Detection Framework for Content Based Cybercrime in Online Social Networks

*A **Thesis***

submitted in partial fulfillment of the requirements for the award of degree of

Doctor of Philosophy

by

Amanpreet Singh

(901411012)

under the guidance of

Dr. Maninder Kaur

Associate Professor

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology (Deemed to be University)

Patiala -147004, INDIA

June 2021

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
API	Application Program Interface
AUC	Area Under the Curve
BLSTM	Bidirectional Long Short-Term Memory
BoW	Bag of Word
CNN	Convolutional Neural Network
CS	Cuckoo Search
DNN	Deep Neural Network
EBoW	Embedding-enhanced Bag of Word
EC	Evolutionary Computation
FPR	False Positive Rate
HITS	Hyperlink-Induced Topic Search
IBk	Instance Based Classifier
K-FSVM	Kernel-Fuzzy Support Vector Machine
KNN	K-Nearest Neighbor
LR	Logistic Regression
LSTM	Long Short-Term Memory
LSVM	Linear Support Vector Machine
NB	Naïve Bayes
NLP	Natural Language Processing
OSN	Online Social Network
PCA	Principal Component Analysis
PLSA	Probabilistic Latent Semantic Analysis
PoS	Part of Speech

PVC	Participant Vocabulary Consistency
RBF	Radial Basis Function
RF	Random Forest
RGA	Refined Genetic Algorithm
ROC	Receiver Operating Characteristic
SMO	Sequential Minimal Optimization
SMOTE	Synthetic Minority Over-sampling TechniquE
SMP	Segmented Max Pooling
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
TPR	True Positive Rate

Certificate

I hereby certify that the work which is being presented in this thesis entitled "**Detection Framework for Content Based Cybercrime in Online Social Networks**", in partial fulfillment of the requirement for the award of degree of "**Doctor of Philosophy**" submitted in Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed to be University), Patiala (India), is an authentic record of my own work carried out under the supervision of **Dr. Maninder Kaur** and refers other research works which are duly listed in the reference section.

The matter presented in this thesis has not been submitted for the award if any other degree of this or any other university.

Amanpreet Sip

(**Amanpreet Singh**)
Regn. No. 901411012

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Maninder Kaur

(**Dr. Maninder Kaur**)
Associate Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology (Deemed to be University)
Patiala, 147004
Punjab, INDIA.

Acknowledgements

First and foremost, I express my deep sense of gratitude to God, the almighty who gave me the patience and strength to complete this work. He has given me an opportunity to believe in my passion and pursue my dreams. I could never have done this without his divine blessings.

At this moment of accomplishment, I would like to give my special appreciation and sincere thanks to my honorable Supervisor, Dr. Maninder Kaur, Associate Professor, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed to be University), Patiala (India), for being a pillar of support and encouragement throughout my research work. Her guidance helped me in all the time of research and writing this thesis. Her experience, strength, tenderness and willfulness, has taught me valuable lessons of life, which are going to be of immense help to me in taking decisions in going forward. I could not have imagined having a better advisor and mentor for my Ph.D. study.

I convey my sincere expression of thanks to Dr. Maninder Singh, Professor and Head, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed University), Patiala (India), for providing me the necessary administrative assistance in completion of the work. I am thankful to my Ph.D. committee members, Dr. Inderveer Chana, Professor and Associate Head, Dr. Sanmeet Bhatia, Associate Professor, Computer Science and Engineering Department, and Dr. Rajesh Khanna, Professor, Electronics and Communication Engineering, Thapar Institute of Engineering and Technology (Deemed to be University), Patiala (India), for their constructive comments and regularly ensuring the progress of my research work. I am

thankful to all the faculty and staff members of Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed University), Patiala (India), for their support.

A very special gratitude goes out to University Grants Commission (UGC), Government of India, New Delhi for providing financial support for the work, in the form of UGC NET-JRF/SRF.

I am forever indebted to all my family members, my loving mother, Mrs. Ranjit Kaur and my father, Mr. Bharpur Singh whose love and affectionate blessings have been a constant source of inspiration in making my vision a reality. I am also thankful to my wife Mrs. Kamalpreet Kaur, my brother Mr. Jatinder Singh and my sister-in-law Mrs. Daljit Kaur for their love, motivation, encouragement and their confidence in me. Lastly, it is my lovely and lively nieces Harleen Kaur and Inderjeet Kaur who have been the source of my liveliness and always cheering me up during the entire phase of Ph.D. I also acknowledge the cooperation and encouragement extended to me by my friend Dr. Sukhpal Singh Gill, Lecturer, Queen Mary University of London.

Patiala

February, 2020 (Amanpreet Singh)

Abstract

The recent development of social media poses new challenges to the research community in analyzing online interactions among people. Social networking sites offer great opportunities for connecting people with each other, but also increase the vulnerability of young people to undesirable phenomena, such as content-based cybercrime. This may cause many serious and negative impacts on a person's life and even lead to committing suicide. Cybercrime has emerged as a money-driven industry with malicious intent towards online social networks. Cyber-criminals aim to manipulate vulnerable areas in cyber-space by playing on human understanding and making a profit. They threaten minors, especially adolescents, who are not adequately overseen whilst online.

In the recent past, the issues of Content-based Cybercrime have gained considerable attention. Social media providers seek for accurate and efficient way of recognizing offensive content for shielding their users. Content-based Cybercrime detection is one of the conspicuous area of data mining that deals with the recognition and examination of bully contents usually presented in social media. The current work emphasizes on cyberbullying, one of the prominent problems that arose due to the increasing fame of social network and its fast acceptance in our day-to-day survives. The social network provides a convenient platform for the cyber predators to bully their preys especially targeting young youth. In severe cases, the victims have attempted suicide due to humiliation, insult, and hostile messages left by the predators.

To address this issue, there is an urgent need for a robust content based cybercrime detection framework. This thesis proposes three techniques for efficient detection of content-based cybercrime in online social networks. First one, cuckoo inspired SVM approach, aims to concurrently optimize the parameters and feature selection with a target to build the quality of SVM. This chapter proposes a novel hybrid model that is the integration of Cuckoo Search and SVM, for feature selection and parameter optimization for efficiently solving the problem of content-based cybercrime detection. In second approach, multiconfiguration detection technique, has been proposed to explore possible combinations of various preprocessing, feature selection and classification methodologies using the cuckoo search metaheuristic approach. This approach seeks to improve the performance of content based cybercrime detection system. In third approach, a novel cuckoo inspired stacking ensemble framework has been proposed that is the integration of cuckoo search and several machine learning models. The proposed framework automatically seeks for near-optimal combinations of classification techniques along with their tuning parameters for efficiently solving the detection problem of content-based cybercrime in multimedia platforms.

The performance of the proposed approaches has been evaluated by testing on four different datasets obtained from Twitter, ASKfm and FormSpring to identify bully terms. The results of the proposed approaches demonstrate significant improvement in the performance of classification on all the datasets in comparison to recent existing models. The experimental results demonstrate the high efficiency and effectiveness of the proposed approaches. These approaches outperformed other recent techniques on all the datasets, giving high predictive recall value via 10-fold cross-validation.

Chapter 1

Introduction

Advancement in Internet technology has impacted the relationship among people and their communication to a greater extent (Kraut *et al.* (1998)). Internet services are a growing part of human interaction, and progress in mobile and network technology has enhanced the ability for individuals to connect (Shih (2007)). The Internet has changed the majority aspects of a human lifetime: education, entertainment, politics, and so on. It affects someone's mood: they feel connected, happy, loved, lonely, depressed, scared, and so forth. Maybe not willingly, but undoubtedly our lives have become interwoven with the Internet. Nowadays, online social networking platforms allure people with a wide range of age groups. In the past few years, online social networking has become very popular. Social networking provides humanity a vast and convenient platform for exchanging their ideas, perceptions all around the world. There is a continuous growth in the size of social networks. Figure 1.1 depicts the count of active users (in millions) on various social networking sites worldwide as of May 2019. A thousand millions of users throughout the globe are using one or more social networking sites, with the count increasing rapidly at a fast pace. This count includes every age group, whether young or adult, and both males and females. Out of this huge count, nearly millions of individuals are habituated to these social networking sites. Different social networking sites like ASKfm, Formspring, and Twitter, etc. offer today's youth a platform for entertainment and pleasure (Harridge-March (2010)).

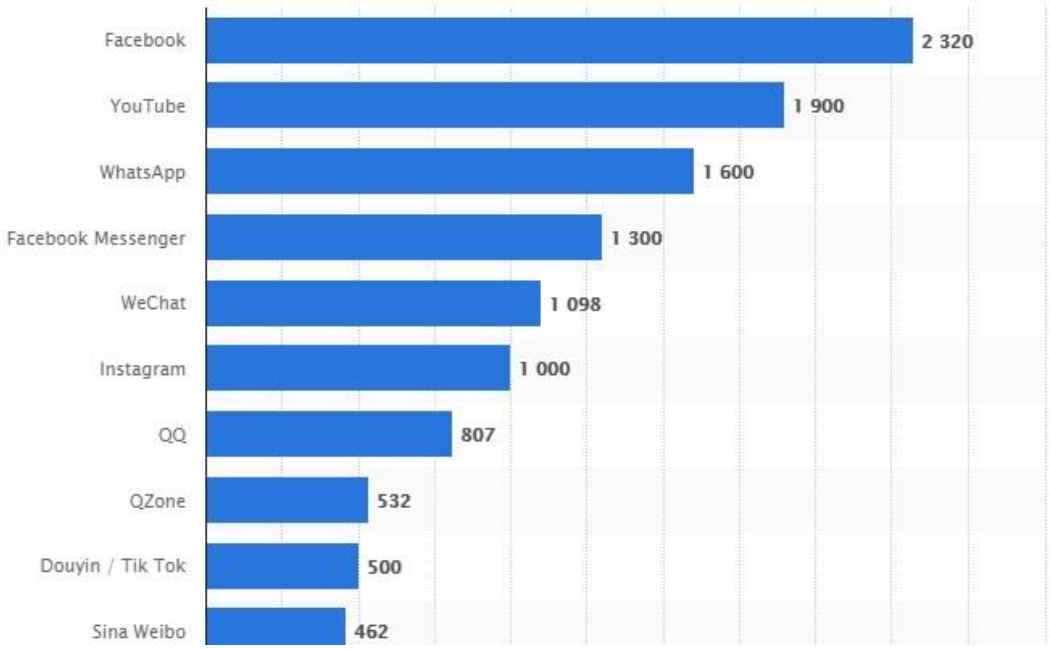


Figure 1.1: Social networking sites worldwide, ranked by number of active users (in millions) as of May 2019

At present, the friendships and relationship networks are shaped through a wide array of digital devices. The majority of daily greetings, friendly get-togethers, and family chitchats take place from behind a screen. A large number of users use online social networking sites as modernistic communication tools as well as a dynamic repository of data in real-time, providing users with the convenience of creating their profiles and communicate with their peers despite geographic sites. Online social networking platforms like Twitter, Facebook, WhatsApp, Tumblr, Forums, message boards, and blogs have become an important communication medium, especially for young people (Cooke and Buckley (2008)). Online social networking data can provide with innovative visions into the building of human societies and networks, earlier considered impossible in terms of extent and scale. Additionally, these online social networks can transcend the physical world's boundaries in studying human behaviors and relationships. Most of the time, people reach out to others for help, love, and friendship, but on the other side, hostility and hatred have also always been part of human culture, and they have a detrimental impact on societal history (Loia *et*

al. (2018)). Besides convenience and extreme openness, social media can be effortlessly utilized for spreading uncivilized and unsocial activities. The offensive wrongdoings and patterns of behavior driven by the darker sides of human nature can be observed in these virtual settings. The social media prompt the youth into a globe of harmful threats such as '*Cybercrime*'. These social media sites have gained significant proliferation and popularity with increased incidents of cybercrime. These social platforms have become a central hub for cybercrime victimization (Yar and Steinmet (2019)), escalating the coverage of cybercriminals to previously unreachable places and countries.

Social media has evolved into a prestigious platform that allows people to express their feelings and has the advantage of being able to hide and perpetrate acts of violence against peers such as dating aggression, harassment, bullying, and gang-related crimes. Social media has also been used as a self-harming agent, especially cyber-suicide (Hinduja and Patchin (2010)). Young people of the 21st century cannot live without the Internet and social media - for example, Facebook, Twitter. Over ninety percent of young people use the Internet every day, and about seventy percent have active accounts on one of the social networking sites (Sampasa-Kanyinga and Hamilton (2015); Finklea and Theohary (2015)). Face-to-face and verbal violence are still more widespread than virtual violence. Research suggests that most children and adolescents (65–91%) report little or no involvement in violence through social media platforms (Griffin and Gross (2004)). However, violence in the virtual world is a growing problem that requires additional research and prevention. Research (Patchin and Hinduja (2006)) shows that violence in social media and cyberbullying are becoming more and more common. The study included a random sample of 4,441 teenagers aged 10-18 from 37 schools. It was found that nearly 20 percent of teenagers in 2010 claimed to have been victims of cyberbullying, and 20 percent said that at one point in their lives, they experienced bullying in cyberspace. The number of cybercrimes on social

media platforms such as Facebook, Twitter, etc. is increasing at a fast pace. A statistical investigation of cybercrimes around the world shows that America, China, Germany, Britain, and Brazil are at the top of this list (Park *et al.* (2017)). Communication on social media is characterized by a higher degree of anonymity, which provides honest opportunities to increase hostility in interpersonal interactions (Kowalski *et al.* (2012)). Cyber-attackers are always curious about making the best use of online networking sites to fulfill their craving by the act of online bullying, phishing, spreading malware, etc. As the Internet becomes more accessible, the number of social media users is increasing day by day. Due to this rise in the online population, cybercriminals have become hidden in the depths of the Internet and feel safe in the crowd.

1.1 Cybercrime

Cybercrime is viewed as one of the most hazardous fears for the expansion of any vulnerable situation; it has a severe influence on every facet of the development of a state. Administration bodies, non-profit governments, remote industries, and residents are all probable targets of the cyber-criminal crowd. The "*cybercrime industry*" operates correctly as authentic administrations working on an international level, with safety experts approximating the standard measure of misfortunes to be processed in the demand of billions of bucks each year. The term '*Cybercrime*' encompasses any illegal action that hurts victims using computer, transmission, data, or applications (Jahankhani *et al.* (2014)). Cybercrime can expand crime in the physical world. It also appeared purely as a result of the convenience and efficiency of the Internet.

1.1.1 Categorization of Cybercrime

Cybercrime is classified into two categories: technology based and content based cybercrime. Any particular terrorist group associated with sexual harassment, fear, child pornography, national

security, demoralization, etc. accomplishes the content based cybercrime. The technology based cybercrime includes hacking, incidents of espionage, injecting malicious code (Thangiah *et al.* (2012)). Figure 1.2 shows the taxonomy of cybercrime with some examples. The folks tangled in both categories should have some skilled consciousness. The cyber crooks generally inclined to live in different types of the globe and relish receiving the honor of several nations.

For executing different types of content based cybercrimes, cybercriminals have exploited social media to its fullest extent in the form of spamming, cyberbullying, the spread of malware and phishing (Kim *et al.* (2011); Thangiah *et al.* (2012); Dewang and Singh (2018)). Out of these cybercrimes, content based cybercrime has been a significant and dire issue ever since the emergence of the Internet. One of the main content based cybercrime problems is cyberbullying, where the target subjects are under-age victims. Specifically, cyberbullying has risen as a noteworthy issue alongside the ongoing improvement of online correspondence and social networks.

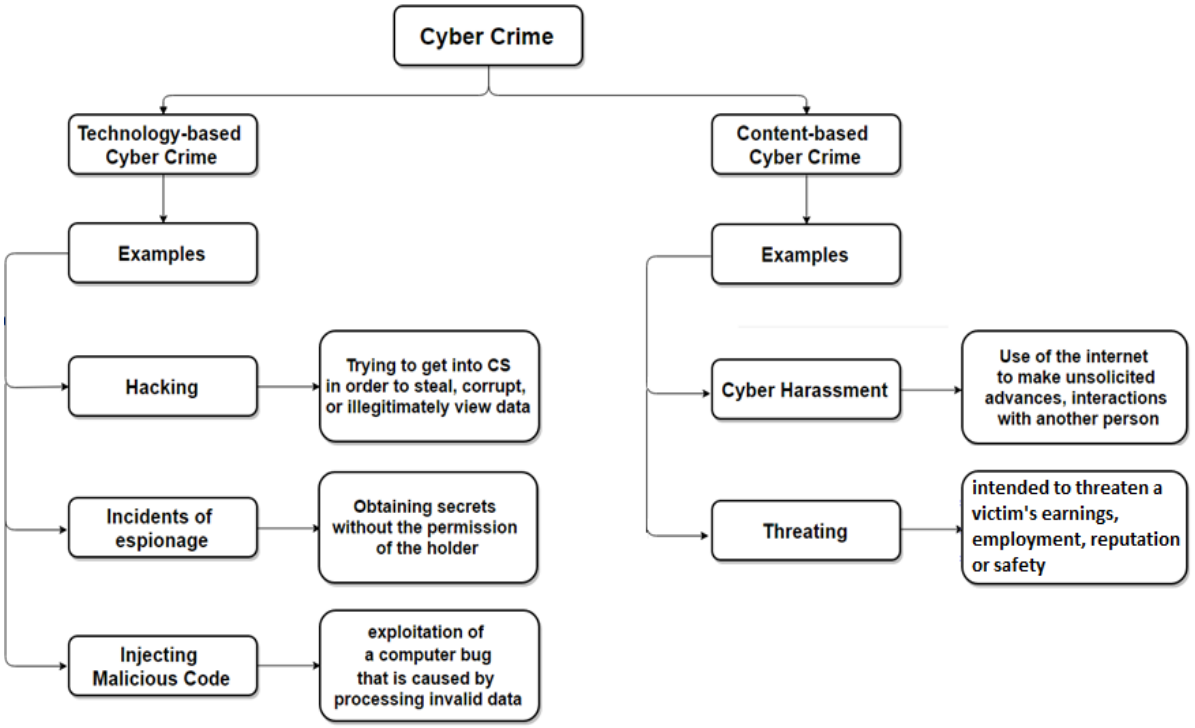


Figure 1.2: Taxonomy of cybercrime with examples

1.2 Content Based Cybercrime: Cyberbullying

In recent years, cyberbullying is identified as a severe nationwide health issue, in which victims considerably raise the grave threat of suicide (Aboujaoude *et al.* (2015)). Cyberbullying is one form of online misbehavior that has deeply affected society with harmful consequences. Traditional bullying used to be a demonstration of dominance and consolidation of social status by making use of physical power and creating fear and discomfort for those who were weaker and vulnerable. With the development of online technology, bullying has also emerged in cyber societies. Cyberbullying can be stated as an intentional action that is conducted through digital technology to hurt someone (Patchin and Hinduja (2006)). Unlike traditional bullying, which was inherently limited to streets and schoolyards, the wide variety of technological devices used in daily lives has also brought cyberbullying into people's homes and bedrooms. The authors (Hinduja and Patchin (2014)) stated the definition of cyberbullying as "the activity of harming or harassing someone via the internet or social networks in a repeated and deliberate manner". These may incorporate transferring messages bearing the hurtful text, harassing the victim in an online community. Cyberbullying has become a prevalent problem among young people on the Internet, affecting their emotional and psychological well-being. Victims may also suffer from severe neurotic tendencies such as self-harm, suicidal ideation, and depression (Hinduja and Patchin (2010)). A study revealed online social networking sites like Twitter, ASKfm, Formspring, etc. being turned into a "*cyberbullying playground*" (Xu *et al.* (2012)). On these sites, platforms are also provided where users experience bullies, bystanders, or bullying as victims. A survey conducted by the ditchthelabel.org, an anti-retaliation charity nationwide in 2013, revealed that there are two in three of the victims of cyberbullying who are in the age group of 13-22 years old. According to the literature (Juvonen and Gross (2008)), 43% of the young people were once bullied

through social media. Also, it has been shown that the rate of cyberbullying victim cases lies between 10-40%. The more damage it causes is the bullying messages that have been kept for a long time on the Internet, resulting in more severe and far-reaching consequences of cyberbullying. Cyberbullying can have a more rigorous and sustained impact than physical bullying; physical bullying is restricted to geographic and temporal constraints, but the Internet allows cyberbullying to go far beyond these limitations in specific online environments where cyberbullying occurs. There is growing evidence of adolescents and children using online social networks for bullying (Sourander *et al.* (2010)). Online bully content spreads quickly and has a broader audience (Sticca and Perren (2013)).

There has been an evident number of life-threatening experiences due to cyberbullying, especially among youths throughout the world (Li (2006)). The recent studies show that this problem is more exaggerated in the USA, where about 43% of teens are the targets of cyberbullying (Bonanno and Hymel (2013)). It is consequently apparent that the readiness of tools that can distinguish potential practices categorized as cyberbullying can be extremely valuable to avert circumstances of "threat" to the prey. Regardless of whether the issue is presently vigorously reflected from a social perspective, computational examinations in this field are, to a great extent, yet unexplored and just a couple of explores on cyberbullying are available.

1.2.1 Background and Motivation

Bullying is an acutely commonly observed act in adolescents. Bullying increased drastically in all age groups when it got exposure to technology. In cyberspace, to bully someone, a person does not have to be physically strong. A person who has never bullied anyone in real life can also anonymously starts bullying on social media. Cyberbullying is sufficiently threatening and

destructive that it can lead victims to contemplate suicide, and, in the worst scenarios, it can result in suicidal attempts (Hinduja and Patchin (2010)) and cause life-long mental damage to victims. Numerous deadly cyberbullying encounters have been accounted for globally, consequently drawing consideration towards its harmful effect.

Academics have recognized the importance of the early detection of such activities. On average, 20% to 40% of all teenagers have been mistreated online, as suggested by research reports (Nixon (2014)). With appropriate detection of possible harmful messages, successful prevention can be achieved. There is a critical necessity to dig into cyberbullying in the context of its prevention and detection. The current study focuses on a detailed literature review of the detection of content-based cybercrime comprising cyberbullying. However, there is a requirement for intelligent systems to identify possible risks automatically. This is what encouraged this study to control bullying by detecting it in different scenarios so that the people can take initiatives to end it. The proposed system efficiently detects bullying in online social media like Twitter, Formspring, and ASKfm that can help people to diminish bullying and help the victims to get rid of it.

1.2.2 Definition of Cyberbullying

There is a lack of conceptual clarity in the definition of cyberbullying, and the distinction among different types of cyberbullying is often vague (Vandebosch and Van Cleemput (2008)). Several definitions of cyberbullying are suggested in the literature, and all of them somehow refer to an aggressive and harmful act that is conducted through an electronic device. However, these definitions can be distinguished through their details, such as those who are involved in the incident (groups and individuals) and requirements for being deliberate and repeated overtime (Tokunaga (2010)). Table 1.1 presents some of the definitions of cyberbullying suggested in the literature.

However, the authors (Dehue *et al.* (2008)) indicate that a situation must meet three conditions to be considered as cyberbullying; the act should be intentional, be repeated over time, and should involve psychological torment.

Table 1.1: Definition of cyberbullying in several studies

Literature	Definition
Slonje and Smith (2008)	*An aggressive, intentional act carried out by a **group** or **individual**, using electronic forms of contact, **repeatedly** or **overtime** against a victim who cannot easily defend him or herself.*
Tokunaga (2010)	*Cyberbullying is any behavior performed through electronic or digital media by **individuals** or **groups** that **repeatedly** communicates hostile or aggressive messages intended to inflict harm or discomfort on others.*
Patchin and Hinduja (2006)	*Willful and **repeated harm** inflicted through the medium of electronic text.*
Juvonen and Gross (2008)	*The use of the Internet or other digital communication devices to insult or threaten someone.*

However, this definition has aspects which cannot be fully covered when it is considered in experimental settings for studying cyberbullying from a technical perspective. Cyberbullying can happen through different modalities. It can happen through posting nasty videos about someone or publicly uploading private pictures without having the consent of their owner. Cyberbullying through text is one of the most common mediums in which vulgar comments are posted, and threatening and foul messages are sent to the victim. Some various algorithms and tools can automatically detect and remove bullying posts, or trigger some administrator's follow-up action

in response to online bullying incidents. For example, the repetitiveness of the act cannot always be determined as part of events that may happen in private conversations that are not accessible. Moreover, the balance of the power between the victim and the bully cannot be easily verified by just analyzing the content of the bullying incidents.

1.2.3 Categorization of Cyberbullying

Bullying is usually defined as a subcategory of aggressive behavior. It is characterized by repetition over time and an imbalance of power between bully and victim (Smith *et al.* (1999)). In the 1980s, bullying was mostly seen as direct face-to-face physical (such as hitting) and verbal (such as teasing) attacks (Slonje and Smith (2008)). During the 1990s, the scope of bullying has also been broadened to include indirect aggression, such as spreading rumors, and relational aggression, for example, by damaging someone's relationships. In recent years, with the development of technologies and the growth of Internet use, a new form of bullying has emerged, called cyberbullying. Cyberbullying is a general term that also refers to similar constructs, such as online bullying and Internet harassment. There are different categories of common cyberbullying (Willard (2007); Beran and Li (2008)), as shown in Figure 1.3:

- Flaming: Sending rude and vulgar messages to a group or person.
- Outing: Posting private information (picture, phone number, etc.) or manipulated/photo-shopped personal materials of an individual without her or his consent.
- Harassment: Repeatedly sending insulting messages or emails to a person.
- Exclusion: Excluding someone from participating in an online group.

- Impersonation: Fantasizing to be another person for directing out materials on her or his behalf.

- Cyberstalking: Terrorizing someone by sending threatening and intimidating messages.

- Denigration: Spreading online gossips about a person.

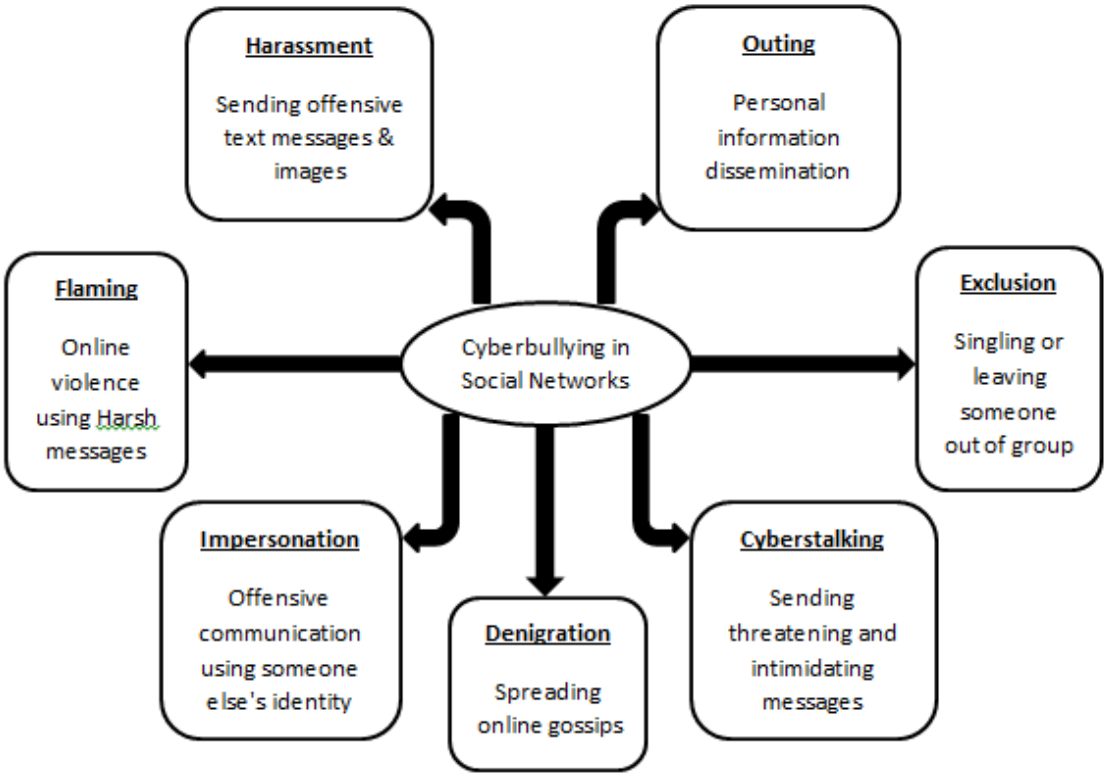


Figure 1.3: Various ways of cyberbullying in online social networks

1.2.4 Components of Cyberbullying

Cyberbullying consists of several components. These components affect how the bullying takes place, and consequently, the studies conducted on cyberbullying differ depending on the components involved. Table 1.2 illustrates the status of the components distinguished in each phase.

- The fundamental component is the people, called **actors**, involved in the incident. The actors can be grouped into the following three categories:

  - **Bully**: the person who intentionally uses obscenity, threat, or aggression to impose domination or cause fear and distress in others.

  - **Victim**: the person who is targeted by the bully. Victims cannot easily defend themselves and are usually vulnerable to the imbalance of power between them and the bully.

  - **Bystander**: the person who witnesses the incident but is not directly involved in the process. The bystanders can provide support for the victim by posting positive feedbacks for the victim and reacting against the bullies. They can also escalate the distress caused by the bullies by supporting their actions.

- The **platform** in which cyberbullying takes place is another influential component in the process, and therefore it should also be taken into consideration in the studies. Online social networks are the primary communication platforms. An online social network is a web-based platform to build social relations among people with similar interests and activities. Social networks introduce each of their members through her/his page (profile), which mostly contains personal information and benefits of the user. Networks also provide means for users to interact over the Internet, for example, through email and instant messaging. Social network sites are varied, and they offer different activities such as photo and video sharing, posting comments, and following the actions of others in the network. In some cases, part of the dynamics comes from the presence of a monitoring function that could help to discourage bullying behavior.

- Another component is the **content** and the modality through which the bullying takes place. As explained earlier, cyberbullying can happen through videos, pictures as well as through posting hurtful and offensive textual contents.

