propagation of the pooling layer and the fully connected WCNN layer is identical to that of CNN.

During the back propagation phase of WCNN, which follows forward propagation, the MSE for each training sample is computed using the loss function. This MSE is then utilized to adjust the predicted values for the training samples.

$$H = \frac{1}{2}\sum_{n=1}^{N}\left(\hat{f} - f_n\right)^2 \tag{3.24}$$

Changes in weights and biases in the fully connected layer, convolutional layer, and pooling layer require backpropagation of errors. However, the pooling layer differs from the fully connected and convolutional layers in CNNs in terms of how backpropagation is applied. The input error of the pooling layer is given by Equation (3.25) if d is the pooling layer and layer $(p+1)$ is the convolutional layer.

$$\delta_{ab}^{p} = \frac{1}{bc^{p}} PoolExoand\left(\delta_{ij}^{p+1}\right) \cdot \chi^{'}\left(\frac{net_{mn}^{p} - ac^{p}}{bc^{p}}\right) \tag{3.25}$$

$$\chi^{'}(a) = -1.75\sin(1.75a) \cdot e^{-\frac{a^2}{2}} - a \cdot \cos(1.75a) \cdot e^{-\frac{a^2}{2}} \tag{3.26}$$

The world cup optimization (WCO) optimization algorithm is presented as a means of reducing the loss function. In WCO [67], a team of eighteen members is chosen at random. Here, a team's rank plays a significant role in selecting where they are seeded. Following the initialization of random teams, the cost function is ascertained. The teams' ranks are used to determine the order of seeding. Without any opposition, the team with the higher rank is positioned as the first seed. Following the sowing, the contest starts. First, choose the teams and the continents. Assign the team members to nations that are spread throughout C number of continents.

$$Continent = \left\lfloor Country_1, Country_2, ..., Country_{C_{\text{var}}} \right\rfloor \tag{3.27}$$

$$Country_i = \left\lfloor b_1, b_2, ..., b_{C_{\text{var}}} \right\rfloor \tag{3.28}$$

Where, b_i indicates the i^{th} team and C_{var} represents the dimension of the optimization issue. A continent's points total are utilized to calculate ranking.

$$rank = e_q\left(Continent\right) = e_q\left\lfloor b_1, b_2, ..., b_{r_{\text{var}}} \right\rfloor \tag{3.29}$$

$$R = C \times N \tag{3.30}$$

Where, C stands for the variable's dimension and N for the number of continents. The cost function is assessed in step two. The team with the higher-ranking advances to the next round, although rankings are unfair because a stronger team might have

faced off against another stronger team on the same continent. To get around this problem, need to find the average value and standard deviation of the continents.

$$\overline{B} = \frac{1}{f}\sum_{i=1}^{f} B_i \tag{3.31}$$

$$\beta = \sqrt{\frac{1}{f-1}\sum_{i=1}^{f}\left(B_i - \overline{B}\right)^2} \tag{3.32}$$

Where, f represents the member quantity in B and $\overline{B}$. The symbol β represents the mean value and standard deviation of continent B. The teams are ranked throughout the ranking phase.

$$\begin{aligned} B_1 &= \left[B_{11},....,B_{1n}\right]^S \\ B_2 &= \left[B_{21},....,B_{2n}\right]^S \\ &\cdots \\ B_5 &= \left[B_{51},....,B_{5n}\right]^S \end{aligned} \tag{3.33}$$

$$B_{Total} = \left[B_{11},....,B_{1n},B_{21},....,B_{2n},....,B_{51},....,B_{5n}\right]^S \tag{3.34}$$

Where, S is the transposition operator and n is the number of teams. In this step, the top two teams are chosen and added to vector $\left(B_{Rank}\right)$ for upcoming competitions, with the team with the highest B_{Total} value being named the first-ever cup champion.

$$B_{Rank} = \left[B_{11},B_{12},B_{21},B_{22},.....,B_{51},B_{52}\right]^S \tag{3.35}$$

$$B_{Chapions} = \min\left(B_{Total}\right) = \min\left(\left[B_{11},....,B_{1n},B_{21},....,B_{2n},....,B_{51},....,B_{5n}\right]^S\right) \tag{3.36}$$

Where, $B_{champion}$ champion stands for the lowest possible solution value. New teams are generated for the next round of the competition based on the previous ranking. WCO operates by taking into account a two-part vector.

$$pop = B_{Total} = \left[B_{Best},B_{Rand}\right] \tag{3.37}$$

Where, B_{Rand} shows a vector as follows, $pop(B_{Total})$ indicates regenerated teams with size $(C \times N)$, and B_{Best} represents a random value inside a specific interval.

$$L < B_{Best} < U \tag{3.38}$$

$$U = \frac{1}{2} \times ac \times (U_b + L_b) \tag{3.39}$$

$$U = \frac{1}{2} \times ac \times (U_b - L_b) \tag{3.40}$$

Where, ac is the representation of a coefficient in the interval $[L_b; U_b]$. B_{Best} denotes the optimal location of the previous search space, and B_{Rand} denotes new genesis numbers in the search space, in terms of exploitation and exploration improvement. If the stop requirement is met, the algorithm is finished; if not, iterations are performed. The data is categorized into two phases, such as crime and non-crime, following the categorization method.

3.3 Results and Discussion

The Twitter user dataset is utilized in the PYTHON framework to implement the suggested tasks. The performance parameters used for the performance evaluation are AUC, recall, accuracy, precision, and F1-score. The suggested approach is contrasted with several current approaches, including GSCA, HANNCC, HMNCC, and HREACC [68]. Users will be able to enjoy superior security as a result. When compared to other comparable methods, the performance outcomes, such as crime detection, are superior.

3.3.1 Dataset description

This dataset, which has 20,000 rows, is made up of Twitter user data. It includes the user's name, account profile, location, arbitrary tweet and image, and so forth. Various constraints, like unit ID, gender, profile image, text, tweet generated, and so on, are present in the dataset. 30% percent of the dataset (https://data.world/data-

society/twitter-user-data) is used for testing, while the remaining 70% percent is used for training.

3.3.2 Performance metrics

Performance is assessed using a variety of evaluation criteria, including accuracy, recall, F1-score, and precision. The following is an explanation of the performance metrics:

- **Precision:** It can be defined as the proportion of positive samples classified correctly out of the total number of samples.

$$\Pr e = \frac{tp}{tp + fp} \tag{3.41}$$

- **Recall:** It can be defined as the proportion of positive samples classified correctly as positive, relative to the total number of positive samples.

$$rec = \frac{tp}{tp + fn} \tag{3.42}$$

- **F1-score:** The weighted consonant means of recall as well as accuracy is tested using the F1 score, often known as the F-measure.
  - $f - measure = \frac{2 \times pre \times sensitivity}{pre + sensitivity}$ (3.43)

- **Accuracy:** The value close to the true value is defined as the accuracy

$$acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.44}$$

- **AUC:** For classification problems, AUC is the performance metric in relation to various threshold values. It demonstrates how the classification process can distinguish between classes. The model performs better when the AUC value increases.

The terms tp and tn in Equations (3.41)– (3.44) signify the number of true positive and true negative values, respectively. fp and fn, correspondingly, represent the corresponding false positive and false negative values. The suggested method's confusion matrix is shown in Figure 3.2.

3.3.3 Performance evaluation

The proposed approach has been compared with methodologies such as HREACC, HANNCC, HMNCC, and GSCA. Table 3.1 presents the performance metric values obtained by both the suggested and existing approaches.

Table 3.1: Comparison of performance metrics

Classifier	Accuracy	Precision	Recall	F-measure	AUC
HMNCC	0.81	0.843	0.778	0.809	0.592
GSCA	0.84	0.881	0.811	0.845	0.548
HANCC	0.89	0.928	0.871	0.898	0.653
HREACC	0.93	0.961	0.901	0.930	0.752
Proposed	0.989	0.988	0.989	0.989	0.989

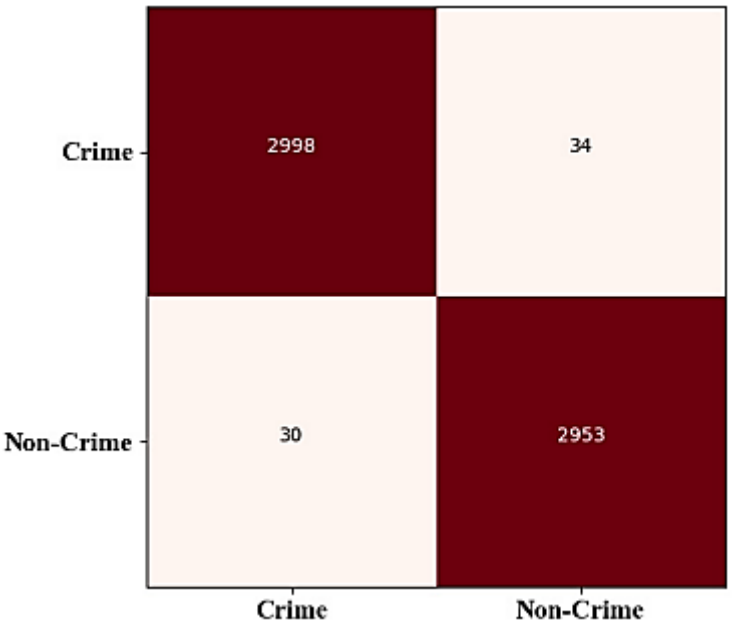

Figure 3.2: Confusion matrix.

Both the actual and the anticipated values are included in the confusion matrix. Another name for the confusion matrix is the mistake matrix. According to the

statistics, 2998 criminal data and 2953 nonprime data were appropriately classified. Figure 3.3 displays the precision performance evaluation.

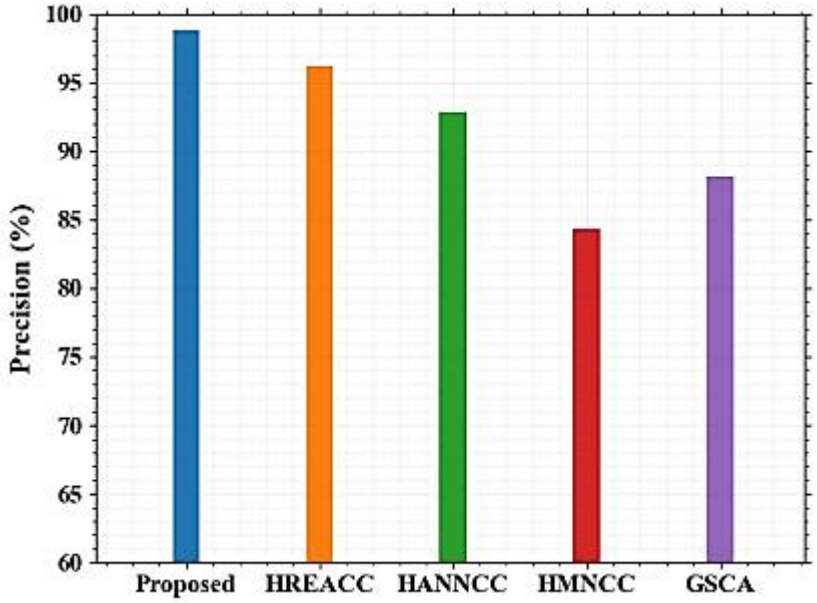

Figure 3.3: Performance analysis of precision.

The accuracy is compared to other methods such as GSCA, HANNCC, HMNCC, and HREACC. The suggested approach outperformed previous methods with a precision of 98.8%. HREACC, HANNCC, HMNCC, and GSCA have precision rates of 96%, 92%, 84%, and 88%, in that order. The recall performance evaluation is displayed in Figure 3.4.

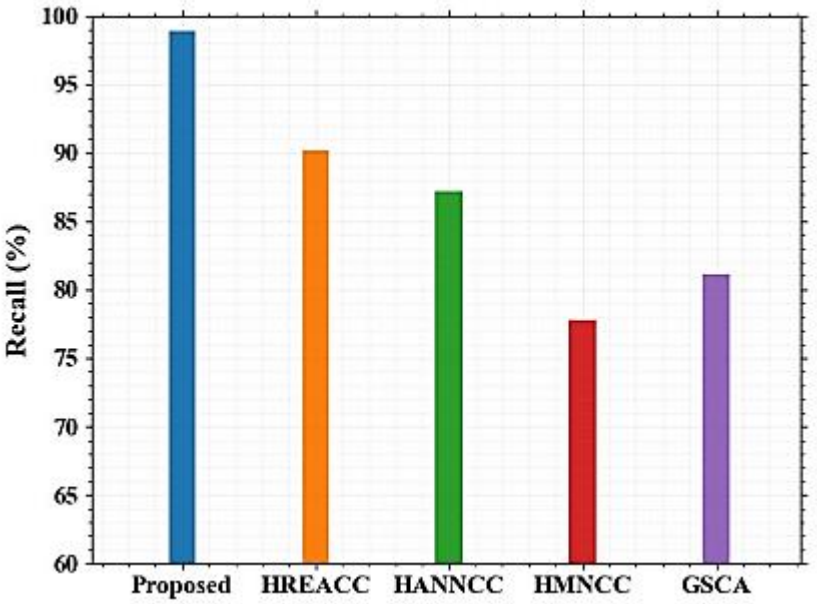

Figure 3.4: Performance analysis of recall.

The recall is compared using a number of methods, including GSCA, HANNCC, HMNCC, and HREACC. Recall values for the procedures HREACC, HANNCC, HMNCC, and GSCA were 90%, 89%, 77%, and 81%, respectively, whereas the suggested method yielded a recall value of 98.9%. Together with other current approaches, the suggested method's F1 measure is displayed in Figure 3.5.

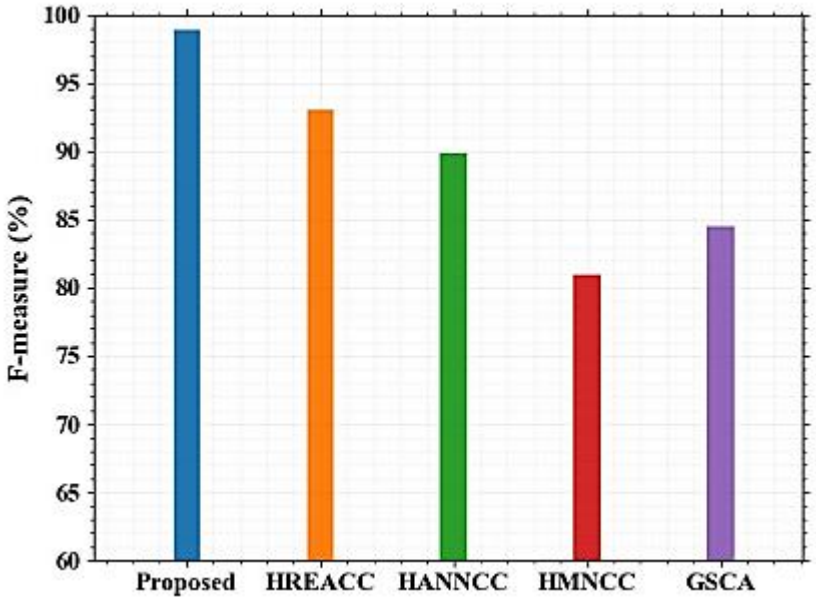

Figure 3.5: Performance analysis of F1 measure

The proposed approach has an F1-score of 98.9%, while the methods such as HREACC, HANNCC, HMNCC, and GSCA have respective F1-scores of 93%, 89.8%, 80.9%, and 84.5%. The suggested strategy yields the highest I1-score value. The accuracy's performance evaluation is displayed in Figure 3.6.

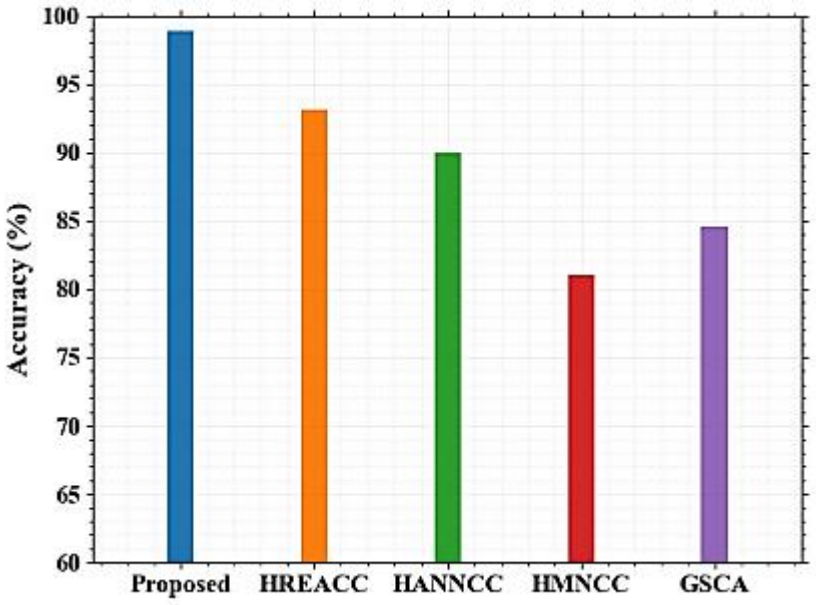

Figure 3.6: Performance analysis of accuracy.

The accuracy of the suggested technique was approximately 98.9%, surpassing that of other algorithms already in use. The accuracy of the approaches such as HREACC, HMNCC, HANNCC and GSCA were 93%, 81%, 89% and 84%, according to that order. The suggested technique's performance evaluation is displayed in Figure 3.7 using a variety of performance criteria.

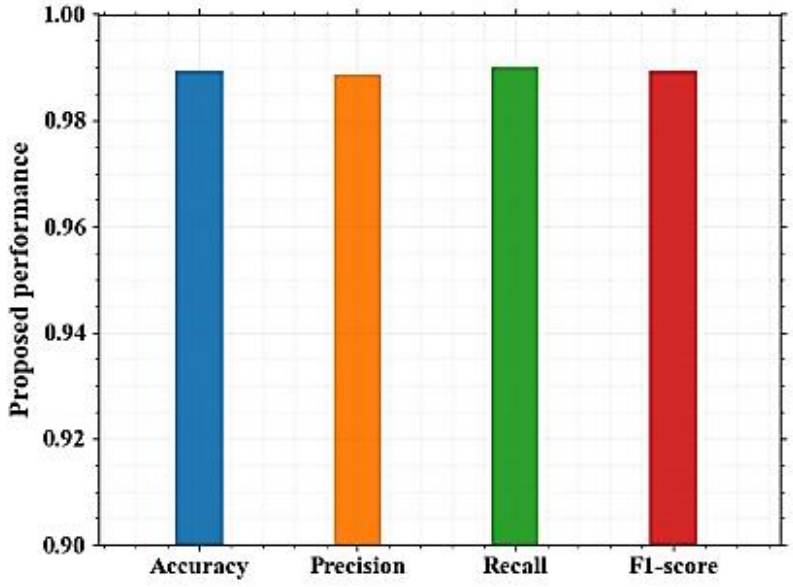

Figure 3.7: Performance analysis of proposed method

The performance criteria considered include F1-score, recall, accuracy, and precision. The suggested approach achieved impressive metrics, with 98.9% accuracy, 98.8% precision, 98.9% recall, and 98.9% F1-score. Figure 3.8 illustrates the Area Under the Curve (AUC) for the suggested method.

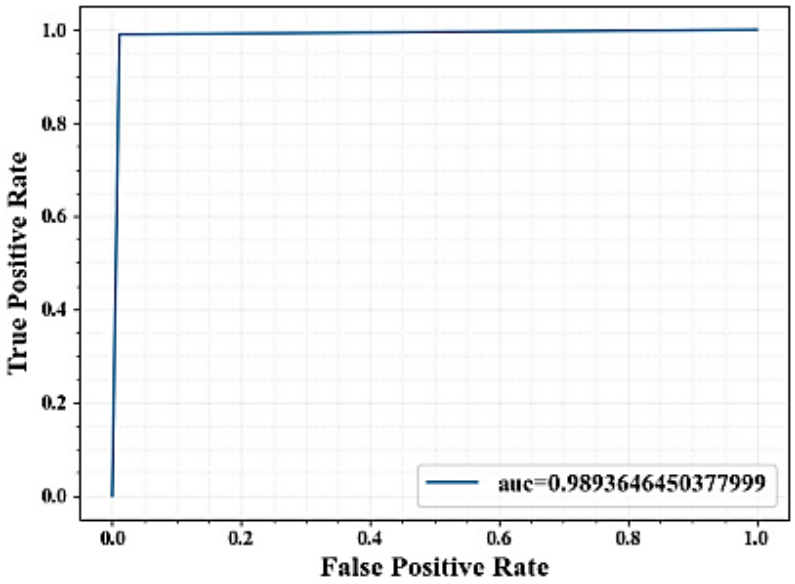

Figure 3.8: AUC

The suggested strategy yielded an AUC of 0.989%. The vertical axis represents the true positive rate, while the horizontal axis represents the false positive rate. The suggested strategy performs better, as evidenced by the higher AUC value.

3.4 Conclusion

This study examines several approaches to Twitter crime detection. Millions of people use Twitter as a venue for many forms of online communication. Consequently, this is where the Twitter data's criminal detection mechanism is executed. This research proposes a hybrid WCNN in addition to WCO. The dataset undergoes initial preprocessing steps, including stop word removal, tokenization, and stemming. Following this, various techniques such as feature hashing, TF-IDF, Glove, and BoW are employed to extract features. Feature selection is then performed using MTGA. Subsequently, a CNN integrated with wavelet transformation is utilized for crime data classification, with performance maximized using the WCO algorithm. Furthermore, data clustering is conducted using FMRF. The proposed approach is compared with

several existing algorithms, including GSCA, HANNCC, HMNCC, and HREACC. Implementation is carried out using Python programming language. The experimental results demonstrate exceptional performance metrics, with an accuracy of 98.9%, precision of 98.8%, recall of 98.9%, F-measure of 98.9%, and AUC value of 98.9%. These results outperform those obtained by other algorithms significantly. While the proposed algorithm is currently finalized, future adaptations may involve leveraging real-time Twitter data streaming to forecast future crimes. possible to add kinds of crimes to the system to increase its resilience and efficiency.