Table 1.2: Cyberbullying components and actions in pre- and post-bullying phases

		Pre-Bullying	**Bullying**	**Post-Bullying**
Actors	Bully	To be monitored		To be identified/ To be warned or to be excluded from the network
	Victim	To be trained To be educated		To be identified/ To receive support
Bystanders		To be alerted To be monitored		To be alerted/ To be monitored
Platform		Exclusion of risky user profiles		Identification of bullies and victims. Follow-up actions, e.g., organizing help after incident, alerting of bystanders, removing offensive
Content		Previously analyzed content to be used to identify risky user profiles		Bullying content to be detected, offensive content to be deleted

Some of the examples of cyberbullying incidents are given below.

Example 1. Tweet: "*the scariest thing is that you seem to have little idea of what you are talking about and yet people actually listen*"

In the above tweet there is no swear-keyword used by the user (sender), but still this message is an insult to a 2nd person as indicated by "*you*". The aim of the proposed system is to identify all such messages which may or may not contain swear-keywords.

Example 2: Examples from bullying traces datasets are given below (Xu *et al.* (2012)):

- *Reporting a bullying episode: "some tweens got violent on the n train, the one boy got off after blows 2 the chest... Saw him cryin as he walkd away :(bullying not cool"*
- *Cyberbullying direct attack: "Lauren is a fat cow MOO BITCH"*

1.2.5 Phases of Cyberbullying

In traditional bullying, the moment at which the bullying takes place can be recognized. The kicking, cursing, and biting are evident indicators that signal the moment of bullying. Therefore, the social studies on the bullying problem can easily be divided into those who propose preventive training and awareness-raising programs for the stages before the bullying happens, and those who provide support and guidance for the consequences of bullying after an incident. Unlike what is the case in traditional bullying, it isn't straightforward to determine the exact moment in which cyberbullying takes place. Therefore, in technical studies on cyberbullying, such divisions have not been considered. The problem of cyberbullying can be split up (possible solutions and precautions related to cyberbullying) according to the two main phases of the entire chain of activity and reaction: the pre-bullying phase and the post-bullying phase. In the study of measures addressing the pre-bullying phase, the main concentration is on prevention strategies. In contrast, in the study of measures addressing the post-bullying phase, the focus is on the detection of bullying incidents after they have happened. Computational models for the detection of risky user-profiles typically require information on previous cyberbullying events. Note that to come up with alerts suggesting an action that could be taken to stop or decrease future harmful acts by a bully, the pre-bullying models need input from the models for the detection of cyberbullying incidents, which are applied in the post-bullying phases.

1.2.6 Impact of Cyberbullying

Studies show that in European countries, about 18% of the children have been involved in cyberbullying via the Internet or mobile phones (Görzig and Frumkin (2013)). A survey conducted in Britain shows that 25% of adolescents between 12 to 19 years old have experienced cyberbullying (Park *et al.* (2004)). The National Crime Prevention Council reported in 2011 that cyberbullying is a problem that affects almost half of all-American teens.

The consequences of cyberbullying are similar to traditional bullying and have been shown to include depression, low self-esteem, and in cases, even ending up to suicide attempts (Schenk and Fremouw (2012); Bottino *et al.* (2015); Sabella *et al.* (2013); Pham and Adesman (2015)). However, in some cases, the consequences of cyberbullying can be more severe and longer-lasting due to some specific characteristics of cyberbullying. Cyberbullying can be undertaken 24 hours a day, every day of the week, and unlike traditional bullying, it is independent of place and location (Nocentini *et al.* (2010)). Moreover, online bullies can stay anonymous, and being bullied by an unknown person can be more distressing than being bullied by someone familiar (Görzig and Ólafsson (2013)). Furthermore, anonymity triggers cyberbullying behavior for people that would not bully face-to-face (Dooley *et al.* (2009)). Online materials spread very fast, and in a couple of minutes, thousands of Internet users can see whatever that goes online. There is also the persistence and durability of online materials and the power of the written word. In the case of cyberbullying through text, the targeted victim and bystanders can read what the bully has said over and over again. Also, in the case of images, the harmful content can stay online for an extended long period, and if tagged with the name or other personal features of the victim, it will keep showing up in the results of searches.

1.3 Thesis Contributions

To tackle with the problem of cyberbullying from different perspectives, a variety of methods have been proposed by researcher worldwide. Among these methods, the researchers are mainly focusing on machine learning approaches to help combat cyberbullying. From the survey, it is concluded that there are a few contemporary types of research in cyberbullying detection, which can provide high precision and recall rates. None of the existing research work used evolutionary computation for content-based cybercrime detection. This work attains an unbiased criterion and assess the efficiency of the classification process in the context of cyberbullying detection. The major contributions of the thesis work are:

- The Cuckoo inspired SVM approach, multiconfiguration detection technique, cuckoo inspired stacking ensemble framework have been proposed and results achieved validate the effectiveness of the proposed work.
- The CS-based models have been used to detect cyberbullying by automatically selecting near-optimal classification schemes.
- Different empirical experiments are conducted on four real-world, publicly available online social network datasets to verify the effectiveness of the proposed frameworks.
- The proposed framework is an up-and-coming model for detecting content-based cybercrime in terms of evaluation metric of classification i.e., recall. The proposed framework exhibit increased recall rate in given various datasets by determining the optimal values of SVM tuning parameters / classification schemes.

1.4 Thesis Organization

The thesis has been organized in the following chapters:

Chapter 1: Introduction: This chapter provides an overview of content-based cybercrime in online social networks as well as the effects it can have on victims, along with its challenges and related applications.

Chapter 2: Literature Review: This chapter provides a comprehensive literature review of research efforts in the field of content-based cybercrime detection in online social networks. A comparative study of the existing content-based cybercrime detection models has also been provided in this chapter.

Chapter 3: Cuckoo Inspired SVM Approach: This chapter discusses a novel hybrid model that is the integration of Cuckoo Search (CS) and Support Vector Machine (SVM), for feature selection and parameter optimization to efficiently solve the problem of content-based cybercrime detection. The proposed model aims to concurrently optimize the parameters and feature selection with a target to build the quality of the SVM classifier.

Chapter 4: Multiconfiguration Detection Technique: This chapter discusses a novel technique that explores possible large combinations of various preprocessing, feature selection, and classification methodologies using the cuckoo search metaheuristic approach. The model seeks to improve the performance of a content-based cybercrime detection system.

Chapter 5: Cuckoo Inspired Stacking Ensemble Framework: This chapter proposes a novel cuckoo inspired stacking ensemble framework that is the integration of CS and several machine learning models. The proposed framework automatically seeks for near-optimal combinations of

classification techniques along with their tuning parameters for efficiently solving the problem of content-based cybercrime detection.

Chapter 6: Conclusion and Future Scope: The thesis concludes with this chapter by highlighting the contributions made by the proposed research work. It also provides an insight into the future directions for working in this area.

Chapter 2

Literature Review

This study includes research efforts on content-based cybercrime detection in online social networks. The review incorporates articles printed over the last decade, beginning through the innovative initiative of Yin *et al.* (2009). The extent of the work included in the survey highlights the increasing concern received in recent years by cyberbullying prevention. Although supervised learning approaches dominated the methods considered by many studies, are scholars have shown readiness to use new effort from different fields of Natural Language Processing (NLP) to improve the performance. In this section, we review existing research in the field of content-based cybercrime detection techniques.

2.1 Cyberbullying Detection: A systematic review

The section discusses previous studies from eight perspectives that are: used methodology, conclusions or findings, demerits, the dataset used, preprocessing steps used, content-based features used, models or technique, and evaluation metrics used, as illustrated in table 2.1 and table 2.2. For searching, electronic literature is explored through Scopus, IEEE Xplore virtual library, and the ACM Digital Library. The key search method was related to the subject “detection of content-based cybercrime, unsocial behavior and harassment” without considering publication year as a filter.

More papers were discovered using the citations in observed articles via the article's references as an initial topic, thereby exploring 58 academic papers in total as a result of the search. At the very first stage, the titles, abstract, and concluding arguments of the discovered papers were reviewed for assessing their relevance, and 18 papers were discarded, as they were not found to apply to the survey. In the next phase, the detailed review of the whole text of the remaining 40 articles was carried out, and further six more papers were filtered out that did not focus on cyberbullying detection. This process led to 34 papers in the final list of papers for the survey. These papers under consideration dealt with different subjects such as story coordinating to recognize upset teenagers, youth violence participation recognition, cyberbully inhibition strategies.

The survey targeted on the abstract view of the work done in terms of data sources, availability of the datasets, detection techniques, features extracted, evaluation parameter used, and pre-processing steps. Tables 2.1 presents a summary of key statistics abstracted from the reviewed research. It delivers a concise outline of the methodology used, conclusion/findings, demerits and dataset used, types, and recognition tasks for each of the 30 papers. It is revealed from the survey that the most common task executed in cyberbullying detection is the binary classification. In this context, the text containing bully terms is entitled as the member of "bullying" class, and the message without bully terms categorized as "non-bullying" type. The significant job after this is the recognition of documents that own the essential features of the "bullying" type. The 30 research papers out of 34 research reviewed, targeted binary classification solely as a cyberbullying detection task or as a hybrid with other tasks. Herein, the text classification is generally expedited by supervised learning systems.

Table 2.1: Various methodologies for detecting content based cybercrime

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Yin *et al.* (2009)	Used supervised machine learning approach in which features (local, contextual, and textual) of Documents are utilized to learn an SVM Classifier.	Addition of the sentiment and contextual features provide significant performance to basic model which uses only Local feature.	Dataset is of stand-alone posts, pragmatics of conversation are not considered, only supervised learning techniques. Predators and victims were not identified.	Datasets of Slashdot, Kongregate, MySpace websites.
Dinakar *et al.* (2011)	Used supervised machine learning approach in which binary & multiclass Classifiers classify bullying sensitive topics.	Label-specific classifiers (binary) are more effective than multiclass classifiers at detecting content-based cybercrime.	Dataset is of stand-alone posts, pragmatics of conversation are not considered, only for supervised learning techniques. Predators and victims were not identified.	Youtube comments from several videos after clustering into sexuality, race & culture.
Bayzick *et al.* (2011)	Proposed a rule-based system called BullyTracer and also developed a truth-set of MySpace threads to check accuracy of proposed system.	Correctly identify 85.3% as cyberbullying posts and 51.91% as innocent posts of MySpace dataset.	Falsely flag a lot of innocent posts as cyberbullying. Only uses rule-based system, unsupervised or supervised learning technique was not used.	Thread-style forum transcripts crawl from MySpace.
Reynolds *et al.* (2011)	Supervised machine learning approach in conjunction with labelled data was used to learn the system to identify bullying content.	Model was capable to recognize 78.5% posts in Formspring dataset that have cyberbullying in a small sized sample.	Only for supervised learning techniques, dataset is of stand-alone posts, predators and victims were not identified, pragmatics of conversation are not considered.	18554 user's data which contain 1 to 1000 posts is used.

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Dinakar *et al.* (2012)	Used common-sense knowledge base with associated reasoning techniques in addition to machine learning classifiers.	In the task of detection of textual cyberbullying, binary classifiers outperform multiclass classifiers.	Other aspects of the problem like dialogue and pragmatics of conversation did not considered.	Manually Labelled corpus of Formspring & youtube data.
Nahar *et al.* (2012)	Proposed a sentimental analysis technique for cyberbullying content detection by using PLSA (feature selection method).	Finds the Most Influential persons (predator or Victims) using HITS Algorithm.	Not focused on Indirect Cyberbullying.	Kongregate, MySpace, Slashdot datasets
Nahar *et al.* (2013)	Proposed a session-based framework which incorporated an ensemble of one-class classifier and addressed the real-world scenario where just minimal set of positive instances were given.	Effectively classifies bully instances using session-based one-class ensemble classifier which uses small set of labelled data and huge unlabelled data.	Baseline swear-keywords method can be incorporated along to improve the accuracy.	Datasets of Twitter, Kongregate, MySpace, and Slashdot
Dadvar *et al.* (2013)	Proposed effective technique which used a combination of user-based, content-based and cyberbullying-specific features.	Evaluation Parameter gives best result when all features used in combination.	Some features such as user profile, channel subscribed can be taken into account.	Youtube comments on 3 top videos for variety of topics.
Nahar *et al.* (2014a)	Proposed semi-supervised learning approach using fuzzy SVM classifier.	This technique was suitable in real-world situation handling noisy, imbalanced or streaming data and outperformed all other methods.	Severity levels of bullying messages were not taken into consideration. Also feature space used is static.	Kongregate, Slashdot, MySpace datasets

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Nahar *et al.* (2014b)	Proposed semi-supervised learning in the session-based framework that incorporates an ensemble of one-class classifiers.	Results indicated that in real world situations, the proposed approach performed very well, where for initial training only a few positive instances of cyberbullying are available.	Not focused on Indirect Cyberbullying, user-based features are not used.	Datasets of Kongregate, Slashdot, MySpace and Twitter websites
Huang *et al.* (2014)	Proposed a technique by integrating social network features with the textual features to detect cyberbullying.	Outcomes showed that new attributes (social) are beneficial in identifying cyberbullying. Proposed model also detects the most influential persons.	Not focused on Indirect Cyberbullying, Only for Supervised Learning techniques.	Datasets of Twitter website
Mangaonkar *et al.* (2015)	Proposed collaborative paradigm that used different machine learning techniques for bully or non-bully data classification.	Without much tuning of algorithms, collaboration techniques worked better than sequential methodology in terms of time consumed and accuracy.	Lack Very less features are used to train machine learning algorithm.	Datasets of Twitter website
Van Hee. *et al.* (2015)	Presented the annotation and construction of a corpus of posts from Dutch social media (ASKfm) and explored the feasibility of automatic cyberbullying detection.	By exploring the automatic cyberbullying detection model's feasibility, the results presented that when more fine-grained categories are taken into consideration the detection of cyberbullying is not a trivial task.	Some features such as syntactic patterns, semantic information can be taken into account. Also, author role information can improve in cyberbullying detection.	Datasets of ASKfm website

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Hosseinmardi *et al.* (2015)	Collected an Instagram data set sample consisting of comments associated with images, and developed a tagging study for image content in addition to cyberbullying using human labellers at the crowd-sourced Websites (like Crowdflower).	Proposed model identified that there is significant class of media sessions of Instagram that exhibits cyber aggression but not cyberbullying.	More features and detailed labelling surveys can improve accuracy.	Datasets of Instagram website
Al-garadi *et al.* (2016)	Suggested a feature-based classifier for detecting cyberbullying using supervised machine learning in the Twitter media.	SMOTE + Random forest using proposed features, showed the best results in detecting cyberbullying.	Other social media data and social networking graph can be used to investigate cyberbullying behaviour.	Datasets of Twitter website
Galán-García *et al.* (2016)	Proposed a methodology for detecting forged Twitter profiles and offered an efficacious real-world use case.	PolyKernel SMO model outperformed in terms of AUC.	More NLP techniques can be used to improve the accuracy, further data from other social media can be used.	Datasets of Twitter website
Singh *et al.* (2016a)	Proposed a framework (probabilistic information fusion) that utilizes interdependencies associated with different textual and social features, their confidence score, and uses those for better cyberbullying predictors.	Proposed fusion approach provides better results in detecting cyberbullying using heterogeneous textual and social features.	Only emphases on a specific social network (Twitter), more sophisticated features can be used.	Datasets of Twitter website

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Zhao and Mao (2016)	The authors used semantic extension of deep learning model stacked denoising autoencoder for developing Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA).	Proposed method was capable of utilizing the concealed attributes from the bully data and train the distinguishing and strong illustration of text.	By considering order of words in messages could improve the robustness of learning model.	Datasets of Twitter, MySpace websites
Zhao *et al.* (2016)	A learning method was proposed for detection of cyberbullying by concatenating bullying, latent semantic and BoW features together.	Capture Semantic Information behind words. LSVM is capable of detecting text containing bullying.	Dataset is of stand-alone posts, Only for Supervised Learning techniques.	Datasets of Twitter website
Dani *et al.* (2017)	Proposed a framework based on sparse learning by integrating user-post relationships and sentiment information.	Experimental results showed the impact of sentiment information on two real-world datasets as well as effectiveness of the proposed model.	Other languages can be incorporated as well. The effect of the sarcasm facts concealed in the posts can be investigated.	Datasets of Twitter and Myspace website
Raisi and Huang (2017)	Proposed a weakly supervised Participant Vocabulary Consistency (PVC) model using machine learning model for simultaneously inferring new vocabulary indicators of bullying and user roles in molestation-based bullying.	Proposed model was analysed on datasets from diverse social media based on qualitative and quantitative evaluation.	Network features or the sequence of conversations can be considered to improve accuracy.	Datasets of Askfm, Instagram, and Twitter website

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Singh *et al.* (2017)	Designed a predictive classifier based visual features (e.g. portrayed emotions, nudity, race, gender etc.) for automatic cyberbullying detection.	The usage of visual features outperformed textual features in detecting cyberbullying by improving accuracy.	Huge extent of pictorial processing APIs in addition to visual content can be considered for better detection of cyberbullying.	Datasets of Instagram website
Agrawal and Awekar (2018)	Developed four models based on DNN (i.e. BLSTM, LSTM, CNN and BLSTM) for cyberbullying detection in online social sites.	Proposed model systematically analyses detection of cyberbullying based on different themes through numerous SMPs by means of transfer learning and deep learning-based classifier.	Additional information such as data about the users' social graph and their profile could further improve model performance.	Datasets of Formspring, Twitter and Wikipedia websites
Dadvar and Eckert (2018)	Proposed deep learning-based models also evaluated and transferred the model's performance trained on one platform to another platform.	The proposed models based on deep learning outperform the models based on machine learning models by applying on Wikipedia, Twitter, Formspring and YouTube dataset.	Profile info of the social sites' users can also improve the model's accuracy.	Datasets of Wikipedia, Twitter, Formspring and YouTube
Rafiq *et al.* (2018)	Developed a cyberbullying detection system for media-based social networks, consisting of a dynamic priority scheduler, a novel incremental classifier, and an initial predictor.	The proposed system drastically reduces the time to raise alerts and the classification time. Without sacrificing accuracy, it was very receptive in raising alarms and greatly scalable.	Investigate the plateauing effect that limits the effectiveness of adding more memory.	Datasets of Vine

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Van Hee *et al.* (2018)	Proposed an automatic detection model for cyberbullying in social sites by modelling texts written by bystanders' preys, and bullies of online bullying.	Experiments reveal that proposed method was a promising approach for detecting signals of cyberbullying automatically in social sites data and outperforms a word n-gram and keyword-based baseline.	Fine-grained categories related to cyberbullying such as hate, racism expressions, curses and threats can be detected.	Datasets of ASKfm Posts in English and Dutch
Cheng *et al.* (2019)	Proposed XBully, a framework for detecting cyberbullying, that initially re articulates multi-modal data from social network and then targets to train node-embedding illustrations upon it.	Broad experimental outcomes on real-life datasets validate the efficacy of the proposed framework and find that multi-modal data can suggest valuable visions for depicting and detecting cyberbullying behaviours.	Building a deeper understanding of various modalities in depicting cyberbullying behaviours will advance detection of cyberbullying.	Datasets of Instagram and Vine
Yao *et al.* (2019)	Developed efficient timely and scalable semi supervised methods that extrapolate from a small seed set of expert annotations as and also seeks to drastically reduce the number of features used while maintaining high classification accuracy.	The experiments showed the feasibility of the proposed framework, with up to 62.17% of improvement in accuracy and 67.86% reduction in terms of timeliness comparing to baseline approaches.	Evaluate the performance of proposed approach on additional datasets from diverse platforms.	Dataset of Instagram

Table 2.1 Continued…

Author(s)	Methodology	Conclusion/Findings	Open Issues	Datasets Used
Balakrishnan *et al.* (2019)	Proposed a cyberbullying detection model using users' personalities. Random Forest was used to classify users into one of four roles, namely, bully, aggressor, spammer or normal, based on 9484 tweets.	Detection of cyberbullying improved when all the personalities were incorporated. The identification of the key personalities from Big Five and Dark Triad show that, although every one of these traits impact cyberbullying collectively.	Detecting cyberbullying victims as well as bystanders. The proposed study can be improved by incorporating more users' personalities.	Dataset of Twitter
Zhang *et al.* (2019)	Develop an automatic cyberbullying detection model by extracting multiple textual features and investigating their effects with multiple machine learning models.	Multiple textual features and multiple machine learning algorithms significantly improve the classification quality.	An emotion dictionary specialized for cyberbullying and corresponding to the expressions on Twitter or newly emergent words can be constructed.	Dataset of Twitter

Table 2.2: Summary of related work in the field of content based cybercrime detection (NA: means Not Applicable)

Literature	**Preprocessing**	**Content-based Features Used**	**Dataset Used**	**Models/ Methodology**	**Evaluation Metrics**
Yin *et al.* (2009)	Stemming, Tokenizing, Positive instance replication	TF-IDF	MySpace, Slashdot, Kongregate	SVM	Recall, F-score, Precision
Dinakar *et al.* (2011)	Stop words removal, Stemming, Unimportant character removal	N-gram, TF-IDF weighted unigrams, Profanity	YouTube	J48, SVM, NB, JRip	Accuracy, Kappa
Bayzick *et al.* (2011)	NA	Second person pronouns, Swear word, Insult word	MySpace	NA	True/ False (Positive/ Negative)
Reynolds *et al.* (2011)	Unimportant words removal, Frequent words removal	BoW	Formspring	IBk, JRip, J48, SMO	Recall, Precision, Accuracy
Dinakar *et al.* (2012)	Tokenizing, Removal of stop words	Profanity, TF-IDF, Weighted unigrams, BoW	Formspring, Youtube	NB, Tree-based learner, SVM	Accuracy, Kappa
Nahar *et al.* (2012)	Web addresses, Re-tweet Swear-keywords, Hash tag, Least or most frequent words, Stop words removal	TF-IDF unigrams	MySpace, Slashdot, Kongregate, Twitter	Ensemble	Recall, Precision, F-score

Table 2.2 Continued… (NA: means Not Applicable)

Literature	**Preprocessing**	**Content-based Features Used**	**Dataset Used**	**Models/ Methodology**	**Evaluation Metrics**
Nahar *et al.* (2013)	Stemming, Removal of stop words	BoW	MySpace, Slashdot, Kongregate	SVM	F-score, Accuracy
Dadvar *et al.* (2013)	Stemming, Stop words removal	Profanity, Emoticons, Pronouns, Message length, N-gram, Bully keywords	Youtube	SVM	Recall, Precision, F-score
Nahar *et al.* (2014a)	NA	Special Characters, Profanity, Pronouns, Capitalization	MySpace, Slashdot, Kongregate	K-FSVM, LR, RF, NB	Recall, F-score Precision, Accuracy
Nahar *et al.* (2014b)	Lower casing letters, Stop word, Most and least frequently used words	TF-IDF unigrams	MySpace, Slashdot, Kongregate, Twitter	Ensemble	Recall, Precision, F-score
Huang *et al.* (2014)	NA	Capitalization, Punctuation, Profanity, Emoticons	Twitter	ZeroR, SMO, NB, J48,	True positive rate, ROC
Mangaonkar *et al.* (2015)	Tokenization	N-gram	Twitter	SVM, LR, NB	Recall, Precision, Accuracy

Table 2.2 Continued… (NA: means Not Applicable)

Literature	**Preprocessing**	**Content-based Features Used**	**Dataset Used**	**Models/ Methodology**	**Evaluation Metrics**
Van Hee *et al.* (2015)	Lemmatization, Tokenization, PoS-tagging	Character trigram BoW, Bigram BoW, Word unigram	ASKfm	LSVM	F-score
Hosseinmardi *et al.* (2015)	Unimportant character removal, Stop words removal	Number of posts within interval less than one hour, Number of comments for the image, n-gram	Instagram	LSVM	Recall, Precision, Accuracy
Al-garadi *et al.* (2016)	Spelling correction, Lowercase conversion, Removal of white spaces	First and Second person, Profanity	Twitter	KNN, RF, SVM, NB	F-score, AUC, Recall, Precision
Galán-García *et al.* (2016)	Tokenization	TF–IDF, N-gram	Twitter	NB, KNN, RF, J48, SMO	AUC, TPR, FPR, Accuracy
Singh *et al.* (2016a)	Tokenization	PoS tags, Question marks, Density of uppercase letters, count of exclamation points, count of smileys, Density of bad words	Twitter	SMOTE	Recall, Precision, Accuracy, F-score
Zhao and Mao (2016)	Tokenization	Profanity, Ensemble BoW	Twitter	LSVM	Recall, Precision, F-score

Table 2.2 Continued… (NA: means Not Applicable)

Literature	**Preprocessing**	**Content-based Features Used**	**Dataset Used**	**Models/ Methodology**	**Evaluation Metrics**
Zhao *et al.* (2016)	URLs replaced by predefined characters, Tokenization	Cyberbully keywords, Profanity, BoW	MySpace, Twitter	LSVM	Accuracy, F-score
Dani *et al.* (2017)	Stopwords removal, Stemming	Bigrams, List of profane words	MySpace, Twitter	SICD	F-score, AUC
Raisi and Huang (2017)	Stop word, Retweets, punctuation removal, Duplicate tweet, URL emojis	N-gram (n=1, 2)	Twitter, Instagram, Ask.fm	PVC	Precision
Singh *et al.* (2017)	Unimportant character, Stopwords removal	Number of dashes, Tentativeness, Sadness, Anger in comments, Positive emotions, Word count	Instagram	SMOTE	Accuracy, ROC Area
Agrawal and Awekar(2018)	NA	NA	Formspring, Twitter, Wikipedia	Deep learning	Recall, Precision, F-score
Dadvar and Eckert (2018)	Stopwords removal, Punctuations	NA	YouTube, Formspring, Wikipedia, Twitter	Deep learning	Recall, Precision, F-score
Rafiq *et al.* (2018)	Tokenization	Negative comments, Total negative words, Unigrams	Vine	AdaBoost, LR	Recall, Precision, F-score

Table 2.2 Continued… (NA: means Not Applicable)

Literature	**Preprocessing**	**Content-based Features Used**	**Dataset Used**	**Models/ Methodology**	**Evaluation Metrics**
Van Hee *et al.* (2018)	Abbreviations, White spaces, Tokenization removal, Hyperlinks	Character n-gram BoW, Word n-gram BoW	Posts of ASKfm in Dutch and English	LSVM	AUC, Precision, F-score, Recall, Accuracy
Cheng *et al.* (2019)	NA	NA	Vine, Instagram	LR, LSVM, RF	Micro F1, Macro F1
Yao *et al.* (2019)	Tokenization	Unigrams, profane unigrams and bigrams	Instagram	KNN, SVM, RF	Recall, Precision, Accuracy, AUC
Balakrishnan *et al.* (2019)	NA	Number of (user mentions, hash tags, followers, following/friends, status), Popularity	Twitter	RF	Precision, Recall, F-measure
Zhang *et al.* (2019)	URLs, newline characters, unimportant twitter-specific terms removal	Word n-gram, Character n-gram, Word2Vec, Doc2Vec, Emotion values of tweets, Twitter-specific characteristics, Character n-gram	Twitter	LSVM, LR, Decision Tree, RF	Recall, Precision, Accuracy, F-score

2.2 Review outcomes

There are various electronic media for cyberbullying– such as SMS, MMS, chat rooms, forums, Email, and social networking sites (YouTube, Twitter, Snapchat, and Facebook, etc.). The major work in the literature targeted social as a primary source of data because of its free access in the community zone. The SMS, emails, chatrooms, and MMS are the private ways of correspondence, and communications through these e-media are more reluctant to be freely accessible. The graph shown in Figure 2.1 illustrates the measurable level of the utilization of different online social media datasets for content-based cybercrime detection practices in social media. It is clear from the figure that the majority of work targeted Twitter and MySpace as the most common data sources in cyberbullying detection. Various studies including the work of Nahar *et al.* (2013); Nahar *et al.* (2014b); Mangaonkar *et al.* (2015); Singh *et al.* (2016); Zhao *et al.* (2016); Agrawal and Awekar (2018); Dadvar and Eckert (2018), and many others used data from twitter whereas the work of Bayzick *et al.* (2011); Raisi and Huang (2017); Van Hee *et al.* (2018), among others, utilized MySpace. Slashdot and Kongregate are in the third position with Yin *et al.* (2009); Nahar

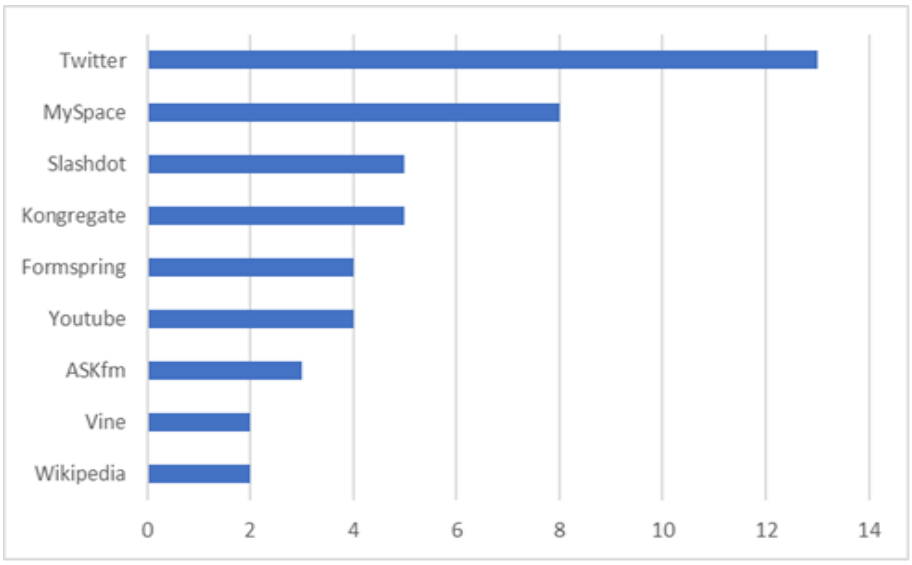


Figure 2.1: Usage of various datasets

et al. (2013); Nahar *et al.* (2014b); Dadvar and Eckert (2018) using messages from YouTube data. Table 2.2 depicts the insights of various methodologies used for detecting content based cybercrime in terms of main characteristics briefly described in the following subsections.

2.2.1 Content based features

Due to the extremely subjective aspect of cyberbullying detection, it is possible to have diverse effects of the same text on different folks, and it is a challenging task to find the results these during detection time. In this direction, the literature heavily used content based attributes like the occurrence of spelling, pronouns, document length, profanity, etc. The corpus and detection techniques contribute to the effectiveness of these features. This study emphasized the work utilizing content-based features that are extractable vocabulary terms of a corpus such as punctuations, profanity keywords, pronouns, etc. Figure 2.2 portrays the quantifiable range of the usage of several content based features for content based cybercrime recognition on online social networks.

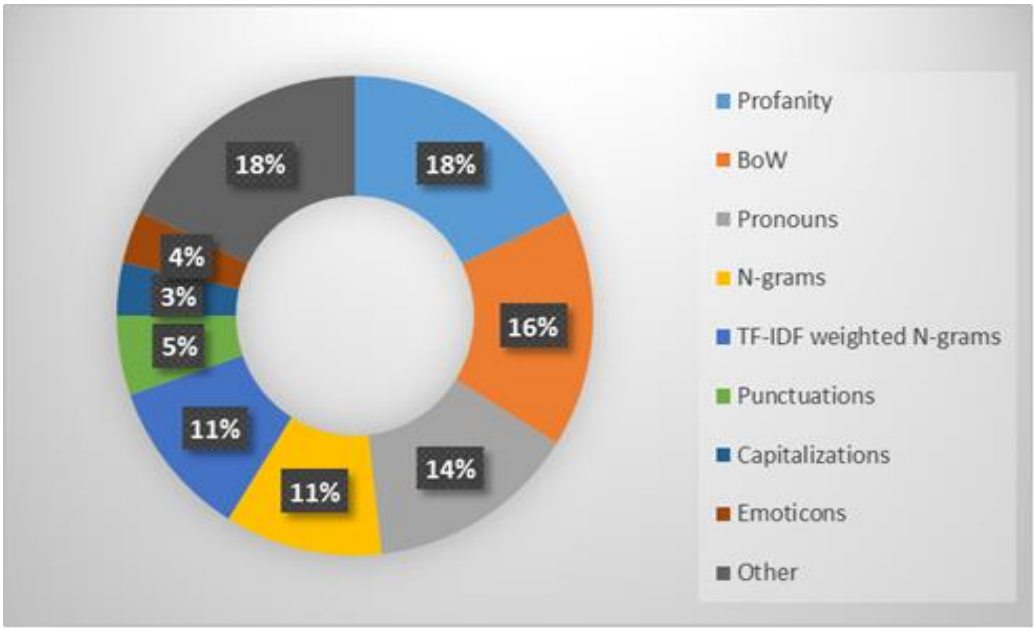


Figure 2.2: Exploitation (as %) of various content based features

In this review, the content based Features are grouped as profanity, pronouns, cyberbullying keywords, n-grams, TF-IDF, BoW, spelling, and document length. The cyberbullying texts generally incorporate insulting and abusive language. In this study, a total of eight papers out of 30 contains profanity term in the text as a sign of cyberbullying.

2.2.2 Data preprocessing

It is the initial step to lessen noisy data, thus enhancing the correctness of the system. It may be a two-edged weapon as the useful content may vanish from the corpus during the preprocessing phase. For example, converting uppercase text to lowercase; inadvertently may result in losing the context as capitalization is generally used to signify uproar in written communication. In the articles reviewed, 22 papers from the considered sample accomplish the preprocessing step. Herein, stemming and tokenization are the most frequently used preprocessing phases. Stemming is generally performed on a text with BoW and n-gram TF-IDF as a feature set.

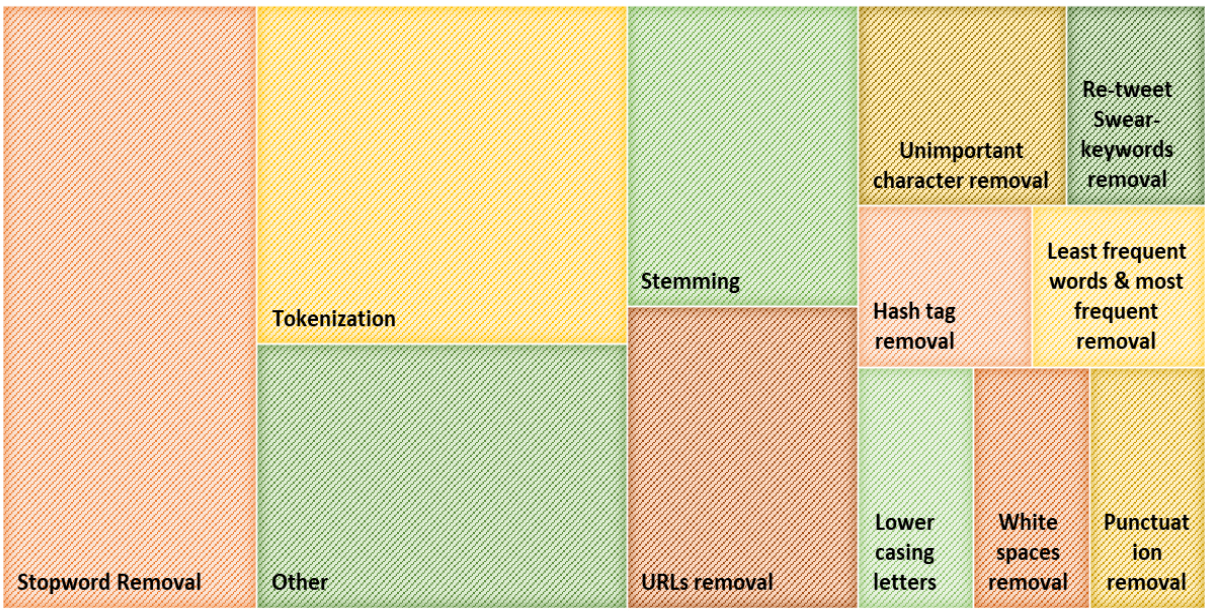


Figure 2.3: Quantitative extent of the use of various preprocessing methods

Through stemming, the importance of stemmed words within the datasets is highlighted by collapsing stemmed words into a term/stem. The tokenization step splits the text into an order of distinct words, thereby representing a document as a set of its phrases. Figure 2.3 depicts quantitative extent of the use of various preprocessing methods. The studies done in our survey include other essential preprocessing tasks such as stopword removal that seem, by all means, to be of little importance to the area being referred to. In some cases, the stopword elimination can also unintentionally remove significant terms. It would be better to initially find whether the stopwords are utilized infrequently used sentences before their removal.

2.2.3 Techniques for cyberbullying detection

From the survey, it is revealed that the majority of the work involved supervised learning techniques in the area of detection of bully content. The work of Yin *et al.* (2009) was the original work founded during topic exploration in cyberbullying detection. Figure 2.4 represents the quantitative extent of the use of various evaluation parameters. The key evaluation parameters in supervised learning are accuracy, precision, recall, and f-score. Nevertheless, the work in many papers provided experimental results using these metrics; still, these studies cannot be compared directly due to the usage of a diverse set of datasets. Moreover, the research works that conducted their experimental work on the same dataset are inclined to use to extract different samples from the same dataset.

Figure 2.5 depicts the application of several models/techniques organized year wise for content-based cybercrime detection in online social sites. It is clear from the figure that SVM and NB are the commonly used classification techniques for attaining effective outcomes for the detection of bully content on social networks by LSVM, RF, LR, J48, SMO, deep learning, and

other techniques. The graph shown in Figure 2.5 portrays the quantifiable range of the usage of several practices for cyberbullying detection in social media platforms.

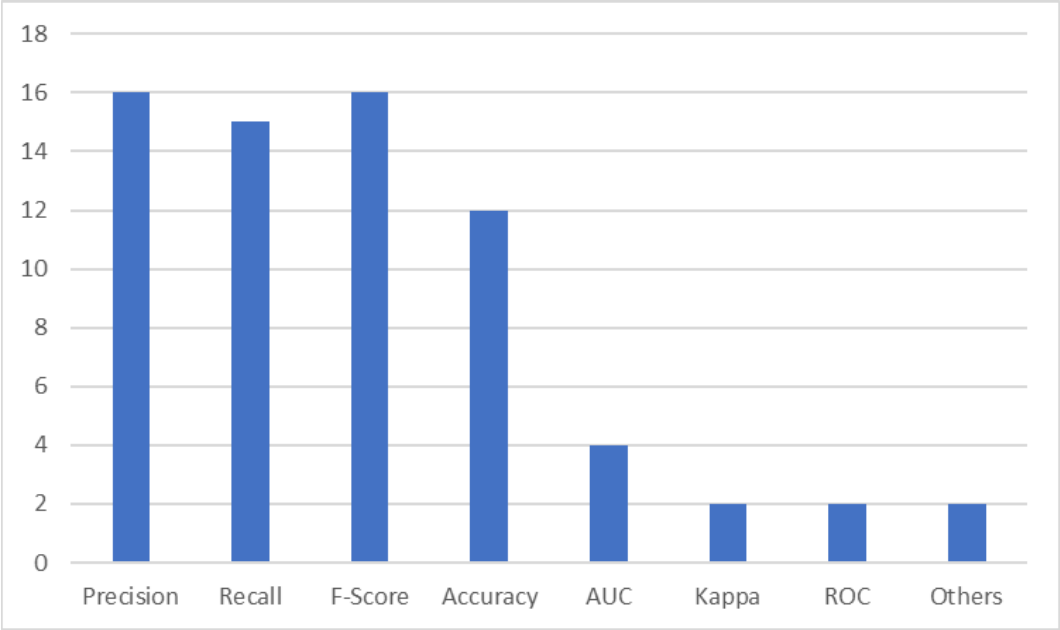


Figure 2.4: Usage count of various evaluation parameters

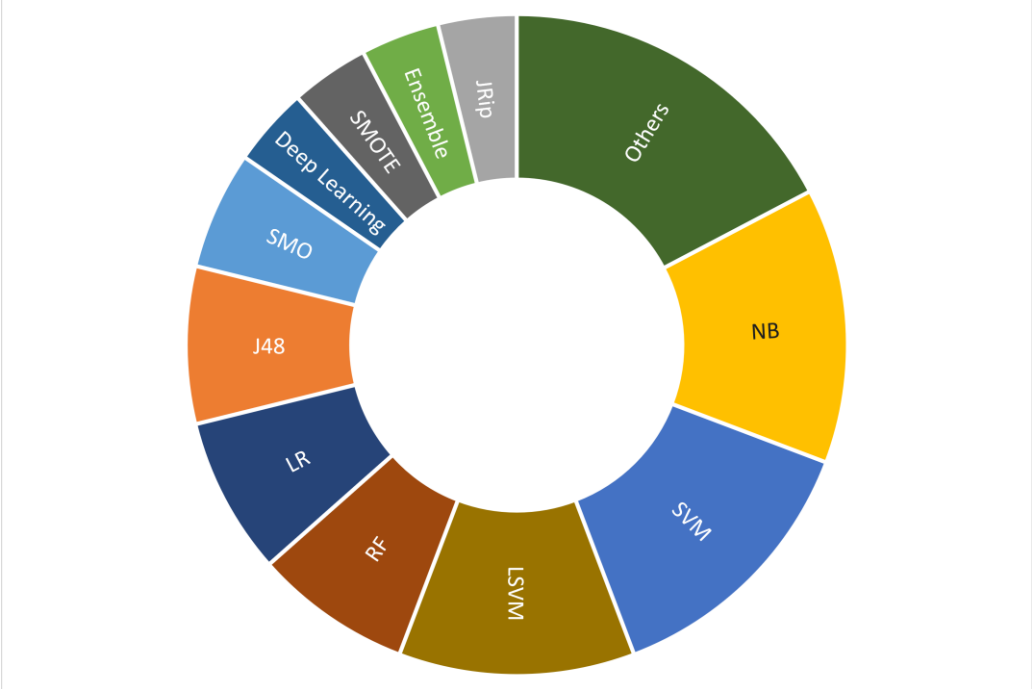


Figure 2.5: Proportion of usage of various machine learning methodologies

Figure 2.6 depicts the year-wise count of published work in the field of content-based cybercrime detection in online social media. It is clear from the figure that there is progress in the research work in the field of cyberbullying detection from 2009 onwards, and this growth is increasing at a faster pace.

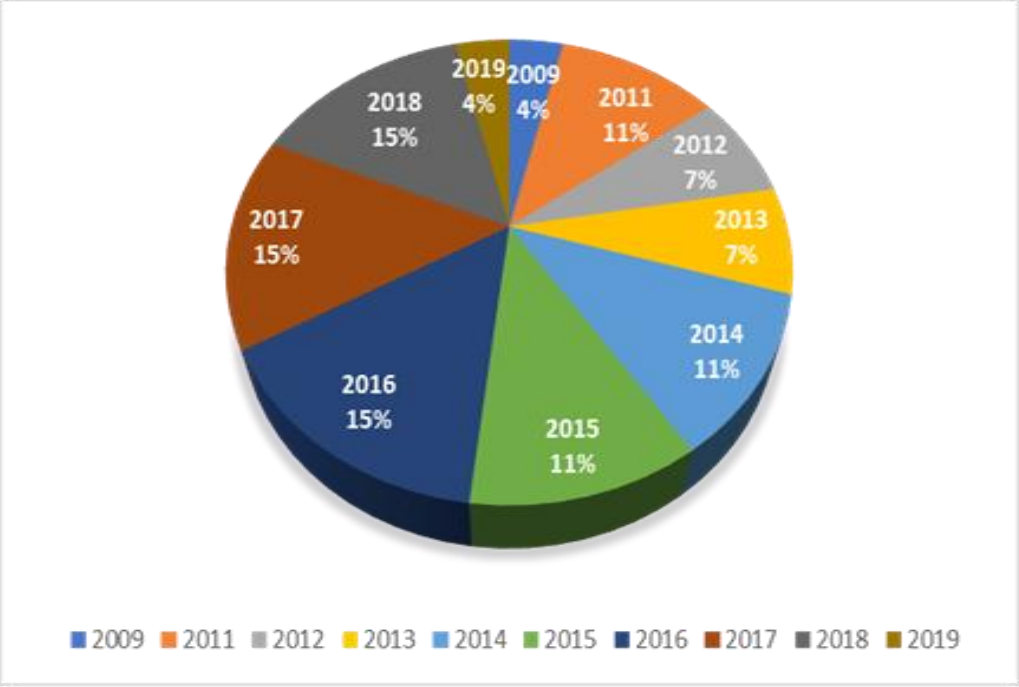


Figure 2.6: Year-wise cumulative assessment of published work

2.3 Research Gaps

To tackle the problem of cyberbullying from different perspectives, there are numerous works available in the literature in the field of content based cybercrime detection involving a varied number of techniques such as SVM, NB, J48, Decision Tree, Bagging, DNN, etc. Further, in this area, the primary focus is on feature engineering, i.e., finding features that can separate bullying comments from non-bully ones. Finding suitable features is still a difficult and challenging issue.

The existing techniques considered a classification model with a peculiar set of preprocessing steps and feature selection techniques, which may not lead to the best solution. The reason for the same is that these schemes are technique-dependent. The automatic selection and combination of techniques in alternative ways could improve the overall accuracy of the prediction models for specific datasets. Various researchers have identified a bunch of problems in this area of content based cybercrime detection and have tried to offer the answers as well. However, the proposed solutions have not been accepted universally and are still left unanswered. The features that work well for Youtube comments may not work well for comments on ASKfm due to different social media platforms being likely to have different vocabulary and expressions caused by restrictions on communication, different age groups, and users' interests. The current work identifies and provides a brief overview of such research issues as follows:

- Several cyberbullying detection models have been evaluated in the literature, and the results of these studies are often highly technique-dependent. The issue of which modeling technique to use for cyberbullying detection remains an open research question.

- There is a need for an exploration operation that must be able to identify promising features from a large set of features. The exploration operation must be computationally efficient to avoid performance degradation of the entire classification process.

- It has been shown in the literature that changing the tuning parameters of classifier affects classification accuracy. For this purpose, an optimization algorithm is required that seeks an optimal or near-optimal value for the classifier parameters. To address this issue, a metaheuristic optimization algorithm is used in the proposed methodology.

- A few research works have been done in the field of classification ensemble, and one crucial issue while building ensemble is how to select classifiers such that the decision-making quality of the ensemble is superior to that of any individual classifier. But no one considered this issue in their research work.

- How to assign reasonable parameters to each base-classifier considered in the ensemble learning and how to combine "good and different" base-classifiers is still an open research issue in the current field.

2.4 Problem Statement

The problem relies on identifying the presence of content based cybercrime and classifying cybercrime activities in a social network such as Flaming, Harassment, Racism, Terrorism, and Bullying using learning and classification techniques, which helps subsequently to take preventive measures to combat cybercrime.

The effectiveness of the proposed method to detect content-based cybercrime activities on social media is computed based on the following evaluation parameters and on obtaining a precise type of cybercrime activity. Various evaluation parameters are (i) Precision: The total number of correctly identified true bullying posts out of retrieved bullying posts. (ii). Recall: Number of correctly identified bullying cases from the total number of true bullying cases. (iii). F-1 measure: the harmonic mean of precision and recall.

2.5 Objectives

The objectives of this research work are:

1. To study, explore & analyze various approaches for detecting Content based Cybercrime in Online Social Networks (OSNs).
2. To design and develop a framework for detection of Content based Cybercrime in Online Social Networks.
3. To verify and validate proposed work in context with existing approaches targeting precision, recall & F-1 measure.

Chapter 3

Cuckoo inspired SVM Approach

This chapter discusses a novel hybrid model that is the integration of Cuckoo Search (CS) and SVM, for feature selection and parameter optimization for efficiently solving the problem of content-based cybercrime detection. The proposed work aims to concurrently optimize the parameters and feature selection with a target to build the quality of the SVM classifier.

3.1 Introduction

The detection of cyberbullying is generally formulated as a binary classification problem that involves the distinction of posts containing bully and non-bully text. The post containing bully text is represented as the positive class, while the negative one comprises the post containing non-bully text. Only a few recent datasets for the detection of content based cybercrime have been made publicly available (Zhao *et al.* (2016)). Subsequently, several studies work together with the former or created their social media corpus that are prone to content of bullying, such as ASKfm (Yin *et al.* (2009); Bayzick *et al.* (2011); Van Hee *et al.* (2018)), Formspring (Reynolds *et al.* (2011); Dinakar *et al.* (2012); Agrawal and Awekar (2018)), and YouTube (Dinakar *et al.* (2011); Dadvar *et al.* (2013)). Various machine learning methods are applied to detect cyberbullying, such as PCA (Zhao *et al.* (2016); Zhao and Mao (2016)), fuzzy logic (Dinakar *et al.* (2012)), decision tree (Dinakar *et al.* (2011); Dadvar *et al.* (2013)), KNN (Reynolds *et al.* (2011)) and SVM (Zhao

et al. (2016); Zhao and Mao (2016); Van Hee *et al.* (2018)). The literature (Zhao *et al.* (2016); Van Hee *et al.* (2018)) reports that SVM is an effective approach among these methods due to following reasons:

(i) SVM is a margin-based classifier, which has admirable generalization capabilities for learning a small number of samples due to which it is frequently used in real-world classification applications (Yu and Kim (2012); Maldonado *et al.* (2020)).

(ii) SVM has demonstrated its efficiency and robustness in classifying cyberbullying messages, and it is widely used in the methods of detecting cyberbullying as a popular method. However, the SVM parameters have a significant impact on classification accuracy (Hsu *et al.* (2003)).

The SVM classification accuracy can be enhanced by the proper setting of tuning parameters (C and γ) for the Radial Basis Function (RBF) kernel. For this purpose, the grid search algorithm may be used to find the best C and γ. The method exhaustively calculates the accuracy of n-fold cross-validation for each combination from a specific area of these two parameters. In the case of feature selection, the grid search algorithm utilizes the exhaustive feature selection approach for brute-force evaluation of feature subsets. The process subsequently leads to extensive calculations, making the algorithm more time consuming for parameter optimization and feature selection ask. Various search techniques such as heuristic search, greedy search, random search, complete search, etc. have been applied to feature selection (Bhattacharya and Das (2010); Reynolds *et al.* (2011); Singh *et al.* (2016b); Yadav *et al.* (2018); Zhao and Mao (2016)). However, most existing techniques of feature selection still suffer from high computational cost and stagnation in local optima (Liu *et al.* (2011)). Therefore, to better solve these problems, an adequate global search method is needed. Evolutionary Computation (EC) methods are recognized for their global search

capabilities. The CS, an optimization algorithm, has a powerful ability to search for an optimal solution in the space of candidate solutions.

The motivation of this work is to design an intelligent model using the Cuckoo Search algorithm and SVM classifier (CS-SVM) to detect content-based cybercrime in online social networks. The current work is a first step towards the application of the CS-SVM model in the field of content-based cybercrime detection. The CS-SVM model is a novel and efficient classification scheme in terms of parameter optimization of SVM and feature selection. The principle goal of current work is to enhance the performance of SVM by optimizing its C and γ parameters using the CS algorithm because CS outperforms other meta-heuristic algorithms as proved in the existing literature (Yang (2010)). The right choice of the parameters of SVM classifier and optimal utilization of different training dataset subsets by feature selection are the main issues targeted by the CS algorithm. In the proposed work, the recall evaluation parameter is considered for the design of the objective function to explore the maximum generalization ability of SVM. The performance of CS-SVM model is compared with three recent models applied on four different existing recent datasets from Twitter (Zhao *et al.* (2016); Agrawal and Awekar (2018)), ASKfm (Van Hee *et al.* (2018)) and FormSpring (Agrawal and Awekar (2018)). As seen from literature survey that major research work has been done by using SVM as classification model. The main difference between the CS-SVM model and others using SVM model is that the CS-SVM explores optimal combinations of features and SVM tuning parameters, rather than assessing only pre-established combination. The novelty of this research is to assess the efficiency of the classification process and attain an unbiased criterion in the context of content-based cybercrime detection.

3.2 Preliminaries

3.2.1 Support Vector Machine

The SVM method targets to discover the decision boundaries in the search space, separating one class from another class. Great generalization ability and global optimization are the main advantages of SVM (Gunn (1998)). Further, it offers better solutions compared to existing techniques such as Refined Genetic Algorithm (RGA) and Artificial Neural Network (ANN) in classification problems. This method seeks to find an optimum direction of discrimination in the feature space; thereby giving excellent performance for high dimensional feature space. The kernel function inside SVM classifier, named RBF has two tuning parameter γ and C.

The value of γ parameter describes how far the specific training instance is affected (Chapelle *et al.* (2002); Yu and Kim (2012)). A high value means "close" and low means "far". The C value controls the tradeoff between maximizing gross profit and reducing training errors (Chapelle *et al.* (2002); Yu and Kim (2012)). It maintains the balance between the misclassification of the training samples and the simplicity of the decision surface. It assigns all instances of more extensive training as support vectors to model the degrees of freedom that determine lower values in an appropriate manner and select more samples for selection.

The performance of the SVM is mainly determined by values of γ and C parameters. The choice of SVM parameters is significantly important. It is not identified in advance what values of γ and C are best for a given problem; therefore, some parameter explorations should be carried out (Hsu *et al.* (2003)). For this reason, to select the optimal values of SVM tuning parameters, the CS algorithm has been used.

3.2.2 Cuckoo Search Algorithm

Parasitic activities of a certain cuckoo class are exceptionally fascinating. These birds put their eggs in the host bird's nest, and these eggs mimic the external uniqueness such as color and place of host eggs. If the host bird recognizes that these eggs are alien eggs, the host birds spit these alien eggs, or leave the nest, or make new nest elsewhere.

The authors (Yang and Deb (2009)) proposed a new meta-heuristic optimization algorithm based on parasite behavior of cuckoos called Cuckoo Search. In this algorithm within each nest, every egg signifies a solution. The target of the algorithm is to generate the new and significantly improved solution (cuckoo egg) for replacing not up to par solutions. CS algorithm is based on three basic principles.

- There is one egg per cuckoo that is laid in an arbitrarily selected nest.
- Distribute high-quality eggs (solutions) in nests to the following generations.
- There is a predetermined probability number ρ, the discovery of alien egg by the host, $\rho \in [0, 1]$.

For the sake of simplicity, to have new random solutions, the final statement can be found by substituting a percentage ρ of the 'n' nests with new nests. According to the probability ρ, the nests with the lowest quality eggs are changed with new ones. In each iteration, the solution will be updated by using a random walk. In CS, random searching is attained by Lévy flight. The Lévy flight comprises taking a succession of repeated random steps (Yang (2010)). From the mathematical perspective, it takes two consecutive steps to make random values using the Lévy flight (Sharma *et al.* (2013)).

- Steps Generation.
- Random Direction Choice.

For this purpose, Mantegna algorithm is used where L, the step length can be determined as shown in (3.1).

$$L = \frac{x}{|z|^{1/\omega}} \tag{3.1}$$

where ω is the scale parameter. In the current work the ω value is set to 1.5. x and z are obtained from normal distribution as shown in (3.2).

$$x \sim N(0, \sigma_x^2), \text{ and}$$

$$z \sim N(0, \sigma_z^2) \tag{3.2}$$

where σ_x and σ_y are calculated using the (3.3) and Γ denotes gamma function.

$$\sigma_x = \left\{ \frac{\Gamma(1+\omega) \times \sin(\pi\omega/2)}{\Gamma[(1+\omega)/2] \times \omega \times 2^{(\omega-1)/2}} \right\}^{1/\omega}, \text{and}$$

$$\sigma_z = 1 \tag{3.3}$$

The pseudocode of the cuckoo search algorithm is presented in Figure 3.1.

3.3 Proposed CS-SVM Approach

The proposed approach for the detection of content based cybercrime comprises mainly of three components.

- Preprocessing of datasets.

- Feature extraction by PCA.
- Application of CS for finding best features and optimal SVM Parameter values that will be further used to train the classifier.

BEGIN
Define objective function $P(n), n = (n_1, \ldots, n_d)^t$;
Initialize cuckoo's population of size s i.e. n_k $(k = 1, 2, \ldots, s)$;
WHILE (Stopping Criterion achieved **OR** $g <$ Maximum Generation)
Get a cuckoo (say k) and using Lévy flight, randomly generate a new solution
Evaluate fitness/quality, Q_k;
Randomly choose a nest/cuckoo's population among s.
IF $(Q_k > Q_l)$
Replace l by the new solution;
END
Use Lévy flight to generate new solution and Abandon a portion (P_a) of worse solutions
Keep the best quality Solutions and Rank them, Save Current best;
END WHILE
Post process results and visualization;
END

Figure 3.1: Pseudocode of the CS algorithm

Preprocessing is the initial stage of the proposed work that incorporates stopword elimination, stemming and tokenization. Thereafter, tokenization is performed using bi-gram model. The next step involves application of PCA for dimensionality reduction. After this feature selection and SVM kernel parameter optimization is performed using CS algorithm and then selected features and generated kernel parameters are fed into SVM model for further classification. The detailed description of various components of the CS-SVM is as follows.

3.3.1 Preprocessing Stage

The preprocessing phase is applied before feature extraction to preprocess the noisy raw data procured from social media, containing stopwords, unwanted words, etc. It includes the following components.

- Stopwords removal: The precision of dataset is improved by removing all stopwords e.g., the, a, as, who, is, etc. by matching them with acronym and deployed stopword dictionary.
- Stemming: It is done using porter stemmer to replace derivationally related forms of words with root words. The words not having suffix symbol as alphabet are eliminated.
- Tokenization: In the current work, the bi-gram model is used for splitting the document into independent terms.

3.3.2 Feature extraction by PCA

The dataset after the preprocessing phase contains a vast collection of inter-correlated features that may significantly add on to the complexity of data analysis. To get rid of this problem, the current work involves the application of PCA method (Wold *et al.* (1987)) that reduces feature space dimensionality by transforming the existing correlated features variables to linearly uncorrelated features (more compact) while keeping as much information as possible (Chawla *et al.* (2010)).

3.3.3 Application of CS for finding best features and optimal SVM tuning parameters

In the current work, SVM classifier with RBF kernel has been used. The RBF kernel works well in the case of non-linear relationship between the features and the class labels. The main idea of performing feature selection in the input data is to give the SVM classifier function with different input space having different dimensions of the problem. The underlying CS-SVM approach exploits the advantages of CS for feature selection, looking best feature subset in the search space. Another application of the CS-SVM approach is to search for optimal internal parameters of SVM classifier. The research conducted by authors (Valentini and Dietterich (2004)), revealed that

detailed tuning of the parameters C and γ of RBF kernel also provides more diversity and prevents overfitting. For this purpose, various combinations of C and γ are explored by CS algorithm. The detailed pseudocode of the proposed CS-SVM approach for feature selection and parameter optimization for content based cybercrime detection is given in Figure 3.2. The main sections of the pseudocode are as follows.

Population Representation

The underlying model starts with a population of solutions generated randomly. For this purpose, the solution representing a cuckoo egg in a nest is encoded in the format that contains all information pertaining to feature selection and SVM tuning parameters.

Each solution (cuckoo's egg) in the CS population is represented by a vector of n+2 size, wherein the first n patterns signify the selected features for the problem that are further utilized by the SVM classifier at its learning stage. The two remaining patterns represent the tuning parameters (C, γ) of the SVM model. The structure of solution is represented as below.

f_1	f_2	.....	f_n	c	g

where (f_1, f_2... f_n) substring showcase feature selection with (f_1 ...f_n) ∈ {0, 1}. The value at the respective index of the substring decides the involvement or elimination of that feature in the learning process. The value 1 state that the respective feature is selected and 0 represent the elimination of that feature from the final feature subset. Equation (3.4) is used to construct this binary vector, which provides Boolean lattice binary values that restrict the new solutions to binaries only:

$$Z(X_{l,k}) = \frac{1}{1 + e^{X_{l,k}}}, \text{and}$$

$$X_{l,k} = \begin{cases} 1, & \text{if } Z(X_{l,k}) < \hat{u} \\ 0, & \text{otherwise,} \end{cases} \tag{3.4}$$

where $\hat{u} \sim U(0,1)$ & $X_{l,k}$ represents the solution at any generation.