CHAPTER 4

TWITTER CRIME DETECTION USING DAC-BiNET

This chapter provided a thorough analysis of the methods and components of the suggested DAC-BiNet linked with the Aquila optimal network. To improve the model's efficiency, extensive pre-processing, sophisticated TF-IDF feature extraction, feature hashing, and glove modeling were all used. The streamlined performance of the model was further enhanced by the implementation of a unique strategy for feature reduction that utilized Possibilistic Fuzzy LDA-based clustering. The chapter examined the DAC-BiNet's practical use in Twitter crime detection and classification, providing insight into the model's performance in actual use cases. The outcomes were carefully examined and discussed, providing insightful information about the model's advantages and any shortcomings. In closing, a brief synopsis consolidated the chapter's contributions to the field of crime data prediction by summarizing the main conclusions from the investigation of the proposed DAC-BiNet with Aquila optimal network.

4.1 Overview

The number of criminals and crimes is continuously increasing, and security is now a big worry nowadays. An offensive conduct that causes bodily or psychological injury to a person or piece of property and is sanctioned by criminal law is referred to as a "crime" [69-70]. As a result of technological advancements, a number of crimes are committed on social media sites such as Facebook, Instagram, and Twitter [71-72]. Social media platforms are becoming a significant source of sentiment-rich data that users may use to convey their thoughts, ideas, and feelings through tweets, blog posts, reviews, and other forms of content. One of the more well-known social media platforms, Twitter serves as a global platform for short text messages known as tweets [73].

Due to a lack of predicted criminal characteristics, daily crimes on Twitter, a major microblogging service with 217 million subscribers, are on the rise [74]. Fraudsters are creating phony accounts to propagate spam by pretending to be genuine users. The quality of life and economic progress are severely impacted by cybercrimes. It is difficult to identify these crimes, hence it is important to create efficient criminal pattern detection [75-76]. Data mining (DM) approaches can be used to uncover hidden patterns and extract knowledge from massive amounts of data [77-78]. Examples of these techniques are classification, rule mining, and clustering. Text mining (TM) is the process of transforming unstructured text material into structured form by applying Natural Language Processing (NLP) techniques [79-80]. The goal of crime detection is to find crimes on social media before they seriously affect society. AI's subfield of computer vision (CV) teaches robots to perceive and comprehend the visual environment [81]. Automated solutions are required for the efficient classification of criminal and non-crime tweets due to the technological expansion of Twitter crimes. Police officers and victims both gain from AI-based detection of crime, which reduces the amount of time needed to resolve crimes.

The number of crimes on social media is growing daily, which causes users to face severe issues. The Twitter platform has garnered increased attention in recent years owing to its efficacious offerings. However, the crime that is displayed in the information collected by Twitter raises a number of concerns for society. Thus, this work presents a resilient Deep Attention Convolutional Bi-directional Aquila Optimal Network (DAC-BiNet) to detect various illicit actions on Twitter. First, during the pre-processing stage, the noise in the twitter data is eliminated by segmenting the sentences, removing punctuation, tokenizing, eliminating stop words, fixing acronyms and slang, lemmatizing, lower casing, stemming, and removing hashtags and URLs. The second step involves feature extraction, which uses techniques like feature hashing, glove modeling, and ITF-IDF (Improved Term Frequency-Improved Document Frequency) to extract the useful features. The best attributes that are suitable for detecting crimes are selected from this for the following step. The next step is to minimize the size of the feature set using a probabilistic fuzzy LDA (Latent Dirichlet Allocation) based clustering algorithm, which increases prediction capability while requiring less processing time. Ultimately, DAC-BiNet method is used to detect

and classify provided tweet. The experimental setup for this proposed study is carried out using the Python platform, and the effectiveness of the suggested model is assessed by comparing it to a number of performance indicators using some of the methods that are currently in use. The enhanced results obtained by the suggested model are 98.23% accuracy, 83.86% precision, 90.05% recall, 98.86% specificity, and 86.84% F1 score.

4.2 Proposed methodology

This suggested study presents a DAC-BiNet based criminal tweet detection to provide the safest and most secure online services. Data collection, pre-processing, multi-feature extraction, clustering, and crime twitter detection are processes utilized to find criminal tweets. First, the web source is used to gather the tweet data. Because the online tweet data is unstructured, pre-processing phase must be completed before detection procedure. Sentence segmentation, punctuation removal, tokenization, stop word removal, slang correction and acronym, lemmatization, lower casing, stemming, hashtag and URL removal are the processes utilized in the pre-processing. The goal of feature extraction is to convert any type of data such as text or images into a numerical format. In this case, multi-feature techniques like Glove Modelling, Feature Hashing, and ITF-IDF are used to extract the tweet features. The optimal feature extraction procedure is examined, and the clustering model receives the best feature extraction result. Next, Possibilistic Fuzzy LDA (Latent Dirichlet Allocation) is used to cluster the contents of unstructured tweets. The ability of prediction models is increased by this clustering technique, which reduces the initial collection of features to a smaller set. Lastly, DAC-BiNet can be used for identification and categorization of criminal tweets. Here, the Aquila Optimization (AO) Algorithm is used to minimize the losses in the deep learning model. Nonetheless, the deep learning classifier model's effectiveness is increased by using this optimization technique to determine parameters from provided input data. As a result, suggested DAC-BiNet model can greatly benefit from the application of an optimization method to fine-tune the network's parameters and accurately classify data on Twitter crimes. The work flow of the suggested DAC-BiNet approach is shown in Figure 4.1.

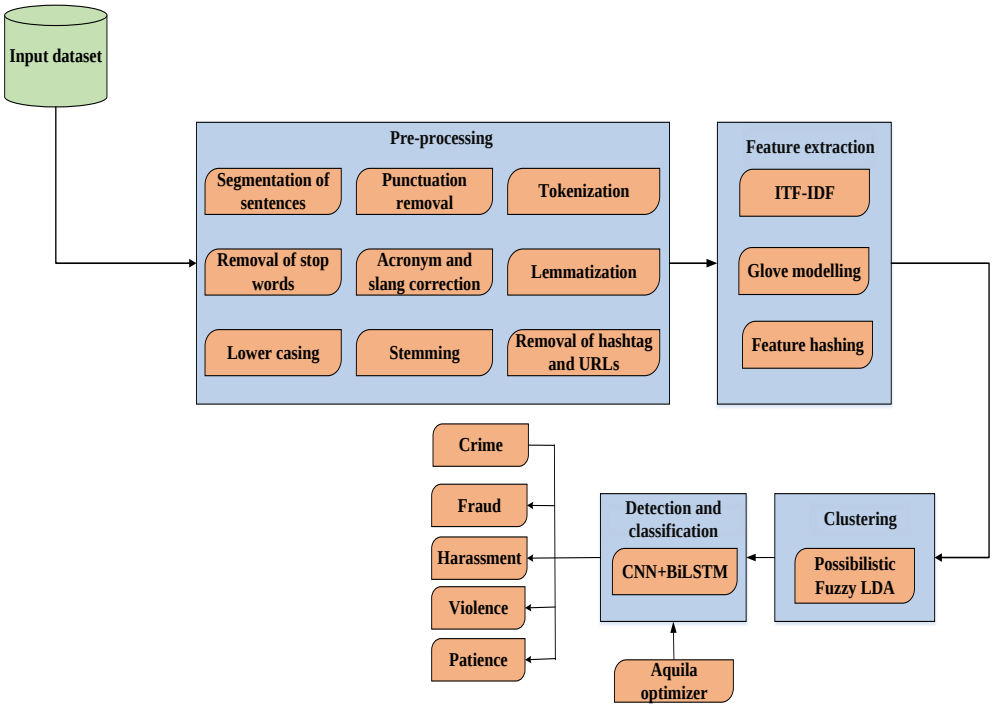

Figure 4.1 Structure of proposed DAC-BiNet model

4.2.1 Pre-processing

Pre-processing is done in order to reduce noise in the provided Twitter data and aid in the transformation of unstructured data into structured data. The suggested work used a variety of pre-processing techniques, including sentence segmentation, tokenization, punctuation removal, stop word removal, slang correction and acronym, lemmatization, lower casing, stemming, hashtag and URL removal etc.

Sentence segmentation: In this stage, the data derived from Twitter is separated into discrete words or tokens. This phase, to put it simply, breaks the sentences up into words. The line "Rosy is very beautiful" on Twitter is an example of sentence segmentation; following the process, it becomes "Rosy," "is," "very," and "beautiful."

Punctuation removal: This stage is crucial for dividing the provided input material into sentences, paragraphs, and phrases. This step's primary goal is to eliminate various punctuation from the Twitter data, such as question marks, colons, exclamation points, and commas.

Tokenization: This stage divides the text data into smaller pieces, like words or distinct phrases, from paragraphs, sentences, phrases, or the entire text data. The input Twitter data's words are separated into tokens, and the tokenization method mentions the text's meaning by confirming the word order.

Elimination of stop words: This stage concentrates the data that provide crucial information by eliminating low-level information from the data. Pronouns and articles that are present in provided data are typically classed as stop words. Stop words include "the," "what," "is," and so on. These phrases are eliminated in this step because they do not provide enough information to detect criminal activity on Twitter.

Correction of acronyms and slang: This step broadens the range of comparable acronyms and slang terms found in Twitter data. One example given is that the term "bf" substituted for "beautiful flowers," while terms like "2day" are enlarged to "today."

Lemmatization: This stage determines the word's lemma based on the meaning of the word in the Twitter data. This stage removes the inflectional endings and helps to explain the morphological analysis of the Twitter word.

Lower casing: In this stage, the words from Twitter are converted to lower case, such as BOOK to book.

Stemming: This process strips a word of its prefix and suffix and returns it to its root stem. This is a crucial step in NLP and NLU that speeds up system performance.

URL and hashtag removal: In this stage, the URL is deleted from the input data along with all of the hashes that are currently present in the Twitter data. For instance, The Proud Family replaces #TheProudFamily. The example URL for URL removal is https://www.computerhope.com; this kind of URL is eliminated from the provided Twitter data.

By following the aforementioned procedures, the noisy data is removed from the provided Twitter data and the work is pre-processed. Additionally, this stage creates structured data based on Twitter, which facilitates the subsequent phases of crime detection.

4.2.2 Feature extraction

To lower the computational difficulty of the suggested model, the key features are retrieved from pre-processed data. The key elements deliver information that is helpful in detecting crimes. The suggested work makes use of glove modelling, feature hashing, and ITF-IDF for this purpose.

4.2.2.1 Improved Term Frequency-Improved Document Frequency (ITF-IDF)

The basic text-based feature extraction technique, known as the ITF-IDF method, extracts a variety of required features. An alternative name for the ITF-IDF approach is the inverse frequency document weight feature. The weight value calculated to reduce longer text length and the repeated characteristics in provided Twitter data are denoted as TF. So, a weight assessment technique used to start feature weight addition process. Each Twitter data set's feature weight is provided as follows:

$$\omega_i = \frac{TF_i \times \log_2\left(\frac{S}{s_i} + 0.01\right)}{\sqrt{\sum\left(TF_i \times \log_2\left(\frac{S}{s_i} + 0.01\right)\right)^2}} \tag{4.1}$$

Here, the quantity of data in the content of the features f_i is denoted as s_j, feature value of the pre-processed data from twitter represented as ω_i, the total volume of the training sample is marked as S, and the number of occurrence that present in features item f_i is denoted as TF_i

The suggested TF-IDF is reliable for classifying various types of information if number of times a word appears in provided tweet is significant while the volume of times the same term appears in a different tweet is low. Due to the tweet data's location information not being determined, the TF-IDF method's classification performance is not very good. Furthermore, semantic information cannot be extracted using the TF-IDF approach. As a result, the recall and precision performance declines. In essence, the terms are verbs, adjectives, phrases, or nouns. However, stop words can also refer to conjunctions, quantifiers, and function words. Because of the superfluous material, removing such terms cannot yield improved results. Filtering is thus included in this

sort of word to improve recall and precision capabilities. Let's assume that the weight construction formula for computing feature items takes into account of word's length, position, and speech component. This revised TF-IDF proposal improves the performance of the traditional TF-IDF. It is stated as,

$$\omega_i = \frac{TF_i \times \log_2\left(\frac{S}{s_i} + 0.01\right)}{\sqrt{\sum\left(TF_i \times \log_2\left(\frac{S}{s} + 0.01\right)\right)^2}} \times \gamma \times \sum_{i=1}^{n} \lambda \times \xi \tag{4.2}$$

Where ω_i stands for the feature weight of the tweet data, γ for feature length parameter, λ for volume many instances presented in position i, and ξ for the word's speech location parameter.

4.2.2.2 Feature hashing

Every feature in the data that is displayed is subjected to a hash function in feature hashing process. Various lengths of unstructured text data are converted into features with comparable lengths of numeric tweets using the feature hashing technique. By comparing hash values, the hashing feature can reduce the dimensionality of the data and facilitate rapid feature exploration by altering the string. Since each n-gram is mapped to feature on a tweet, the number of features collected from feature hashing method is improved.

4.2.2.3 Glove modelling

The word embedding method that mentions words in tweets as a matrix of numerical vectors or values is called Glove Modelling. Every word with the same meaning is given the same vector representation using the word embedding technique. In this case, every word in the Twitter data is changed into single vector, and low-dimensional space is mostly used for the mapping process. It is also enabled according to the size of the data. The Glove modelling technique uses the percentage of co-occurrence possibilities to convey the meaning of a word. Word vectors are effectively produced by this glove modelling for word similarity tasks. Semantic characteristics are extracted using this strategy.

The required features are extracted utilizing feature hashing, glove modeling, and ITF-IDF techniques. After that, the characteristics that are most useful for detecting crimes are examined, and the best characteristics are chosen for additional processing. The elements from the ITF-IDF approach are more effective in the planned work; therefore, they move on to the next phase.

4.2.3 Possibilistic Fuzzy LDA based clustering for feature reduction

This clustering method reduces the enormous dimensionality of features derived from feature extraction methods. The suggested work uses possibilistic fuzzy LDA technique to make decreased feature set. Here, possibilistic fuzzy c-means (PFCM) is provided for build a reduced feature set for an effective crime detection method, while LDA is used for topic modeling and identification.

Each page is mentioned using the LDA, a simple probabilistic topic approach, as a random collection of latent themes. Each latent topic in LDA is represented as an average over a fixed set of words, and it used to determine underlying latent topic's design based on the available data. Generally speaking, the entire document's words are produced in two stages. In the first stage, a distribution over themes is randomly chosen for every word in the retrieved feature set. A word is a singular data point from an alphabetical index in the LDA method, and it is represented as $\{1,.....,A\}$, $b=(b_1,b_2.... ....b_m)$ represent a string of w words, and let $d=(b_1,b_2.... ....b_N)$ represent a collection of N documents. The 3D Bayesian graphical technique determines how the LDA method is displayed. In this case, there are three nodes with random variables, and the edges represent possible relationships between the variables. Analysis is done on parameters like ϕ and λ during corpus formation. A joint distribution across random variables is stated for generative process of LDA, and the likelihood density as a function of arbitrary variables at p -dimension is assessed as,

$$q(\zeta/\phi)=\frac{\Gamma\left(\sum_{i=1}^{p}\phi_i\right)}{\prod_{i=1}^{p}\Gamma(\phi_i)}\zeta_1^{\phi_1-1}..... .....\zeta_p^{\phi_p-1} \tag{4.3}$$

The combined distribution of all topics is calculated as follows:

$$q(\zeta,x,y\mid\phi,\lambda)=q(\zeta\mid\phi)\prod_{w=1}^{W}q(x_w\mid\zeta)q(y_w\mid x_w,\alpha) \tag{4.4}$$

The formula for corpus probability is as follows:

$$q(d \mid \phi, \lambda) = \Pi_{f=1}^{N} \int q(\zeta_w \mid \phi) \times \left(\Pi_{w=1}^{w_f} \sum_{x_{fw}} q\left(x_{fw} \mid \zeta_f\right) q\left(y_{fw} \mid x_{fw}, \lambda\right) \right) f\zeta_f \tag{4.5}$$

Where ϕ refers to the dirichlet parameter, the document level topic variables are denoted as ζ, the per-word topic assignment is x, the acquired word is denoted as y, the topics are denoted as λ, the total word count is W, and the document is denoted as N. Key words that are classified as positive or negative are retrieved, each with a unique weight value, and compiled into a dictionary. During the testing phase, the dictionary and the testing twitter data are coordinated to obtain the positive and negative weight values. Once the positive and negative keywords have reached their weight values, the clustering process is carried out in order to reduce the feature size. Using PFCM, the twitter data is clustered to a degree indicated by the membership grade, with each tweet considered a component of the cluster. The following is the PFCM's objective function:

$$G_{PFCM}(q, f, v) = \sum_{i=1}^{k} \sum_{j=1}^{w} \left(q_{ij}^{n} + t^{w} \right) f^{2}\left(x_j, v_i\right) \tag{4.6}$$

The restrictions are stated as follows:

$$\sum_{i=1}^{k} \lambda_{ij} = 1, \forall j \in \{1, \ldots\ldots, w\} \tag{4.7}$$

$$\sum_{j=1}^{w} f_{ij} = 1, \forall i \in \{1, \ldots, k\} \tag{4.8}$$

In this case, the partition matrix is represented by q, the typicality matrix by f, the cluster centroids vector by v, and the objective function by G_{PFCM}. Equations (4.9), (4.10) and (4.11) are used to compute the degree of membership and cluster centre for each iteration of the objective function.

$$\lambda_{ij} = \left[\sum_{p=1}^{k} \left(\frac{ex_j, v_i}{ex_j, v_p} \right)^{\frac{2}{n-1}} \right]^{-1}, \; 1 \le i \le k, \; 1 \le j \le w \tag{4.9}$$

$$f_{ij} = \left[\sum_{p=1}^{m} \left(\frac{ex_j, v_i}{ex_j, v_p} \right)^{\frac{2}{w-1}} \right]^{-1}, \quad 1 \le i \le l \;\; 1 \le j \le n \tag{4.10}$$

$$v_i = \frac{\sum_{p=1}^{w} \left(o_{ip}^{n} + t_{ip}^{w} \right) x_p}{\sum_{p=1}^{w} \left(o_{ip}^{n} + t_{ip}^{w} \right)}, \; 1 \le i \le k \tag{4.11}$$

Where w denotes the number of data points, k is the number of cluster centroid, and the coordinates are indicated by ($\left(x_j, v_i\right)$). This information is used to calculate the spacing between the data points and the cluster center. Using standard cluster centers and prototypes, PFCM creates memberships and opportunities for each cluster. The objective function used in the clustering process determines how well the clustering performs. The two requirements are taken into consideration when developing the powerful objective function. The distance between each cluster should be kept to a minimum initially. Second, there should be less space between the data points that were allocated to the clusters.

The PFCM's objective function, denoted by parameter η, is improved by driven prototype learning. This parameter is updated after every iteration, and η is expressed as,

$$\eta = \exp\left(- \min_{i \ne p} \frac{\left\| v_i - v_p \right\|}{\alpha} \right) \tag{4.12}$$

If the variation in the sample is represented by the letter α and is thought to be,

$$\alpha = \frac{\sum_{j=1}^{w} \left\| x_j - \bar{x} \right\|^2}{w} \tag{4.13}$$

$$\text{Where, } \bar{x} = \frac{\sum_{j=w}^{w} x_j}{w} \tag{4.14}$$

A weight parameter is used to describe the equivalent value of η. This weight function is assessed as follows: it provides better classification results.

$$\omega_{ji} = \exp\left(- \frac{\left\| x_j - v_i \right\|^2}{\left(\sum_{j=1}^{w} \left\| x_j - \bar{v} \right\|^2 \right) \times k / w} \right) \tag{4.15}$$

Where ω_{ji} represents weight function of j^{th} point with the i^{th} class. This procedure is the basis for the clustering, which reduces the feature sets and produces highly accurate classification results.

4.2.4 Twitter crime classification and detection using DAC-BiNet

The DAC-BiNet classification model receives the clustered feature set as an input. To achieve better detection results, the suggested method hybridizes CNN and Bi-LSTM algorithms. First, a clustered feature set is entered into the CNN method. The CNN layers then extract most significant features and send them to Bi-LSTM technique for crime detection. The input layer, convolutional layer, pooling layer, Bi-LSTM layer, attention layer, and output layer are six layers that make up the suggested DAC-BiNet model. The suggested DAC-BiNet's design is shown in Figure 4.2.