The last two dimensions of solution i.e. *c* and *g*, represent the values of *C* and γ (kernel parameters) respectively. Therefore, the parameters for SVM classifier, *C* and γ are mapped to corresponding ranges of $c \in [C_1, C_2]$ and $g \in [\gamma_1, \gamma_2]$, by the formula shown in (3.5).

$$c = C_1 + \text{random}(0,1) \times (C_2 - C_1), \text{and}$$

$$g = \gamma_1 + \text{random}(0,1) \times (\gamma_2 - \gamma_1) \tag{3.5}$$

where $\text{random}(0,1)$ function gives random number in the range of 0 and 1.

Generate New Population using Lévy Flight

New population (set of solutions) is obtained by randomly chosen l^{th} solution through Lévy flight by using (3.6).

$$\text{new\_}X_{l,k} = \text{old\_}X_{l,k} + \alpha \times L \times (\text{old\_}X_{l,k} - \text{optimal\_}X_{l,k}) \tag{3.6}$$

where $\text{new\_}X_{l,k}$ is the fresh solution obtained through Lévy flight; $old\_X_{l,k}$ is randomly chosen solution within the population; $\text{optimal\_}X_{l,k}$ represents the best solution generated so far; α is step size; and L is the Lévy flight vector or step length. After finding the fresh solution, fitness values of two solutions are estimated and the highest quality solution is kept.

Discovery of Alien Eggs

For every cuckoo egg in the population, the probability matrix is utilized for foreign eggs detection. The probability matrix is represented as in (3.7).

$$\rho_{l,k} = \begin{cases} 1, & \text{if random}(0,1) < \rho \\ 0, & \text{otherwise,} \end{cases} \tag{3.7}$$

where $\rho_{l,k}$ is probability of discovering alien eggs in the lth solution for the kth variable of cuckoo's dimension. The value of ρ is compared with the output of a uniform random number generator, $\text{random}(0,1)$, to determine whether local random walk is considered or not. After determining the discovering probabilities, new solutions are obtained using (3.8).

$$\text{new\_X}_{l,k} = \text{old\_X}_{l,k} + \text{LS} \times \rho_{l,k} \tag{3.8}$$

where $\rho_{l,k}$ is the probability matrix and *LS* is the matrix of local step size, which is obtained by using the formula (3.9).

$$LS = \text{random}(0,1) \times (\text{randomperm}_i(\text{Solution}) - \text{randomperm}_j(\text{Solution})) \tag{3.9}$$

where $\text{randomperm}_i()$ and $\text{randomperm}_j()$, shuffle the solution randomly. Finally, the existing and new objective function values are compared for each solution and the best solution enters in the next generation according to the simple rule as given in (3.10).

$$\text{new Solution} = \begin{cases} new\_\text{X}_{l,k}, & \text{if F}(new\_\text{X}_{l,k}) > \text{F}(old\_\text{X}_{l,k}) \\ old\_\text{X}_{l,k}, & \text{otherwise} \end{cases} \tag{3.10}$$

Termination Criteria

The creation of new solutions and the discovering of the strange egg's steps are repeated until a predetermined stopping criterion is fulfilled or maximum generations are achieved.

PSEUDOCODE

***Input*:** Read labelled training dataset *DS* (bigram model), Max Generations(□), No. of Features(F), No. of solutions(N), Loss Parameters(α & ρ), SVM parameter range($[C_1, C_2]$ & $[\gamma_1, \gamma_2]$), No. of Dimensions(D), fitness vector (FT)

***Output*:** Globally best positions (features) of Cuckoos, Set of Classes C = {0, 1} where 0 reflects non bully and 1 reflects bully term

***Begin*:** for each solution (cuckoo's egg), $l = 1{:}N$

for each solution's dimension, $k = 1{:}D - 2$

Cuckoo's egg i.e. solution, $X_{l,k}^{0} \leftarrow Random\ \{0, 1\}$;

end for

$X_{l,D-1}^{0} \leftarrow Random[C_1, C_2]$ & $X_{l,D}^{0} \leftarrow Random[\gamma_1, \gamma_2]$;

$\mathrm{FT}_l \leftarrow -infinity$;

end for

while (Generation no., j < □)

for each solution, $l = 1{:}N$

$optimal\_X_l \leftarrow$ SVM_Fit(DS, X_l^{j-1}, FT_i);

end for

for each solution, $l = 1{:}N$ *// Generate New Solution by Lévy Flight*

$X_l^j \leftarrow$ Cuckoo_Lévy(X_l^{j-1});

$optimal\_X_l \leftarrow$ SVM_Fit(DS, X_l^j, FT_l);

end for

for each solution, $l = 1{:}N$ *// Discovery of Alien Eggs*

$X_l^j \leftarrow$ Update_worse_cuckoo(X_l^{j-1});

$optimal\_X_l \leftarrow$ SVM_Fit(DS, X_l^j, FT_l);

end for

end while

End

Functions:

1. SVM Fitness Function (SVM_Fit)

SVM_Fit(DS, X_l^{j-1}, FT_i)

for each dimension of solution, $k = 1{:}D - 2$

$DS' \leftarrow DS - \{DS : X_{l,k}^{j-1} = 0\}$;

end for

Train and Evaluate SVM using C & γ, $i.e.$ ($X_{l,D-1}^{j-1}$ & $X_{l,D}^{j-1}$) and

K-Fold cross Validation over DS' and store Recall result in Res;

if (Res > FT_l) then

$FT_l \leftarrow$ Res;

for each dimension of solution, $k = 1: D$

$optimal\_X_{l,k} \leftarrow X_{l,k}^{j-1}$;

end for

end if

return $optimal\_X_l$;

2. **Cuckoo Lévy Flight**

Cuckoo_Lévy(X_l^{j-1})

for each dimension of solution, $k = 1: D-2$

$X_{l,k}^{j} \leftarrow X_{l,k}^{j-1} + Lévy\ Flight\ (as\ in\ (3.6))$;

end for

for each dimension of solution, $k = 1: D-2$

if $\left(\hat{u} < \frac{1}{1+e^{X_{l,k}^{j}}}, (\text{as in } (3.4))\right)$ then

$X_{l,k}^{j} = 1$;

else

$X_{l,k}^{j} = 0$;

end if

end for

return X_l^j;

3. **Abandon Worse Build New**

Update_worse_cuckoo(X_l^{j-1})

for each dimension of solution, $k = 1: D-2$

Generate local step size, LS using (9) and $\rho_{l,k}$ using (7)

$X_{l,k}^{j} = X_{l,k}^{j-1} + LS * \rho_{l,k}, (as\ in\ (8))$;

end for

return X_l^j;

Figure 3.2: Pseudocode of the proposed CS-SVM approach

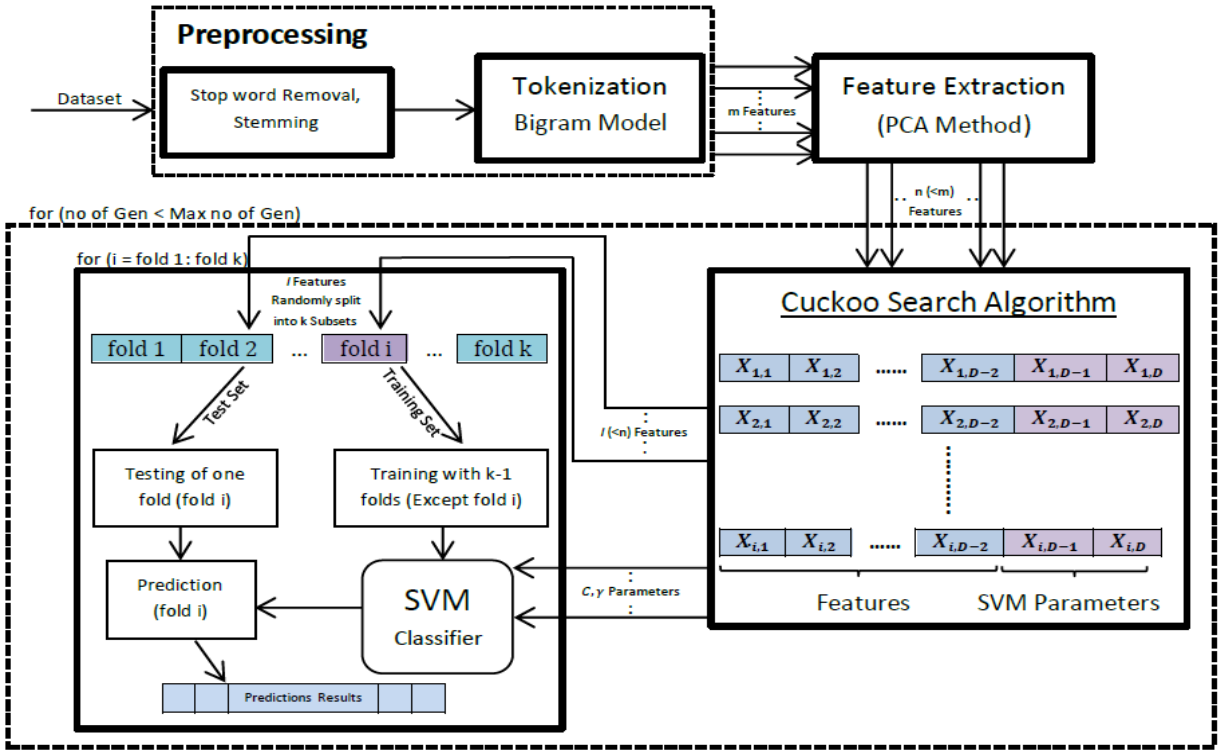


Figure 3.3: Schematic diagram of the proposed CS-SVM approach

The overall process of the proposed CS-SVM approach is showcased in the schematic diagram as shown in Figure 3.3. The flow of CS-SVM approach begins with reading the input dataset. Thereafter, various preprocessing operations are applied such as stopword removal, stemming and tokenization. In the next step bigram model is applied on the tokens. Afterward feature extraction is done using PCA. The next step involves the utilization of CS algorithm to seek for optimal subset of features and kernel parameters of SVM classifier. The reduced feature set is then sent to the classifier to perform a 10-fold cross validation for estimating the recall of the classifier. In the current work, the fitness value of the solution reflects the recall value. These steps of feature selection and SVM parameter optimization by CS are repeated until maximum generations achieved.

3.4 Experimental Analysis of Cuckoo inspired SVM Approach

The CS-SVM approach is implemented in Python with Scikit-Learn in the current work. The configuration of the underlying hardware used for the experiment is Intel processor Core_i7@4.0 GHz with16 GB memory (1867 MHz DDR3), running on Microsoft Windows 7 Professional Edition Operating System. For each dataset, SVM is used as base classifier using LIBSVM Python library (Chang and Lin (2011)) with RBF kernel settings. The choice of parameters values of RBF kernel and feature selection is accomplished using CS-SVM model.

3.4.1 Parameters in underlying CS algorithm

Table 3.1 shows the various parameter values used in the proposed model like: number of solutions (N), maximum generations (g), step size (α), discovery probability (ρ) and Lévy step length (L) of underlying CS algorithm. In the proposed CS-SVM model, CS algorithm is exploited to give different combinations of two important parameters: C and γ whose range is within [0.01, 1000] and [0.0001, 10] respectively.

Table 3.1: Summarizes the parameters used in the experimental analysis in the CS-SVM approach

Parameter	Value
Number of solutions, *N*	20
Maximum Generations, *g*	500
α	0.15
ρ	0.25
L	1.5

3.4.2 Datasets Used

In this work, four different datasets from various social media platforms like Formspring, ASKfm and Twitter are considered for cyberbullying detection. The used datasets have unstructured and imbalanced user comments, collected from freely available social forums and public comment sections. Different chat platforms, forums, Question/Answer, and microblogging, contain those free-format posts of users that are usually more prone to attract bully content. Table 3.2 illustrate each dataset's summary and source information, along with an average length of samples, a total number of samples, and the bully to non-bully sample ratio. As the datasets used are of diverse nature, containing varied number of instances, the proposed model is experimented on varied scales of problem complexity.

The Dataset#1 (Zhao *et al.* (2016)) contains tweets from Twitter social media that are manually labelled as bullying and non-bullying. This dataset in-total contains 1762 tweets with 1078 as non-bully and 684 as bully tweets. The Dataset#2 (Van Hee *et al.* (2018)) is a collection of ASKfm data containing questions and answers of anonymously users. After language filtering, the dataset comprised 15310 posts that are retrieved from April to October, 2013. Dataset#3 (Agrawal and Awekar (2018)), comprises of user answers procured from Formspring social media, in which a sample is a one response per question by each user. Using Amazon's Mechanical Turk services, each instance is labelled as a non-bully or bully. The Dataset#4 (Agrawal and Awekar (2018)) is another Twitter dataset in which there is at least one bullying, bullied or bully keywords in each collected tweet. In this dataset, each sample is annotated by Five skilled annotators.

Table 3.2: Summary of Datasets

Dataset Used	Dataset Source	No. of Instances	Dataset Style	Avg. Length of Instance	Class Distribution (Bully/Non-Bully)%
Dataset#1	Twitter	1762	Micro-Blog	67	26/41
Dataset#2	ASKfm	15310	Q&A	212	21/83
Dataset#3	Formspring	11967	Q&A	530	11/41
Dataset#4	Twitter	16914	Micro-Blog	30	10/33

3.4.3 Evaluation Criteria

The cyberbullying text classification is a crucial matter because of false negatives and false positives. In case of false negatives, the bullying content must not bypass as a regular content, simultaneously in case of false positives, it is a sensitive issue if the classifier identify post with non-bullying content as bullying content. The minimizing false negatives count is more important than that of false positives in cyberbullying detection. Various performance evaluation metric considered in this chapter are precision, recall, and the f-measure as shown in Table 3.3.

Table 3.3: Performance Metrics for Classification

Metrics	Formula	Evaluation Focus
Recall	$\frac{tp}{tp + fn}$	Number of correctly identified bullying instances out of the total count of instances of true bullying.
Precision	$\frac{tp}{tp + fp}$	The net count of correctly identified bullying instances from the retrieved instances of bullying.
F-Measure	$\frac{2 * p * r}{p + r}$	The harmonic mean of recall and precision.

3.4.4 Observations and Analysis

In experiments, the proposed CS-SVM approach uses SVM with the RBF kernel, also called the RBF kernel or Gaussian kernel. The CS algorithm, due to its outstanding capability of global optimization, produces optimal parameters thereby, eliminating the errors resulting from a random selection of relative parameters. For validating the effectiveness of the proposed CS-SVM is compared with several models: EBoW (Zhao *et al.* (2016)), LSVM (Van Hee *et al.* (2018)), and DNN (Agrawal and Awekar (2018)). The EBoW model concatenated latent semantic features, BoW features, and bullying features collectively to improve the SVM accuracy. The LSVM model used linear SVM for bully text classification in conjunction with hyper-parameter optimization and feature selection for improving classification accuracy. The DNN model named BLSTM (Bidirectional LSTM) with attention model with three transfer learning is utilized i.e. Complete Transfer Learning, Feature Level Transfer Learning and Model Level Transfer Learning, and the best results among them are further used. The configuration of model is: epoch = 5, batch_size = 128 and learn_rate = 0.01. The LSVM and EBoW performs better for detecting bullying text, and the DNN model is a relatively better for cyberbullying detection. Table 3.4 presents average and Standard Deviation values of recall, precision, and f-measure for executing the models for ten runs, where bold values are the maximum among different models. In the current work the recall parameter is the main parameter used for evaluating the performance of model.

The table 3.4 depicts that the proposed model outperforms the other models in terms of recall. In case of Dataset#1, the CS-SVM approach presents average recall of 0.8014 that is best one in comparison to other models. Moreover, the CS-SVM is able to be more accurate for filtering out bully content with 0.7971 precision and 0.8418 AUC value. While comparing to other models, the LSVM showcase lowest precision and recall of 0.7565, 0.7677 respectively. For Dataset#2, the

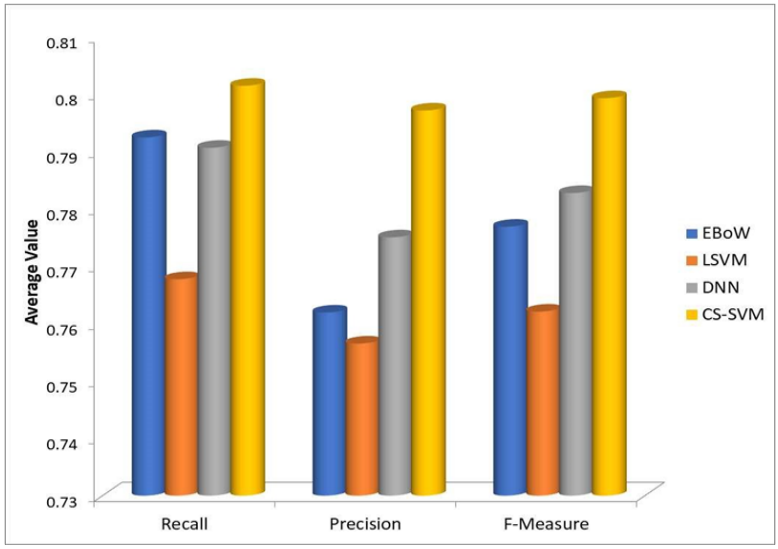
CS-SVM provides best average value of recall i.e. 0.8101 whereas other methods LSVM, DNN and EBoW provide 0.7934, 0.7914, and 0.7717 respectively. Also, the CS-SVM gives 0.7991 precision and 0.8612 AUC value. In case of Dataset#3, the DNN model gives better performance with average value of recall i.e. 95% but not better in comparison to the CS-SVM. The average recall value of LSVM and EBoW method are nearly equivalent i.e. 0.9067 and 0.8996 respectively.

Table 3.4: Comparative performance of CS-SVM with the state-of-the-art approaches

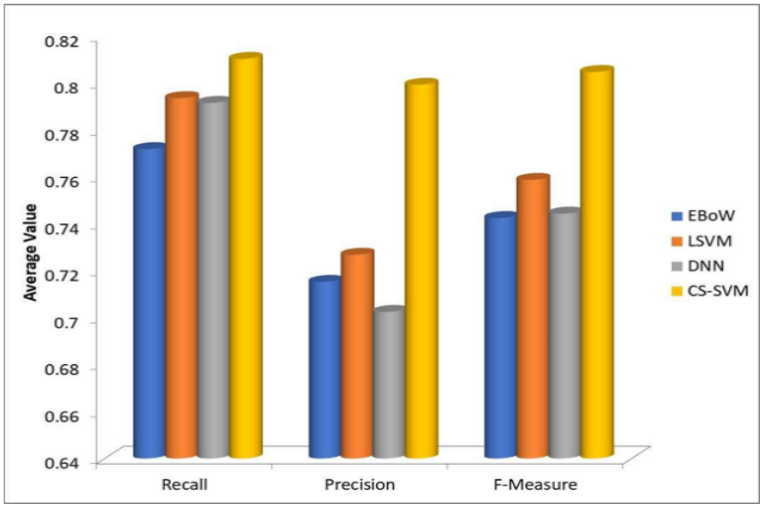
Datasets	**Models**	**Recall**		**Precision**		**F-Measure**		**AUC**
		Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	
Dataset#1	EBoW	0.7924	0.002914	0.7619	0.007059	0.7769	0.004185	0.8155
	LSVM	0.7677	0.005498	0.7565	0.003784	0.7621	0.004468	0.7824
	DNN	0.7906	0.003627	0.7750	0.004250	0.7827	0.003292	0.8117
	CS-SVM	**0.8014**	0.003651	**0.7971**	0.006206	**0.7992**	0.004597	**0.8418**
Dataset#2	EBoW	0.7717	0.005417	0.7152	0.002488	0.7424	0.003457	0.7714
	LSVM	0.7934	0.005621	0.7267	0.006524	0.7586	0.006096	0.8142
	DNN	0.7914	0.004904	0.7024	0.005473	0.7442	0.005819	0.8097
	CS-SVM	**0.8101**	0.005331	**0.7991**	0.004149	**0.8046**	0.004666	**0.8612**
Dataset#3	EBoW	0.8996	0.004088	0.8724	0.003855	0.8858	0.003844	0.9385
	LSVM	0.9067	0.007090	0.8765	0.003518	0.8913	0.004710	0.9451
	DNN	0.9494	0.005016	**0.9093**	0.003173	0.9289	0.003636	0.9778
	CS-SVM	**0.9551**	0.004813	0.9073	0.005819	**0.9306**	0.005268	**0.9785**
Dataset#4	EBoW	0.9253	0.005794	0.8416	0.004875	0.8815	0.005479	0.9645
	LSVM	0.9391	0.004932	0.8320	0.003986	0.8823	0.003773	0.9714
	DNN	0.9572	0.004211	0.8997	0.003751	0.9276	0.004017	0.9791
	CS-SVM	**0.9659**	0.007158	**0.9004**	0.002803	**0.9320**	0.004029	**0.9848**

The best average recall for this dataset is 0.9551, obtained by the CS-SVM model also the precision value of CS-SVM approach is 90.7% along with 97.9% of AUC. Finally, in case of Dataset#4, the CS-SVM again performs best, with 0.9659 as an average recall. However, LSVM gives 0.9391, EBoW shows 0.9253, and DNN presents an average value of recall i.e. 0.9572. The CS-SVM provides the best f-measure, precision and AUC values in case of Dataset#3, i.e., 0.9320, 0.9004 and 0.9848 respectively. Overall, the CS-SVM results are better in comparison to the other methods.

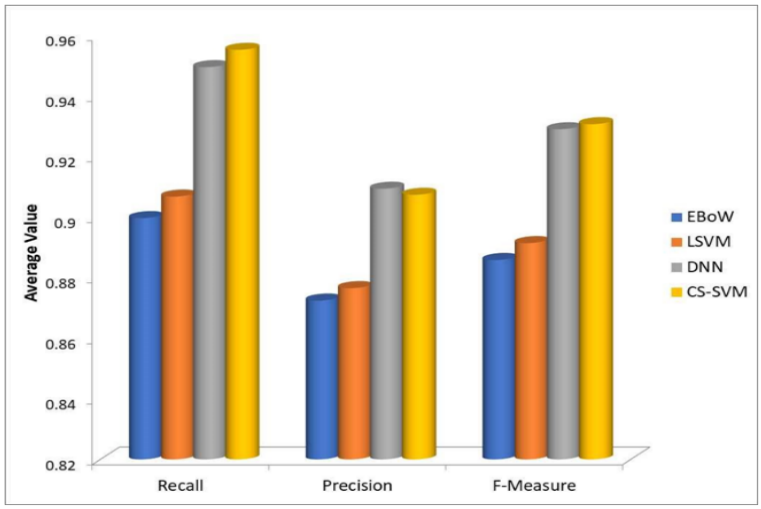
It is clear from the results that the average recall value of the CS-SVM is relatively high and varied in the range from 0.8014 to 0.9659 for given datasets. Well-selected training dataset and a wide range of variation in selected kernel parameters are some of the factors that led to high recall rate of SVM classifier. By analyzing the results, it is concluded that changing the kernel function parameters *c* and *g* using CS algorithm results in the improvement in the recall value of the SVM model by 4 to 7% for datasets in hand in comparison to other models. Based on the measures of precision, recall, f-measure, and AUC, the proposed model outperforms other models for all the experimental datasets.



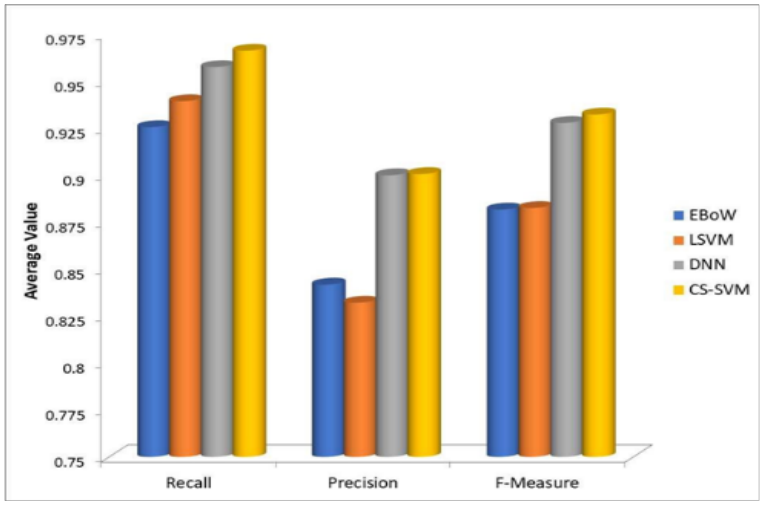
(i)



(ii)



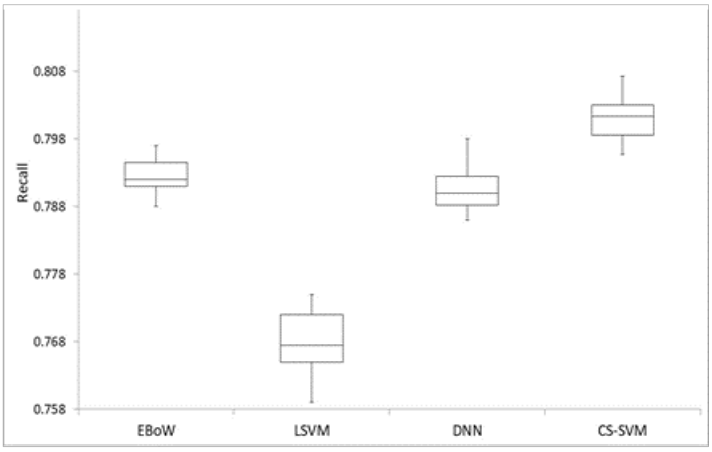
(iii)



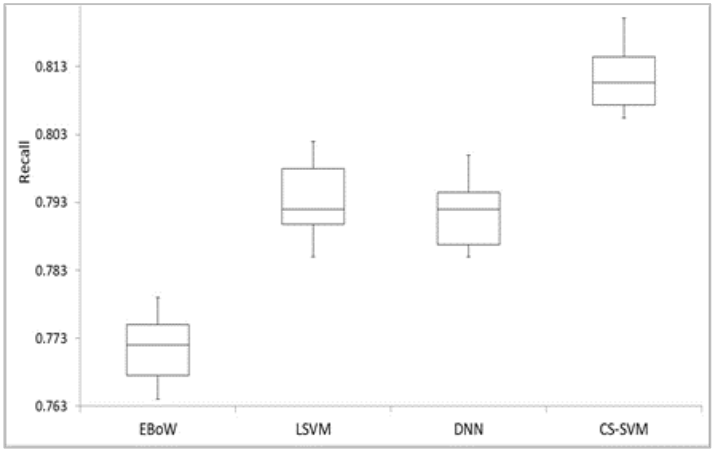
(iv)

Figure 3.4 (i, ii, iii, iv): Average Recall, Precision, and F-Measure value of the Models for Dataset#1, Dataset#2, Dataset#3, and Dataset#4 respectively

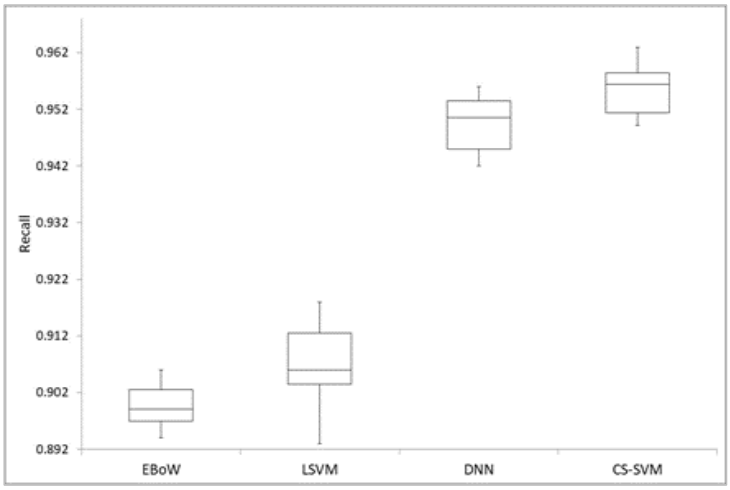
The evaluation results of recall, precision, and f-measure of all datasets are shown in Figure 3.4 (i, ii, iii, iv). The figure depicts the comparison of the average value of precision, recall, and f-measure of the CS-SVM model with other models. It is clear from the figure that the CS-SVM model shows better average precision, recall and f-measure for detecting bully text. It can also be seen that average precision, and f-measure is comparatively better in case of Dataset#3 than other datasets. The performance comparison of all the considered models and the CS-SVM model is graphically visualized by drawing boxplot (Williamson et al. 1989). It depicts the empirical distribution of the data. Figure 3.5 (i, ii, iii, iv) presents the boxplot for LSVM, EBoW, DNN and CS-SVM models. In Figure 3.5, the x-axis represents the model names whereas the y-axis labels the considered evaluation parameter i.e. recall value.



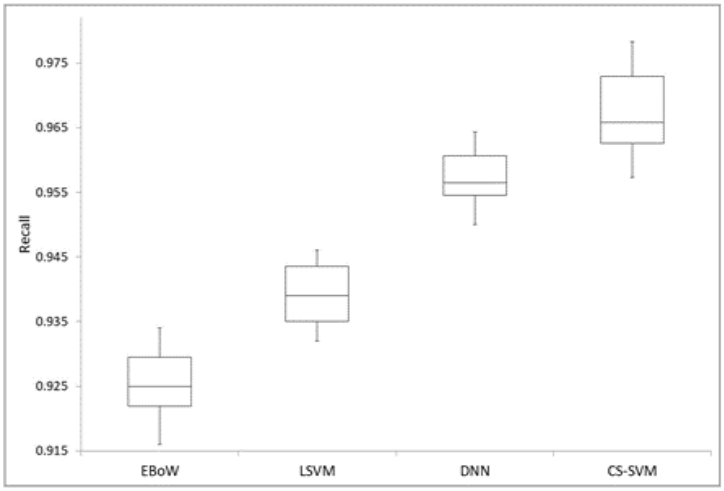
(i)



(ii)



(iii)



(iv)

Figure 3.5 (i, ii, iii, iv): Recall value of the Models for Dataset#1, Dataset#2, Dataset#3, and Dataset#4 respectively

The boxplots clearly reveal the better performance and consistent results of CS-SVM model in comparison to other models. The stability of the models is also achieved as there are no outliers present. It is observed from the boxplot that for some datasets, the minimum recall of CS-SVM is even better than the highest recall value of other models. Moreover, the variance in average recall of the CS-SVM in comparison to other is greater for Dataset#1 and Dataset#2 than remaining one. Whereas, in case of Dataset#3 and Dataset#4, the DNN model gives better performance in comparison to LSVM and EBoW models.