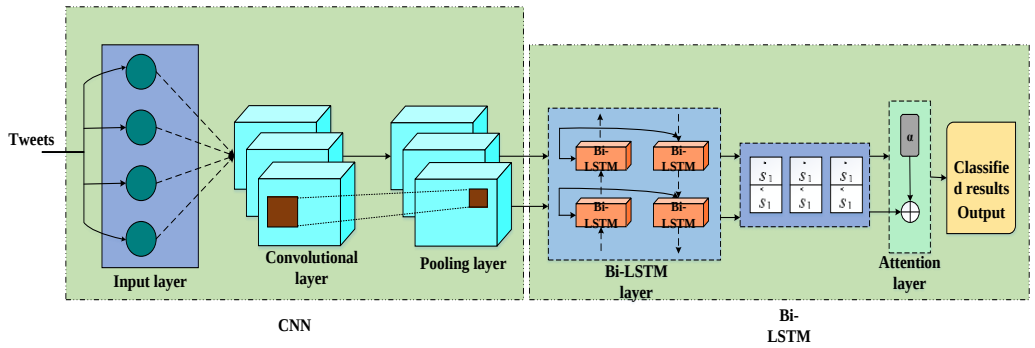

Figure 4.2 Proposed DAC-BiNet model

The input layer receives the feature set that was produced by the clustering technique. The convolutional layer receives the feature set from this input layer. Although the convolutional layer captures important characteristics, the detection performance may be impacted by its excessively large dimension. This means that a max-pooling layer is used in order to decrease the feature dimension. Additionally, this max-pooling layer lessens the problem of over-fitting and facilitates the production of ideal outcomes.

$$U_l = a\left(x_l * m_l + y_l\right) \tag{4.16}$$

Whereas the input vector is denoted as x_l, the weight of the convolutional kernel is mentioned as m_l, the bias of the convolutional kernel appears as y_l, and output value

following convolutional layer procedure is mentioned as $a(x) = \max(0, x)$. The CNN method's condensed features are transmitted to Bi-LSTM Layer. This Bi-LSTM layer uses forward as well as backward LSTMs to handle the input features of varying lengths. The Bi-LSTM layer uses two hidden states to help it flow the data both forward and backward. Encoding the crime data in two directions is the primary goal of this bidirectional encoding. The output of memory cell is assessed as,

$$\vec{S} = \left\{\vec{S}_1, \vec{S}_2, ..., \vec{S}_l\right\} \tag{4.17}$$

$$\overleftarrow{S} = \left\{\overleftarrow{S}_1, \overleftarrow{S}_2, ..., \overleftarrow{S}_l\right\} \tag{4.18}$$

$$S_l = \vec{S}_l \oplus \overleftarrow{S}_l \tag{4.19}$$

Where, mentions the hidden state of forward direction, represents the hidden state of backward direction and the addition operation is indicated as. The Output layer is updated by perform concatenation between $\vec{s}$ and $\overleftarrow{s}$.

Whereas $\overleftarrow{S}$ stands for hidden state of the backward direction and $\oplus$ denotes addition operation, $\vec{S}$ refers to hidden state of the forward direction. Combining between $\vec{S}$ and $\overleftarrow{S}$ is performed to update the output layer. Determining the relationship between the words is crucial for classifying the crimes that are presented in the twitter tweet. In this way, the attention layer receives the hidden sequences. This layer of attention determines each word's significance as well as the connections between them. It detects the significant features according to the higher weight values. The attention weights of each word are assessed as,

$$X_m = \chi_m^F \left(W_m S_l + y_m\right) \tag{4.20}$$

$$X_m = Soft\max \frac{\exp\left(X_m\right)}{\sum_m \exp\left(X_m\right)} \tag{4.21}$$

$$k = \sum_w x_w S_w \tag{4.22}$$

In this case, W_m stands for the weight matrix, y_m for the bias term, χ_m^F for the transposed weight vector, x_m for weights of attention via the SoftMax function, and k

for weighted sum for hidden demonstration. The output layer receives input from attention layer and uses the sigmoid function to anticipate outcome. It is stated as

$$b = \begin{cases} 0 & ep[0, 0.5] \\ 1 & Otherwise \end{cases} \tag{4.23}$$

Where the categorized result is written as and ep stands for the estimated probability. One is shown as presented offenses and zero is shown if there are no crimes accessible. The classifier identifies the various criminal classes that are present in the tweet if the crime is analyzed. The offered loss function reduces the detection performance in the suggested classification stage. By using the AO technique to update the weight values, the loss function is minimized [82]. It states that the categorization stage's loss function is,

$$C_t = \frac{1}{S}\sum_{i=1}^{S}\left(x_i - q_i^{'}\right)^2 \tag{4.24}$$

Where x_i the real is amount and $q_i^{'}$ is the estimated value, and S denotes the maximum number of iterations. The fitness function is used to examine the offenses on Twitter data, and the results are as follows:

$$Fitness = Minimize\left[C_t\right] \tag{4.25}$$

Reducing the loss function through weight parameter updates for efficient crime detection is the aim of the AO technique. The population-based approach used in this optimizer was motivated by the hunting habits of Aquila's. The AO approach's feature initialization procedure is shown as,

$$Tweet\ features = \begin{bmatrix} h_{1,1} & \ldots & h_{1,j} & \mathrm{h}_{1,\mathrm{L}\text{-}1} & h_{1,C} \\ h_{2,1} & \ldots & .h_{2,j} & \ldots & h_{2,C} \\ \ldots & \ldots & .h_{i,j} & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{N-1,1} & \ldots & .h_{N-1,j} & \ldots & h_{N-1,C} \\ h_{N,1} & \ldots & h_{N,j} & h_{N,L-1`} & h_{N,C} \end{bmatrix} \tag{4.27}$$

The present feature solution is mentioned in the equation (4.26), where the size of the problem is represented by C, the total number of features is indicated by N, and the location of text feature in the j^{th} result is represented by h_j. Following the startup process, features are produced at random, such as

$$T_{ij} = rand \times (o^y{}_j - c^y{}_j) + CY_j,\ i = 1,2,...,N \quad j = 1,2,...,C \tag{4.27}$$

In this case, the random number is represented by $rand$, the upper bound is denoted by $o^y{}_j$, and the lower bound is indicated by $c^y{}_j$ in position j. The AO algorithm goes through four distinct steps, including expanded exploration, narrowed exploration, expanded exploitation, and narrowed exploitation, to produce the best results.

Expanded exploration: During this phase, the search agent looks at important features and chooses the best ones to help with effective crime detection. This procedure is described as

$$H(v+1) = H_{best}(v) \times \left(1 - \frac{l}{V}\right) + \left(H_Q(v) - H_{best}(v) \times rand\right) \tag{4.28}$$

Where, $(1-\frac{l}{V})$ term is used to achieve the big exploration process, and $H(v+1)$ denotes next iteration in the v^{th} result. The most effective approach that gives the position of the useful feature is shown as $H_{best}(v)$. V Stands for the maximum iteration, v indicates the current iteration number, and $rand$ stands for a random distribution function with a range of 0 to 1.

$$H_Q(v) = \frac{1}{N}\sum_{i=1}^{N} H_i(v),\ \forall j = 1,2,...,C \tag{4.29}$$

Where N the total amount of features is, C is the issue size, and v is the iteration that is being discussed at that time.

Narrowed exploration: Where the search agent first focuses on the required feature locations. Following that, the search agent selects the feature of crime detection by circling the important features. Here is how the search agent's position is updated:

$$H_2(v+1) = H_{best}(v) \times \xi(C) + H_r(v) + (x - y) \times rand \tag{4.30}$$

Where the feature position is indicated by C, the function of the levy flight distribution is denoted by $\xi(C)$, the randomly generated solution is represented by

$H_r(v)$, and $H_2(v+1)$ shows subsequent iteration of the v^{th} result. The formula used to calculate the levy flight distribution is,

$$\xi(C) = u \times \frac{k \times \beta}{|\eta|^{\frac{1}{\chi}}} \tag{4.31}$$

The random numbers from 0 to 1 are signified by k and β, while o denotes a constant value of 0.7. β, the constant value, is assessed as

$$\beta = \left[\frac{\gamma(1+\chi) \times \sin\left(\frac{\pi\chi}{2}\right)}{\gamma\left(\frac{1+\chi}{2}\right) \times \chi \times 2^{\gamma-1/2}} \right] \tag{4.32}$$

Expanded exploitation: To find the feature qualities, the search agent descends vertically, which widens the exploitation process. Next, given the Twitter data, it begins to select the qualities that are appropriate for identifying crimes. Based on size and quality, search agents can select features. Additionally, traits that meet the criteria for fitness are chosen and are provided as follows:

$$H_3(v+1) = (H_{best}(v) - H_o(v)) \times \varpi - rand + ((o^y - l^y) \times rand + c^y) \times \vartheta \tag{4.33}$$

If the future iteration of solution v is denoted by $H_3(v+1)$, the parameters are represented by ϖ and ϑ, random function is denoted by rand, spans from 0 to 1, and upper and lower bounds are denoted by o^y and c^y.

Narrowed exploitation: The search agent can choose most helpful features for categorization based on how well a feature works. The search agent selects features based on the last location in this case. This conduct is described as

$$H_4(v+1) = Quality\ function \times H_{best}(v) - (vf_1 \times H(v) \times rand) - vf_2 \times \xi(C) + rand \times J_1 \tag{4.34}$$

Where vf_1 stands for variable feature qualities, vf_2 for lowering value, and $H(v)$ for the current answer. The searching process of the last iteration is represented by $H_4(v+1)$. The term $Quality\_function$ aids in controlling the strategy of exploration. The calculation of the quality function is,

$$Quality\_function(v) = v^{\frac{2\times rand-1}{(1-V)^2}} \tag{4.35}$$

$$vf_1 = 2\times rand - 1 \tag{4.36}$$

$$vf_2 = 2\times\left(1-\frac{v}{V}\right) \tag{4.37}$$

Where, $Quality\_function(v)$ indicates the random number between 0 and 1 and indicates the range of the quality function in each iteration. The fitness function is calculated for each iteration, and results are related between iterations. The best fitness function is then analyzed in order to determine the optimum solution. Therefore, the suggested DAC-BiNet model effectively detects the crimes that are provided from the Twitter data. Table 4.1 illustrates the AO approach's pseudocode.

Table 4.1 Pseudocode of AO algorithm

Algorithm: AO approach
Set the features to initial.
Set up settings such as $\varpi, \chi, ...$
if (The requirement is not met) **do**
Apply equation (**4.**25) to evaluate the fitness function.
$H_{best}(v)$= Determine which solution is optimal depending on fitness function.
for $(i = 1,2,...., N)$do
Revise the existing resolution. $H_Q(v)$
Update p, r, vf_1, vf_2 and $\xi(C)$
if $rand \le 0.5$ then
Utilizing equation (6.28), update the solution
if fitness $(H_1(v+1)) < fitness(H(v))$ **then**

$H(v) = H_1(v+1)$

end if

else

Utilizing equation (6.30), update the solution

if fitness $(H_2(v+1)) < fitnessH(v))$ **then**

$H(v) = H_2(v+1)$

end if

else

if $rand \leq 0.7$ **then**

Utilizing equation (6.33), update the solution

if fitness $(H_3(v+1)) < fitness(H(v))$ **then**

$H(v) = H_3(v+1)$

end if

end if

else

Utilizing equation (6.34), update the solution

if fitness $(H_4(v+1)) < fitness(H(v))$ **then**

Using equation (6.35), evaluate quality function

end if

return

The many classes of crimes on Twitter, such as harassment, fraud, police, crimes, and violence, are discovered from given input data by suggested DAC-BiNet classification model. The provided loss function is lowered with the help of Aquila tuning, improving classification performance.

4.3 Results and discussions

Using a dataset of Twitter users, the suggested DAC-BiNet crime detection model is investigated in this section. The suggested study uses a Python tool to conduct the experimental setup, and it compares the suggested techniques' performance to other known techniques using a number of performance indicators. The simulation investigation of the suggested DAC-BiNet model is carried out utilizing a dataset of Twitter user data. The 20,000 rows in this collection comprise user names, several account profiles, location data, random tweets, and photos. This dataset has been used in a number of previous publications, with improved results. For the purpose of identifying crimes on Twitter, the suggested study prefers this dataset. The suggested work's hyperparameter settings are mentioned in Table 4.2.

Table 4.2 Hyper parameters

Parameters	**Values**
Number of epochs	100
Learning rate	0.1
Activation	Relu
Kernel size	5
Batch size	128
Pool size	2

4.3.1. Performance metrics

The primary presentation indicators employed to calculate efficacy of suggested model are processing time, F1-score, accuracy, specificity, recall, and precision. The model's efficacy is demonstrated by the classification accuracy determination. A comparison is made between the suggested model's obtained results and those of other deep

learning techniques, such as CNN, Bi-LSTM, DBN (Deep Belief Network), and DNN (Deep Neural Network).

- ***Accuracy***

In order to assess how well a classification system performs in categorizing different classes of crimes from Twitter data, accuracy is crucial criterion. The definition of it is the proportion of accurately classified data to total amount of data. The definition of accuracy is as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{4.38}$$

Where, fp mentions false positives, fn represents false negatives tp , indicates true positives, and tn represents true negatives.

- ***Specificity***

The specificity metric calculates how well the classification model can classify each sample's true negatives. Additionally referred to as genuine negatives, specificity is assessed as

$$Specificity = \frac{tn}{tn + fp} \tag{4.39}$$

Where, true negatives represent tn and false positives signifies fp

- ***Recall***

The amount of correctly classified data from total number of positive predictions made by suggested DAC-BiNet method is calculated using recall metrics. Recall, also referred as sensitivity, is a measure that shows how many precisely categorized data points were missed. It is quantified as,

$$\mathrm{Re}call = \frac{tp}{tp + fn} \tag{4.40}$$

Where, fn indicates false negatives and tp represents true positives.

- ***Precision***

The number of positive, accurate predictions the suggested classifier made is shown by the precision metric. Its definition is the ratio of the total number of positive data that the classifier predicted to the number of perfectly categorized positive samples. The prediction is stated as follows:

$$\Pr ecision = \frac{tp}{tp + fp} \tag{4.41}$$

- ***F1-score***

The statistic known as the F1-score represents the test's accuracy measure and is calculated as the harmonic average of recall and precision. Better classification is indicated by the F1-score with value one, which is mentioned as,

$$F_1 - score = 2 \times \frac{P \times r}{P + r} \tag{4.42}$$

Where, precision and recall mentioned as P & r respectively.

4.3.2 Performance evaluation

Using a deep learning model, a performance evaluation is carried out using Twitter data that was collected from the dataset. The dataset includes Twitter data pertaining to a variety of classes, including violence, harassment, fraud, crime, and police. The suggested study made use of several potent strategies to detect crimes on Twitter. By extracting significant features during feature extraction stage, computational complexity is decreased, and features are grouped into low-dimensional feature sets through the clustering process. This makes it easier for the classifier to identify crimes in the Twitter data. A comparative examination employing several deep learning techniques is presented in this section. The suggested work compares performance using standard methods such as CNN, DBN, Bi-LSTM, and DNN. Figure 4.3 displays the suggested DAC-BiNet classification's confusion matrix.

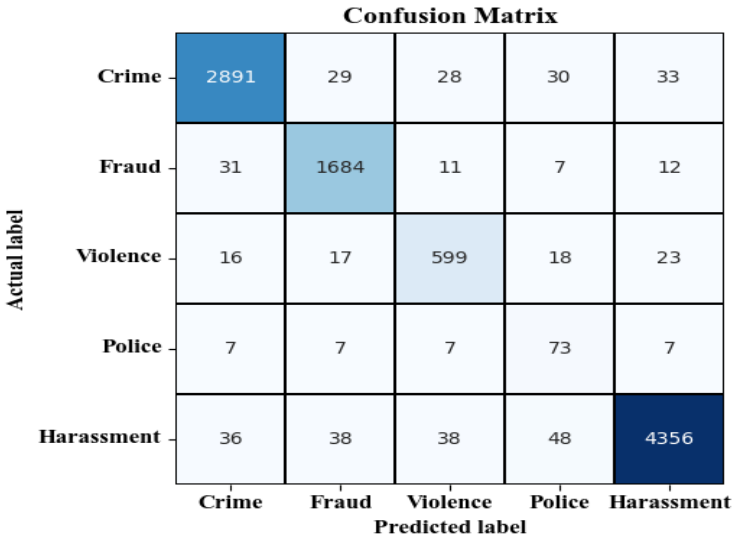

Figure 4.3 Confusion matrix of proposed DAC-BiNet model

The suggested model's capacity for prediction is demonstrated by the confusion matrix. Different types of offenses are discovered in each testing session. Of the 3011-crime data, 2891 are properly classified as crimes, whereas the other data, including 29, 28, 30, and 33, are incorrectly labeled as harassment, assault, fraud, and police. Of the 1745 fraud data, 1684 are correctly labeled as fraud, with just a small number of data being misclassified. For example, 31 data are incorrectly classified as crimes, 11 as violent crimes, 7 as police crimes, and 12 as harassment. Out of 673 data points on violence, 599 are correctly labeled as such, 16 are incorrectly classed as crimes, 17 are incorrectly classified as fraud, 18 are incorrectly classified as police, and 23 are incorrectly projected to be harassment. Of the 101 police data sets, 73 are correctly classed as police, while the remaining data sets are incorrectly classified as other classes. In a similar vein, of the 4516-harassment data collected, 4356 are accurately identified as such; the other data are misclassified. This demonstrates the efficacy of the suggested classifier in terms of detecting crimes using various classes. The precision and efficacy comparison of the suggested and current procedures is shown in Figure 4.4.

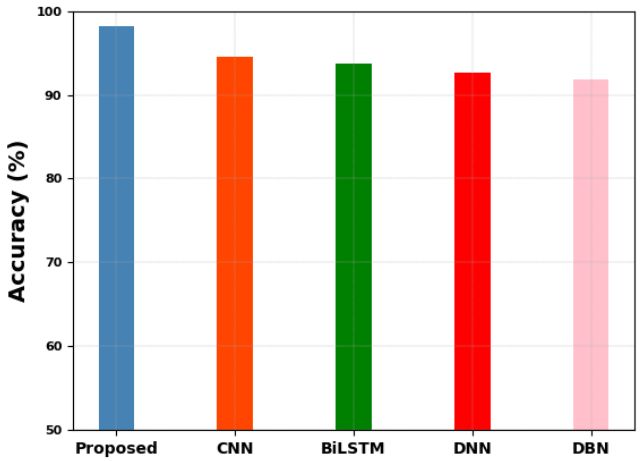

Figure 4.4 Comparison analysis using accuracy performance

The suggested model's accuracy value is contrasted with that of several other methods now in use, including CNN, Bi-LSTM, DNN, and DBN. The comparative analysis unequivocally demonstrates that the suggested classifier outperforms the others in terms of accuracy. In the proposed study, 10 distinct processes are used in the preliminary processing phase to reduce noisy data from the dataset. It enhances the quality of the data and facilitates the classifier's performance. Additionally, by updating the weight values, the optimization technique immediately raises the classification stage's accuracy. By lowering the feature size, the hybrid deep learning technique of CNN and Bi-LSTM extracts key characteristics and increases the accuracy of crime detection. However, due of their increased computational complexity, the current methods are unable to boost the accuracy of crime detection. Additionally, the characteristics extracted by the current approaches are primarily higher dimensionality vector features. This restriction has an impact on the classification's performance and reduces accuracy. Robust strategies are employed in the proposed work to address these constraints. The accuracy of the suggested DAC-BiNet is 98.23%, while that of CNN, Bi-LSTM, DNN, and DBN is 94.508%, 93.724%, and 91.838%, respectively. The specificity performance contrast of suggested and current approaches is shown in Figure 4.5.

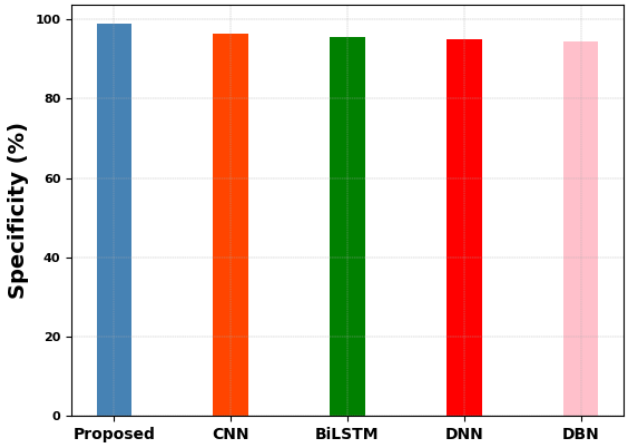

Figure 4.5 Comparison analysis using specificity performance

It appears from the accuracy examination of above figure that suggested method outperforms alternative deep learning methods. The proposed work's reached specificity value is 98.86%, while CNN, BiLSTM, DNN, and DBN have values of 96.53%, 95.03%, 94.46%, and 95.65%, respectively. Several problems cause the specificity of traditional techniques to decrease. The large dimensionality of characteristics in the methods now in use causes a number of issues with crime detection. However, the CNN method's reduced feature extraction size helps to achieve good specificity performance. The recall evaluation between the suggested and several current techniques is shown in Figure 4.6.

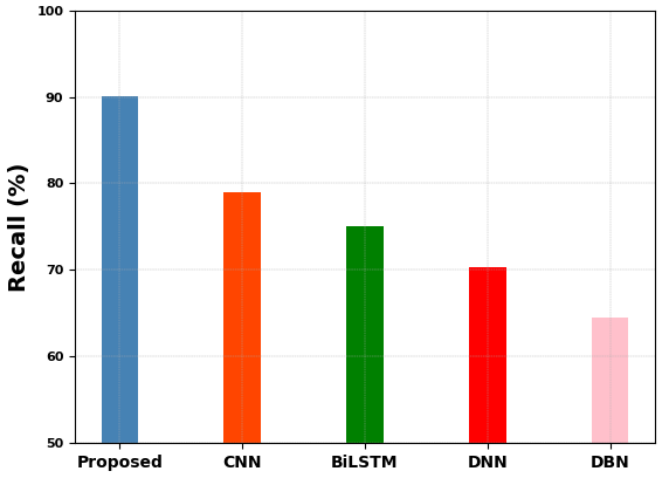

Figure 4.6 Comparison analysis in terms of recall performance

The suggested classification technique's recall performance is contrasted with a number of traditional deep learning methods. According to the analysis outcomes, suggested DAC-BiNet model performs improved than the others in terms of recall. 90.05% is the achieved recall value of the proposed work, compared to 79% for CNN, 75% for BiLSTM, 70.29% for DNN, and 64.55% for DBN. The system recall performance is significant because to the efficient CNN and higher training capabilities of Bi-LSTM. This investigation shows that the suggested methods are more suited for detecting crimes on Twitter involving a variety of classifications. The accuracy performance of the suggested method is shown with several current techniques in Figure 4.7.