The following points can be emphasized based on the results of the current work:

- The high recall and precision of the CS-SVM approach reveal that the CS-SVM approach gives a good differentiation between the classes, bully and nonbully.
- The classification results show that the CS-SVM approach is comparable to the best existing model in the classification of content-based cybercrime detection.
- From the results, it is concluded that the CS-SVM approach provides the highest predictive performance (in terms of Recall) for all cyberbullying datasets whereas the DNN model gives the second-best performance.

3.5 Discussion

There are only a few contemporary types of research in cyberbullying detection, which can provide high precision and recall rates. None of the existing research work used evolutionary computation for content-based cybercrime detection. In this chapter, a novel hybrid Cuckoo Search (CS) and Support Vector Machine (SVM) based on feature selection and optimum parameters optimization approach has been discussed for cyberbullying detection in four different datasets taken from Formspring, ASKfm and Twitter. The feature set has been reduced using PCA, and 10-fold cross-validation has been applied to learn the proposed model. To improve the performance of SVM classifier, the CS approach has been used as an optimizer, that select an appropriate feature subset and kernel parameters. The key contributions of this research are:

- First, the CS-SVM approach is capable of handling very high dimensional data despite other techniques for feature selection that a user needs to configure the number of features required. The CS-SVM model approach is able to determine the most useful features

automatically in terms of classification accuracy within an acceptable processing time without requiring the user to set the required number of features.

- Second, The CS-SVM approach is an up-and-coming model for detecting content-based cybercrime in terms of evaluation metric of classification i.e., recall. The proposed model exhibits increased recall rate in given various datasets by determining the optimal values of SVM tuning parameters.
- Finally, the CS-SVM approach provides better results in terms of precision, recall, and f-measure when applied on the above four datasets and compared to three recent models.

The proposed CS-SVM approach has been successfully applied for cyberbullying detection. The achieved high accuracy rate in classifying bully text 97.1% demonstrates greater effectiveness over existing models.

The next chapter elaborate the second proposed technique, that exploits the best configurations depicting optimal combinations of preprocessing steps, feature selection models and classification techniques, thereby improvising classification recall.

Chapter 4

Multiconfiguration Detection Technique

In this chapter, a novel technique that explores possible combinations of various preprocessing, feature selection and classification methodologies using the cuckoo search metaheuristic approach is discussed. This technique seeks to improve the performance of content-based cybercrime detection system.

4.1 Introduction

In the proposed work, an intelligent content-based cybercrime detection technique is developed using a Cuckoo Search (CS) metaheuristic. This technique is designed specifically to detect content-based cybercrime in online social networks. The present work is a first step towards developing an efficient classification model based on the CS algorithm in the field of content-based cybercrime detection. The proposed technique classifies new texts using the most effective known combination of traditional schemes (i.e., preprocessing steps, feature selection techniques, and classification models). The present work proposes a novel CS-based machine learning model that is superior to the state-of-art models from literature in detecting bully content. The typical cyberbullying detection process is viewed as a generic classification problem, wherein features are extracted from the preprocessed data to classify the social media posts. The underlying technique exploits the best configurations depicting optimal combinations of preprocessing steps, feature

selection models and classification techniques thereby improvising classification recall and optimizing processing time (in comparison to the exhaustive approach).

The main objective of the present study is to enhance the performance of the proposed model. This objective is to be met by using the CS algorithm to determine which configuration of classification processes yields the best performance. The motivation of this research is to attain an unbiased criterion and evaluate the effectiveness of the possible configurations in the context of cyberbullying detection. The literature showcases several models for cyberbullying detection that generate different results in different contexts using various classification processes. The overall performance of the detection models can be improved by automatic selection and combination of techniques in alternative ways. The same can be achieved efficiently with the help of nature-inspired algorithms. Among the existing meta-heuristics, the CS algorithm is proven to outperform other meta-heuristic algorithms as revealed in the literature (Yang and Deb (2014); Fister *et al.* (2014)). In the proposed model, the recall value (as a fitness function) has been evaluated across different CS configurations. The recall parameter is considered as the objective function evaluation in exploring the maximum generalization ability of classifier.

The proposed technique has been compared with existing cyberbullying detection techniques based on machine learning classification schemes (preprocessing steps, feature selection techniques and classification models) as described in Table 4.1. As seen in the table, the existing techniques considered a classification model with a peculiar set of preprocessing steps and feature selection techniques, which may not lead to the best solution. The reason for the same is that these schemes are data dependent. The aim of this research is to attain an unbiased criterion and assess the efficiency of the classification process in the context of cyberbullying detection. The key

difference between this proposed model and others is that this model will explore optimal combinations of learning scheme, rather than estimating only pre-established combinations.

For the task of content-based cybercrime detection, only a few datasets have become available publicly in recent past, such as the datasets provided in the context of the CAW 2.0 workshop for training purpose and more recently the bullying traces of Twitter dataset (Chen *et al.* (2017)). Consequently, many studies have collaborated with the former or made their own datasets from online social sites that are susceptible to bullying content, like Formspring, YouTube and ASKfm. The proposed model is applied to four different recent datasets procured from Twitter, ASKfm and Formspring (Agrawal and Awekar (2018); Van Hee *et al.* (2018); Zhao *et al.* (2016)). The analysis and the performance of the proposed model are compared with three existing recent models. Efforts to address the detection of cyberbullying issues vary, and the proposed technique is tried to test diverse datasets. The proposed technique emphasizes the selection of models but, more specifically, selection from a large number of schemes such as text preprocessing, feature selection and classification methodologies for cyberbullying detection. The preprocessing phase includes dataset cleaning (by eliminating punctuation, hashtags, URLs, etc.), tokenization and stemming. The next phase includes feature engineering that detects profanity including rude, nasty, bad, abusive or hateful words, pronouns, unigram/bigram/trigram BoWs, etc. In the final phase, the messages are classified as bullying or non-bullying using machine learning classification techniques.

Table 4.1: Comparative list of different Configurations and Datasets exploited in the Proposed & Existing Techniques

Work	Preprocessing									Feature Selection						Classification			Dataset Used				
	Punctuation Removal	Tokenization	Lower Casing	Stemming	Hashtag Removal	Number Removal	URL Removal	Duplication Removal	Diacritic Removal	BoW	TF-IDF	Character n-gram	Word n-gram	Profanity	Pronoun	SVM	NB	DNN	Twitter	Formspring	ASKfm	Kongregate	Youtube
Bayzick *et al*. (2011)		✓												✓	✓	✓					✓	✓	
Dinakar *et al.* (2012)	✓			✓					✓		✓		✓	✓		✓	✓			✓			
Nahar *et al.* (2012)	✓			✓						✓						✓					✓	✓	
Dadvar *et al.* (2013)	✓			✓									✓	✓	✓	✓							✓
Nahar *et al.* (2013)	✓				✓		✓	✓			✓		✓			✓			✓		✓	✓	
Nahar *et al.* (2014a)		✓												✓	✓		✓				✓	✓	
Mangaonkar *et al.* (2015)		✓											✓			✓	✓		✓				
Van Hee. *et al*. (2015)		✓		✓								✓	✓	✓		✓					✓		
Al-garadi *et al.* (2016)			✓	✓										✓	✓		✓		✓				
Galán-García *et al.* (2016)	✓	✓									✓		✓				✓		✓				
Singh *et al.* (2016)		✓												✓					✓				

Table 4.1 Continued…

Dataset Used	Youtube							↘	↘			
	Kongregate											
	ASKfm				↘					↘		↘
	Formspring						↘	↘				↘
	Twitter	↘	↘	↘	↘		↘	↘			↘	↘
Classification	DNN						↘	↘				
	NB								↘			↘
	SVM	↘	↘			↘				↘		↘
Feature Selection	Pronoun					↘						↘
	Profanity	↘	↘									↘
	Word n-gram			↘	↘		↘		↘	↘		↘
	Character n-gram						↘			↘		↘
	TF-IDF			↘								↘
	BoW	↘	↘			↘		↘		↘	↘	↘
Preprocessing	Diacritic Removal				↘							↘
	Duplication Removal						↘					↘
	URL Removal	↘			↘					↘		↘
	Number Removal						↘			↘		↘
	Hashtag Removal											↘
	Stemming			↘						↘		↘
	Lower Casing						↘					↘
	Tokenization		↘		↘	↘	↘		↘	↘	↘	↘
	Punctuation Removal	↘		↘	↘		↘	↘			↘	↘
Work		Zhao and Mao (2016)	Zhao *et al.* (2016)	Dani *et al.* (2017)	Raisi and Huang(2017)	Singh *et al.* (2017)	Agrawal and Awekar(2018)	Dadvar and Eckert (2018)	Rafiq *et al.* (2018)	Van Hee *et al.* (2018)	Balakrishnan *et al.* (2019)	Proposed Technique

4.2 The Proposed Multiconfiguration Detection Technique

The aim of the proposed technique is to find a competitive model for the task in the (possibly large) candidate models set. A model is represented by the parameters that along with the input dataset determine the functionality of the model. The final model must be competitive among the defined space of models, so the search for the required model should be performed accurately and efficiently. A model is capable of classifying new texts, using preprocessing steps, feature selection techniques, and classification models. Various configurations like preprocessing steps, feature selection along with classification techniques have a significant impact on the performance of the learning model. In (Huang *et al.* (2015)), An estimate of many data preprocessing techniques has been empirically tested in terms of the effectiveness of machine learning methods to estimate the effort. The results show that data processing techniques can significantly affect predictions, but can also sometimes negatively impact predictive performance. The authors concluded that there is a need for accurate selection of data preprocessing steps according to the characteristics of datasets along with machine learning methods. A large configuration space has been selected to tackle the different datasets of different natures that are defined in the following paragraphs. Figure 4.1 illustrates the general flow of the proposed methodology. The following paragraphs include the description of each machine learning task for content-based cybercrime detection.

4.2.1 Preprocessing

The data provided as input to machine learning task first undergoes several preprocessing steps for reducing the noisy data that further aids in improving overall accuracy. Various preprocessing steps used in proposed methodology are (a) Punctuation removal: To remove punctuation symbols in the text like (,), [,], {, }, <, >, ?, !, ;, :, ., -, ’, ". (b) Tokenization: To break phrases and sentences

in a document into a sequence of words (c) Lower casing: To lower all casing of text. (d) Stemming: To reduce words to their stems like abusing reduced to abuse (e) Hashtag removal: To remove all hashtags into a single tag. (f) Number removal: To remove numbers in the text. (g) URL removal: To remove URLs in the text. (h) Duplication removal: To remove duplicate contiguous symbols in the text. (i) Diacritic removal: To remove diacritic symbols in the text like à, á, â, ã or ä to simply a.

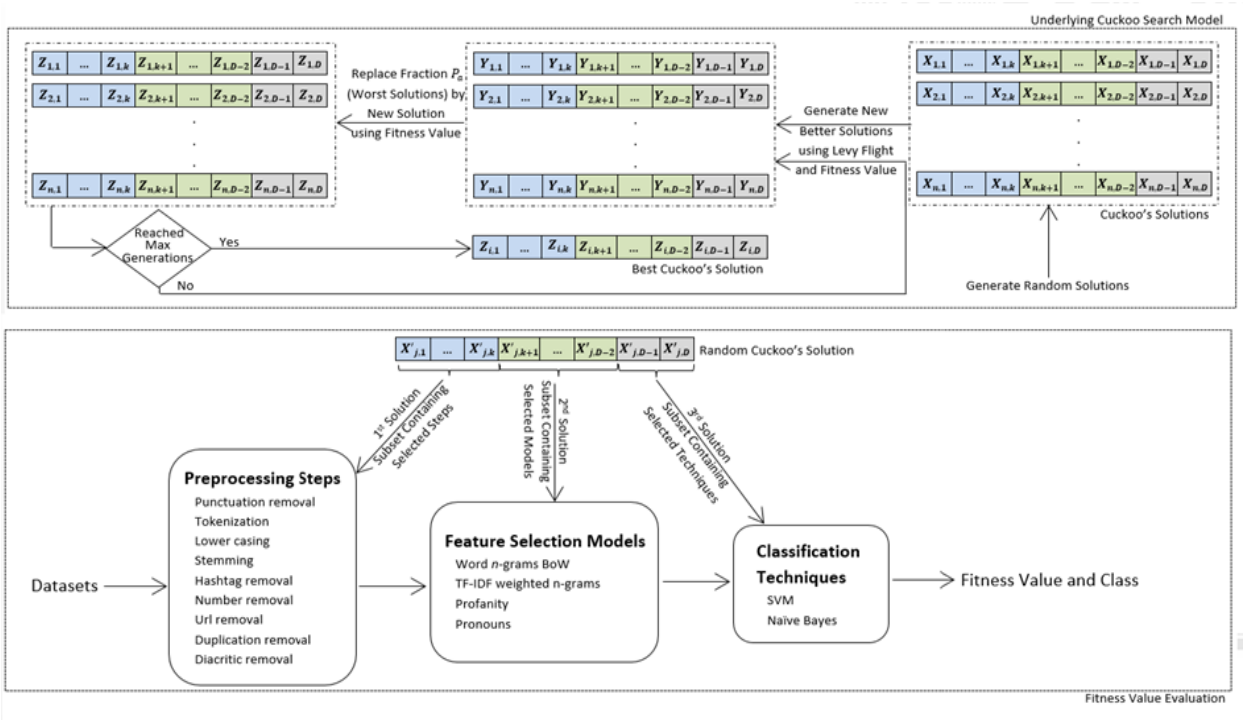


Figure 4.1: Flow of proposed multiconfiguration detection technique

In some cases, a bully can use URL, punctuation or diacritic symbols to bully victims. For example, URL names (porn sites are an obvious target) may contain abusive language or using punctuation/diacritics (eg. Sh!t, @ss, f*ck, bïtch) to make some abusive sign. In this scenario, removing URLs, punctuation or diacritics may lead to missing important features/information, leading to more false negatives. The choice of preprocessing steps should therefore be accomplished very carefully.

4.2.2 Feature Engineering

After preprocessing the data, features engineering is applied for further model training. The proposed technique considers three features from word n-grams BoW (n = 1, 2, 3 i.e. unigram, bigram, trigram), three of Term Frequency-Inverse Document Frequency (TF-IDF) weighted n-grams (n = 1, 2, 3), profanity and pronouns features. (a) Word n-grams BoW: Within a document, the occurrence of words is described by BoW, and in this family of feature engineering, the text is firstly tokenized into words, and afterward, from m words, m−n+1 tokens (a sequence of n consecutive words) are produced, i.e., BoW with word n-grams. For example, Bullying does not happen in a vacuum creates the following 3-grams: Bullying_does_not, does_not_happen,not_happen_in, happen_in_a, in_a_vacuum. (b) TF-IDF weighted n-grams: In TF-IDF a word's importance to a document is measured with in a collection of documents. Using TF-IDF alongside with n-grams sometimes provide better results (Yin *et al.* (2009)). Because of this reason, it is generally used in conjunction with n-gram to enhance detection accuracy. (c) Profanity: The existence of profanity in the text can be a sign for cyberbullying. This feature is created either from external libraries such as urbandictionary.com and noswearing.com or from the word lists compiled by the scholars. The use of this only feature can lead to failure in detection of cyberbullying. The same has been cautioned in the work (Rafiq *et al.* (2015)) and authors argued that not all usage of profanity feature constitutes bullying. The use of profanity feature may also fail to lead good recall value for some cases. (d) Pronouns: For an improvement in profanity feature, this feature in close proximity to profanity is potentially more indicative of cyberbullying than profanity alone, for e.g. the phrase "The f**king flight was late again" does not reflect cyberbullying although it enclosed profanity term however the sentence "You f**king idiot" could

be. Features like word n-grams BoW, TF-IDF weighted n-grams along with pronoun are used for posts that do not contain swear words but still flagged as bullying.

4.2.3 Classification Techniques

The next machine learning task is selecting a classification technique. There are many techniques for classification created over the past years. SVM and Naïve Bayes are predominantly used in the literature of text classification. (a) SVM: The SVM method targets to discover the decision boundaries in the search space, separating one class from another class. High generalization ability and global optimization are the main advantages of SVM (Joachims (1998)). The method seeks to find an optimum direction of discrimination in the feature space; thereby giving excellent performance for high dimensional feature space. The literature reveals that SVM gives the best performance for large dimensional input (Joachims (1998); Ben-Hur and Weston (2010)) (as in the proposed approach), and the linear kernel also gives better results under these conditions. The proposed approach uses the linear SVM classifier (where the parameter C equals to 1). (b) Naïve Bayes: These are a family of simple probabilistic classifiers that are based on application Bayes' theorem with naive (strong) independence assumptions between the features (Chen *et al.* (2009)). These classifiers are very scalable, demanding the parameters count to be proportional to the count features (variables) in learning.

4.2.4 Solution Space

In the proposed work the solution space, C is represented by a set of all possible configurations of, i) preprocessing steps, L_p, ii) feature engineering, L_f, iii) classification techniques, L_c. Where:

$$C = 2^{L_p} \times 2^{L_f} \times L_c$$

Even in the simplest case, the configuration space increases exponentially with the count of possible preprocessing, feature engineering and classification techniques. The job of finding the best model for configuration space is an np-hard problem. In the proposed approach the L_p and L_f are considered equal to nine and eight respectively, and L_c has a value equal to two. Using this configuration, the count of possible combinations of preprocessing, feature engineering and classification techniques excluding the empty sets are $(2^{L_p} - 1) \times (2^{L_f} - 1) \times L_c$, lead to 260610 possible configurations. For instance, if configuration involves a few seconds for evaluation. It will require huge amount of time (in months) for exhaustive evaluation of configuration space. Herein metaheuristic approaches can be utilized to find the solution of the problem by seeking for the model with a near optimal solution in a fair amount of time.

4.2.5 Model selection using Cuckoo Search

The detailed pseudocode of the proposed model for content-based cybercrime detection is given in Figure 4.2. The main sections of the pseudocode are as follows.

Solution Representation and Initialization

The proposed model begins with a population of randomly generated solutions. For this purpose, the solution representing a cuckoo egg in a nest is encoded in the format that contains all information pertaining to preprocessing steps, feature selection techniques, and classification models.

Each cuckoo's egg (i.e. solution) in the population is represented by a vector of $L_p + L_f + 1$ size, wherein the first L_p and subsequent L_f patterns signify the chosen preprocessing steps and features selection model for the problem in hand that are further utilized by the classifier (one bit, $L_p + L_{f+1}$ for selecting classifiers among two) at its learning stage.

PSEUDOCODE

Input: Read labelled training dataset DS, Max number of Generations (G), No. of Host Nest or Cuckoo's Solutions(N), Loss Parameters (α & ρ), No. of Dimensions of Cuckoo's Solutions (D), Pool of preprocessing steps, Feature selection models and classification techniques (P)

Output: Globally best Cuckoo's Solution (Set of pre-processing steps, feature selection models and classification techniques), Set of Classes C = {0, 1} where 1 reflects bully and 0 reflects non-bully term

Begin: for each Host Nest, $i = 1{:}\,N$

for each dimension of egg in Host Nest, $j = 1{:}\,D$

Host nest containing egg i.e. Solution, $X_{i,j}^{0} \leftarrow Random\ \{0, 1\}$;

end for

Quality i.e. fitness of i^{th} Solution, $\text{FT}_i \leftarrow -infinity$;

end for

while (Generation no., $\varepsilon < G$)

for each Host Nest, $i = 1{:}\,N$

Cuckoo's egg, i.e. new better Solution, $New\_X_i \leftarrow$ Cuckoo_Fit (DS, $X_i^{\varepsilon-1}$, FT_i);

end for

for each Host Nest, $i = 1{:}\,N$

for each dimension of egg in Host Nest, $j = 1{:}\,D$

$X_{i,j}^{\varepsilon} \leftarrow X_{i,j}^{\varepsilon-1} + Lévy\ Flight\ (as\ in\ (2))$;

end for

for each dimension of egg in Host Nest, $j = 1{:}\,D$

if $\left(\hat{u} < \frac{1}{1+e^{X_{i,j}^{\varepsilon}}}, (\text{as in } (1))\right)$ then

$X_{i,j}^{\varepsilon} = 1$;

else

$X_{i,j}^{\varepsilon} = 0$;

end if

end for

$New\_X_i \leftarrow$ Cuckoo_Fit (DS, X_i^{ε}, FT_i);

end for

for each Host Nest, $i = 1: N$

for each dimension of egg in Host Nest, $j = 1: D$

Generate local step size, LS using (5) and $\rho_{i,j}$ using (3)

$X_{i,j}^{\varepsilon} = X_{i,j}^{\varepsilon-1} + LS * \rho_{i,j}, (as\ in\ (4))$;

end for

$New\_X_i \leftarrow$ Cuckoo_Fit (DS, X_i^{ε}, FT_i);

end for

end while

End

Cuckoo's Fitness Function (Cuckoo Fit)

Cuckoo_Fit (DS, $X_i^{\varepsilon-1}$, FT_i)

for each dimension of egg in Host Nest, $j = 1: D$

$P' \leftarrow P - \{P : X_{i,j}^{\varepsilon-1} = 0\}$;

end for

Use P' for pre-processing, feature selection & classification on DS and store Recall result in Res;

if (Res > FT_i) then

$FT_i \leftarrow$ Res;

for each dimension of solution, $j = 1: D$

$optimal\_X_{i,j} \leftarrow X_{i,j}^{\varepsilon-1}$;

end for

end if

return $optimal\_X_i$;

Figure 4.2: Pseudocode of the proposed multiconfiguration detection technique

The structure of a solution is represented as follows.

X_1	X_2	…	X_{L_p}	X_{L_p+1}	X_{L_p+2}	…	$X_{L_p+L_f}$	$X_{L_p+L_{f+1}}$

Where (X_1, X_{L_p}), $(X_{L_p+1}, X_{L_p+L_f})$ and $X_{L_p+L_{f+1}}$ substring showcase selected preprocessing steps, feature selection models and classification models respectively with (X_1, X_{L_p}), $(X_{L_p+1}, X_{L_p+L_f})$ and $(X_{L_p+L_{f+1}}) \in \{0, 1\}$. The value at the respective index of the substring decides the involvement (value 1) or elimination (value 0) of that respective component in the learning process. Equation (4.1) is used to restrict new solutions to binary only.

$$X_{i,j}^{k+1} = \begin{cases} 0, & \text{if } \dfrac{1}{1+e^{X_{i,j}^{k}}} > \theta \\ 1, & \text{otherwise} \end{cases} \tag{4.1}$$

where $X_{i,j}^{k}$ represents the solution at generation k and θ has the standard uniform distribution with minimum 0 and maximum 1.

Generate New Solution

After initialization, CS uses a Lévy flight random walk to search a new solution using (4.2).

$$X_{i,j}^{k} = X_{i,j}^{k-1} + \alpha \times L\text{évy} \times (X_{i,j}^{k-1} - Optimal\_X_{i,j}) \tag{4.2}$$

where $X_{i,j}^{k-1}$ is randomly chosen solution within the population; $X_{i,j}^{k}$ is the new solution generated using Lévy flight; α is step size; $\text{Optimal\_X}_{i,j}$ represents the best solution generated so far; and Lévy is the step length or Lévy flight vector. After finding the new solution, the fitness values of two solutions are evaluated, and the highest quality solution is kept in the pool for further processing.

Discover Worst Solution

The high-quality eggs (optimal) which are more similar to the host bird's eggs have more chance to grow (next generation) and became a mature cuckoo. Some eggs (not optimal) are recognized by host bird with a probability $P_a \in [0,1]$. For every egg in the population, the probability matrix is utilized for foreign eggs detection. The probability matrix is represented as in (4.3).

$$\rho_{i,j} = (\text{random}(0,1) < P_a)?\,1:0 \tag{4.3}$$

where $\rho_{i,j}$ is a probability of discovering alien eggs in the i^{th} solution for the j^{th} variable of cuckoo's dimension. The comparison of P_a is performed with the outcome of $\text{random}(0,1)$ (a uniform random number generator), to check if local random walk is given consideration or not. After determining the discovering probabilities, new solutions are obtained using (4.4).

$$X_{i,j}^{k} = X_{i,j}^{k-1} + \rho_{i,j} \times \text{LS} \tag{4.4}$$

where LS is the matrix of local step size, which is obtained by using the formula (4.5).

$$\text{LS} = \text{random}(0,1) \times \Delta X_i^{k-1},$$

$$\Delta X_i^{k-1} = (X_{i,m}^{k-1} - X_{i,n}^{k-1}) \tag{4.5}$$

Finally, the objective function values are compared for existing and new solutions and the best solution enters in the next generation according to the simple rule as given in (4.6).

$$\text{new\_Solution} = \left(\text{F}\left(X_{i,j}^{k}\right) > \text{F}\left(X_{i,j}^{k-1}\right)\right)?\,X_{i,j}^{k} : X_{i,j}^{k-1} \tag{4.6}$$

Calculate the fitness value of new_Solution and judge whether condition of termination is attained. If it is satisfied, new_Solution is the optimal solution, otherwise return to Candidate Solution Generation step.

4.3 Experimental Results

The proposed work is implemented on Intel Core_i7@4.0 GHz processor with 16 GB memory (1867 MHz DDR3), running on Windows 10 using Python language. To evaluate the performance of the proposed model and comparing it to the state-of-the-art approaches, the current work utilizes several popular datasets from the literature (Agrawal and Awekar (2018); Van Hee *et al.* (2018); Zhao *et al.* (2016)). The implementation incorporates 10-fold cross-validation to avoid overfitting.

Four different social media datasets from Twitter, ASKfm, and Formspring are considered for cyberbullying detection. These datasets contain imbalanced and unstructured user comments gathered from freely accessible user comment sections and social media message boards. Various micro-blogging, Question/Answer, forums, and chat platforms contain free-format user posts that are generally more vulnerable to entertain offensive content. Due to diversity in these datasets, the proposed model is tested on different scales of problem complexity because of a varied number and average length of instances.

The classification of cyberbullying text is a critical problem due to false negatives and false positives cases. On the one hand, the bullying text should not get bypassed as a normal text (false negatives) and, on the other hand, to identify the non-bullying post as bullying itself (false positives), is also a sensitive issue. But minimizing false negatives is more vital than false positives in cyberbullying detection (Reynolds *et al.* (2011)). In this work recall, precision and the f-measure metrics are considered for performance evaluation.

Table 4.2: Comparative results of Performance Metrics for Datasets

Datasets	Models	Recall		Precision		F-Measure		AUC
		Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation	
Dataset#1	EBoW[1]	0.7924	0.002914	0.7619	0.007059	0.7769	0.004185	0.8155
	LSVM[2]	0.7677	0.005498	0.7565	0.003784	0.7621	0.004468	0.7824
	DNN[3]	0.7906	0.003627	0.7750	0.004250	0.7827	0.003292	0.8117
	Proposed Model	0.8142	0.005884	0.8023	0.005418	0.8082	0.005802	0.8529
Dataset#2	EBoW[1]	0.7717	0.005417	0.7152	0.002488	0.7424	0.003457	0.7714
	LSVM[2]	0.7934	0.005621	0.7267	0.006524	0.7586	0.006096	0.8142
	DNN[3]	0.7914	0.004904	0.7024	0.005473	0.7442	0.005819	0.8097
	Proposed Model	0.8248	0.005534	0.7954	0.005453	0.8098	0.004912	0.8637
Dataset#3	EBoW[1]	0.8996	0.004088	0.8724	0.003855	0.8858	0.003844	0.9385
	LSVM[2]	0.9067	0.007090	0.8765	0.003518	0.8913	0.004710	0.9451
	DNN[3]	0.9494	0.005016	0.9093	0.003173	0.9289	0.003636	0.9778
	Proposed Model	0.9619	0.002885	0.9146	0.004778	0.9377	0.003964	0.9841
Dataset#4	EBoW[1]	0.9253	0.005794	0.8416	0.004875	0.8815	0.005479	0.9645
	LSVM[2]	0.9391	0.004932	0.8320	0.003986	0.8823	0.003773	0.9714
	DNN[3]	0.9572	0.004211	0.8997	0.003751	0.9276	0.004017	0.9791
	Proposed Model	0.9702	0.006925	0.8975	0.006421	0.9324	0.006901	0.9875

[1] EBoW (Zhao *et al.* (2016))
[2] LSVM (Van Hee *et al.* (2018))
[3] DNN (Agrawal and Awekar (2018))

4.3.1 Results and Analysis

To validating the efficiency of the proposed model, it is compared with LSVM, EBoW, and DNN models. The EBoW model concatenates BoW features, latent semantic features and bullying features together to improve the SVM classifier accuracy. The LSVM model uses linear SVM classifier for classifying bully text in conjunction with feature selection and hyper-parameter optimization to improve classification accuracy. The DNN model named BLSTM (Bidirectional LSTM) with attention model with three transfer learning is utilized i.e. Complete Transfer Learning, Feature Level Transfer Learning and Model Level Transfer Learning, and the best results among them are further used. The configuration of model is: *epoch* = 5, *batch_size* = 128 and *learn_rate* = 0.01. The literature reveals that EBoW and LSVM have good performance for cyberbullying detection and the DNN model is a relatively advanced model for cyberbullying detection. The following paragraph discusses the comparative results of the proposed approach with above mentioned state-of-the-art methods.