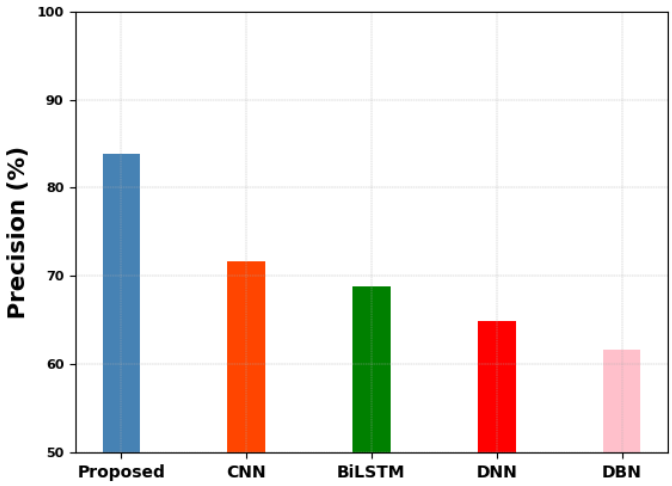

Figure 4.7 Comparison analysis in terms of precision

It is evident that the suggested method performs with more precision than alternative deep learning techniques when comparing its precision performance to that of the others. The precise performance of the suggested criminal detection model has enhanced by 83.86%. However, the suggested method's precision differs slightly from the ones that are already in use. The precision values achieved for CNN, BiLSTM, DNN, and DBN are 71.62%, 68.82%, 64.90%, and 61.70%, respectively. The outcome demonstrates how much more successful the suggested model is than the alternative strategies. The outcome of the contrast in terms of F1-score is shown in Figure 4.8.

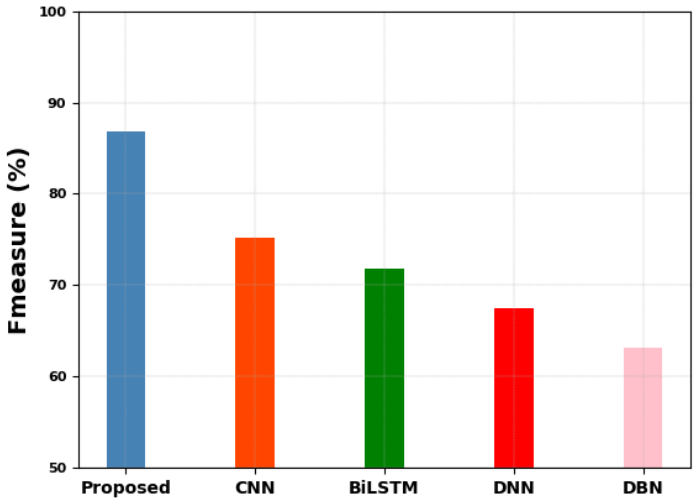

Figure 4.8 Comparison analysis in terms of F1-score

The proposed model's obtained F1-score is contrasted with that of other traditional methods. This investigation seems to show that compared to the other methods, the suggested DAC-BiLSTM classifier receives a higher F-measure value. The suggested F1-score is 86.84%, whereas CNN, BiLSTM, DNN, and DBN are 75.13%, 71.78%, 67.49%, and 63.09%, respectively. This indicates that the suggested methods work better than alternative strategies that are currently in use. Figure 4.9 shows the suggested model's processing time.

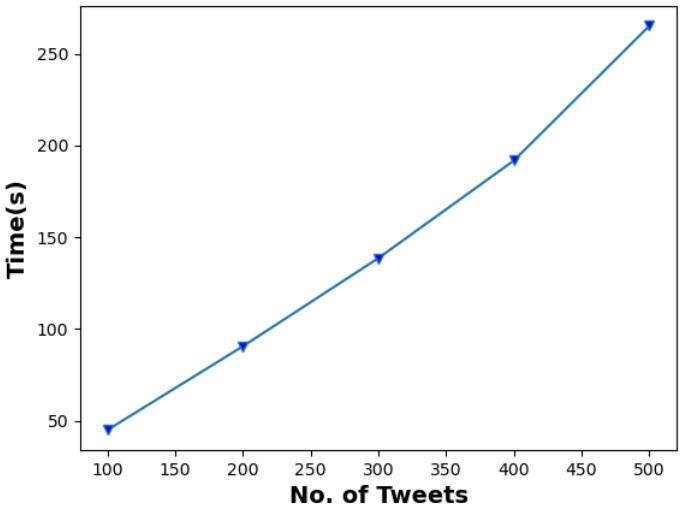

Figure 4.9 Processing time of proposed crime detection model

By changing the quantity of tweets that are supplied, the analysis time of the suggested work is examined. The suggested model achieves a shorter processing time for criminal detection. The clustering technique in the suggested model helps to shorten the system's running time by grouping the feature set with the least amount of size. The system performs better as a result of the shorter processing time, which also shows that the suggested model is suitable for crime detection. Table 4.3 displays the proposed work's total performance.

Table 4.3 Overall performance

Methods	**Recall (%)**	**Precision (%)**	**F1 score (%)**	**Accuracy (%)**	**Specificity (%)**
DBN	64.55	61.70	63.09	91.838	94.46
DNN	70.29	64.90	67.49	92.698	95.03
BiLSTM	75	68.82	71.78	93.724	95.65
CNN	79	71.62	75.13	94.508	96.53
Proposed	90.05	83.86	86.84	98.23	98.86

Figure 4.10 shows the ROC curve of the suggested model using several kinds of twitter data.

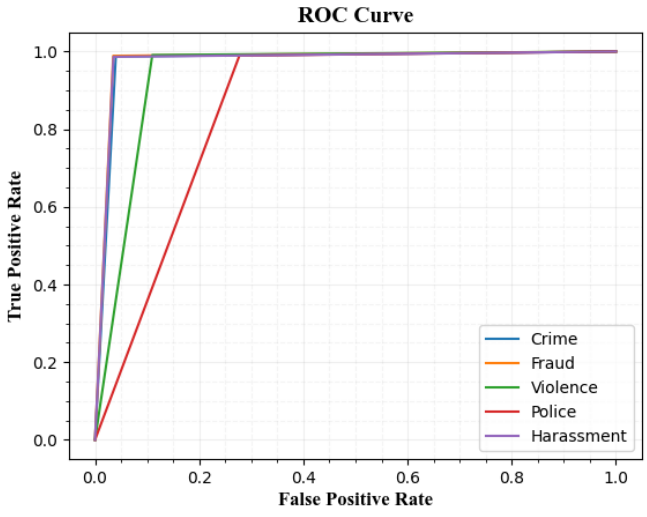

Figure 4.10 ROC curve of proposed model

Through adjustment of the true positive rate and false positive rate, the suggested model's ROC curve is examined. The ROC value for each class is one, and it indicates that the suggested categorization performs better in terms of class-specific crime detection. In the suggested approach, noisy data are first eliminated to produce high-quality data for processing later on. Following that, three distinct methods are used to extract the features. The features that are appropriate for detecting crimes are then moved on to the following step. By extracting features, the system's complexity is reduced. Subsequently, the clustering stage consolidates the characteristics into a reduced set, hence enhancing the classifier's predictive power. Lowering the feature size makes computation much simpler and expedites the time it takes to detect crimes. Due to overfitting and high computational complexity, the current CNN technique is unable to accurately detect crimes. Furthermore, there is a huge feature dimensionality issue with the prior methods. However, the suggested effort improves the results of the criminal detection mechanism by utilizing very effective strategies.

4.4. Conclusion

In order to detect crimes on the Twitter network, this study presents the DAC-BiNet hybrid deep learning technique. First, pre-processing is done on the twitter data from the given dataset in order to reduce the noise that is presented. This is done by segmenting sentences, removing punctuation, tokenizing, removing stop words, fixing acronyms and slang, lemmatizing, lower casing, stemming, and removing hashtags and URLs. Following the conclusion of pre-processing, ITF-IDF, feature hashing, and glove modeling are used in the feature extraction stage to extract the tweet features. To facilitate the detection procedure, the text data are converted into numerical form in this instance. Clustering is carried out using the Possibilistic Fuzzy LDA technique, which reduces the number of clusters from the features in order to improve prediction ability. Subsequently, the proposed DAC-BiNet model is used to detect and classify the crimes, and Aquila optimization is used to reduce the presented loss function. According to the simulation findings, the suggested methods produce better results with respect to accuracy (98.23%), precision (83.86%), recall (90.05%), specificity

(98.86%), and F1-score (86.84%). Future research will introduce several word embedding strategies throughout the feature extraction phase, perhaps yielding more features and better outcomes. In addition, the suggested model is expanded to handle real-time data processing in subsequent studies.

CHAPTER 5

An Analysis of Crime Data under Apache Pig on Big Data

This chapter delves into the comprehensive analysis of crime data by leveraging Apache Pig on big data. It not only elucidates the design framework and implementation process but also conducts a thorough examination of various aspects. Specifically, it scrutinizes the frequency of crimes targeting girls and women over a span of four years, dissects the occurrence of crimes in India categorized by types, and investigates the frequency of crime accusations in specific states, accompanied by the corresponding outcomes. Culminating in a succinct summary, this chapter provides a holistic accepting of the intricate patterns and trends discerned through the analysis of crime data

5.1 Overview

Crime data analysis holds significant importance for national security, as well as for safeguarding the development of children and the socioeconomic status of adults. The National Crime Records Bureau's Crime Statistics offer an overview of both IPC and SLL crimes, categorized by crime head and states/union territories, providing valuable insights into crime patterns and trends. The process of analyzing crime data involves identifying hotspots, observing patterns in crime, and visualizing crime [83]. Decision-makers, the police department, and the government can all utilize the crime data analysis system to find out their actual intentions for reducing crime and to establish security strategies related to it. As a result, the main focus of our work was on large-scale analysis of crime data using Apache Pig and a variety of grunt shell commands in conjunction with the Hadoop distributed file system. Large-scale crime data loading, processing, storing, and analysis can be accelerated using big data analytics techniques. Every industry, including education, retail, social media, banking and securities, transportation, finance, and manufacturing, has seen an increase in data size over the past year, month, or even day. Due to the massive increase in data volume, it is now imperative to examine, research, evaluate, and comprehend proximate trends with other patterns in order to enhance decision-making. Because of this inefficiency, the term BDA was coined. BDA is a laborious process that involves analyzing patterns

and hidden information within large data sets while effectively managing their enormous size, complexity, and high dimensionality [84].

Every industry, including education, retail, social media, banking and securities, transportation, finance, and manufacturing, has seen an increase in data size over the past year, month, or even day. Due to the massive increase in data volume, it is now imperative to examine, research, evaluate, and comprehend proximate trends and other patterns in order to enhance decision-making. Because of this inefficiency, the term BDA was coined. BDA is a laborious process that involves analyzing patterns and hidden information within large data sets while effectively managing their enormous size, complexity, and high dimensionality [85]. Huge in size does not equate to large scale data. Gives a five-vector definition of fattening. Doug Laney explained the three Vs: variety, velocity, and volume. These are the following:

- Volume: Petabytes are used to store massive amounts of data that are larger than terabytes in bulk.
- Velocity: It explains the rate at which data enters and exits devices, including computers, SM, mobile phones, and other devices.
- Variety: describes a variety of data types and sources, including semi-structured, framed, and disordered data.
- Veracity: This pertains to the presence of inconsistent and conflicting data, which can often lead to challenges in managing data quality and accuracy.
- Value: This aspect aligns with the concept of the "5Vs" when a substantial volume of data holds importance for business operations, yet its significance remains untapped until meaningful insights are derived from it.

Big data analysis of large-scale criminal data is conducted using Apache Pig and the Hadoop distributed file system. The Hadoop file system is aimed for efficient utilization of hardware resources, enabling rapid and cost-effective identification of patterns in criminal activities. Large data sets can be processed quickly and easily using Hadoop's versatility. It is made up of several parts, such as the Map Reduce and Hadoop Distributed File System, which are utilized to store, process, and analyze data more effectively [86]. Crime has an impact on adult socioeconomic status, public safety, and children's development. In an effort to decrease crime and progress the

excellence of life for civilians, government officials and policy makers must be astute about the elements that influence the crime rate. In this work, analyze large data pertaining to crime rates and offenses and use Apache Pig and Hadoop to become acquainted with the social problem of crime. Specifically, to involve the selective finding of vicious crime occurrences that are similar to earlier incidents, utilizing incident-level crime data.

5.2 Proposed Work

Big data is a new weapon in the fight against crime, which has been on the rise as the population grows. Therefore, crime analysis serves a legislative purpose by effectively identifying disorder, patterns, and trends in crime. The population is growing, and with it are crimes and crime rates. These crimes include robberies, rape attempts, theft, extortion, and dowry killings. As a result, crime surveys are becoming more and more difficult for the government to interpret and make decisions about maintaining law and order. Crime surveys are crucial to protecting the citizens of our nation from these kinds of crimes. The best method to enhance this is to apply BDA to the raw, disordered data that is generated every day from a variety of sources, including social media, the manufacturing industry, the education sector, and so on. Big Data Analytics (BDA) encompasses tools utilized to extract actionable and meaningful insights from raw data. These instruments contribute to decision support systems, empowering governmental bodies such as councils and judiciaries to make informed and impactful decisions aimed at reducing crime.

Figure 5.1 illustrates the key stages of the large-scale crime data analysis methodology. This framework initiates with data loading utilizing the Sqoop Hadoop copy command, followed by data storage in the Apache Pig grunt shell using Hadoop's distributed file system. Subsequently, the data undergoes pre-processing before being analyzed using various graphs generated within Apache Pig.

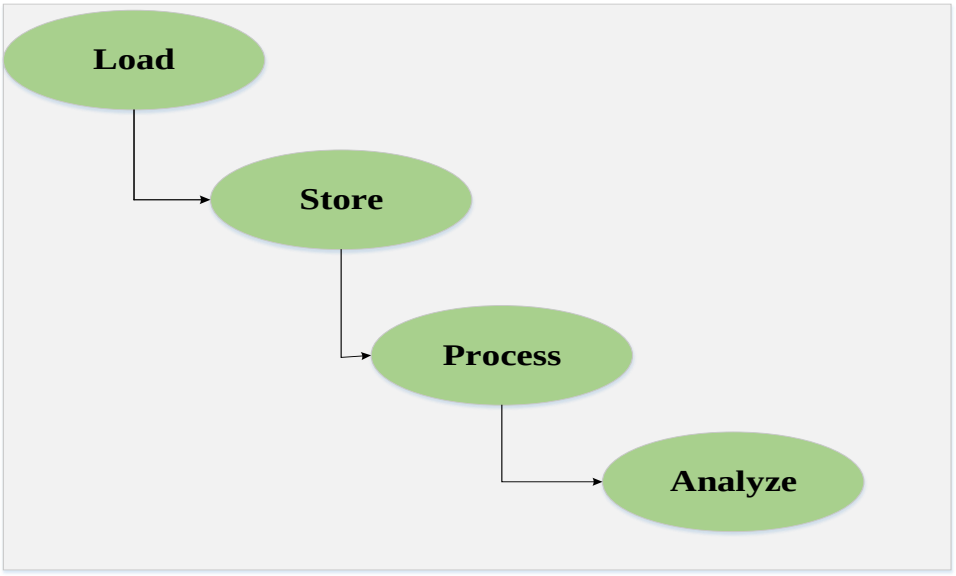

Figure 5.1: Designed Framework for Analysis

In figure 3.1 data is loaded with the use of Sqoop and the Hadoop copy command. DATA is loaded using the LOAD command to the Apache Pig storage. And analyzed by utilizing Apache Pig.

LOADING: Three types of Apache: Flume, Kafka, and Pig has been developed to manage a variety of data formats, including disorderly, semi-framed, and framed data. These connections are all compatible with HDFS, a distributed file system for Hadoop, and RDBMS. All of these transfer data to HDFS. These are all useful for storing semi-structured data that is a mess.

STORING: Map reduce has completed the job reduce task with the use of intermediary key-value pairs.

$Map(T_1, u_1) \rightarrow list(T_2, u_2)$ Process the key/value that serves as a reduction function step by step. $\text{Re}\,duce(T_2, list(u_2)) \rightarrow list(u_3)$ Cut down on the steps takes the map reduce step's output and gathers the output's outline. The output is ordered in each aspect.

ANALYZE: The method that collected data is processed determines the usefulness of the information that is obtained, which is crucial for the government to make informed decisions. The final result, displayed in the paper under Apache Pig, was obtained by storing, loading, and preparing the data using the Hadoop distributed file system using a variety of commands in the Grunt shell.

5.2.1 Frequency of crimes against girls and women in four years (2016-2019)

During the four-year period from 2016 to 2019, the incidence of crimes against girls as well as women provides important insights on the dynamics of gender-based violence during this era. Examining crime data from this time frame can give a thorough grasp of the tendencies, patterns, and variances in incidences that impact women and girls. Analyzing shifts in the number of crimes in various categories such as trafficking, harassment, sexual assault, and domestic abuse may provide insight into how successful social and legal interventions are [87]. In addition, policymakers, law enforcement, and advocacy groups can benefit from knowing about any noteworthy variations or recurring trends over the designated years. This information can help them create focused strategies and preventive measures that will address and lessen crimes against girls and women in the future. The foundation for creating a more secure and safe environment for women and girls in society is provided by this analysis.

A glimpse of the trends and patterns in gender-based offenses during this time period can be obtained from the frequency of crimes against females over the four-year period from 2016 to 2019. A thorough grasp of the variations and possible shifts in the occurrence of such crimes can be obtained by analyzing the data for each year. These patterns may be influenced by elements including shifts in legislative frameworks, law enforcement tactics, and public awareness. Analyzing this data could provide light on the efficacy of programs and laws meant to safeguard and advance women's and girls' safety. Taking into account the distinct categories of criminal activity, geographical variances, and any developing trends is crucial when formulating focused approaches for both intervention and mitigation. The data obtained from examining the incidence of crimes against women and girls within the designated timeframe is beneficial for legislators, law enforcement organizations, and advocacy groups who strive to promote a more secure atmosphere for this population.

5.2.2 Frequency of crimes accusing in particular states

States differ greatly in how frequently crimes occur, depending on a wide range of criteria including socioeconomic status and law enforcement tactics. Certain states have greater rates of crime because of things like unemployment, poverty, and limited access to high-quality healthcare and education. These socioeconomic problems may

create an atmosphere that is more conducive to criminal activity. Additionally, because urban environments can present a greater concentration of possible targets and offenders, states with densely populated urban areas may have higher crime rates than those with more rural populations. Policies and practices related to law enforcement are also very important in determining crime rates. Since their proactive efforts can dissuade criminal activity and result in quicker reactions to occurrences, states with strong and efficient law enforcement agencies may have lower crime rates. On the other hand, states with underfunded or ineffective law enforcement may find it difficult to deal with criminal matters quickly, which could lead to a persistence or escalation of crime. The legal system and legislation may also have an effect on crime rates. States with strong legal systems may have reduced crime rates because prospective perpetrators are scared of the repercussions. Conversely, states with lax or unclear laws may unintentionally promote criminal activity. Moreover, state-specific crime rates are influenced by cultural and demographic variables. States with diverse populations could encounter particular difficulties because of cultural differences, whereas those with strong social cohesion might have lower crime rates. To sum up, the incidence of criminal activity in specific states is a multifaceted phenomenon that is shaped by a range of elements including legislative frameworks, socioeconomic situations, law enforcement strategies, and cultural influences. For states to effectively reduce crime and improve public safety, it is imperative to comprehend and handle these diverse factors.

A proactive approach is desperately needed to confront and reduce the rising crime rates, as seen by the increasing number of accusations in certain states. The steady rise in charges within specific states indicates an expanding issue that needs to be addressed right away. It is necessary to increase police presence and put in place extra surveillance mechanisms in these states in response to this worrying trend. Authorities are able to rapidly respond to new threats, effectively restrict criminal activity, and dissuade potential offenders by utilizing advanced surveillance technologies and augmenting law enforcement resources. Public safety and crime prevention can be greatly enhanced by providing more resources for law enforcement officers, as well as by strategically placing them and increasing their visibility in high-crime areas. Furthermore, in order to develop a complete plan for addressing the underlying causes

of criminal behavior and promoting a safe environment for the citizens of these states, a comprehensive approach that incorporates community participation, crime prevention initiatives, and focused law enforcement activities is necessary. For the purpose of creating and executing long-term solutions that address current issues while simultaneously promoting long-term crime reduction and the welfare of the impacted areas, law enforcement agencies, local communities, and legislators must collaborate.

5.2.3 Frequency of crimes in India by their Types

Like many other nations, India faces a wide variety of criminal activities, each of which poses unique difficulties for the criminal justice system and the general public [88]. The total crime rate is largely influenced by violent crimes, which include rape, kidnapping and abduction, murder, dowry deaths, and other crimes. There are regional variations in the incidence of these crimes, which are impacted by many circumstances. Murder is a horrible crime that takes a person's life. It happens more frequently in some places than others, and it's frequently impacted by cultural norms, socioeconomic differences, and the efficiency of the legal system. In some areas where dowry customs are still widespread, there are alarming cases of gender-based violence known as "dowry deaths." The frequency of kidnappings and abductions, whether for ransom or other reasons, can vary depending on elements including the state of the economy and the existence of criminal networks. These crimes represent a major threat to public safety. The crime of rape is a matter of great concern as it not only causes serious physical and emotional harm to its victims but also reflects cultural attitudes about gender. The prevalence of rape incidents in India has drawn attention from both domestic and foreign audiences, igniting debates on the necessity of legislative changes, more law enforcement, and increased public awareness in order to solve this widespread problem.

A multifaceted approach is necessary to confront these violent crimes thoroughly and limit their recurrence. This entails enhancing the capacity of law enforcement, guaranteeing prompt and equitable judicial proceedings, advocating for gender parity, and cultivating a cultural transition towards a climate of safety and respect. Campaigns for awareness and educational programs can also be quite effective in upending deeply rooted social norms that support these crimes. For all Indian residents to live in a better

and more secure environment, it is imperative that targeted measures be implemented in recognition of the interdependence of these challenges.

5.2.4 Data set Description

Data in a CSV file format. The file is made up of numerous lines with data records, each with multiple fields divided by commas. Data sets included numerous fields in the file, such as

- Victims under various crime categories;
- Crime incidence and rate
- How crimes against women and girls are handled by the police and courts

Tabular data is often exchanged and stored in a structured data format called CSV. This file contains multiple lines that each represent a data record in the context of crime data, with fields separated by commas. Every line in the CSV file represents a distinct item in the dataset, offering a methodical and well-structured approach to storing data pertaining to different facets of criminal activity. The dataset covers a wide range of fields that provide information on many aspects of crime. A quantitative summary of the prevalence of different criminal activities can be found in the "Incidence & Crime rate of the Crime" field, which probably contains data on the frequency of particular crimes and the accompanying crime rates. Understanding the overall state of crime and seeing long-term patterns or trends requires this data. Details regarding the victims impacted by various crime kinds are probably available in the "Victims under Crime types" box. These details might include their demographics, the kind of victimization they endured, and the particular crimes they were victims of. Understanding the effects of different crimes on various population segments is made easier by analyzing this data. Information regarding how the legal system handles crimes against women and girls is probably contained in the "Police and court Disposal of Crime against Women and Girls" area. This could contain information about court cases, police investigations, and case outcomes. Comprehending the disposal procedure is crucial in assessing the efficacy of law enforcement and judicial systems with respect to tackling crimes that are targeted specifically towards women.

In general, the CSV file functions as an organized database of crime-related information, allowing scholars, analysts, and decision-makers to carry out

comprehensive examinations and make knowledgeable judgments regarding the frequency, targets, and legal ramifications of distinct offenses. The data can be easily parsed and manipulated using spreadsheet applications or programming languages thanks to the comma-separated structure, which makes data exploration and interpretation more productive.

5.3 IMPLEMENTATION

Grunt Shell

An essential part of the Hadoop ecosystem, the Grunt Shell is particularly linked to Apache Pig, a high-level platform designed to analyze massive amounts of data on top of Hadoop. Grunt Shell is a command-line interface that allows users to enter Pig Latin scripts to be translated into a sequence of MapReduce tasks that are executed on a Hadoop cluster. With the help of this shell, developers may work in an interactive, iterative environment and input Pig Latin instructions in real time. Pig Latin's ease of use abstracts away the difficulties involved in constructing low-level MapReduce applications, facilitating the expression of complicated data transformations by engineers and data analysts. Pig scripts can be executed with the help of the Grunt Shell, which also offers an environment for debugging, optimization, and data analysis. Users have the ability to load datasets, filter entries, do aggregations, and engage in interactive data structure analysis. Because it allows users to repeatedly tweak and debug their scripts before deploying them on large datasets, the interactive feature of the Grunt Shell is especially beneficial during the development and testing phases. All things considered, Apache Pig and the Grunt Shell in Hadoop work hand in hand to simplify and improve the creation and implementation of data processing workflows on distributed computing platforms.

Apache Pig storage

Big Data is represented by pigs as data flows. Pig is an advanced platform or instrument that is employed in the processing of big datasets. Compared to MapReduce, Pig offers a higher level of abstraction for processing data. It provides Pig Latin, a high-level scripting language used to develop data analysis code. Programmers utilize Pig Latin to create scripts for handling data stored in HDFS.

These scripts are internally translated into specific map and lessen tasks by Apache Pig's Pig Engine. However, these lower-level tasks are abstracted from view to provide programmers with a high level of abstraction

Features of Apache pig

- Rich sets of operators, like as filtering, sorting, joining, aggregating, and others, are delivered by Apache Pig for executing various operations.
- Simple to read, write, and learn. Particularly for SQL programmers, Apache Pig is quite helpful.
- In Apache Pig, joining operations are simple.
- Users can take advantage of the capabilities of other Apache Hadoop ecosystem components, like Apache Hive, Apache Spark, and Apache ZooKeeper, while transforming data with Apache Pig's integrations.
- This data structure is richer, hierarchical, and multivalued.
- Pig has the ability to analyze data that is both structured and unstructured.

Applications of Apache pig

- Pig Scripting facilitates the exploration of large datasets by providing support for ad hoc queries across extensive data collections.
- When huge data sets processing techniques are being prototyped necessary to handle time-sensitive data loads.
- For gathering a significant number of datasets using site crawls and search logs.
- Used when sampling is required to provide analytical insights.

5.3.1 Frequency of Crimes against Girls and Women in Four Years (2016-2019)

Table 5.1: Pseudo code for Frequency of crimes against Girls and Women

Pseudo code: Hand-me-down in Grunt Shell Provided Input: Data of Crime Provided Output: Frequency of crimes against Girls and Women

1) Type the command into the PIG Grunt Shell

2) L: DATA is loaded onto the Apache Pig storage utilizing LOAD command

3) M: For every L beget the crime by year and find the figure of Girls and women victim

4) N: Merge by year

5) Data: For every N perform merge, Summate (L.Year);

6) Examine the outcome

The following pseudo code describes how to use Apache Pig in a Grunt Shell environment to analyze the incidence of crimes against women and girls. The procedures include loading crime data, organizing it according to victim gender and year, integrating the data, totaling incidents by year, and analyzing the result. The first step of the pseudo code is to enter the required instructions into the PIG Grunt Shell, which is a command-line interface for working with Apache Pig. Using the LOAD command, the input data which contains information on crimes is loaded into the Apache Pig storage (Step 2). The code seeks to classify the incidents by year and count the proportion of female victims for each line of the loaded data (Step 3). This stage probably entails combining the counts by year and filtering the data according to the victim's gender. Step 4 of the code involves merging the data based on the shared 'year' characteristic after grouping and counting the data. Consolidating the data and preparing it for additional analysis require this step. The code seeks to classify the incidents by year and count the proportion of female victims for each line of the loaded data (Step 3). This stage probably entails combining the counts by year and filtering the data according to the victim's gender. Step 4 of the code involves merging the data based on the shared 'year' characteristic after grouping and counting the data. Consolidating the data and preparing it for additional analysis require this step.

The percentage change of several categories of crime against women and girls per crime head from 2016 to 2019 was shown in Figure 5.2. Pig Latin, the assistance language used in Apache Pig, is used for all of these tasks. Pig Grunt Shell select, reconstruct, and store commands.

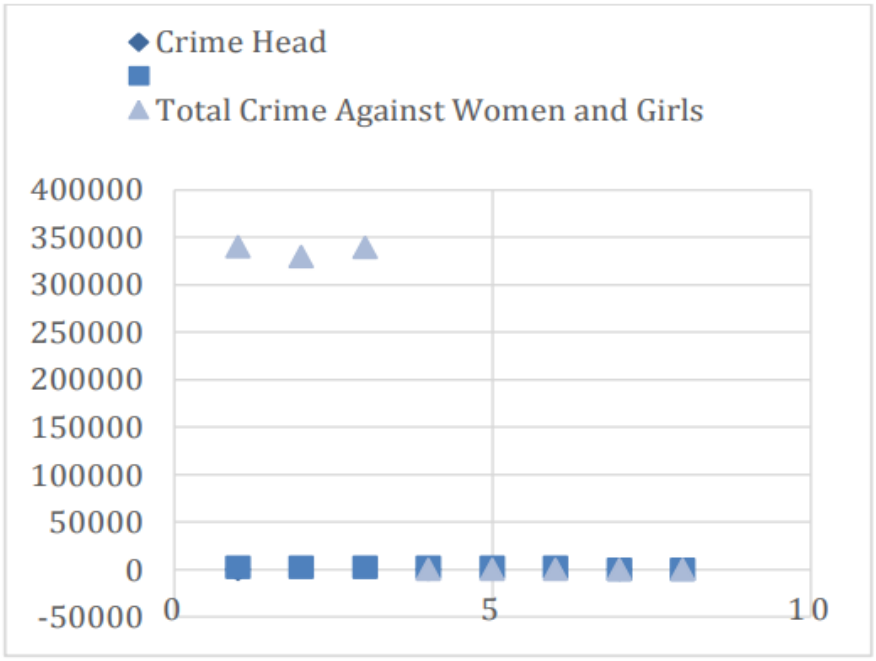

Figure 5.2: percent variation of Crime against Women and Girls

Figure 5.3 showed a decline in the number of crimes against women and girls that were recorded between 2016 and 2019. The number was 11,295 when it was first recorded in 2016. Since then, it has been declining. However, enact a number of legislations to safeguard women's and girls' extremely precarious situations. This government needs to improve by taking significant action to end crime against women and girls.

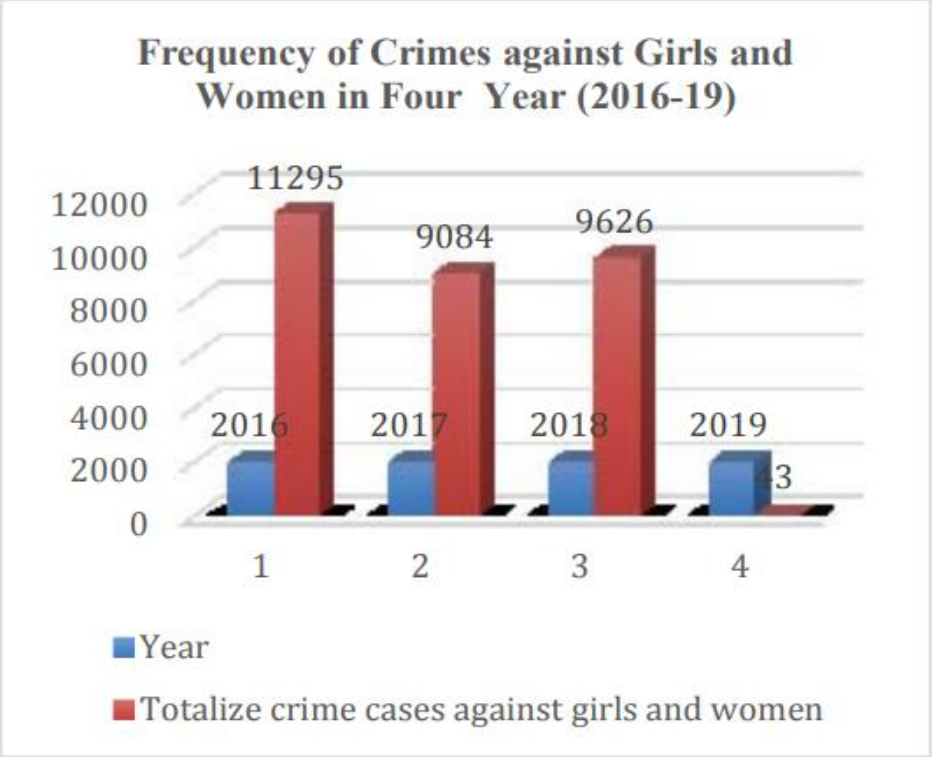

Figure 5.3: Frequency of Crimes against Girls and Women in Four Years (2016-2019)

5.3.2 Frequency of crimes in India by their Types

Algorithm 5.2: Pseudo code for the Frequency of crimes in India by their Types

India's crime frequency by category Pseudo code: Hand-me-down in Grunt Shell Provided Input: Data of Crime Provided Output: India's crime rate broken down by category 1) Type the command into the PIG Grunt Shell

2) L: DATA is loaded onto the Apache Pig storage utilizing

LOAD knowledge

3) M: For every L beget the crime type by corruption head

4) N: Merge by crime head

5) Data: For every N perform Combine, Summate (Year)

6) Examine the outcome

The provided pseudo code describes a basic procedure that uses Apache Pig in a Grunt Shell environment to analyze the frequency of crimes by category in India. The process entails loading crime data, classifying it according to crime type, integrating the data, totaling the incidents by year, and then assessing the result. The provided pseudo code describes a basic procedure that uses Apache Pig in a Grunt Shell environment to analyze the frequency of crimes by category in India. The process entails loading crime data, classifying it according to crime type, integrating the data, totaling the incidents by year, and then assessing the result. The provided pseudo code describes a basic procedure that uses Apache Pig in a Grunt Shell environment to analyze the frequency of crimes by category in India. The process entails loading crime data, classifying it according to crime type, integrating the data, totaling the incidents by year, and then assessing the result. The code uses the 'crime head' as a key to extract the crime type information for each line of the imported data (Step 3). These crime categories are then combined according to the shared "crime head" (Step 4). Consolidating the data and getting it ready for additional analysis depend on this stage. After merging, the algorithm executes a data operation (Step 5) that combines and totalizes the number of crimes that have occurred throughout time. This stage probably compiles the information to give a thorough picture of how frequently each kind of crime occurs over time. Analysis of the result is the last step. This could entail producing data, charts, or any other kind of study to learn more about the trends and patterns pertaining to various crime categories in India. Grunt Shell and Apache Pig together offer a

parallelized and scalable data processing method that highlights the effectiveness of managing big datasets. This pseudo code describes a set of instructions for processing and analyzing crime data in India with an emphasis on the frequency of crimes broken down into several categories. To obtain valuable insights on the crime scene in the area, a series of steps including loading, grouping, merging, and aggregating data are required. The IPC and SLL crime incidence per crime head were combined to create Figure 5.4, which showed the percentage difference of crime in India.

Figure 5.4: Crime in India comprising IPC (Indian Panel Code) and SLL (Special and local Laws)

Figure 5.5 showed how, as the population grows, so do crimes and crime rates. These crimes include robberies, rape attempts, theft, extortion, dacoity, and other crimes. Consequently, the government finds it difficult to make important decisions about maintaining law and order as a result of crime surveying.

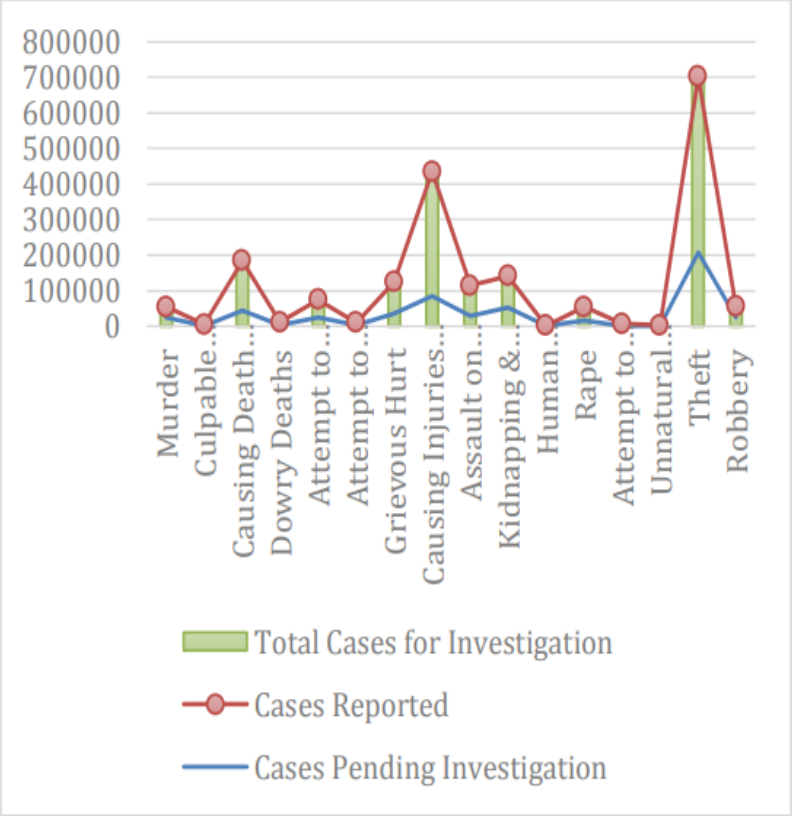

Figure 5.5: Frequency of Crime in India by their type

5.3.3 Frequency of crime accusing in particular states

Algorithm 5.3: Pseudo code for crimes accusing in each state

Pseudo code: Hand-me-down in Grunt Shell Provided Input: Files of Crime Provided Output: No. of corruptions accusing in each state 1) Type the command into the PIG Grunt Shell 2) L: DATA is loaded onto the Apache Pig storage utilizing LOAD command

3) M: For every L beget the crime by state

4) N: Merge by STATE

5) Data: For every Z perform merge, Summate (X. Year);

6) Examine the outcome

The pseudo code starts by entering commands into the PIG Grunt Shell, which is an interface that runs commands on the command line to work with Apache Pig. Using the LOAD command, the criminal data is overloaded into the Apache Pig storage (Step 2). The code attempts to classify the crimes according to the state in which they happened, for each line of the loaded data (Step 3). This technique of grouping makes sure that the analysis that follows will concentrate on how crimes are distributed throughout the various states. The method then merges the data based on the shared 'STATE' characteristic after grouping the data state-wise (Step 4). Consolidating the data and getting it ready for additional analysis depend on this stage. The algorithm next executes a data operation (Step 5), which combines and totalizes the number of crimes that have occurred in each state throughout time. This probably compiles the information to give a thorough picture of the total number of charges in each state. Analyzing the result is the last step (Step 6). In order to obtain insights into the patterns and trends pertaining to the distribution of crimes throughout the many states, this could involve producing statistics, visualizations, or any other type of analysis. The combination of Apache Pig with the Grunt Shell highlights the effectiveness of managing big datasets and proposes a scalable and parallelized data processing method. By choosing, recreating, and storage the appreciation into Pig Grunt Shell, Figure 5.6 showed the percentage variance of offenses accused in each state.

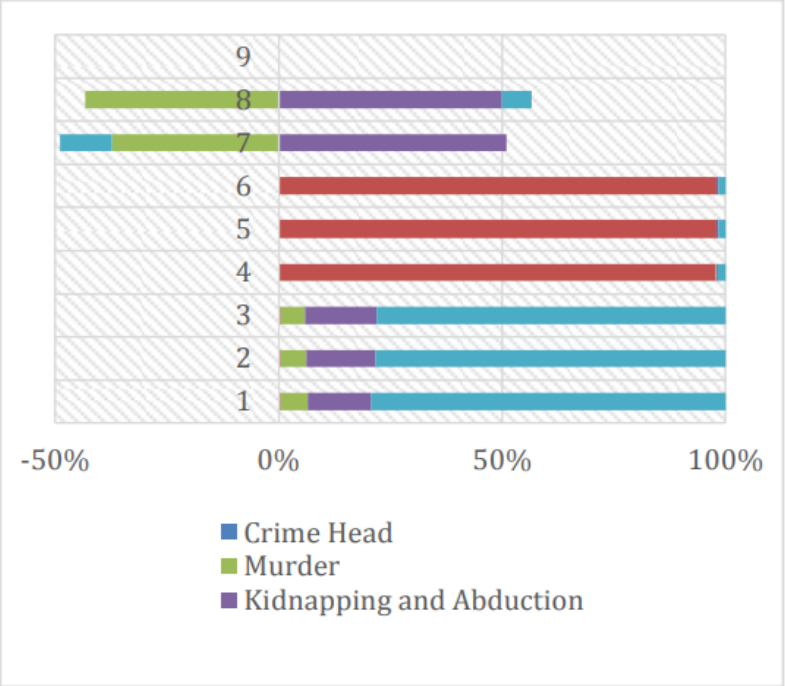

Figure 5.6: Beget the crime type by crime head (Murder Kidnapping and Abduction)

Figure 5.7 illustrates crime-related incidents in India, depicting various segments of crime scenes or incidents as percentages. It highlights sectors more susceptible to crime, including those vulnerable to psychic crimes, and presents the percentage distribution of each type of crime incident within these sectors. Initially, the figure contained a diverse range of characteristics associated with the crime scene. However, we were able to streamline this information by utilizing the select, rebuild, and store commands within the Pig Grunt Shell under Apache Pig with Hadoop.

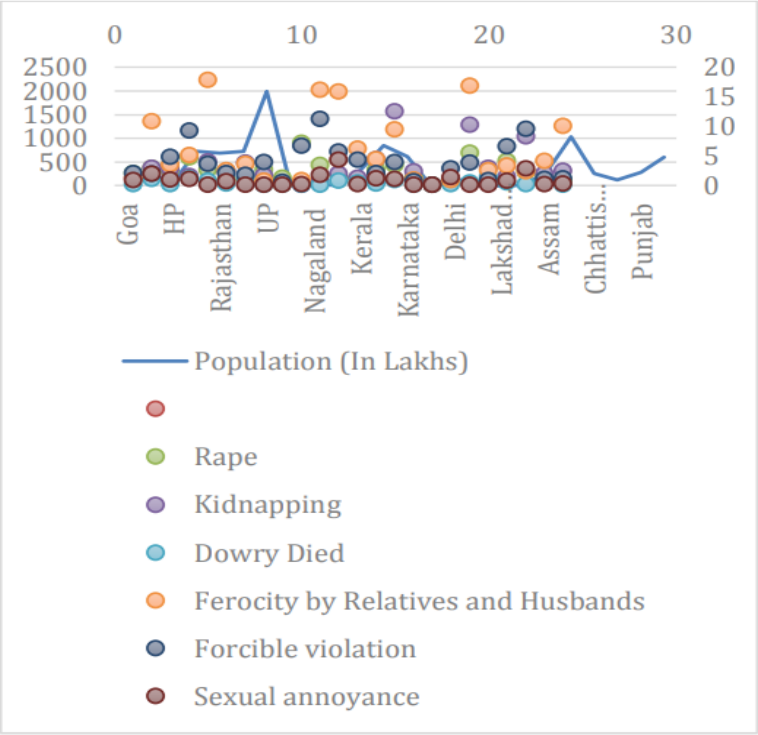

Figure 5.7: Frequency of crimes accusing in particular states

Figures 5.6 and 5.7 show that the number of accusations from each state has been rising daily. Therefore, in order to reduce crime, there has to be more police presence and monitoring in these states.