Table 4.2 presents the average and the standard deviation values of recall, precision, f-measure and AUC (Area Under ROC Curve) values for executing various models for ten runs. The primary parameter considered in the current work is recall value. For Dataset#1, the proposed model gives the average recall value of 0.8142 which is the best value as compared to other models. Furthermore, the same model performs exceedingly precise in filtering bully text from the dataset with a precision and AUC of 0.8023 and 0.8529 respectively. The LSVM model has the lowest recall, precision and AUC value of 0.7677, 0.7565 and 0.7824 respectively in comparison to other models. In the case of Dataset#2, the proposed model provides the best average recall of 0.8248 where other models EBoW, LSVM and DNN provide 0.7717, 0.7934 and 0.7914 respectively. The

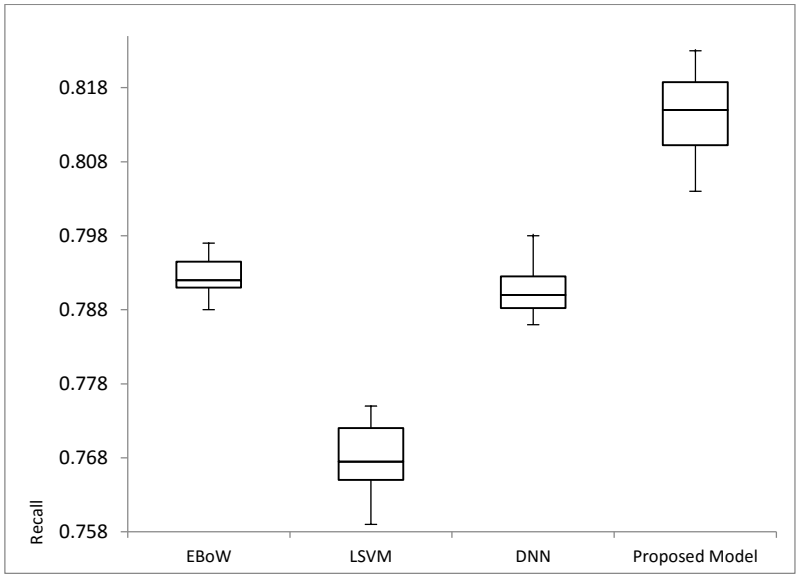
precision and AUC value of the proposed model is 0.7954 and 0.8637 respectively. For Dataset#3, DNN model also performed well with 95% average recall value but not better than the proposed model. The results (recall value) of EBoW and LSVM models are almost similar, i.e. 0.8996 and 0.9067 respectively. For this dataset, the proposed model obtains the best average recall value of 0.9619. The precision and AUC value of the proposed model is also 91.7% and 98.4% respectively. Lastly, for Dataset#4, the proposed model is again performing best, with an average recall value of 0.9702, whereas EBoW gives 0.9253, LSVM gives 0.9391 and DNN gives an average recall of 0.9572. The proposed model provides the best values of precision and f-measure for Dataset#3, i.e., 0.9011 and 0.9367 respectively. These results are better than those achieved by the other models. The results given in Table 4.2 conclude that the proposed model outperforms other models. While implementation, the configuration consumes time close to few seconds to a minute to be evaluated, i.e., a cyberbullying detection dataset with some hundred to few thousand instances on our Intel Core_i7@4.0 GHz processor with 16 GB memory. The underlying cuckoo search algorithm has been set to maximum number of generations equal to 100. The proposed model took few minutes to an hour depending on datasets (1762 instances to 16914 instances) and selected configuration.

Table 4.3 gives an indication of the combinations of learning scheme (pre-processing steps, feature selection techniques and classification models) i.e. cuckoo's solution that score best and hence contribute most to the detection process. It presents the recall, precision, f-measure and AUC values of best three and one worst cuckoo's solutions in the last generation. First nine bits of cuckoo's solution represents the choice of pre-processing steps (punctuation removal, tokenization, lower casing, stemming, hashtag removal, number removal, URL removal, duplication removal and diacritic removal respectively) representing bit '1' for selection and '0'

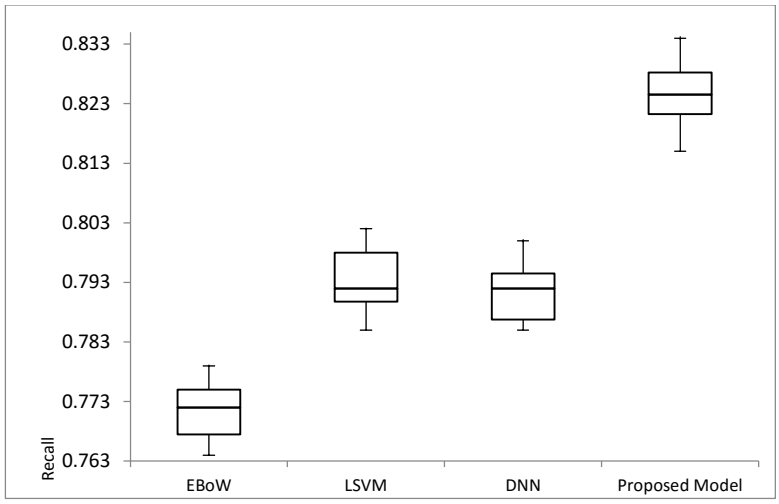
Table 4.3: Results of Performance Metrics for Datasets (depicting three Best and one Worst configuration)

Datasets	Configuration	Cuckoo's Solution	Recall	Precision	F-Measure	AUC
Dataset#1	Best 1	110010100100110110	0.8201	0.8077	0.8139	0.8524
	Best 2	111011000110110100	0.8119	0.7824	0.7969	0.8417
	Best 3	111000101110100110	0.7915	0.7713	0.7813	0.8197
	Worst	010010001000000101	0.7047	0.6248	0.6623	0.7241
Dataset#2	Best 1	100010100110110110	0.8303	0.8009	0.8153	0.8637
	Best 2	111000100110100101	0.8157	0.7824	0.7987	0.8427
	Best 3	001111010100100110	0.7928	0.7715	0.7820	0.8196
	Worst	010001001100000001	0.7641	0.6845	0.7221	0.7589
Dataset#3	Best 1	011100100011110101	0.9648	0.9194	0.9416	0.9841
	Best 2	111110100110100110	0.9597	0.9014	0.9296	0.9792
	Best 3	111010001100111111	0.9508	0.9002	0.9248	0.9765
	Worst	100100000000000100	0.8373	0.8147	0.8258	0.8546
Dataset#4	Best 1	110010000111000110	0.9771	0.9039	0.9391	0.9875
	Best 2	010010100101100110	0.9604	0.8956	0.9269	0.9801
	Best 3	110110100100010100	0.9512	0.8814	0.9150	0.9779
	Worst	010010000010100000	0.8535	0.7854	0.8180	0.8824

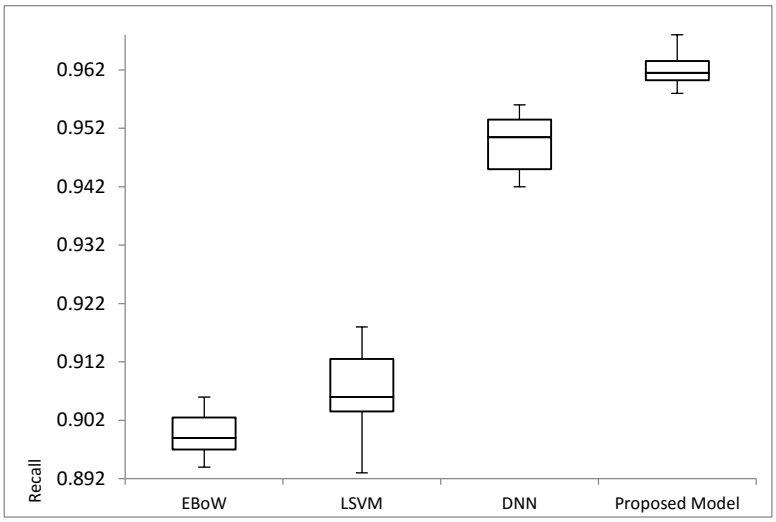
otherwise. Similarly, next eight bits are utilized (word unigrams BoW, word bigrams BoW, word trigrams BoW, TF-IDF weighted unigrams, TF-IDF weighted bigrams, TF-IDF weighted trigrams, profanity and pronouns respectively) for feature selection techniques (1 for selection/ 0 for ignoring). The last bit is for selection of classification models (0 for SVM and 1 for NB). The maximum recall values obtained for all datasets is far better highlighting the model gaining benefits from a variety of learning schemes. The results show that the trained systems generalise well on unseen data. Even for the worst-case scenario, it obtains better recall for all four datasets. As in Table 4.3 for Dataset#1 (Best 1, Best 3), Dataset#2 (Best 1, Best 3), Dataset#3 (Best 2, Best 3) and Dataset#4 (Best 1, Best 2), it has been seen that profanity with pronouns (feature selection techniques) suggest that profane language used with pronouns, characterises many cyberbullying posts. The group of feature selection techniques, combination of pre-processing steps and suitable classification model provides a considerable performance increase over other un-optimized models. The top-scoring systems for each dataset do not differ a lot in performance, except the best 3 for Dataset#4, when compared to the runners-up.



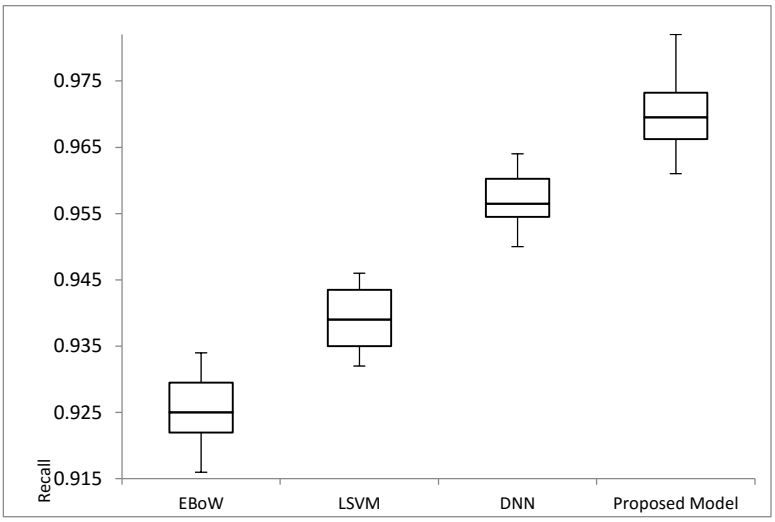
(i)



(ii)



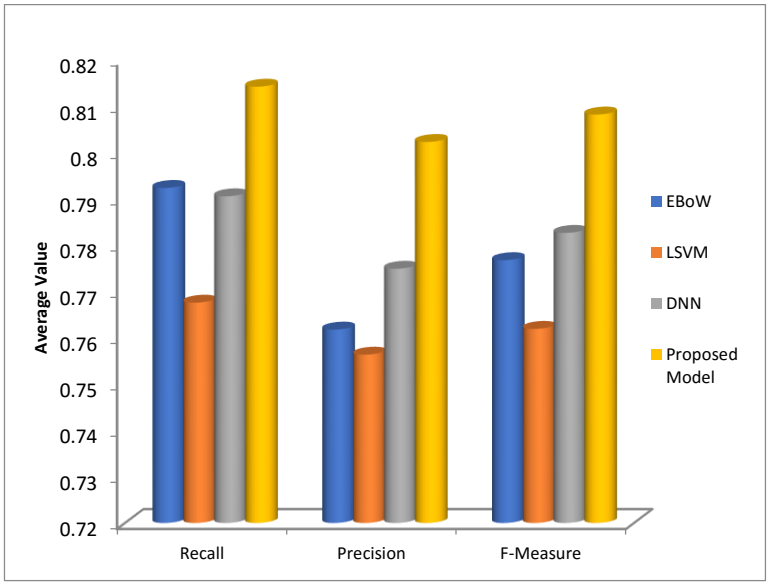
(iii)



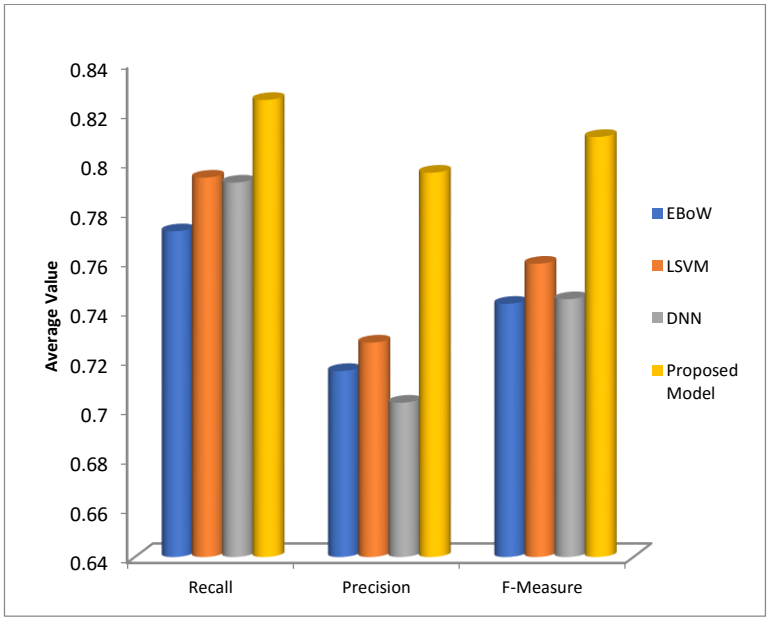
(iv)

Figure 4.3 (i, ii, iii, iv): Recall value of the Models for Dataset#1, #2, #3 and #4 respectively

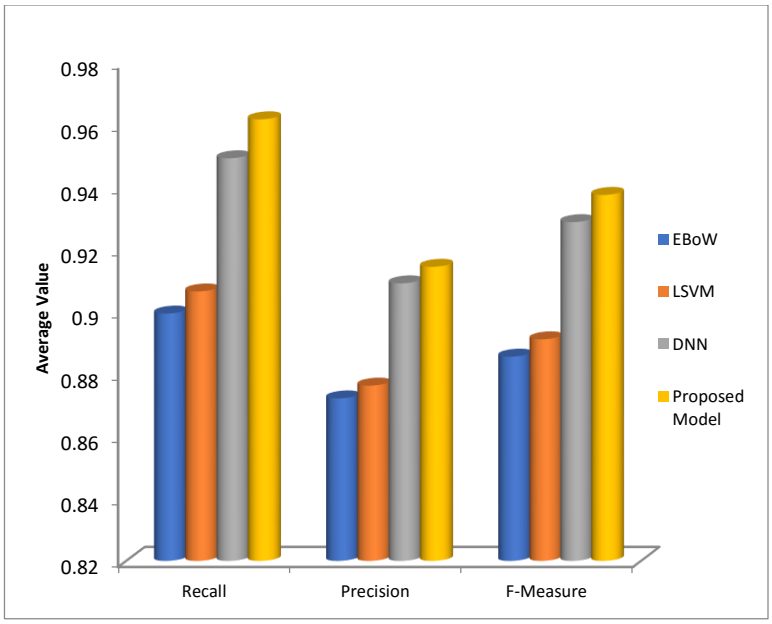
The analysis of the combined pre-processing steps and feature selection techniques reveals that the word unigram BoW and profanity with pronouns prove to be strong features for this all datasets. Also, TF-IDF weighted unigram is good for datasets (except the fourth one). The performance comparison of all the considered models and proposed model is graphically visualized by drawing boxplot (Williamson *et al.* (1989)). It depicts the empirical distribution of the data. Figure 4.3 (i, ii, iii, iv) represents the boxplot for existing and proposed models.



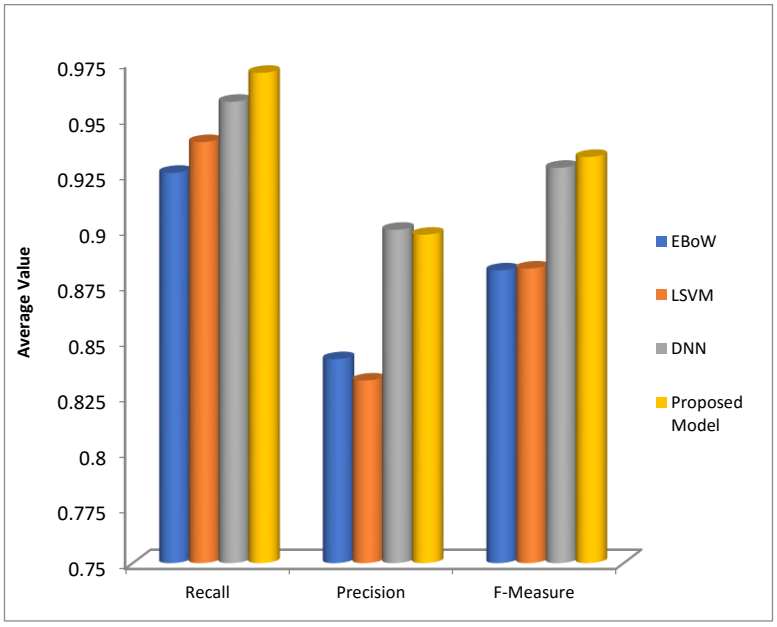
(i)



(ii)



(iii)



(iv)

Figure 4.4 (i, ii, iii, iv): Average Recall, Precision, F-Measure value of the Models for Dataset#1, #2, #3, #4 respectively

In this plot, the name of the models is represented by the x-axis and y-axis labels the corresponding parameter under consideration, i.e. recall value. It is observed from the boxplots

that the proposed model gives better and consistent results in comparison to other models for the considered performance metrics. The reason is that the proposed model explores optimal combinations of learning scheme, while EBoW, LSVM and DNN models estimate only on a pre-established combination. The stability of the models is also achieved due to the absence of outliers. The boxplot reveals that even the minimum recall value achieved by the proposed model is higher than the maximum recall values of other models for some datasets. It is also depicted that the difference in average recall value of the proposed model to others is more in case of Dataset#1 and Dataset#2 than other datasets. In the case of other models, DNN model performed better than EBoW and LSVM models for Dataset#3 and Dataset#4. The evaluation results of recall, precision, and f-measure of all datasets are shown in Figure 4.4 (i, ii, iii, iv).

4.4 Discussion

In this chapter, a novel model is proposed based on the selection of right choice of preprocessing steps, feature selection techniques and classification models to detect cyberbullying text in four different datasets from Twitter, ASKfm, and Formspring. The CS meta-heuristic approach has been used as an optimizer for selecting the best combination of preprocessing steps, feature selection techniques, and classification models. The proposed model showcased great effectiveness in classifying the bully text by achieving the highest recall value of 97.03%.

Next chapter presents a novel stacking ensemble framework that chooses a best selection of classification models along optimal set of tuning parameters using Cuckoo Search meta-heuristic.

Chapter 5

Cuckoo Inspired Stacking Ensemble Framework

This chapter proposes a novel cuckoo inspired stacking ensemble framework that is the integration of Cuckoo Search and several machine learning models. The proposed framework automatically seeks for near-optimal combinations of classification techniques along with their tuning parameters for efficiently solving the problem of content-based cybercrime detection.

5.1 Introduction

The current work proposes a cuckoo inspired stacking ensemble model to further increase predictive recall to identify content-based cybercrime in online social media datasets. The stacking ensemble (Wolpert (1992)) comprises of a set of base classifiers (Stage1) whose output is subsequently exploited by the meta-classifier (Stage2) to generate final model which in turn predicts the class. In the first stage, eight types of ML algorithms are utilized to create base learners. The proposed approach seeks for the best combination of base learners and their configurations. The output from the Stage1 classifiers is fed as input for Stage2 meta-classifier for final predictions. It is clear from the experimental results that the proposed framework is a viable and effective way for precisely identifying content-based cybercrime in online social media. The present work is a first step towards developing an efficient classification model based on the Cuckoo Search (CS) algorithm in the field of content-based cybercrime detection. The following

are the main research work contributions: (i). The underlying work utilizes cuckoo search inspired framework for automatically seeking near-optimal combinations of classification techniques. (ii). The proposed work intelligently seeks for the optimal set of tuning parameters of these classification techniques to further improve the classification recall. (iii). Different empirical experiments are conducted on four real-world, publicly available online social network datasets to verify the effectiveness of the proposed framework. The key target of the current work is to improve the performance of the proposed framework. The same is accomplished by the utilization of the CS metaheuristic yielding the best performance by generating near-optimal configuration of classification schemes.

The overall performance of the detection models can be improved by automatic selection and combination of techniques in alternative ways. The same can be achieved efficiently with the help of nature-inspired algorithms. Among the existing meta-heuristics, the CS algorithm is proven to outperform other meta-heuristic algorithms as revealed in the literature (Gandomi *et al.* (2013); Civicioglu and Besdok (2013)). In the proposed CS framework, the recall value (as a fitness function) has been evaluated across different CS configurations. The recall parameter is considered as the objective function evaluation in exploring the maximum generalization ability of classifier. The motivation of this research work is:

(i) To the best of our knowledge, the literature survey reveals, no work has been done so far in the field of content based cybercrime detection utilizing evolutionary metaheuristic and stacking ensemble for improving the performance of classification detection models.

(ii) There was a need to explore the effect of different configurations (tuning parameters) of classification techniques on the performance of the detection model.

Table 5.1: Comparative list of different Classification Models used in the Proposed & Existing Approaches

Literature	Classification Models							
	SVM	NB	DNN	CNN	RF	AdaBoost	LR	Bagging
Yin *et al.* (2009)	✓							
Reynolds *et al.* (2011)						✓		
Nahar *et al.* (2012)	✓							
Nahar *et al.* (2014b)	✓							✓
Huang *et al.* (2014)		✓				✓		
Mangaonkar *et al.* (2015)	✓	✓					✓	
Van Hee. *et al.* (2015)	✓							
Hosseinmardi *et al.* (2015)	✓							
Singh *et al.* (2016)					✓			
Zhao *et al.* (2016)	✓							
Raisi and Huang (2017)					✓			
Singh *et al.* (2017)	✓		✓					
Agrawal and Awekar (2018)			✓	✓				
Dadvar and Eckert (2018)			✓					
Rafiq *et al.* (2018)						✓	✓	
Van Hee *et al.* (2018)	✓							
Cheng *et al.* (2019)	✓				✓		✓	
Proposed Model	✓	✓	✓	✓	✓	✓	✓	✓

The proposed model has been compared with existing cyberbullying detection techniques (i.e. Support Vector Machine (SVM) (Van Hee *et al.* (2018)), Naïve Bayes (NB) (Mangaonkar *et al.* (2015)), Deep Neural Network (DNN) (Dadvar and Eckert (2018)), Convolutional Neural Network (CNN) (Agrawal and Awekar (2018)), Random Forest (RF) (Cheng *et al.* (2019)), AdaBoost (Rafiq *et al.* (2018)), Logistic Regression (LR) (Cheng *et al.* (2019)) and Bagging (Nahar *et al.* (2014b)) based on machine learning classification models as described in Table 5.1. As seen in the table, the existing literature considered a particular classifier or some classifier but not combined their results, which may not lead to the best solution. The aim of this research is to attain an unbiased criterion and assess the efficiency of the classification process in the context of cyberbullying detection. The key difference between this proposed framework and others is that this framework will explore near-optimal combinations of classification models and their tuning parameters, rather than estimating only pre-established ones.

5.2 Proposed Framework

The work proposes a novel stacking ensemble model that chooses a best selection of classification models along optimal set of tuning parameters using CS meta-heuristic for the detect cyberbullying text in four different datasets from Formspring, ASKfm, and Twitter. This section showcases the theoretical background along with the flow of the proposed methodology for content-based cybercrime detection.

5.2.1 Preliminaries

Before proceeding, this sub-section describes the keywords used in this study along with their utilization in the existing work.

Stacking Ensemble

Stacking ensemble approach proposed by Wolpert (1992), involves the ensemble of base classifiers whose predictions are used as input to the meta-classifier. The Stacking generates an ensemble from a pool of diverse set of classifiers to learn the relationship between the ensemble outputs and the actual class/labels. The prediction of these base classifiers when combined generates a model that is more accurate in comparison to each individual classifier. Stacking uses the concept of meta-classifier, to combine the individual predictions of these base classifiers. One of the challenges in Stacking is obtaining the appropriate combination of base-level classifiers and the meta-classifier, especially in relation to each specific dataset. Herein the CS metaheuristic is utilized to explore and exploit the search space to provide best stacking configuration in a reasonable amount of time. The strategy in ensemble systems is to create a set of accurate and separate classifiers and combine their results so that the combination surpasses all individual classifiers. Figure 5.1 graphically represents the stacking ensemble considering a set of four base classifiers C_1, C_2, C_3 and C_4 that classifies the dataset into two classes.

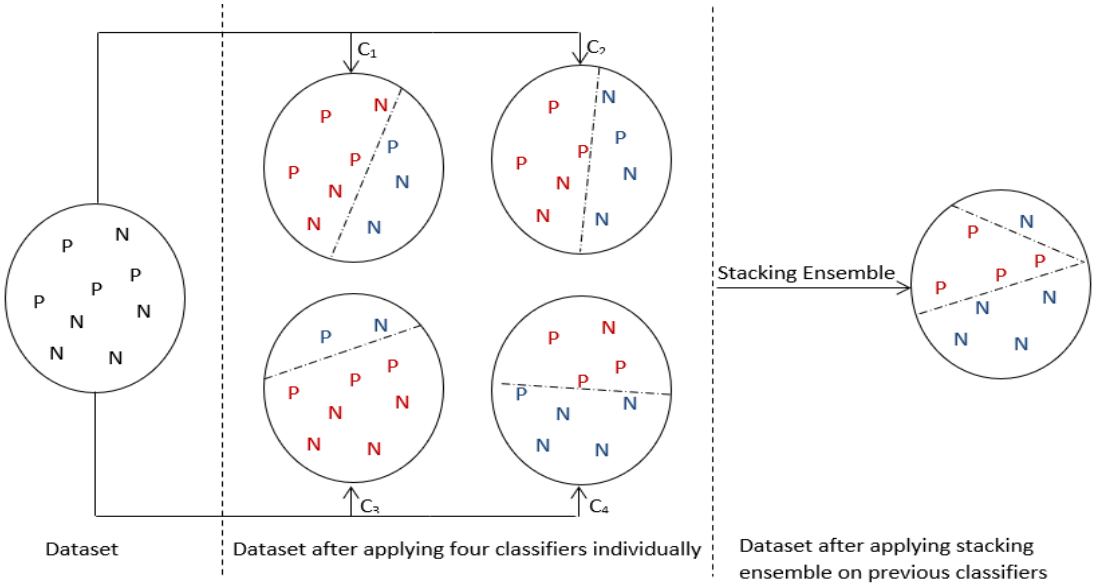


Figure 5.1: An example of stacking ensemble learning considering four base classifiers

5.2.2 Cuckoo Search for Proposed Stacking Ensemble

The proposed approach aims to find a model that is competitive among a broad set of candidate models. Different classifiers with their different optimization parameters represent a model, which, together with the input dataset, determine the functionality of the model. The final model must be competitive within the defined space of models, so the required model must be searched efficiently and accurately. Figure 5.2 illustrates the general flow of the proposed methodology. The following paragraphs include the description of each machine learning task for content-based cybercrime detection.

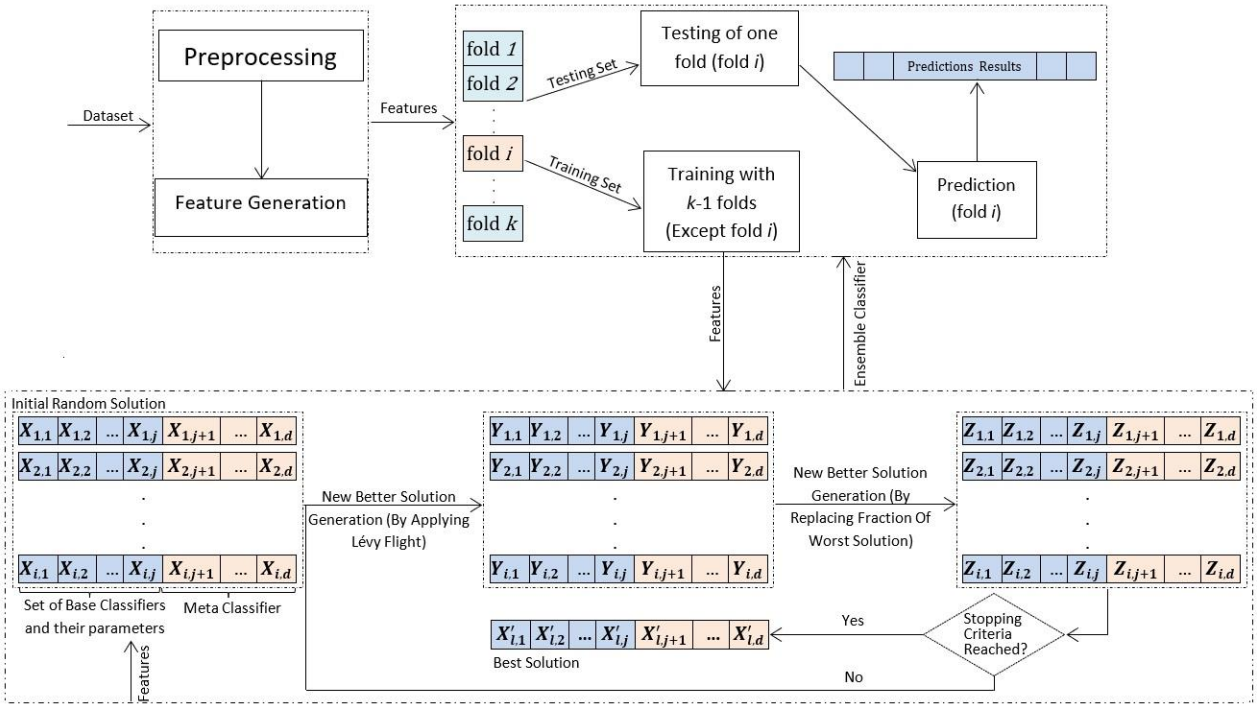


Figure 5.2: A pipeline of Cuckoo Inspired Stacking Ensemble Framework

The data provided as input to the machine learning task is first subjected to a series of pre-processing steps to reduce noisy data, which improves overall accuracy. Various pre-processing steps used in proposed methodology are URL removal, hashtag removal, stemming, lower casing,

tokenization and punctuation removal. After pre-processing the data, features engineering is applied for further model training. The proposed approach considers profanity, pronouns features, TF-IDF weighted bigrams and word bigrams BoW. Figure 5.3 illustrated the detailed pseudocode of the proposed content-based cybercrime detection model. The proposed model is mainly divided into five sections for optimizing the proposed model as follows.