5.4 Summary

This paper provides a comprehensive analysis of crime-related incidents in India, presenting statistics on various types of crime scenes and incidents along with their respective percentages. It highlights the sectors that are more susceptible to criminal activities, including psychic crimes, and delineates the percentage distribution of different crime types within each sector. With the aim of aiding strategic decision-making by law enforcement agencies and assisting the government in evaluating existing crime reduction measures, the paper also presents data on the percentage variation of accused crimes across different states, the frequency of various crime types, and incidents of crimes against women and girls. Through this information, stakeholders can gain valuable insights to enhance crime prevention strategies and improve overall security measures

CHAPTER 6

Comparative Analysis

This chapter delves into a comprehensive exploration of predictive analytics techniques applied to crime data extracted from online social media, conducting a meticulous comparative analysis of various methodologies. The discourse encompasses a detailed presentation of results and an in-depth discussion of the effectiveness and limitations associated with each technique employed. Culminating in a concise summary, the chapter not only provides valuable insights into the nuanced landscape of predictive analytics for online social media crime data but also offers a discerning overview of the implications and potential advancements in this burgeoning field.

6.1 Overview

The internet had expanded by the 20th century's end. tremendously and had profoundly altered our social and economic lives [89]. This change has had a substantial impact on the development of online social networks (OSNs). Because of the wonderful means of communication that the internet offers. These communication alternatives significantly boosted the effectiveness of the OSN, a facet of the digital revolution, and gave rise to it [90]. Even though different academics have varied definitions of what an OSN is, [91] characterized it as an online community made up of people who like the same activities, have similar friends, and share common interests. As of March 2019 [78], there were more than 2.38 billion active Facebook users monthly. The social microblogging site Twitter is used by 330 million individuals per month [92]. These websites serve as new communication channels and sources of dynamic, real-time data for millions of users. enabling them to engage with each other and create unique profiles without any restrictions. By March 2019, there were more than 2.38 billion

active Facebook users each month, according to reports [91]. Every month, 330 million users of the social microblogging site Twitter [93] log on. These websites offer real-time dynamic data sources and new communication routes to millions of users. enabling individuals to engage with each other and create personal online profiles regardless of geographical boundaries or other physical constraints [94]. For the creation of communities and networks of friends that are today enormous in scope and scale, as well as for the study of these networks, interpersonal data from OSNs may offer fresh perspectives and opportunities [95]. All 34 countries have a sharp increase in criminal activity every 34 years [96]. Tough action is required to put an end to these illicit activities. Rate of crime tracking is essential to monitoring these illegal activities and improving public safety. The criteria or constraints on crime used in this work pertain to incident-level crime data, which is kept up to date as a crime dataset and comprises the type of crime, the criminal's ID, the date and place of the occurrence, and the location [97]. Crime statistics can be greatly lowered because social media is useful for measuring crime rates in various nations and locations. Social media serves as both a communication tool and an information source. With over 300 million users, Twitter emerges as a viable option for data analysis. Users of this social networking platform share their ideas, emotions, and even rage. It is difficult to obtain data from Twitter for crime detection because Tweets are user-initiated. Tweets can be in many different formats, including symbols. These issues mean that every Tweet needs to be properly thought through. The study makes mention of the application of text-based data science to the gathering and display of information from several news sources. Researchers have created a number of models utilizing deep learning and machine learning to forecast crime data. Machine learning models were formerly employed to forecast crime data, however in the recent past, deep learning models have gained significant momentum as their performance has surpassed that of machine learning models. In contrast, deep learning models take less time, data, and resources to forecast crime than machine learning algorithms, which also demand more data and are less accurate. Comparing crime data prediction models is prompted by the fact that numerous studies fail to yield the distinctive impact. To forecast crime statistics on social media, one needs datasets, and each dataset contains copious amounts of data about crimes that have occurred in various places and at various periods. Thus, a large number of datasets that are frequently used to forecast crime rates are evaluated.

6.2 BENEFITS OF ONLINE SOCIAL MEDIA

Social media networks make it easier for people to share and discuss material online. Users are able to interact with one other and exchange content by having the option to post text, images, and videos on various social media platforms and to like, share, and comment on each other's posts. Posts and profile images can be made public or private, depending on the platform and the user's choices. Since many social media platforms provide geotagging of postings, social media data can be compared to other forms of geographic data. Social media has both positive and negative effects on day-to-day living. Positively, social media platforms such as Facebook, Instagram, Snapchat, and Twitter offer chat rooms that let users stay in contact with friends and family that live far away. video and audio chats, and a range of other services. The general public may now keep up with events taking place worldwide using only their mobile devices. People use social media platforms to make it easier for them to do so whether looking for work or for any other purpose, like education. Since social media lacks privacy, there's a good potential that someone will use your personal information against you. In today's world, privacy of the individual is a key worry. Cyberbullying and cyber thievery are both based on a third party's illegal access to another person's personal information. The abundance of offensive stuff on social media puts people at risk. They spend the entire day chatting on the internet. Disinformation could therefore be used for a number of objectives, including deceiving individuals, encouraging hate crimes online, or stoking racial or religious hostility. Consequently, social media has become an indispensable part of our daily existence. As a result, the problems that virtual space raises with regard to human rights and accountability become largely unknown.

6.3 PREDICTION PERFORMANCE COMPARISON OF MACHINE LEARNING METHODS USED FOR ANALYSING CRIMES IN SOCIAL MEDIA

Comparing the prediction performance of machine learning techniques used to analyze crimes on social media entails a methodical assessment of different algorithms to determine how well they perform in tackling the intricate problems associated with online crime detection. A clear issue statement outlining the nature of crime analysis—

whether it be identifying possible criminal activity, predicting crime types, or assessing public opinion in relation to criminal incidents is established early in the process. The assessment is based on a meticulously selected dataset that was taken from social media sites and includes relevant crime-related data. A comprehensive preparation is performed on this dataset, which includes categorical variable encoding and handling of missing values. Subsequently, the text data is subjected to feature engineering techniques in order to extract relevant information. This includes chronological context, user profiles, emotion ratings, and keywords. For training and evaluation, a wide range of machine learning models—from cutting-edge methods like neural networks to more conventional algorithms like NB and SVM—are chosen. To measure the prediction performance of the models, certain criteria are applied, including accuracy, precision, recall, F1-score, and area under the ROC curve. with order to ensure a complete understanding of the benefits and drawbacks of each model with respect to the crimes investigated in social media, this comparison analysis also considers bias, fairness, interpretability, and ethical implications. The results are methodically documented, offering insightful information on the effectiveness and applicability of several machine learning techniques for this crucial application domain.

ML generally offers three main approaches to learning: supervised and unsupervised UL and SL, and RL. Supervised learning is advantageous when a label property is given for a particular dataset. Decision trees, regressions, random forests, logistic regressions, KNNs, and other algorithms are a few instances of SL algorithms. The UL can be useful when it's challenging to identify latent correlations in a particular unlabeled dataset. UL is exemplified by clustering techniques such as Kmeans and Apriori algorithms. While it does not provide a label or error message, RL falls somewhere between supervised and unsupervised machine learning. It offers open feedback for every action or prediction made. Genuine input/output pairings are no longer rewarded by RL, which also alters suboptimal operations directly. The Markov Decision Process algorithm respects the Reinforcement technique.

Complex computing systems known as machine learning models are created to identify patterns, gain knowledge from data, and generate predictions or judgments without the need for explicit programming. These artificial intelligence-based models

make use of techniques that allow them to extrapolate knowledge from particular instances, making it easier to pull insightful information from large, intricate datasets. Machine learning models aim to enhance their performance over time by continuously improving their comprehension of patterns by being exposed to additional data. In order to enable models to make predictions on fresh, unseen data, supervised learning entails training them on labelled datasets where inputs are associated with matching outputs. On the other hand, unsupervised learning reveals latent structures in the data by identifying patterns and relationships from unlabeled data. The main goal of reinforcement learning is to teach models how to make decisions in a certain order by giving them feedback in the form of incentives or penalties. Neural networks, decision trees, support vector machines, ensemble approaches, and other techniques are only a few of the many methods that make up machine learning models. Each has advantages and disadvantages. Phases of pre-processing data, feature engineering, training, validation, and testing are all involved in the development and implementation of these models. Machine learning approaches are becoming more and more applicable in a variety of fields, from banking and healthcare to natural language processing and autonomous cars, thanks to advancements in hardware and algorithms. In the field of machine learning, ethical concerns about bias, interpretability, and accountability highlight the significance of responsible development and deployment processes.

Below are several graphics that show recall, accuracy, precision, and F1-score:

Accuracy is defined as the ratio of correctly predicted occurrence to all incidences in the dataset. It provides a broad indication of the model's performance in all classes.

$$Accuracy = \frac{No.of\ correct\ prediction}{Total\ no.of\ predictions} \tag{6.1}$$

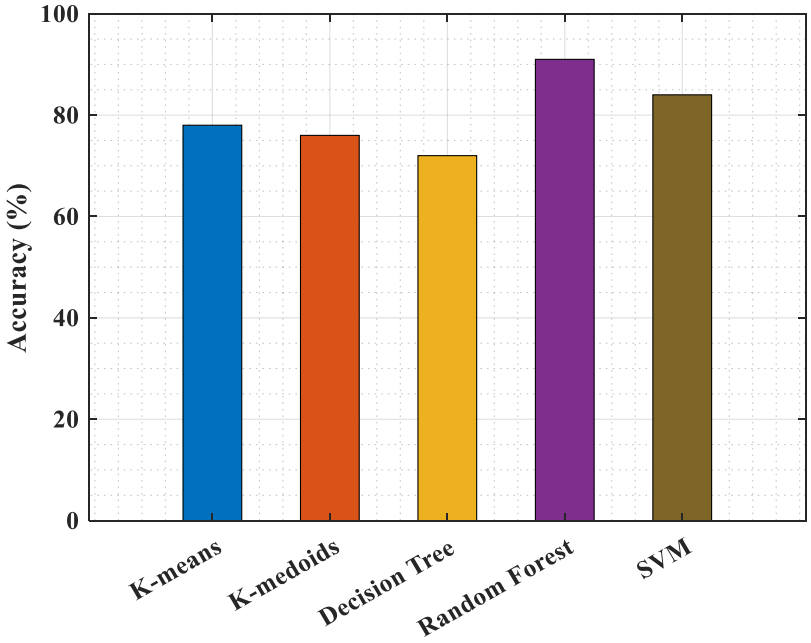

Figure 6.1: Accuracy of the various ML models

The accuracy of five distinct machine learning models K-means, K-medoids, Decision Tree, Random Forest, and SVM is displayed in the Figure 6.1. The different models are listed on the x-axis, and the accuracy is shown as a percentage on the y-axis. The model is more accurate the higher the point on the graph. The graph indicates that Random Forest is the most accurate model, with SVM, Decision Tree, K-medoids, and K-means following in order of accuracy.

Precision gauges how well the model predicts the good outcomes.

$$\text{Precision} = \frac{True\ positive}{True\ positive + False\ positive} \tag{6.2}$$

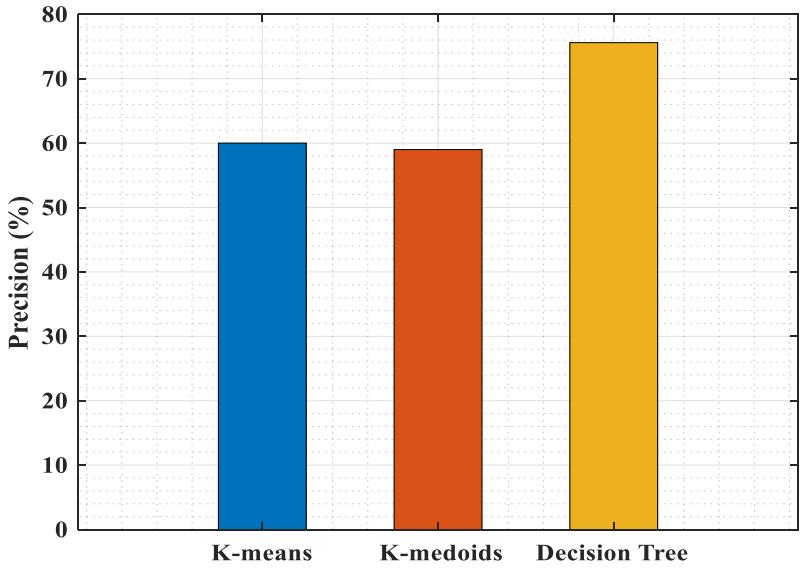

Figure 6.2: precision of various ML model

A graph comparing the accuracy of five distinct machine learning models K-means, K-medoids, Decision Tree, Random Forest, and SVM. The graph's x-axis lists the various machine learning models, while the y-axis is labeled "Precision (%)". A colored bar graph is used to depict each model. The K-means model's precision is represented by the blue bar, the K-medoids model's precision by the red bar, the DT model's precision by the green bar, the RF model's precision by the purple bar, and the SVM model's precision by the orange bar. The height of each model's matching bar indicates how accurate it is. The model is more accurate the higher the bar. The Random Forest model seems to be the most accurate, followed by the SVM, Decision Tree, K-medoids, and K-means models, according to the graph. The precision of the Random Forest model is roughly 79% that of the SVM model is roughly 75% that of the Decision Tree model is roughly 68%, that of the K-medoids model is roughly 62%, and that of the K-means model is roughly 55%.

Recall quantifies the model's capacity to identify each positive occurrence in the dataset.

$$\text{Re}call = \frac{True\ positive}{True\ positive + False\ Negative} \tag{6.3}$$

The F1-score is defined as the recall and precision harmonic means. It presents a middle ground between recall and precision, which is especially helpful in situations when the distribution of classes is not uniform.

$$F_1 - score = \frac{2*(precision \times \mathrm{Re}call)}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{6.4}$$

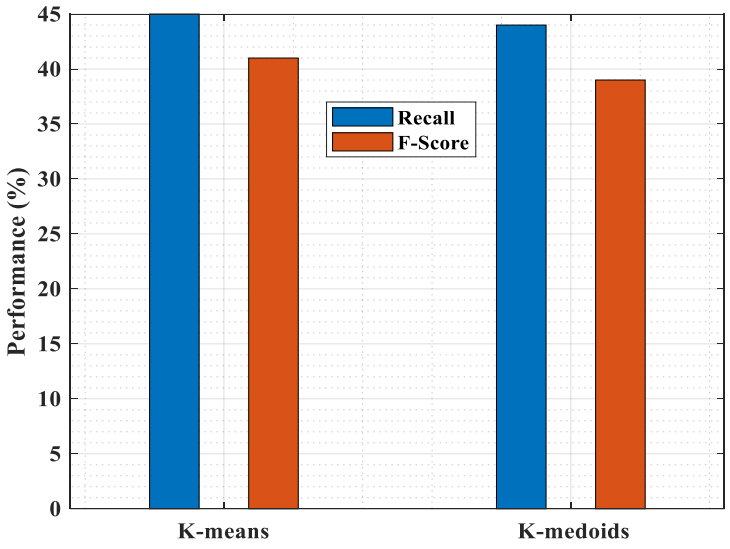

Figure 6.3: comparison of recall and F1-score of various ML model

Five machine learning models' recall and F1-score performance The K-means, K-medoids, Decision Tree, Random Forest, and SVM are compared in the figure. Models are shown on the horizontal axis, and the vertical "Recall" axis spans 0% to 45%. A pair of data points connected by a line represents each model. The orange square denotes the F1-score, whereas the blue circle shows the recall score. Out of the five models, the best recall and F1-score are produced by the Random Forest model. Its F1-score is approximately 38%, and its recall score is approximately 40%. SVM is not far behind, with an F1-score of about 35% and a recall of about 37%. Recall and F1-scores for the Decision Tree, K-medoids, and K-means models decrease with increasing time. Notably, K-means has the lowest F1-score (around 15%) and recall (approximately 20%).

6.4 PREDICTION PERFORMANCE COMPARISON OF DEEP LEARNING MODELS EMPLOYED FOR ANALYSING CRIMES IN SOCIAL MEDIA.

"Deep learning" is a relatively new field of computer science research in the machine learning arena. To get machine learning closer to artificial intelligence's primary objective, deep learning is being developed. By addressing a wide range of challenging pattern recognition problems and enabling machines to mimic human abilities like reasoning and audio-visual perception, deep learning increases artificial intelligence. Moreover, it builds complex function models by stacking many nonlinear layers and using a deep network to extract nonlinear features from input. Examples of modern deep learning applications are CNNs, recurrent neural networks, and latent feed forward networks. Deep learning has been beneficial to a number of domains, including computer vision, image identification, visual arts processing, natural language processing, drug discovery, genomics, cyber security, emotion recognition, and speech recognition.

Artificial neural networks (ANNs) are a subset of machine learning techniques that are referred to as deep learning models because they mimic the structure and functions of the human brain. Owing to their remarkable capacity to acquire intricate hierarchical data representations, these models are especially well-suited for challenging assignments such as natural language processing, photo and audio recognition, and even strategic gaming. Multiple-layer neural networks, often known as deep neural networks or deep learning architectures, are the fundamental building blocks of deep learning. These networks are so deep that they can automatically extract and alter features from unprocessed data, which helps them identify complex patterns and subtleties. While recurrent neural networks (RNNs) are better at handling sequential data, like voice or time-series data, convolutional neural networks (CNNs) are particularly good at jobs involving images. The emergence of architectures such as Transformer and Long Short-Term Memory (LSTM) has improved the ability of deep learning models to capture long-range dependencies and enable parallelized processing. Using methods like back propagation and stochastic gradient descent, millions or even billions of parameters must be adjusted during the training of deep learning models. Deep learning models have achieved amazing success due to their autonomous feature extraction and representation learning capabilities, but their

interpretability is still a problem, which has prompted continuous research into making them more accountable and transparent. The exceptional performance of deep learning models has transformed a number of industries, including healthcare, banking, and autonomous systems, in spite of these obstacles, and has accelerated the development of artificial intelligence.

Below are several graphics that show recall, accuracy, precision, and F1-score:

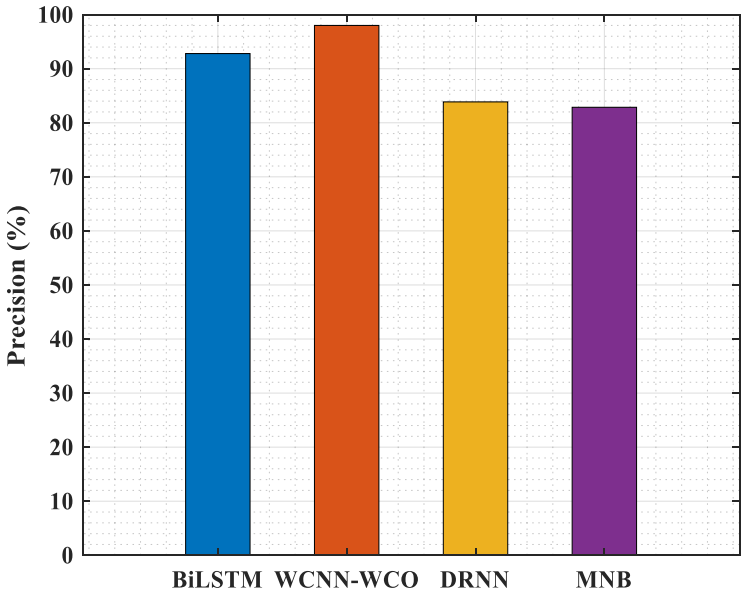

Figure 6.4: Precision of the various DL models

A graph comparing the recall and F1-score of five distinct machine learning models—K-means, K-medoids, Decision Tree, Random Forest, and SVM. The graph's x-axis lists the various machine learning models, while the y-axis is labeled "Recall (%)". A colored data point is used to symbolize each model. The model's recall score is shown by the blue circle, while it’s F1-score is shown by the orange square.

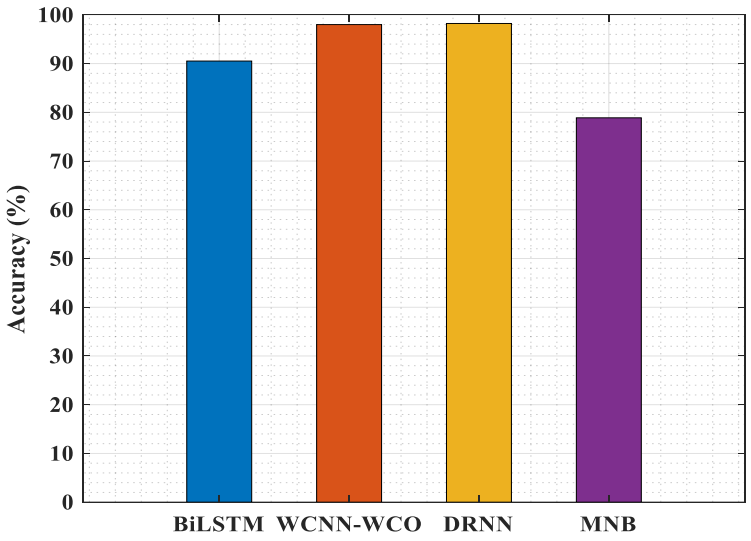

Figure 6.5: Accuracy of the various DL models

Based on the graph, the Random Forest model appears to have the most precise predictions and F1-score among the five models. It has an around 85% accuracy score and an approximate 78% F1-score. With an accuracy of roughly 82% and an F1-score of roughly 75%, SVM is not far behind. The models with decreasing accuracy and F1-scores are the Decision Tree, K-medoids, and K-means. Of all the algorithms, K-means has the lowest accuracy (about 55%) and F1-score (around 48%).

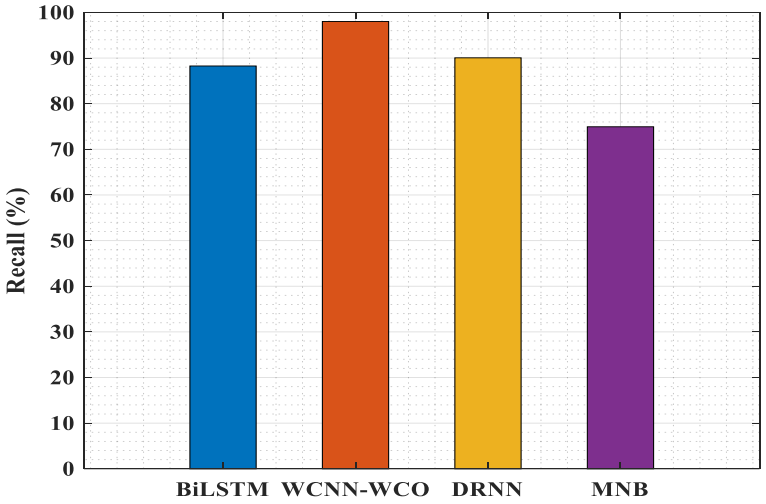

Figure 6.6: Recall of the various DL models

A comparison of recall scores for five distinct deep learning models is shown in the figure: Wavelet CNN with Weighted Overlap Combination (WCNN-WCO), Bi-directional Long Short-Term Memory (BiLSTM), Multinomial Naive Bayes (MNB), Deep Recurrent Neural Network (DRNN), and Logistic Regression (LR). The models are shown on the horizontal axis, while the vertical axis labeled "Recall (%)" goes from 0% to 100%. A colored bar is used to indicate each model. With a score of almost 94%, the BiLSTM model performs the best overall out of the five. At about 90%, WCNN-WCO trails closely behind. Recall ratings for the DRNN, MNB, and LR models decrease gradually. Recall is notably lowest for LR, at roughly 20%.

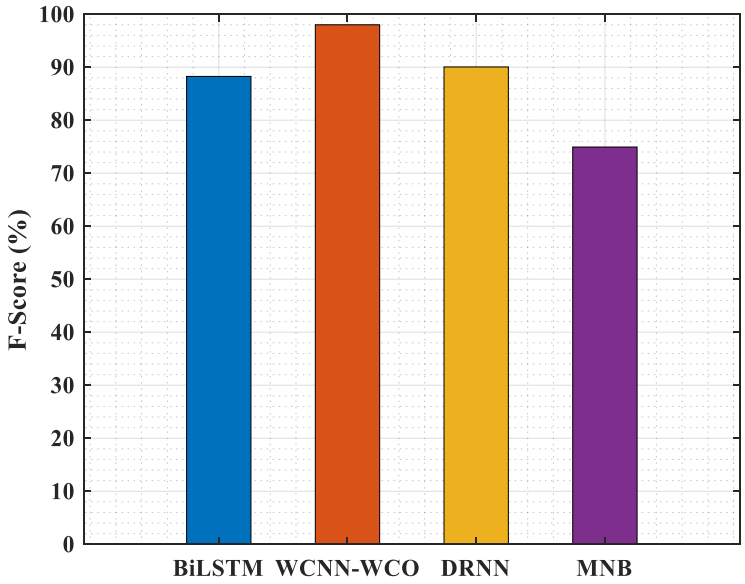

Figure 6.7: F1-score of the various DL models

The location of each model's matching blue circle on the y-axis indicates its F1-score. The model's F1-score increases with the size of the circle. The graph suggests that out of the five models, the BiLSTM model has the greatest F1-score, followed by WCNN-WCO, DRNN, MNB, and LR. The F1-scores of the following models are estimated: DRNN has an F1-score of roughly 78%, MNB has an F1-score of approximately 72%, LR has an F1-score of approximately 65%, and the BiLSTM model has an F1-score of approximately 87%.

6.5 ANALYSIS OF VARIED DATASETS UTILIZED FOR CRIME DATA PREDICTION IN DIFFERENT SOCIAL MEDIA NETWORKS

Predicting crime statistics on social media is mostly dependent on datasets, each of which has a wealth of data on crimes that have occurred in various places and at various times. Regional datasets and common datasets are the two categories of datasets. Regional datasets contain statistics on a specific area, such as assault percentages and various forms of information. Data that is commonly gathered from many social media networks, such as Facebook, Instagram, Twitter, and others, is included in common datasets. The dataset details that were previously discussed are shown in Table 6.1.

Table 6.1: Dataset details

Dataset name	**Year**	**Messages**	**Region**
San Francisco Crime Dataset	January 1, 2003, and May 13, 2015	8,78,049	San Francisco
Formspring.me dataset	2009	13158	Global
Myspace dataset	July 2021	1753 groups	Global
Twitter dataset	October 2018 and August 2019	8,835,016	Global
Crime dataset	January 14, 2014 and January 13, 2017	12384	Europe

San Francisco Dataset

The model is constructed by the study using a Kaggle dataset [98]. The attributes of the dataset, also known as the training set or data, vary and are connected differently. The San Francisco criminal occurrences from Kaggle are included in the training dataset. The period covered by the data is January 2003–May 2015. Nearly 12 years' worth of San Francisco police crime reports are included in the dataset. All crimes in the dataset are categorized into groups that include various types of crimes. There are 878049 observations in the training set and 884263 observations in the testing set. Using freshly unclassified data, the dataset is utilized to assess how accurate the classification methods are. The original training dataset is arbitrarily mixed and split into two subsets, the training dataset and the testing dataset, with 80% and 20% of the original size, respectively.

Formspring.me dataset

13158 messages from the Formspring.me website, posted by 50 distinct users, make up the Formspring.me dataset, an XML file [99]. This dataset was created in 2009 for a research project. "Cyberbullying Negative" and "Cyberbullying Positive" are the two classifications into which the dataset is separated. While negative communications portray messages that do not contain cyberbullying, positive messages do. The Cyberbullying Positive class contains 892 messages, whereas the Cyberbullying Negative class contains 12266 messages. The dataset was split into training and test sets using the "holdout" strategy, which is applied to datasets of similar size. The size and quantity of samples in a few well-known data clusters, such Reuters and 20NewGoups (20NG), are comparable to those indicated in. Thus, the same techniques as in these cases have been applied.

Myspace dataset

The Myspace dataset consists of messages collected from group chats on Myspace [100]. Ten message groups have been created from the annotated group chats in the dataset. For example, if there are 100 messages in a group chat, there will be 10 messages in the first group, 2–11 in the second, and 91–100 in the final group. Labeling is done once for each group of ten messages, and it shows whether bullying is

mentioned in any of the ten messages. This dataset has 1753 message groups total, divided into 10 groups with 1396 negative and 357 positive labels.

Twitter dataset

There are 20,000 rows in this Twitter dataset, and each row includes the user's name, a randomly selected tweet, their account profile, and picture or location data.

Crime dataset

The offenses reported by the police in England and Wales between 1990 and 2011–12, broken down by offense and police force region, make up the crime dataset used for crime analysis [101].

6.6 CHALLENGES IN CRIME DETECTION OVER SOCIAL MEDIA

Numerous techniques for detecting criminal activity on social media networks encounter numerous obstacles, a few of which are listed below.

- Because interpersonal and domestic attacks are rarely concentrated in one location and are not clearly associated with a single victim profile, predictive models are unable to effectively predict any type of criminal activity. Although prediction algorithms can reduce subjective judgments and mitigate certain types of individual bias, systems continue to rely on often-inaccurate crime data that contains systematic reporting problems.
- Several issues have been shown by analyzing the experiences with predictive techniques. It is possible for predictive algorithms to inadvertently reinforce and exacerbate societal stereotypes. Because of the cost of data storage, algorithms are often opaque about the underlying process, and on rare occasions, they have resulted in the infringement of basic rights. Finding bias in a data set is a challenging task that calls for in-depth expertise.
- Like most police technologies, this one needs to be used with a thorough approach to be successful. Law enforcement agencies must possess technological know-how, industrial leadership, and the ability to create

minimal requirements for responsible to successfully integrate predictive technology into their operations, they must undertake development, monitoring, and assessment.

- A number of models, mostly machine learning models, have poor accuracy and perform poorly on recall, precision, and f1-score, among other performance metrics.

6.7 EFFECTIVE FUTURE RECOMMENDATIONS

Predicting and analyzing crimes is a methodical way to find crime. This algorithm is able to predict and pinpoint the sites of criminal offenses. Law enforcement agencies can forecast future criminal activities by using data analysis. These projections, which come in a variety of formats, help organizations allocate resources more effectively and lower crime rates. The prediction is available in other formats, such as:

- Sites of criminal offenses

Criminal offense locales are certain geographic areas or regions where criminal activity is concentrated. They are often referred to as crime hotspots or crime-prone localities. A critical component of criminology and law enforcement is the analysis and comprehension of criminal offense sites. These places frequently display trends and patterns that can be found in a variety of data sources, such as databases maintained by law enforcement agencies, crime reports, and incident reports. Geographic information systems (GIS) and spatial analysis are used to identify locations with higher crime rates or recurrent criminal events in order to identify criminal offense locales. The spatial distribution of criminal activity is studied, and prospective hotspots are identified, using sophisticated analytical tools by researchers and law enforcement organizations. Certain facilities or establishments, environmental features, and socioeconomic circumstances can all have an impact on the concentration of criminal acts in a certain area. In addition to identifying these places, the objective is to create plans of action and methods to stop and lessen criminal activity in those areas. Predictive policing models, which use machine learning algorithms to identify probable crime hotspots based on past data, are frequently used by law enforcement agencies. By taking a proactive stance, law enforcement officers and resources can be

sent to high-risk locations in an effort to improve public safety and discourage criminal activity. Initiatives aimed at community policing may also entail working together with local companies and citizens to address the root causes of crime in certain areas. The study of criminal offense locations is a dynamic and ever-evolving topic that is impacted by data analytics, technological improvements, and cooperative efforts between the community, researchers, and law enforcement.

- The most likely criminal offenses to occur

The most common criminal charges differ depending on a number of criteria, such as socioeconomic status, cultural influences, and geographic location. Some environments appear to be more conducive to the commission of certain sorts of crimes. In metropolitan regions, where valuable things are concentrated and people may be more susceptible owing to high population density, common examples include property crimes like theft and burglary. Different patterns of violent crimes, such as killings and assaults, can be seen; these patterns are frequently impacted by socioeconomic inequality, gang involvement, and other sociocultural variables. As criminals take advantage of digital vulnerabilities, technological improvements have also led to an increase in cybercrimes, such as identity theft and online fraud. Drug-related crimes are more common in places where there is a strong market for illicit drugs or where there are particular socioeconomic problems. Furthermore, crimes against people, including sexual assault and domestic abuse, can happen in a variety of contexts and are frequently impacted by elements like drug or alcohol addiction, conflict within the family, or public perceptions of these problems. A complicated interaction between demographic trends, environmental conditions, and historical data is required to predict individual criminal offenses. To find trends and gauge the chance of specific crimes happening in given situations, researchers and law enforcement organizations use criminological theories, machine learning, and data analytics. To stop or deal with these offenses, proactive measures including community policing, public awareness campaigns, and targeted law enforcement tactics are frequently used. Creating successful crime prevention and intervention plans requires an understanding of the kinds of crimes that are most likely to happen at a specific place. With the use of this information, law enforcement, decision-makers, and community members may better allocate resources, carry out focused interventions, and develop projects that

tackle the underlying causes of particular crimes, thereby improving community safety and well-being.

- During the day, when events are also most probable to occur

Daytime incidents frequently exhibit unique patterns that are impacted by a range of variables, including human behavior, the surroundings, and the types of activities carried out throughout the day. For instance, during the day, when people are out from their homes for work or other reasons, property crimes like theft and burglary may be more common, creating opportunities for illegal entrance. Shoplifting and other criminal activities may be more common in commercial districts when business operations are bustling during the day. Furthermore, due to an increase in vehicular activity throughout the day, traffic-related incidents, such as accidents and violations, occur more frequently at this time. Better visibility during the daytime lowers the risk of accidents brought on by night-time vision impairments. But some crimes against people, such robberies or assaults, can also happen during the day, frequently in public areas where people gather. To determine when events are most likely to occur during the day, law enforcement organizations analyze historical trends and crime statistics. The allocation of resources and patrols to target particular crime patterns during the day is guided by this intelligence. Public awareness and community involvement initiatives may also emphasize situational awareness and preventative actions while promoting safety precautions during the day.

- some people's repeated criminal activity

Recidivism the term used to describe the recurrent criminal behavior of some individuals—presents a serious obstacle to the criminal justice and law enforcement systems. This phenomenon is the result of someone continuing to commit crimes in spite of prior arrests, convictions, or attempts at rehabilitation. A wide range of criminal offenses, from infractions to more violent and serious ones, can result in recidivism. Recurrent criminal conduct is caused by a number of variables, such as socioeconomic circumstances, a lack of educational and career possibilities, problems with substance addiction, and mental health disorders. It can be challenging for people to break free from a pattern of criminal behavior when they find themselves caught in a cycle of illegal activity. The criminal justice system's structural problems, community contexts, and peer relationships can all have an impact on repeat offenders.

When certain people commit crimes repeatedly, a thorough and diversified strategy is needed to address their behavior. In order to interrupt the cycle of recidivism, rehabilitation programs, educational opportunities, vocational training, and mental health services are essential. Reducing recurrent offenses can be accomplished through community-based programs that emphasize reintegration, support systems, and addressing the underlying causes of criminal behavior.

To determine who is more likely to engage in crimes in the future, law enforcement organizations and criminal justice systems frequently use data analysis and risk assessment techniques. To reduce the likelihood of recurrent criminal activity, this data informs the distribution of resources, the creation of focused intervention programs, and the application of supervision techniques. In the end, cooperation between law enforcement, social services, mental health specialists, and community organizations is necessary to break the cycle of recurrent crimes. Reducing recurrent offenses and promoting a safer and more resilient community are more likely when the root causes of recidivism are addressed and people are given the assistance and tools they need for recovery.

- Certain crimes occur in specific places.

A well-established trend in criminology is the incidence of particular crimes in particular locations, highlighting the importance of geographic factors in comprehending criminal behavior. A range of elements, including local infrastructure, environmental features, and socioeconomic situations, can be blamed for the spatial concentration of different sorts of crimes. For example, because of the dense concentration of valuable items and bigger population, metropolitan regions may have higher incidence of property crimes, such as theft and vandalism. On the other hand, certain crimes like agricultural theft or unlawful dumping could be more common in rural areas. For law enforcement organizations and legislators to create focused crime prevention initiatives, they must have a thorough understanding of the spatial distribution of crimes. A well-established trend in criminology is the incidence of particular crimes in particular locations, highlighting the importance of geographic factors in comprehending criminal behavior. A range of elements, including local infrastructure, environmental features, and socioeconomic situations, can be blamed

for the spatial concentration of different sorts of crimes. For example, because of the dense concentration of valuable items and bigger population, metropolitan regions may have higher incidence of property crimes, such as theft and vandalism. On the other hand, certain crimes like agricultural theft or unlawful dumping could be more common in rural areas. For law enforcement organizations and legislators to create focused crime prevention initiatives, they must have a thorough understanding On the geographic distribution of criminal activity.

- Crime trends and hotspots

Crime prevention and law enforcement measures heavily rely on crime trends and hotspots. The patterns and variances in criminal behavior over a certain time frame are referred to as crime trends. Numerous factors, such as changes in law enforcement tactics, population density, and socioeconomic situations, might have an impact on these trends. By offering important insights into the evolving nature of criminal activity, crime trend analysis aids in law enforcement agencies' adaptation and the creation of focused interventions.

On the other hand, concentrated regions where criminal acts happen more frequently than in surrounding areas are known as crime hotspots. Finding these hotspots is crucial for allocating resources as efficiently as possible since it allows law enforcement to concentrate their efforts on areas where there is a greater chance of criminal activity. Advanced data analytics and geographic information systems (GIS) are frequently used to identify crime hotspots, enabling the more intelligent use of law enforcement resources and manpower. Targeted actions in high-crime areas and a proactive approach to new crime patterns are essential components of effective crime prevention. Utilizing a combination of analytical techniques and past crime data, predictive policing models enable law enforcement organizations to identify prospective hotspots and direct resources accordingly. In order to combat crime patterns, community engagement is also essential because working with local companies and individuals makes preventative efforts more successful.

Even though current techniques aren't always precise, police departments can nevertheless benefit from the predictions generated by algorithms.

- The goal of this research is to improve the pre-processing stage model, which can lead to faster processing times and better prediction stages.
- In order to generate more features and achieve better outcomes, this research aims to suggest an efficient procedure for various term implanting strategies throughout the feature extraction stage.
- The aim of this research is to provide a useful classic for the feature selection phase, which is crucial to a successful classification procedure. A metaheuristic algorithm with an enhanced weight update mechanism will be used for feature selection.
- During the classification phase, the study aims to provide a hybrid algorithm that has a functional fitness function. Neural networks, metaheuristic algorithms, or both may be used in the algorithm. The core of the entire model should be an efficient categorization.

These are the future goals that the research hopes to accomplish. This model will reduce computational power consumption and increase classification accuracy.

6.8 Summary

Crime prediction plays a major role in the creation of enforcement tactics as well as the application of crime prevention and control. These days, machine learning is the most used prediction technology. Since existing prediction techniques are unreliable, this study will rather analyze the models and assess their efficacy. This paper examined a number of machine learning and deep learning models' performance indicators, including recall, accuracy, precision, and f1-score. As such, deep learning performs better than machine learning in terms of accuracy and precision when compared. A handful of the datasets and their many types that are utilized to forecast crime numbers are looked at.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This chapter provides a comprehensive outline of the research conducted in the thesis, outlining the key findings, contributions made, and avenues for future exploration. It synthesizes the main outcomes and insights gained from the study, underscores the significance of the contributions made to the field, and identifies potential areas for further investigation and advancement.

7.1 Conclusion

This paper presents a comprehensive analysis of various crime detection methods applied to twitter data, a widely used online platform for diverse communication purposes. The proposed approach introduces a hybrid model combining WCNN with WCO. The process begins with pre-processing the dataset through stemming, tokenization, and stop word removal. Features are then extracted utilizing BoW, TF-IDF, Glove, and feature hashing. Feature selection is conducted using a MTGA, followed by data clustering with FMRF. The classification of crime data is accomplished through the integration of CNN with wavelet technology, and performance optimization is achieved using the WCO algorithm.

The proposed method is benchmarked against existing algorithms, including HREACC, HANNCC, HMNCC, and GSCA, using PYTHON for implementation. The obtained results demonstrate a remarkable precision of 98.8%, accuracy of 98.9%, F-measure of 98.9%, recall of 98.9%, and an AUC value of 98.9%, surpassing the performance of other algorithms. While the current implementation is offline, future modifications may enable real-time crime prediction through Twitter data streaming.

The potential enhancement of the system's efficiency and resilience is suggested through the addition of crime classes.

This research contributes to the ongoing efforts in leveraging advanced technologies for crime detection in the dynamic landscape of online communication platforms. The proposed hybrid model exhibits promising results, emphasizing its efficacy in addressing the challenges posed by criminal activities on social media. The envisioned future developments hold potential for further improving the accuracy and responsiveness of the system in predicting and preventing crimes in real-time.

7.2 Future Scope

- The structure will be improved by including sophisticated monitoring services that prioritize privacy, data consistency, and data authenticity.
- It is anticipated that this next integration would perform better than conventional MapReduce techniques, guaranteeing quicker and more effective data processing for all-encompassing monitoring and crime prevention.
- With the use of real-time Twitter data streaming, it might be altered to forecast upcoming crimes. It is possible to add kinds of crimes to the system to increase its resilience and efficiency.

During the feature extraction step, other word embedding strategies will be implemented, potentially yielding more features and better outcomes. In addition, the suggested model is expanded to handle real-time data processing in subsequent studies

REFERENCES

[1] A. Taei, H. Jönson, and M. Granbom, “Crime, disorder, and territorial stigmatization: Older adults living in deprived neighborhoods,” Gerontologist, vol. 63, no. 5, pp. 910–919, 2023.

[2] A. Di Nicola, “Towards digital organized crime and digital sociology of organized crime,” Trends Organ. Crime, pp. 1–20, 2022.

[3] S. Chen et al., “Exploring the global geography of cybercrime and its driving forces,” Humanit. Soc. Sci. Commun., vol. 10, no. 1, p. 71, 2023.

[4] M. Vivek and B. R. Prathap, “Spatio-temporal crime analysis and forecasting on Twitter data using machine learning algorithms,” SN Comput. Sci., vol. 4, no. 4, p. 383, 2023.

[5] O. Olaniyi, O. J. Okunleye, and S. O. Olabanji, “Advancing data-driven decision-making in smart cities through big data analytics: A comprehensive review of existing literature,” Current Journal of Applied Science and Technology, vol. 42, no. 25, pp. 10–18, 2023.