Encoding of Solution

Input of features (F) is represented by a vector of size f, i.e. $F = (F_1, F_2, \dots, F_f)$, set of '$n$' classifiers are represented as $C = (C_1, C_2, \dots, C_n)$ and the results of these classifiers are stored in a vector $R = (R_1, R_2, \dots, R_n)$. Each classifier has its own vector of hyperparameters of size m. Like for k^{th}classifier the hyperparameter vector is defined as, $H_k = (H_{k1}, H_{k2}, \dots, H_{km_k})$. Here, the hyperparameter vector of each classifier ranges from H^1 to H^2. Like for k^{th}classifier the hyperparameter vector ranges from $(H_{k1}^1, H_{k2}^1, \dots, H_{km_k}^1)$ to $(H_{k1}^2, H_{k2}^2, \dots, H_{km_k}^2)$. The result of k^{th} classifier is represented as, $R_k = C_k(F, H_k)$. The final result of a meta-classifier, $C'\epsilon(C_1, C_2, \dots, C_n)$ is stored in $R' = C'(R, H')$, where $H'$is hyperparameter vector of classifier C'. The population of cuckoos (i.e. solution space) is represented by matrix of size $i \times d$, where i is total number of cuckoos and d is its dimension. In the proposed methodology, j number of bits are required for the selection of base classifiers (with $(2^n - 1)$ possibilities) and their hyperparameters (with $\left(\prod_{k=1}^{n}\left(\prod_{l=1}^{m_k}\|H_{kl}\| + 1\right) - 1\right)$ possibilities) and $(d - j)$ number of bits are required for meta-classifier (with n possibilities). Basic hyperparameter settings have been used for meta-classifier to reduce the complexity of algorithm. The structure of single solution (cuckoo's egg or host bird's nest location) is represented as follows.

X_1	X_2	…	X_j	X_{j+1}	…	X_d

The proposed model randomly generated each solution (i.e. cuckoo's population). Binary encoding is used for the selection of base classifiers, meta-classifier and their hyperparameters i.e. $X_i \in \{0,1\}$. The value at the respective index of the substring decides the participation (value 1) or elimination (value 0) of that respective component in the learning process. To restrict the new solutions to binary, Equation (5.1) is used.

$$\mathrm{X}_{l,s}^{t} = \left(\mathrm{P}(\mathrm{X}_{l,s}^{t}) > \bar{\mathrm{u}} \right)? \, 0 : 1,$$

$$\mathrm{P}(\mathrm{X}_{l,s}^{t}) = \frac{1}{1 + e^{\mathrm{X}_{l,s}^{t}}} \tag{5.1}$$

where $\mathrm{X}_{l,s}^{t}$ represents the l^{th} solution at t^{th} generation for the s^{th} variable of cuckoo's dimension and $\bar{\mathrm{u}}$ follows uniform distribution, $\mathrm{U}(0,1)$.

Adding New Solutions to Next Generation

Equation (5.2) is used for adding new solutions to next generation ($\mathrm{X}_{l,s}^{t+1}$) i.e. searching the host bird's nest location of the next generation in order to gain a new set of host bird's nest locations. After comparing with the last generation nest locations ($\mathrm{X}_{l,s}^{t}$), the best bird's nest location is chosen and entered into the next step. The searching pattern of CS algorithm is Lévy flight ($Lévy$) with step size ($\acute{\alpha}$) and the best solution so far is represented as $\mathrm{Opt\_X}_{l,s}$.

$$\mathrm{X}_{l,s}^{t+1} = \mathrm{X}_{l,s}^{t} + \acute{\alpha} \times Lévy \times (\mathrm{X}_{l,s}^{t} - \mathrm{Opt\_X}_{l,s}) \tag{5.2}$$

Finding Worst Solutions

The probability matrix is created for every cuckoo solution in population, to detect foreign eggs. This matrix is generated as in (5.3).

$$\rho_{l,s} = \begin{cases} 0, & \text{if random\_number}(0,1) > \rho \\ 1, & \text{otherwise,} \end{cases} \tag{5.3}$$

Here, probability of discovering foreign eggs (worst solution) is denoted by $\rho_{l,s}$. A uniform random number between 0 and 1 ($\text{random\_number}(0,1)$) is compared with ρ, to check if local random walk is specified consideration or not. New solutions are generated using (5.4) after determining the discovering probabilities.

$$\mathrm{X}_{l,s}^{t+1} = \text{Step\_L} \times \rho_{l,s} + \mathrm{X}_{l,s}^{t} \tag{5.4}$$

Where matrix of local step size i.e. Step_L is obtained by using (5.5).

$$\text{Step\_L} = (\text{perm\_random}_1(\mathrm{X}_{l,s}^{t}) - \text{perm\_random}_2(\mathrm{X}_{l,s}^{t})) \times \text{random\_number}(0,1) \tag{5.5}$$

Here, solutions are randomly shuffled using $\text{perm\_random}()$. Lastly, the recall values (i.e. objective function) are compared for new solutions and the existing ones using (5.6) and the better solutions enter in the following generation.

$$\text{new\_Solution} = \begin{cases} \mathrm{X}_{l,s}^{t+1}, & \text{if } R(\mathrm{X}_{l,s}^{t+1}) > R(\mathrm{X}_{l,s}^{t}) \\ \mathrm{X}_{l,s}^{t}, & \text{else} \end{cases} \tag{5.6}$$

Last two sections are repeated until the maximum number of generations are completed or a predetermined ending measure is fulfilled.

PSEUDOCODE

***Input*:** Feature i.e. $F = (F_1, F_2, \dots, F_f)$, Number of Cuckoo's Population i.e. Solutions(i), Maximum Generations (Max_Gen), Loss Parameters $(\alpha \ \& \ \rho)$, Dimensions of Solutions (d), Pool of classifiers (C)

***Output*:** Globally best Solution (Ensemble of classification algorithms), Classes' Set, C_S = {0, 1} where 0 reflex Non-bully and 1 reflex bully term

Begin*:** ***FOR (l from 1 to i i.e. Cuckoo's Population (Each Solution))

FOR (s from 1 to d i.e. dimension of Solution)

Solution i.e. $X_{l,s}^0 = Random\_number\ \{0, 1\}$;

END

Fitness of l^{th} Solution i.e. Recall Value, $R_l = 0$;

END

WHILE (Generation_Count i.e. $t < Max\_Gen$)

FOR (l from 1 to i, each Solution)

Better New Solution i.e. New_X_l = Fit_Cuckoo (F, R_l, X_l^t);

END

FOR (l from 1 to i, each Solution)

FOR (s from 1 to d, dimension of Solution)

$X_{l,s}^{t+1} = X_{l,s}^t + Lévy\_Flight\ (Using\ (2))$;

END

FOR (s from 1 to d, dimension of Solution)

IF $\left(\bar{u} < \frac{1}{1+e^{X_{l,s}^t}}, (Using\ (1))\right)$

$X_{l,s}^t = 0$;

ELSE

$X_{l,s}^{t} = 1;$

END

$New\_X_l =$ Fit_Cuckoo (F, R_l, X_l^t);

END

FOR (l from 1 to i, each Solution)

FOR (s from 1 to d, dimension of Solution)

Generate Step_L i.e. local step size $Using\ (5)$ and $\rho_{l,s}\ Using\ (3)$

$X_{l,s}^{t+1} =$ Step_L × $\rho_{l,s} + X_{l,s}^{t}, (Using\ (4))$;

END

$New\_X_l =$ Fit_Cuckoo (F, R_l, X_l^t);

END

END

Fitness Function of Cuckoo Search (Fit_ Cuckoo (F, R_l, X_l^t))

Begin*:** ***FOR (s from 1 to d, dimension of Solution)

$\hat{C} = C - \{C : X_{l,s}^{t} = 0\}$;

END

Use $\hat{C}$ for classification using F and store results in $Recall$;

IF $(Recall > R_l)$

$R_l = Recall$;

FOR (s from 1 to d, dimension of Solution)

$opt\_X_{l,s} = X_{l,s}^{t}$;

END

RETURN opt_X_l;

Figure 5.4: Pseudocode of the Proposed Framework

5.3 Simulation Results

The experiments of the proposed model are conducted using Intel core i7@4GHz with 16 GB of RAM running on windows 10 platform and implemented using Python v2.7. In order to find the optimal combination of base-classifiers in the first stage and the meta-classifier in the second stage, the following eight different machine learning algorithms were exploited: SVM (Van Hee *et al.* (2018)), NB (Mangaonkar *et al.* (2015)), DNN (Dadvar and Eckert (2018)), CNN (Agrawal and Awekar (2018)), RF (Cheng *et al.* (2019)), AdaBoost (Rafiq *et al.* (2018)), LR (Cheng *et al.* (2019)) and Bagging (Nahar *et al.* (2014b)). The parameter settings of these classification algorithms are shown in Table 5.2 and Table 5.3 showcase the empirically adjusted hyper-parameter settings of base-classifiers for different datasets.

Various evaluation parameters used for classification task are, Recall, Precision, F-Measure, AUC, four commonly used metrics. Firstly, Recall can be determined as fraction of the total amount of relevant instances that were actually retrieved. Precision is the portion of relevant instances among the retrieved ones. F-Measure is the harmonic mean of precision and recall. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

$$\text{Recall} = \frac{\text{tp}}{\text{tp}+\text{fn}}, \text{Precision} = \frac{\text{tp}}{\text{tp}+\text{fp}}, \text{F}-\text{measure} = \frac{2*\text{p}*\text{r}}{\text{p}+\text{r}}$$

$$\text{AUC} = \int_{x=0}^{1} \left(\frac{\text{tp}}{\text{tp}+\text{fn}}\right)\left(\left(\frac{\text{fp}}{\text{tp}+\text{fn}}\right)^{-1}(x)\right)dx$$

Where tp, fp, fn and the true positive, false positive, false negative respectively. The following paragraph discusses the comparative results of the proposed approach with above mentioned state-of-the-art methods.

Table 5.2: The default Hyper-Parameter settings of eight different machine learning classification algorithms used in the proposed framework

CLASSIFIER		SVM	NB	DNN	CNN	RF	AdaBoost	LR	Bagging
PARAMETER	#1	Kernel = LINEAR	fit_prior = TRUE	learning_rate = 0.1	embed_size = 300	min_samples _leaf = 3	learning_rate = 1.0	dual = FALSE	max_features = 1.0
	#2	Gamma = auto	alpha = 0.001	Hidden layer = 100	filter_sizes = 300	min_samples _split = 3	base_estimator = None	c = 0.01	max_samples = 1.0
	#3	C = 1.0	-	Hidden_layer_ neurons = 100	max_feature = 10000	max_depth = 12	n_estimators = 50	-	base_estimat or = None
	#4	Max_Iter = 1	-	-	num_filters = 32	warm_start = TRUE	random_state = None	-	oob_score = FALSE
	#5	Probability = FALSE	-	-	Maxlen = 200	n_estimators = 8	algorithm = SAMMER.R	-	boostrap_feat ures = FALSE
	#6	Verbose = FALSE	-	-	batch_size = 256	oob_score = TRUE	-	-	boostrap = TRUE
	#7	cache_size = 200	-	-	Epochs = 3	Verbose = FALSE	-	-	verbose = 0
	#8	class_weight = NONE	-	-	Optimizer = adam	-	-	-	random_state = None
	#9	decision_fun ction_shape = ovr	-	-	Loss = binary_cross entropy	-	-	-	warm_start = FALSE
	#10	Degree = 3	-	-	output_layer _function = Sigmoid	-	-	-	-

Table 5.3: The Hyper-Parameter settings of various machine learning classification algorithms used to classify four different datasets

DATASETS		Dataset#1					Dataset#2			Dataset#3		Dataset#4			
CLASSIFIERS		SVM	NB	LR	Bagging	RF	SVM	LR	Bagging	DNN	CNN	SVM	NB	DNN	RF
PARAMETER	#1	LINEAR	TRUE	FALSE	3	5	LINEAR	TRUE	2	0.1	300	LINEAR	TRUE	0.1	8
	#2	auto	0.001	0.01	1	3	auto	0.001	2	100	200	auto	0.0001	110	3
	#3	1	-	-	NONE	12	5	-	NONE	100	10000	20	-	100	3
	#4	1	-	-	FALSE	TRUE	1	-	FALSE	-	30	1	-	-	TRUE
	#5	FALSE	-	-	TRUE	10	TRUE	-	TRUE	-	200	FALSE	-	-	20
	#6	FALSE	-	-	0	TRUE	FALSE	-	0	-	256	TRUE	-	-	TRUE
	#7	200	-	-	None	FALSE	150	-	NONE	-	3	200	-	-	TRUE
	#8	NONE	-	-	FALSE	-	NONE	-	TRUE	-	adam	NONE	-	-	-
	#9	ovr	-	-	-	-	ovr	-	-	-	binary_ crossentropy	ovr	-	-	-
	#10	3	-	-	-	-	3	-	-	-	Sigmoid	3	-	-	-

Table 5.4: Comparative performance of Cuckoo Inspired Stacking Ensemble Framework with the state-of-the-art approaches

			Models								
	Evaluation Parameters		**SVM[1]**	**NB[2]**	**DNN[3]**	**CNN[4]**	**RF[5]**	**AdaBoost[6]**	**LR[7]**	**Bagging[8]**	**Proposed Model**
Twitter#1 Dataset	Recall	Average	0.7694	0.7145	0.7912	0.7854	0.8184	0.7512	0.761	0.8046	**0.8379**
		Standard-Deviation	0.004785	0.006593	0.005361	0.004094	0.007183	0.005637	0.005681	0.002382	0.003163
	Precision	Average	0.7562	0.7496	0.7752	0.7921	0.7493	0.7109	0.8197	0.7941	**0.8306**
		Standard-Deviation	0.006134	0.003539	0.003734	0.004991	0.006544	0.007003	0.00635	0.007071	0.006887
	F-Measure	Average	0.7627	0.7316	0.7831	0.7887	0.7823	0.7305	0.7893	0.7993	**0.8342**
		Standard-Deviation	0.005376	0.004606	0.004402	0.004498	0.006849	0.006246	0.005997	0.003564	0.004335
	AUC		0.7839	0.7685	0.8147	0.8008	0.7816	0.7621	0.8345	0.8194	**0.8752**
ASKfm Dataset	Recall	Average	0.7934	0.7107	0.7927	0.801	0.7513	0.7452	0.7814	0.7912	**0.8417**
		Standard-Deviation	0.005198	0.003384	0.004568	0.003617	0.004232	0.003232	0.005834	0.005412	0.005332
	Precision	Average	0.7267	0.6542	0.7041	0.7511	0.7124	0.7219	0.7345	0.7115	**0.8029**
		Standard-Deviation	0.004704	0.005297	0.00435	0.005235	0.00489	0.004762	0.003114	0.007235	0.005654
	F-Measure	Average	0.7586	0.6813	0.7458	0.7752	0.7313	0.7334	0.7572	0.7492	**0.8218**
		Standard-Deviation	0.004939	0.004130	0.004456	0.004278	0.004537	0.003851	0.004061	0.006192	0.005488
	AUC		0.8148	0.7416	0.8121	0.8224	0.7925	0.7847	0.8045	0.8111	**0.8731**

[1] SVM (Van Hee *et al.* (2018))
[2] NB (Mangaonkar *et al.* (2015))
[3] DNN (Dadvar and Eckert (2018))
[4] CNN (Agrawal and Awekar (2018))

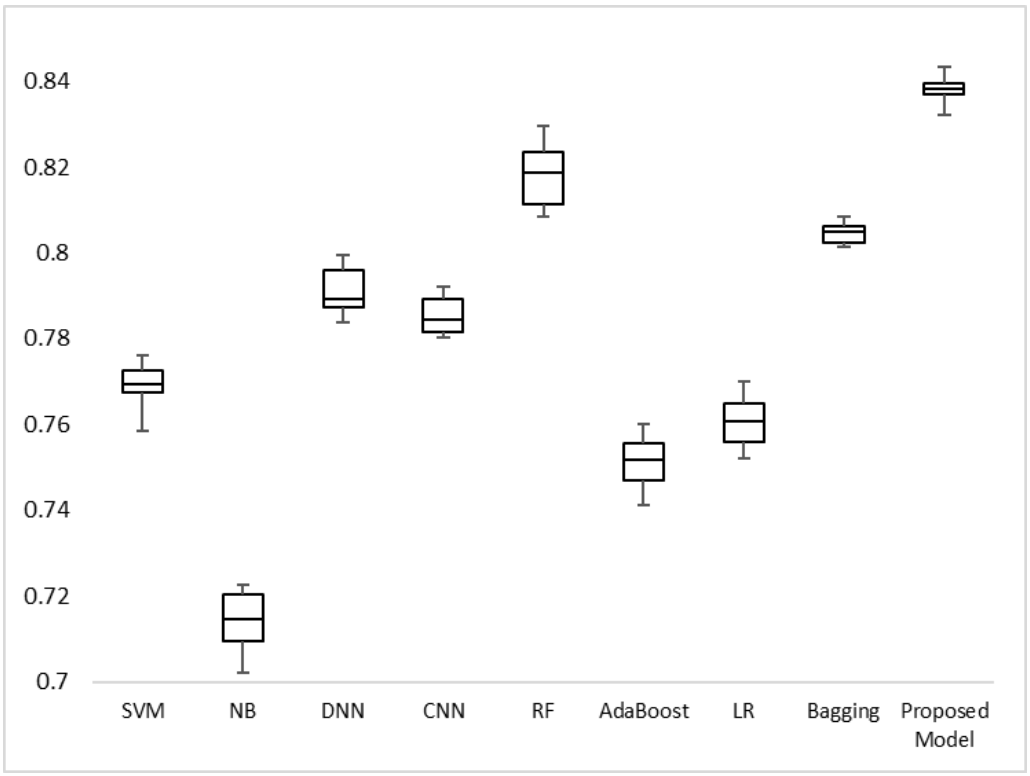
Table 5.4 Continued…

	Evaluation Parameters		SVM[1]	NB[2]	DNN[3]	CNN[4]	RF[5]	AdaBoost[6]	LR[7]	Bagging[8]	Proposed Model
			Models								
Formspring Dataset	Recall	Average	0.9081	0.8156	0.9527	0.9481	0.8954	0.8567	0.8924	0.9214	**0.9724**
		Standard-Deviation	0.004866	0.005557	0.005687	0.003433	0.002049	0.002174	0.003364	0.005356	0.003586
	Precision	Average	0.8796	0.8572	0.9093	0.9126	0.8641	0.8145	0.8478	0.8229	**0.9104**
		Standard-Deviation	0.005371	0.004403	0.005164	0.005334	0.003417	0.002273	0.004805	0.004906	0.007121
	F-Measure	Average	0.8936	0.8359	0.9305	0.9300	0.8795	0.8351	0.8695	0.8694	**0.9404**
		Standard-Deviation	0.005106	0.004913	0.005413	0.004177	0.002562	0.002222	0.003957	0.005121	0.004770
	AUC		0.9498	0.8327	**0.9749**	0.9682	0.9243	0.8745	0.9116	0.9164	0.9745
Twitter#2 Dataset	Recall	Average	0.9452	0.9114	0.9591	0.9234	0.9227	0.8968	0.8842	0.9298	**0.9782**
		Standard-Deviation	0.005047	0.002468	0.003255	0.004175	0.004926	0.005028	0.003846	0.005108	0.003838
	Precision	Average	0.9027	0.8945	0.9001	0.8745	0.9121	0.8421	0.8785	0.8124	**0.9427**
		Standard-Deviation	0.004334	0.003715	0.005203	0.003825	0.004961	0.004138	0.004929	0.004102	0.003452
	F-Measure	Average	0.9235	0.9029	0.9287	0.8983	0.9174	0.8686	0.8813	0.8671	**0.9601**
		Standard-Deviation	0.004663	0.002966	0.004005	0.003992	0.004943	0.004540	0.004321	0.004550	0.003635
	AUC		0.9757	0.9549	0.9814	0.9441	0.9698	0.9145	0.9012	0.9342	**0.9906**

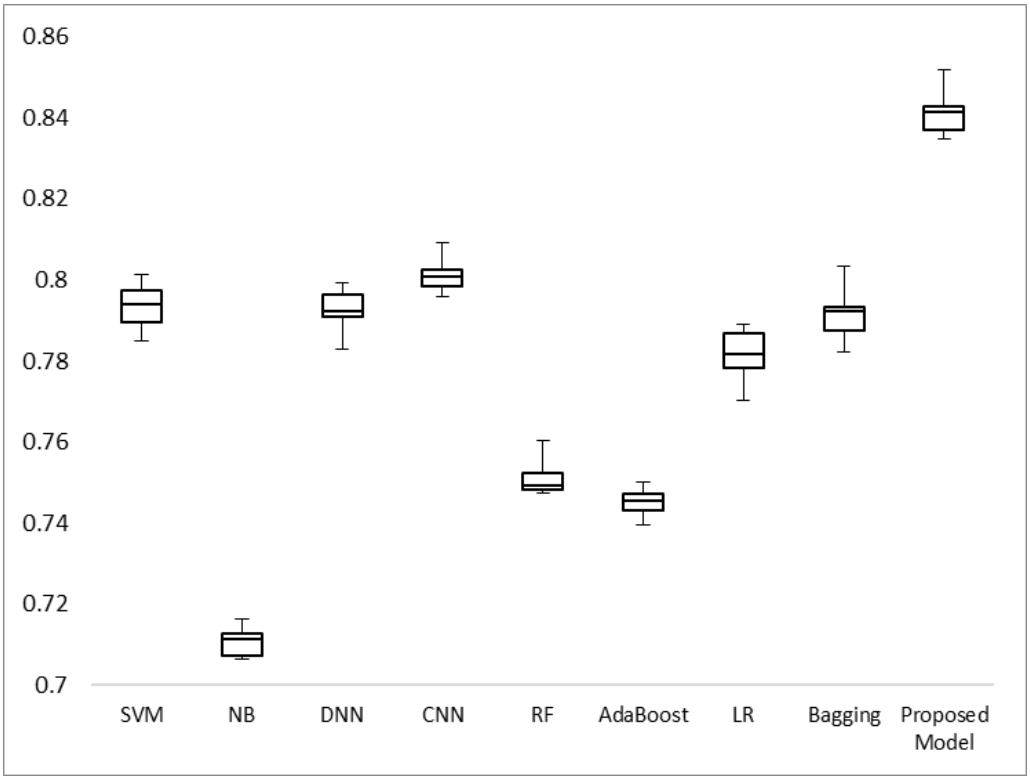
[5] RF (Cheng *et al.* (2019))
[6] AdaBoost (Rafiq *et al.* (2018))
[7] LR (Cheng *et al.* (2019))
[8] Bagging (Nahar *et al.* (2014b))

The table 5.4 depicts the average results of ten runs of various models' execution in terms of different evaluation parameters i.e. precision, recall, f-measure and AUC values along with their standard deviation values. The key parameter applied in this work is recall. The proposed model presents better average recall of 0.8379 for Twitter#1 Dataset in comparison to the other models. Moreover, the model executes remarkably precise in separating bully content from the dataset with a precision of 0.8306 and AUC of 0.8752. For Twitter#1 Dataset, the NB model showcase 0.7145 as the lowest recall value and AdaBoost gives the lowest values of precision and AUC value i.e. 0.7109 and 0.7621 respectively in contrast to other models. The proposed model showcases excellent performance in terms of average recall of 0.8417 for ASKfm Dataset. The model presents values 0.8029 as precision and 0.8731 as AUC value for the same dataset.

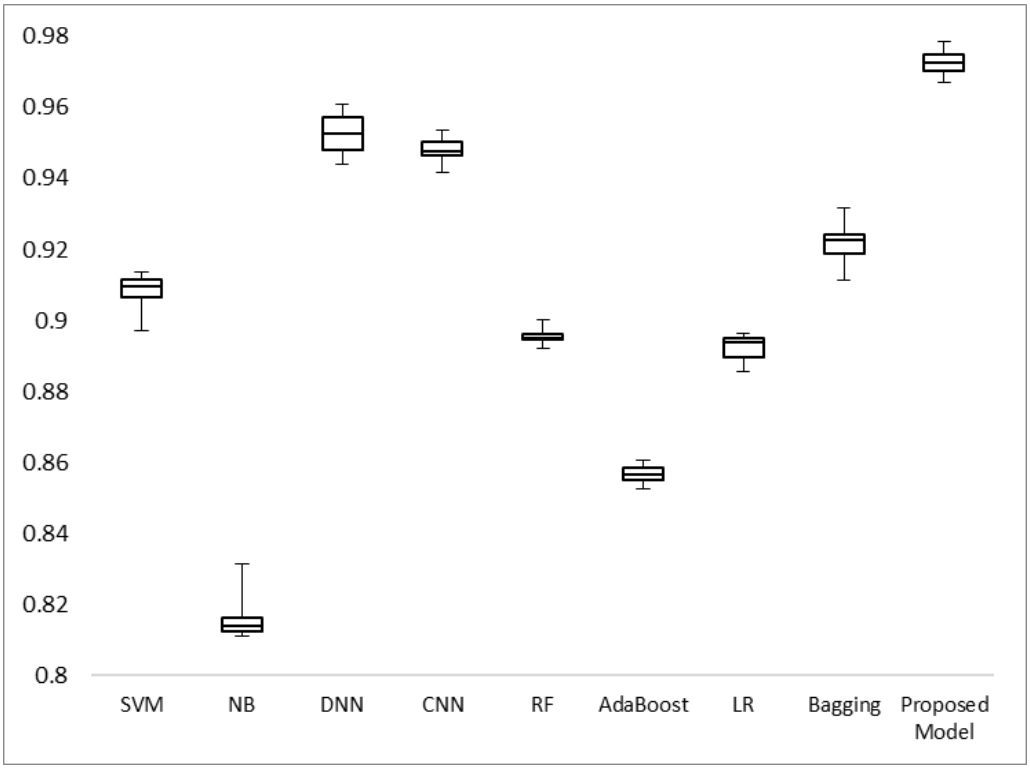
In case of Formspring Dataset, the DNN model achieved better results with 0.95 as average recall but its results are not superior to the proposed model. The RF and LR models give almost similar results of recall value of 0.8954 and 0.8924 respectively. The proposed model gives the best average recall of 0.9724 for this dataset. The model provides 0.91 as precision value and 0.97 as AUC value. Finally, the proposed model again performs best with an average recall value of 0.9782 for Twitter#2 Dataset, though DNN and SVM provides an average recall of 0.9591 and 0.9452 respectively. For Twitter#2 Dataset, the proposed model presents the best values of precision, f-measure and AUC values i.e., 0.9427, 0.9601 and 0.9906 respectively. These results are better than those attained by the other models. The results presented in Table 5.4 deduce that the proposed model surpasses other models.