[6] C. Shenkman, S. B. Franklin, G. Nojeim, and D. Thakur, “Legal loopholes and data for dollars: How law enforcement and intelligence agencies are buying your data from brokers,” 2022.

[7] K. Phillips, J. C. Davidson, R. R. Farr, C. Burkhardt, S. Caneppele, and M. P. Aiken, “Conceptualizing cybercrime: Definitions, typologies and taxonomies,” Forensic Sciences, vol. 2, no. 2, pp. 379–398, 2022.

[8] S. Kumar, Cyber Crimes Against Children in Virtual World an Empirical Study with Specific Reference to. Kangra District Himachal Pradesh, 2023.

[9] S. Ranjan, “VICTIM’S RIGHT AND VICTIMOLOGY UNDER INDIAN CRIMINAL JUSTICE SYSTEM: AN ANALYTICAL STUDY,” AN ANALYTICAL STUDY. Vidhyayana-An International Multidisciplinary Peer-Reviewed E-Journal, vol. 8, no. 5, 2023.

[10] G. Ruffo, A. Semeraro, A. Giachanou, and P. Rosso, “Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language,” Comput. Sci. Rev., vol. 47, no. 100531, p. 100531, 2023.

[11] W. Salem, Chapter two NSAS and the possibility for transnational politics. Non-State Actors in Conflicts: Conspiracies, Myths, and Practices. 2018.

[12] J. Hall-Patton, Great Excitement": Violent Incorporations of the American Southwest (Doctoral dissertation). 2023.

[13] M. van Zomeren, C. d'Amore, I. L. Pauls, E. Shuman, and A. Leal, "The intergroup value protection model: A theoretically integrative and dynamic approach to intergroup conflict escalation in democratic societies," Pers. Soc. Psychol. Rev., p. 10888683231192120, 2023.

[14] M. Bossetta, "The weaponization of social media: Spear phishing and cyberattacks on democracy," Journal of international affairs, vol. 71, no. 1.5, pp. 97–106, 2018.

[15] R. Veresha, "Preventive measures against computer related crimes: Approaching an individual," Informatologia, vol. 51, no. 3–4, pp. 189–199, 2018.

[16] J. Seymour and P. Tully, "Weaponizing data science for social engineering: Automated E2E spear phishing on Twitter," Black Hat USA, vol. 37, pp. 1–39, 2016.

[17] D. Gibert, C. Mateu, and J. Planes, "The rise of machine learning for detection and classification of malware: Research developments, trends and challenges," J. Netw. Comput. Appl., vol. 153, no. 102526, p. 102526, 2020.

[18] M. Kovic, A. Rauchfleisch, M. Sele, and C. Caspar, "Digital astroturfing in politics: Definition, typology, and countermeasures," Stud. Commun. Sci., vol. 18, no. 1, 2018.

[19] M. C. Galdo, M. E. Tait, and L. E. Feldman, "Money mules: Stopping older adults and others from participating in international crime schemes. Dep't of Just," J. Fed. L. & Prac, vol. 66, 2018.

[20] R. Lara-Cabrera, A. Gonzalez-Pardo, and D. Camacho, "Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter," Future Gener. Comput. Syst., vol. 93, pp. 971–978, 2019.

[21] F. J. Cavico and B. G. Mujtaba, "Defamation by slander and libel in the workplace and recommendations to avoid legal liability," Public Organ. Rev., vol. 20, no. 1, pp. 79–94, 2020.

[22] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 US presidential election,” Science, vol. 363, no. 6425, pp. 374–378, 2019.

[23] A. Brown, “Retheorizing actionable injuries in civil lawsuits involving targeted hate speech: Hate speech as degradation and humiliation. Ala,” Ala. CR & CLL Rev, vol. 9, 2018.

[24] M. Arisanty and G. Wiradharma, “The motivation of flaming perpetrators as cyberbullying behavior in social media,” Jurnal Kajian Komunikasi, vol. 10, no. 2, pp. 215–227, 2022.

[25] D. S. Devarakonda, Time series analysis and forecasting of crime data (Doctoral dissertation). Sacramento, 2019.

[26] A. Ristea, J. Kurland, B. Resch, M. Leitner, and C. Langford, “Estimating the spatial distribution of crime events around a football stadium from georeferenced tweets,” ISPRS Int. J. Geoinf., vol. 7, no. 2, p. 43, 2018.

[27] “Adult Online Hate, Harassment and Abuse: A rapid evidence assessment,” Gov.uk, 26-Jun-2019. [Online]. Available: https://www.gov.uk/government/publications/adult-online-hate-harassment-and-abuse-a-rapid-evidence-assessment. [Accessed: 09-Jan-2024].

[28] B. Arora, “A review of sentimental analysis on social media application. Emerging Trends in Expert Applications and Security,” in Proceedings of ICETEAS 2018, 2019, pp. 477–484.

[29] J. V. Mbithi, Impact of social media on National Security in Kenya (Doctoral dissertation). 2022.

[30] C. C. Henry, Evaluating the Effectiveness of an Ensemble Random Forest Machine Learning Algorithm in Detecting Cyberbullying in the 4chan Politically Incorrect Board Social (Doctoral dissertation). 2021.

[31] R. R. Asaad, H. B. Ahmad, and R. I. Ali, “A review: Big Data technologies with Hadoop distributed filesystem and implementing M/R,” Acad. J. Nawroz Univ., vol. 9, no. 1, pp. 25–33, 2020.

[32] A. Palanivinayagam, S. S. Gopal, S. Bhattacharya, N. Anumbe, E. Ibeke, and C. Biamba, “An optimized machine learning and big data approach to crime detection,” Wirel. Commun. Mob. Comput., vol. 2021, pp. 1–10, 2021.

[33] U. Can and B. Alatas, “A new direction in social network analysis: Online social network analysis problems and applications,” Physica A, vol. 535, no. 122372, p. 122372, 2019.

[34] S. Lal, L. Tiwari, R. Ranjan, A. Verma, N. Sardana, and R. Mourya, “Analysis and classification of crime tweets,” Procedia Comput. Sci., vol. 167, pp. 1911–1919, 2020.

[35] M. A. AlGhamdi and M. A. Khan, “Intelligent analysis of Arabic tweets for detection of suspicious messages,” Arab. J. Sci. Eng., vol. 45, no. 8, pp. 6021–6032, 2020.

[36] K. Santhiya, V. Bhuvaneswari, and V. Murugesh, “Automated crime tweets classification and geo-location prediction using big data framework,” Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 14, pp. 2133–2152, 2021.

[37] M. Boukabous and M. Azizi, “Crime prediction using a hybrid sentiment analysis approach based on the bidirectional encoder representations from transformers,” Indones. J. Electr. Eng. Comput. Sci., vol. 25, no. 2, p. 1131, 2022.

[38] A. S. Hissah and H. Al-Dossari, “Detecting and classifying crimes from arabic twitter posts using text mining techniques,” International Journal of Advanced Computer Science and Applications, no. 10, 2018.

[39] K. B. Aljanabi, “Predicting Class Label Using Clustering-Classification Technique: A Comparative Study,” Journal of Kufa for Mathematics and Computer, vol. 10, no. 1, pp. 1–12, 2023.

[40] S. G. Krishnendu, P. P. Lakshmi, and L. Nitha, “Crime analysis and prediction using optimized K-means algorithm,” in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020.

[41] M. B. B. Pepsi and S. N. Kumar, “Supervised Learning Techniques for Classification Of Students’ Tweets,” Turkish Journal of Computer and Mathematics Education, vol. 12, no. 12, pp. 3110–3118, 2021.

[42] S. Al-Saqqa, G. Al-Naymat, and A. Awajan, “A large-scale sentiment data classification for online reviews under Apache spark,” Procedia Comput. Sci., vol. 141, pp. 183–189, 2018.

[43] S. Savaş and N. Topaloğlu, “Data analysis through social media according to the classified crime,” TURK. J. OF ELECTR. ENG. COMPUT. SCI., vol. 27, no. 1, pp. 407–420, 2019.

[44] A. Alqahtani, A. Garima, and A. Alaiad, “Crime Analysis in Chicago City,” in 2019 10th International Conference on Information and Communication Systems (ICICS), 2019.

[45] B. Panja, P. Meharia, and K. Mannem, “Crime Analysis Mapping, Intrusion Detection-Using Data Mining,” in 2020 IEEE Technology & Engineering Management Conference (TEMSCON), IEEE, 2020, pp. 1–5.

[46] A. Baby, J. Jose, and A. Raj, “Psychosomatic Study of Criminal Inclinations with Profanity on Social Media: Twitter,” in Proceedings of International Conference on Data Science and Applications: ICDSA 2022, vol. 1, Singapore; Singapore: Springer Nature, 2023, pp. 611–627.

[47] Y. Sharrab, D. Al-Fraihat, and M. Alsmirat, Deep Neural Networks in Social Media Forensics: Unveiling Suspicious Patterns and Advancing Investigations on Twitter. EasyChair, 2023.

[48] A. M. Alduailaj and A. Belghith, “Detecting Arabic cyberbullying tweets using machine learning,” Mach. Learn. Knowl. Extr., vol. 5, no. 1, pp. 29–42, 2023.

[49] T. Siddiqui, S. Hina, R. Asif, S. Ahmed, and M. Ahmed, “An ensemble approach for the identification and classification of crime tweets in the English language,” Computer Science and Information Technologies, vol. 4, no. 2, pp. 149–159, 2023.

[50] K. Jenga, C. Catal, and G. Kar, “Machine learning in crime prediction,” J. Ambient Intell. Humaniz. Comput., vol. 14, no. 3, pp. 2887–2913, 2023.

[51] S. Saikia, E. Fidalgo, E. Alegre, and L. Fernández-Robles, “Object detection for crime scene evidence analysis using deep learning,” in Image Analysis and Processing - ICIAP 2017, Cham: Springer International Publishing, 2017, pp. 14–24.

[52] R. Jain, A. Nayyar, and S. Bachhety, “Factex: A practical approach to crime detection,” in Data Management, Analytics and Innovation, Singapore: Springer Singapore, 2020, pp. 503–516.

[53] V. Pande, V. Samant, S. Nair, and D. J. Sanghvi College of Engineering, “Crime Detection using Data Mining,” Int. J. Eng. Res. Technol. (Ahmedabad), vol. V5, no. 01, 2016.

[54] S. Nazah, S. Huda, J. Abawajy, and M. M. Hassan, “Evolution of dark web threat analysis and detection: A systematic approach,” IEEE Access, vol. 8, pp. 171796–171819, 2020.

[55] Z. Malik and S. Haidar, “Online community development through social interaction-K-Pop stan twitter as a community of practice,” Interactive Learning Environments, vol. 31, no. 2, pp. 733–751, 2023.

[56] F. Marieke, E. Urry, B. I. Lissenberg-Witte, and S. E. Kramer, “A comparison of the use of smart devices, apps, and social media between adults with and without hearing impairment: cross-sectional web-based study,” Journal of medical Internet research, vol. 23, no. 12, 2021.

[57] S. Ramzan and S. Imran, “Effectiveness of social media platforms in remote learning during lockdowns,” International Journal of Reflective Research in Social Sciences, vol. 3, no. 1, pp. 4–08, 2020.

[58] D. O. Anderez, E. Kanjo, A. Amnwar, S. Johnson, and D. Lucy, “The rise of technology in crime prevention: Opportunities, challenges and practitioners’ perspectives,” arXiv [cs.CY], 2021.

[59] P. Q. Brady, M. R. Nobles, and L. A. Bouffard, “Are college students really at a higher risk for stalking? Exploring the generalizability of student samples in victimization research,” J. Crim. Justice, vol. 52, pp. 12–21, 2017.

[60] O. Babko-Malaya, R. Cathey, S. Hinton, D. Maimon, and T. Gladkova, “Detection of hacking behaviors and communication patterns on social media,” in 2017 IEEE International Conference on Big Data (Big Data), 2017.

[61] L. Almadhoor, “Social media and cybercrimes,” Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 10, pp. 2972–2981, 2021.

[62] H. Jiang, Y. Xiao, and W. Wang, “Explaining a bag of words with hierarchical conceptual labels,” World Wide Web, vol. 23, no. 3, pp. 1693–1713, 2020.

[63] A. Thakkar and K. Chaudhari, “Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks,” Appl. Soft Comput., vol. 96, no. 106684, p. 106684, 2020.

[64] C. I. Eke, A. Norman, L. Shuib, F. B. Fatokun, and I. Omame, “The significance of global vectors representation in sarcasm analysis,” in 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), 2020.

[65] J. Chen and Y. Zu, “Local feature hashing with binary auto-encoder for face recognition,” IEEE Access, vol. 8, pp. 37526–37540, 2020.

[66] J.-W. Liu, F.-L. Zuo, Y.-X. Guo, T.-Y. Li, and J.-M. Chen, “Research on improved wavelet convolutional wavelet neural networks,” Appl. Intell., vol. 51, no. 6, pp. 4106–4126, 2021.

[67] H. Li, K. Li, N. Zafetti, and J. Gu, “Improvement of energy supply configuration for telecommunication system in remote area s based on improved chaotic world cup optimization algorithm,” Energy (Oxf.), vol. 192, no. 116614, p. 116614, 2020.

[68] G. Deepak, S. Rooban, and A. Santhanavijayan, “A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network,” Multimed. Tools Appl., vol. 80, no. 18, pp. 28061–28085, 2021.

[69] J. L. McMullan and A. Rege, “Online crime and internet gambling,” *J. Gambl. Issu.*, no. 24, p. 54, 2010.

[70] J. M. Drew, E. Moir, and M. Newman, “Financial crime investigation: an evaluation of an online training program for police,” *Policing*, vol. 44, no. 3, pp. 525–539, 2021.

[71] L. Almadhoor, “Social media and cybercrimes,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 10, pp. 2972–2981, 2021.

[72] N. Razak, “A Comprehensive Study of Privacy and Security Risk Awareness Among Mobile Internet Users for Social Networks Sites in Malaysia,” *International Journal of Business and Technology Management*, vol. 3, no. 1, pp. 1–20, 2021.

[73] C. Sandagiri, B. T. G. S. Kumara, and B. Kuhaneswaran, “Deep neural network-based crime prediction using Twitter data,” *Int. J. Syst. Serv.-Oriented Eng.*, vol. 11, no. 1, pp. 15–30, 2021.

[74] S. Aslam, “Twitter by the numbers (2023): Stats, demographics & Fun Facts,” *Omnicore Agency*, 09-Mar-2023. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/. [Accessed: 09-Jan-2024].

[75] M. J. Metzger, A. J. Flanagin, P. Mena, S. Jiang, and C. Wilson, “From dark to light: The many shades of sharing misinformation online,” *Media Commun.*, vol. 9, no. 1, pp. 134–143, 2021.

[76] S. C. Thurtell, “When shopping turns scary: Twitter conversations about violent crime in Johannesburg malls,” *Crit. Arts*, vol. 35, no. 1, pp. 85–96, 2021.

[77] R. Asaad and R. M. Rajab, “The Concept of Data Mining and Knowledge Extraction Techniques,” *Qubahan Academic Journal*, vol. 1, no. 2, pp. 17–20, 2021.

[78] D. Ojha, R. P. Singh, and K. Singh Jadon, “An analysis on data mining for sentiment analysis using different classification algorithms,” in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, 2021.

[79] A. Ahadi, A. Singh, M. Bower, and M. Garrett, “Text mining in education—A bibliometrics-based systematic review,” *Educ. Sci. (Basel)*, vol. 12, no. 3, p. 210, 2022.

[80] D. H. P. Benício, J. C. Xavier Junior, K. R. S. de Paiva, and J. D. de A. S. Camargo, “Applying Text Mining and Natural Language Processing to Electronic Medical Records for extracting and transforming texts into structured data,” *Res. Soc. Dev.*, vol. 11, no. 6, p. e37711629184, 2022.

[81] R. Raja, N. Arshath, and N. V. Yuvaraj, “Analyses on Artificial Intelligence Framework to Detect Crime Pattern,” in *Intelligent Data Analytics for Terror Threat Prediction: Architectures, Methodologies, Techniques and Applications*, 2021, pp. 119–132.

[82] L. Abualigah, D. Yousri, M. Abd Elaziz, A. A. Ewees, M. A. A. Al-qaness, and A. H. Gandomi, “Aquila Optimizer: A novel meta-heuristic optimization algorithm,” *Comput. Ind. Eng.*, vol. 157, no. 107250, p. 107250, 2021.

[83] M. Townsley, “Visualising space time patterns in crime: the hotspot plot,” *Crime patterns and analysis*, vol. 1, pp. 61–74, 2008.

[84] L. Yu *et al.*, “Sparse code multiple access for 6G wireless communication networks: Recent advances and future directions,” *IEEE Commun. Stand. Mag.*, vol. 5, no. 2, pp. 92–99, 2021.

[85] S. Darvazeh, R. Iman, and F. M. Vanani, “Big data analytics and its applications in supply chain management,” *New Trends in the Use of Artificial Intelligence for the Industry*, vol. 4, 2020.

[86] S. Maitrey and C. K. Jha, “MapReduce: Simplified data analysis of big data,” *Procedia Comput. Sci.*, vol. 57, pp. 563–571, 2015.

[87] R. Gartner, *The Oxford handbook of gender, sex, and crime*. Oxford University Press, 2014.

[88] J. Van Dijk, *The world of crime: Breaking the silence on problems of security, justice and development across the world*. Sage Publications, 2007.

[89] J. Brown and P. Seely, “The social life of information: Updated, with a new preface,” *Harvard Business Review Press*, 2017.

[90] A. Ferreira and T. Du Plessis, “Effect of online social networking on employee productivity,” *South African Journal of Information Management*, vol. 11, no. 1, pp. 1–11, 2009.

[91] D. C. Rajiv and M.-J. Hambrick, “What is strategic management, really? Inductive derivation of a consensus definition of the field,” *Strategic management journal*, vol. 28, no. 9, pp. 935–955, 2007.

[92] P. Nguyen, *The effect of organic marketing on customer engagement in Social media Channel: Facebook*. 2020.

[93] S. Sohail, M. M. Saquib, and M. Khan, *An Analysis of Twitter Users From The Perspective of Their Behavior, Language, Region and Development Indices--A Study of 80 Million Tweets*. 2021.

[94] D. Kim, Y. Jo, I.-C. Moon, and A. Oh, “Analysis of twitter lists as a potential source for discovering latent characteristics of users,” *ACM CHI workshop on microblogging*, vol. 6, 2010.

[95] E. Borgia, “The Internet of Things vision: Key features, applications and open issues,” *Comput. Commun.*, vol. 54, pp. 1–31, 2014.

[82] J. Heidemann, M. Klier, and F. Probst, “Online social networks: A survey of a global phenomenon,” *Comput. Netw.*, vol. 56, no. 18, pp. 3866–3878, 2012.

[96] P. Fajnzylber, D. Lederman, and N. Loayza, “What causes violent crime?,” *Eur. Econ. Rev.*, vol. 46, no. 7, pp. 1323–1357, 2002.

[97] N. Tasnim, I. T. Imam, and M. M. A. Hashem, “A novel multi-module approach to predict crime based on multivariate spatio-temporal data using attention and sequential fusion model,” *IEEE Access*, vol. 10, pp. 48009–48030, 2022.

[98] M. Khan, A. Ali, and Y. Alharbi, “Predicting and preventing crime: A crime prediction model using San Francisco crime data by classification techniques,” *Complexity*, vol. 2022, pp. 1–13, 2022.

[99] A. C. I. Çiğdem, E. Çürük, and E. S. Eşsiz, “Automatic detection of cyberbullying in formspring. me, myspace and Youtube social networks,” *Turkish Journal of Engineering*, vol. 3, no. 4, pp. 168–178, 2019.

[100] *Gov.uk*.[Online].Available: https://www.gov.uk/government/publications/offencesrecorded-by-the-police-in-england-and-wales-byoffence-and-police-force-area-1990-to-2011-12. [Accessed: 09-Jan-2024].

[101] J. L. McMullan and A. Rege, “Online crime and internet gambling,” *J. Gambl. Issu.*, no. 24, p. 54, 2010.

List of Publications

Papers Published in International Journals:

- Monika, Aruna Bhat, "Automatic Twitter Crime Prediction Using Hybrid Wavelet Convolution Neural Network with World Cup Optimization", World Scientific, International Journal of Pattern Recognition and Artificial Intelligence (2022), Volume 36, No.5, DOI: 10.1142/S0218001422590054. (**SCIE, IF:1.6**)
- A Research Paper entitled "Predictive Analytics of Crime Data in Social media: A Systematic review, incorporating framework, and future investigation schedule" in **SN Computer Science, Springer Publisher, Scopus Indexed (Communicated)**
- A Research Paper entitled "DAC-BiNet: Twitter Crime Detection using Deep Attention Convolution Bi-Directional Aquila Optimal Network" in Multimedia Tools and Applications (2023), DOI: 10.1007/s11042-023-17250-4. (**SCIE, IF: 3.6**)

Papers Published in International Conferences:

- **Monika, Aruna Bhat**. "An analysis of Crime Data under Apache Pig on Big Data", 2019 3rd IEEE International Conference on I-SMAC (IOT in Social, Mobile, Analytics and Cloud), Palladam, Tamil Nadu, India, 11-13 December, 2019. (**Scopus Indexed)**
- **Monika, Aruna Bhat.** "Comparative Review of Different Techniques for Predictive Analytics in Crime Data Over Online Social Media", 5th Springer International Conference on Data & Information Sciences (ICDIS-2023), Bichpuri, Agra, India, 16-17 June, 2023. (**Scopus Indexed)**