0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

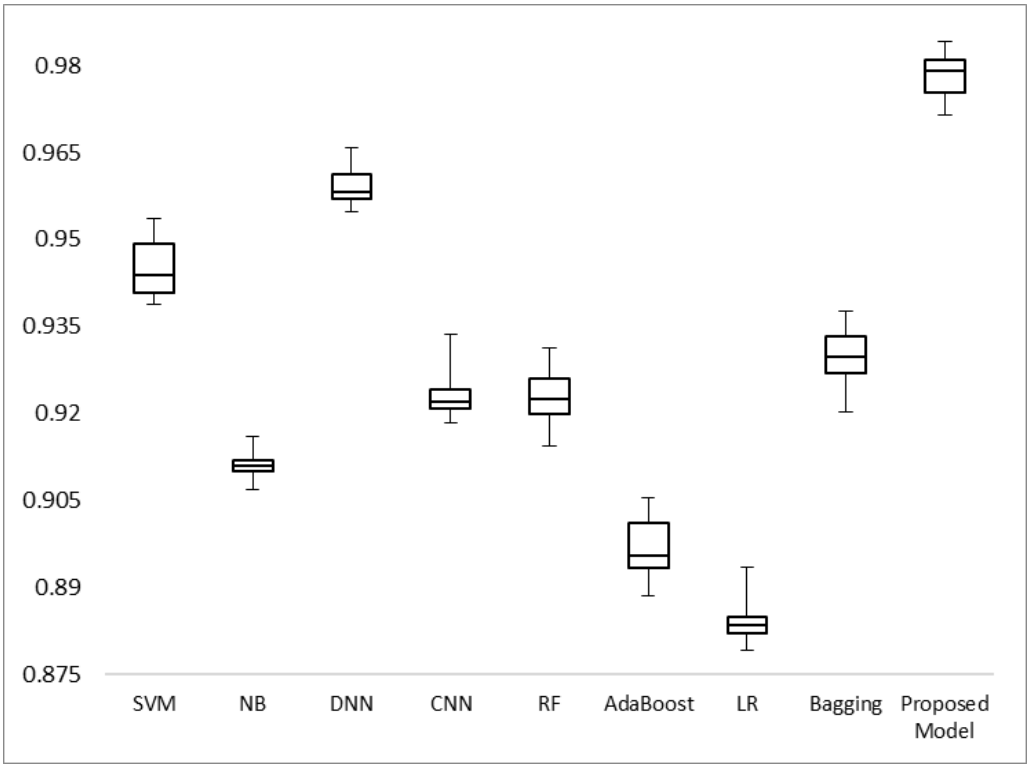
(i)


0.86
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

(ii)

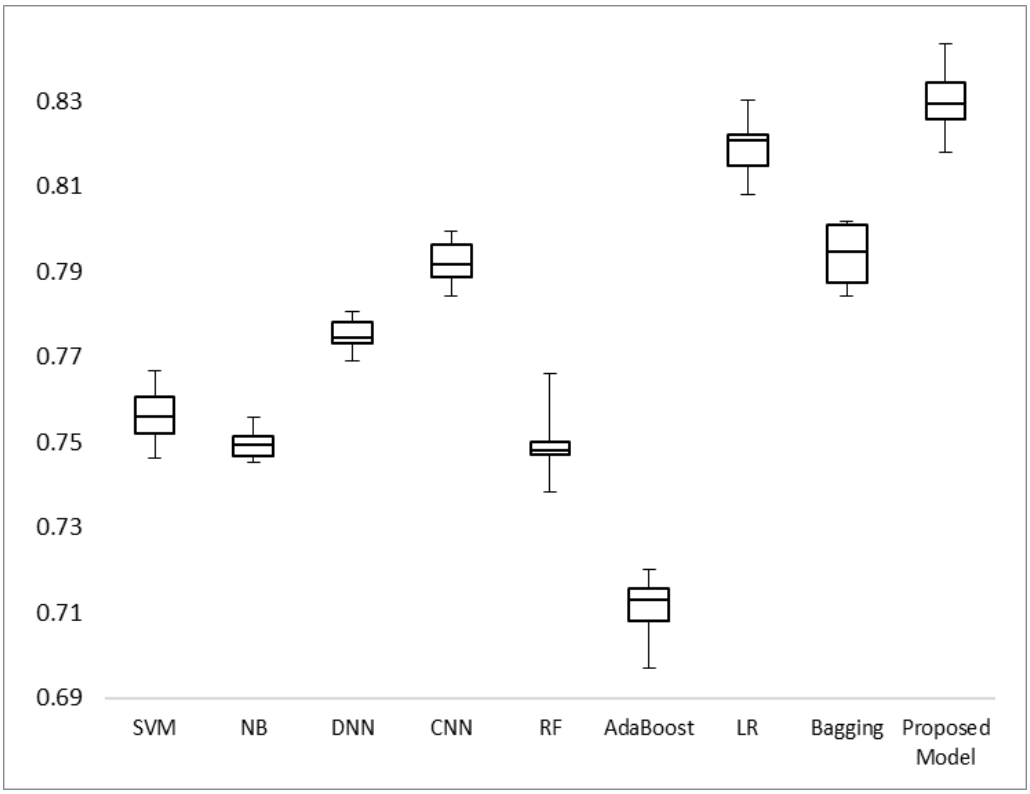


(iii)

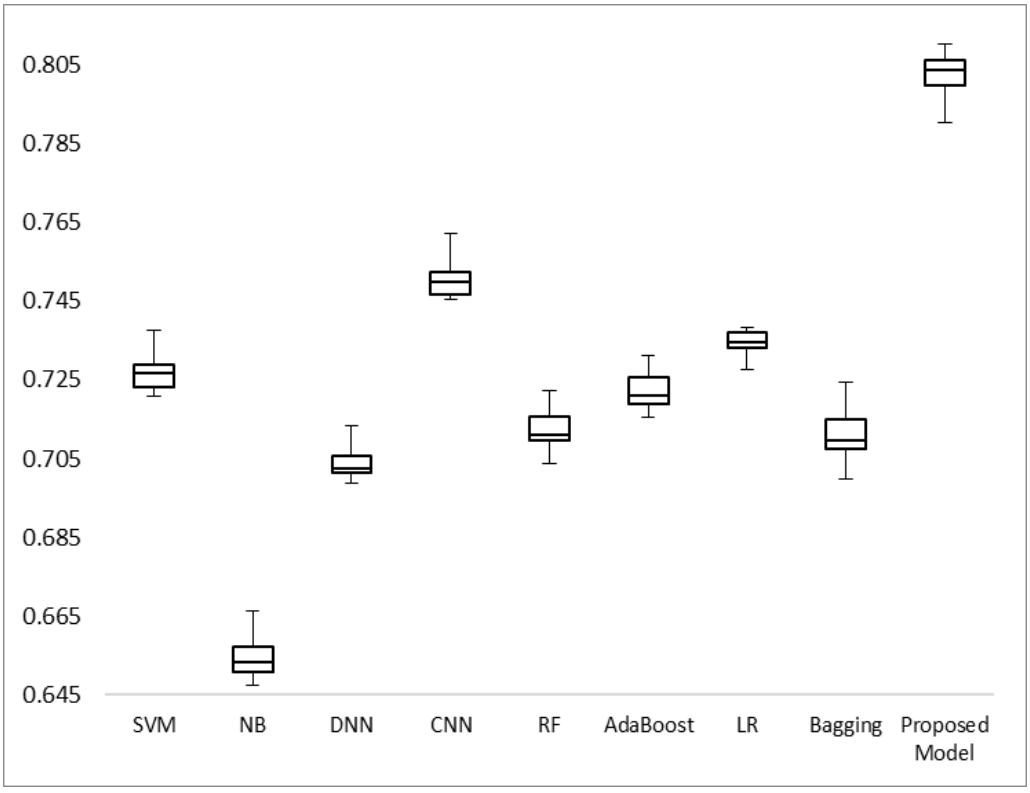


(iv)

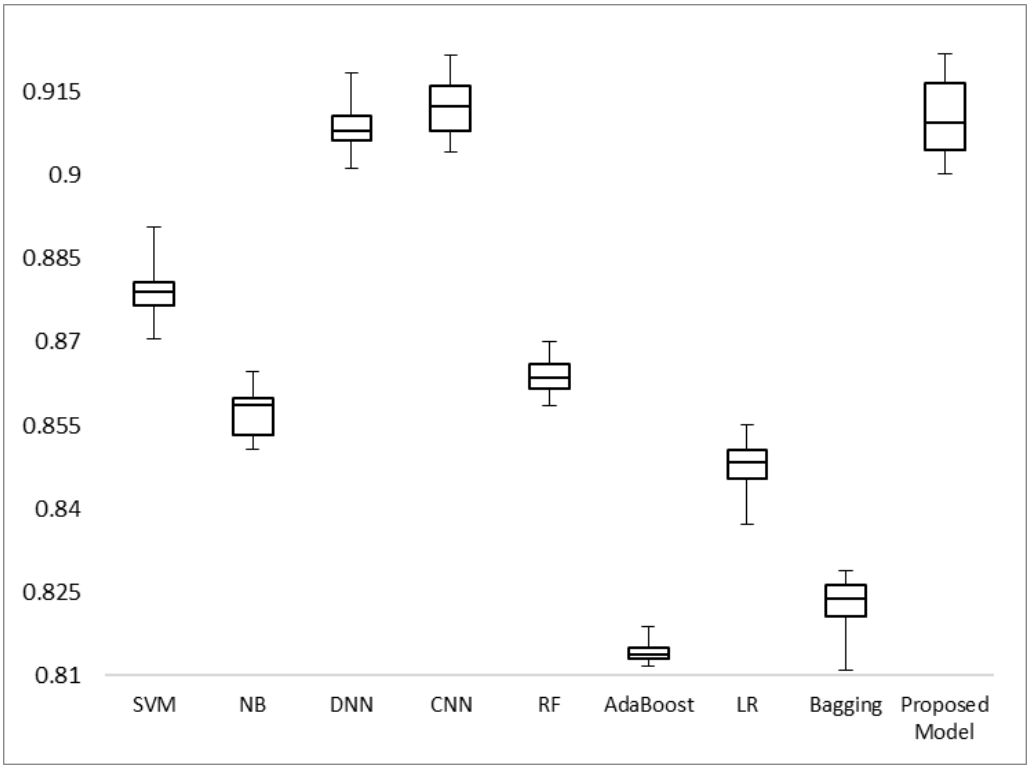
Figure 5.5 (i, ii, iii, iv): Recall value of Models for Twitter#1, ASKfm, Formspring, and Twitter#2 datasets respectively


0.83
0.81
0.79
0.77
0.75
0.73
0.71
0.69
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

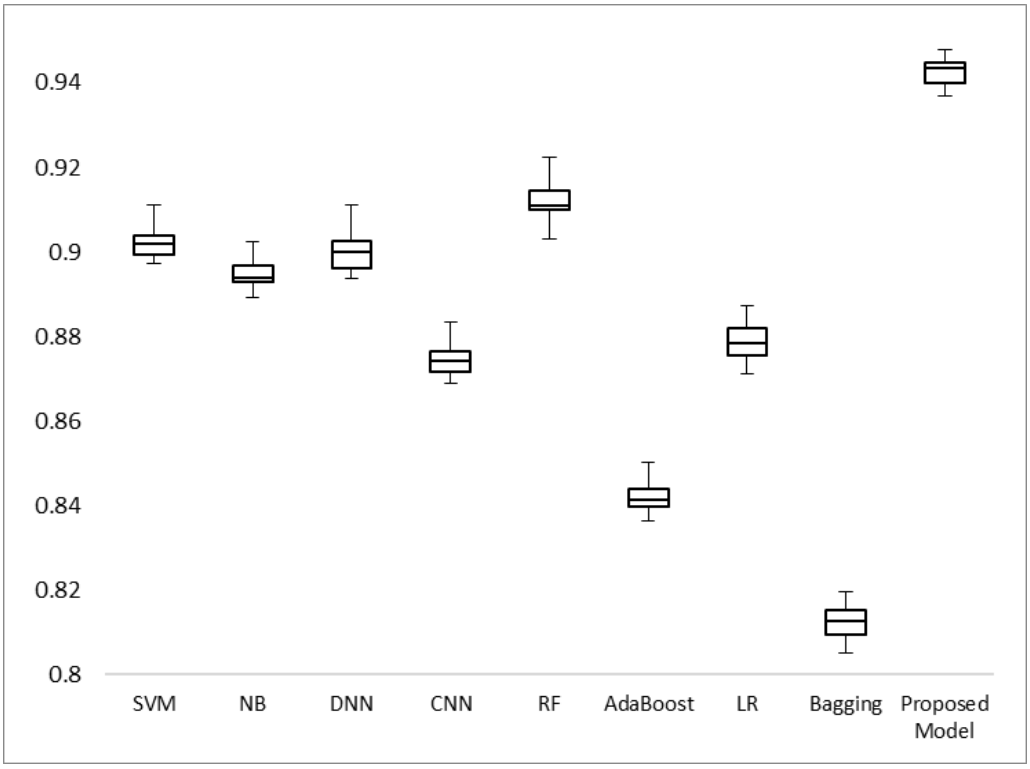
(i)


0.805
0.785
0.765
0.745
0.725
0.705
0.685
0.665
0.645
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

(ii)

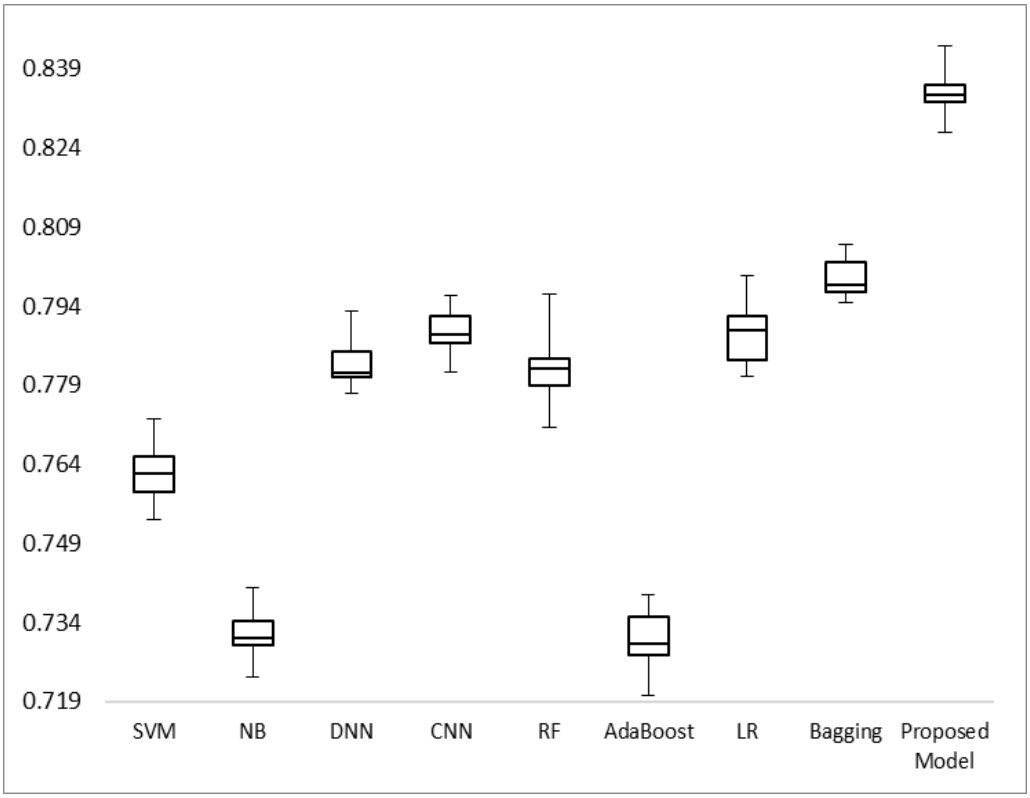


(iii)

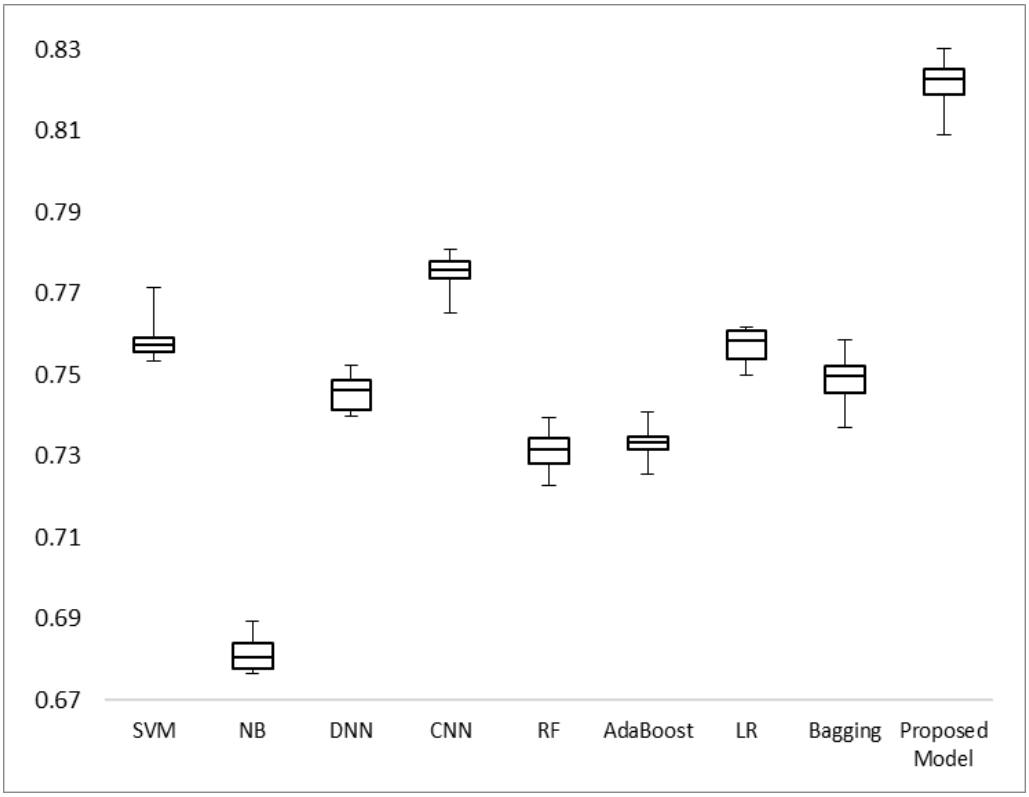


(iv)

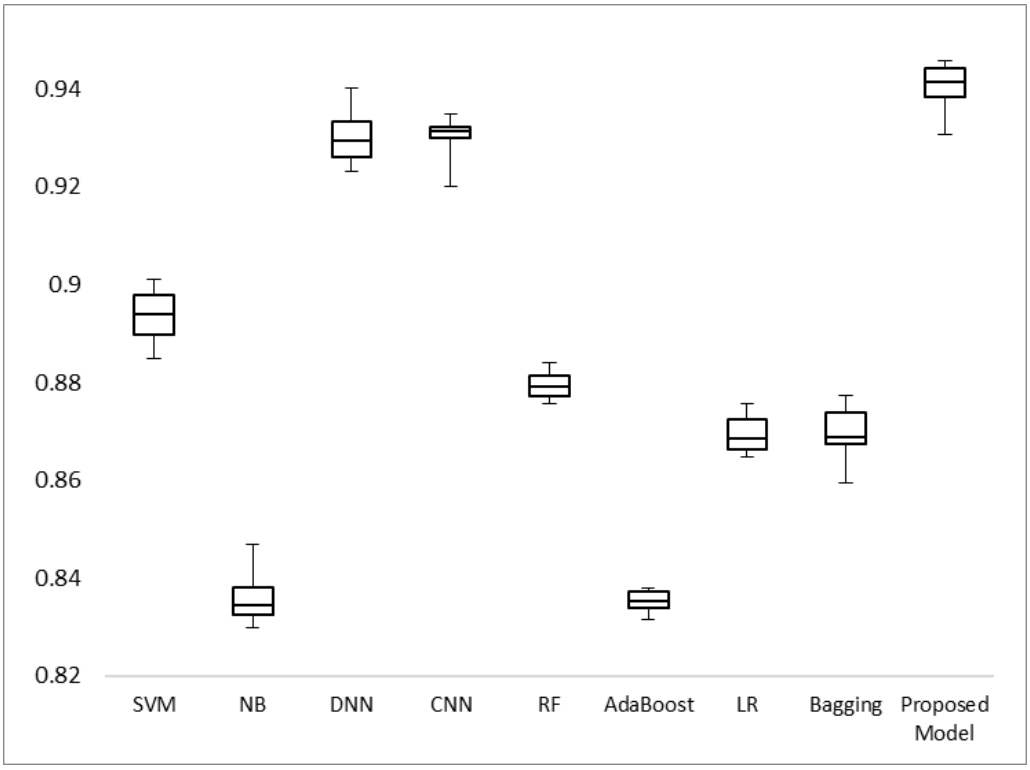
Figure 5.6 (i, ii, iii, iv): Precision value of Models for Twitter#1, ASKfm, Formspring, and Twitter#2 datasets respectively


0.839
0.824
0.809
0.794
0.779
0.764
0.749
0.734
0.719
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

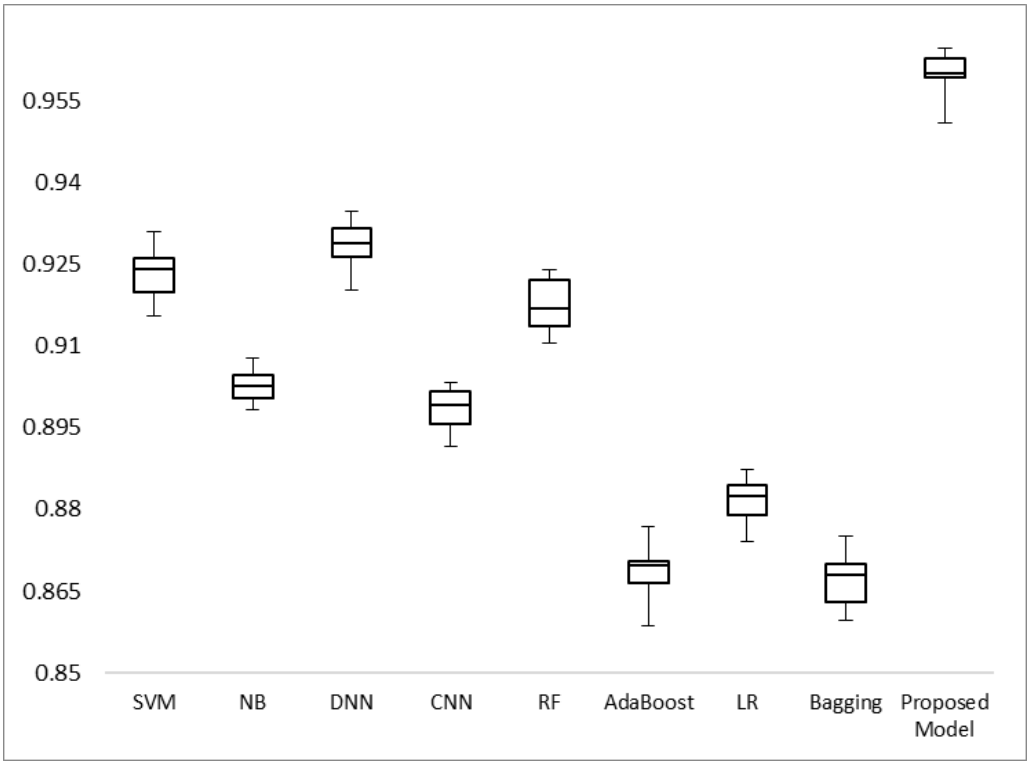
(i)


0.83
0.81
0.79
0.77
0.75
0.73
0.71
0.69
0.67
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

(ii)



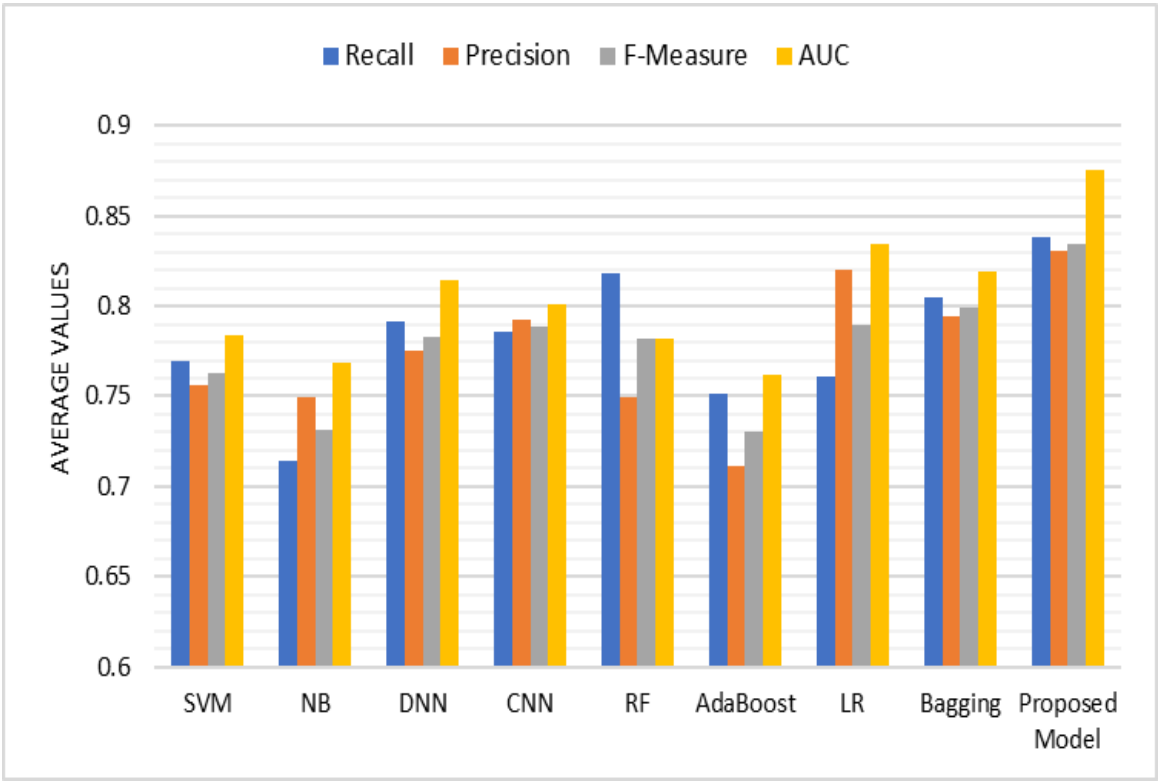
(iii)



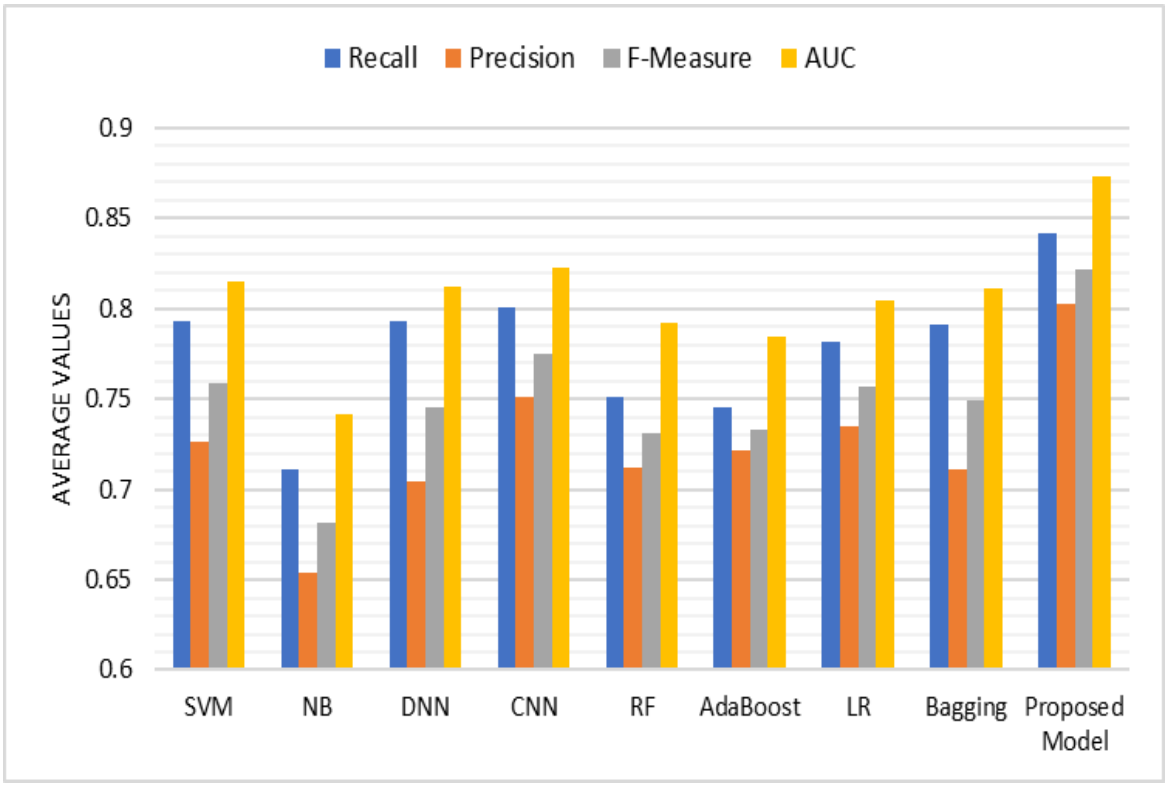
(iv)

Figure 5.7 (i, ii, iii, iv): F-Measure value of Models for Twitter#1, ASKfm, Formspring, and Twitter#2 datasets respectively

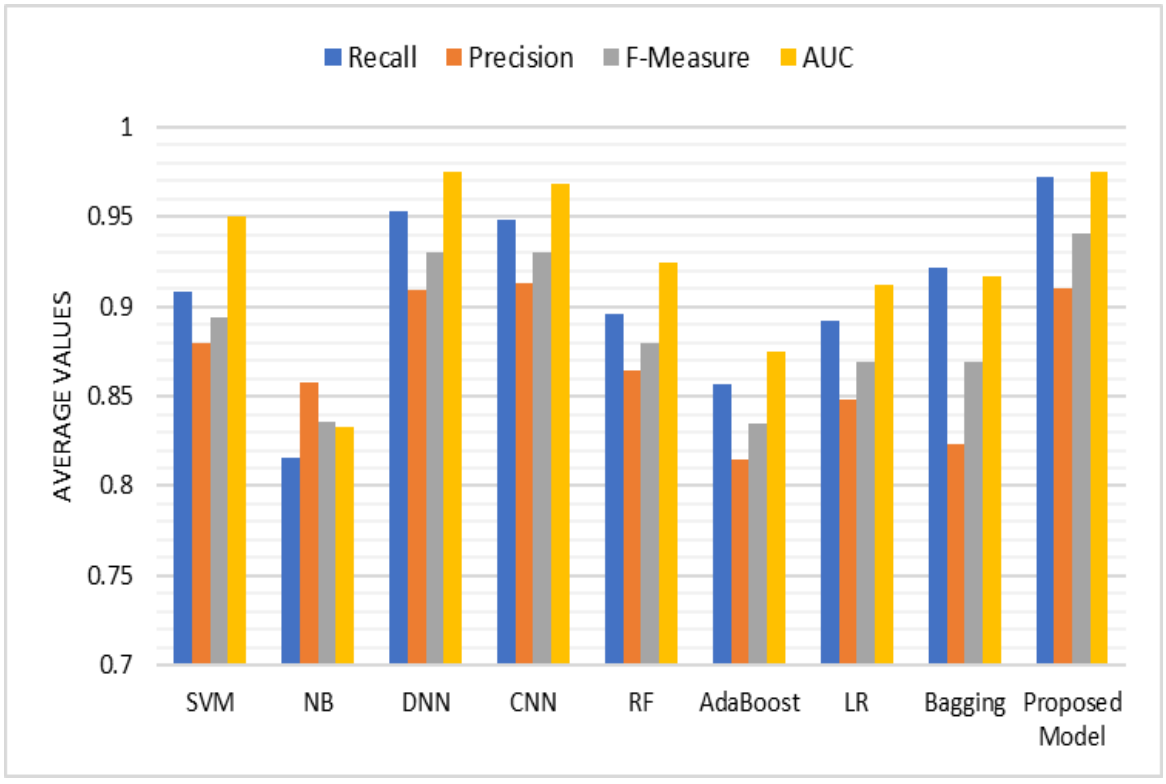
The performance comparison of all the considered models and proposed model is graphically visualized by drawing boxplot (Williamson *et al.* (1989)). It depicts the empirical distribution of the data. Figure 5.5 (i, ii, iii, iv), Figure 5.6 (i, ii, iii, iv), and Figure 5.7 (i, ii, iii, iv) represents the boxplot of recall, precision and f-measure for existing and proposed models respectively. In this plot, the name of the models is represented by the x-axis and y-axis labels the corresponding parameter under consideration, i.e. recall value. It is observed from the boxplots that the proposed model gives better and consistent results in comparison to other models for the considered performance metrics. The reason is that the proposed model explores near-optimal combinations of classifiers and their tuning parameters, while rest models estimate only on a pre-established combination. The stability of the models is also achieved due to the absence of outliers.



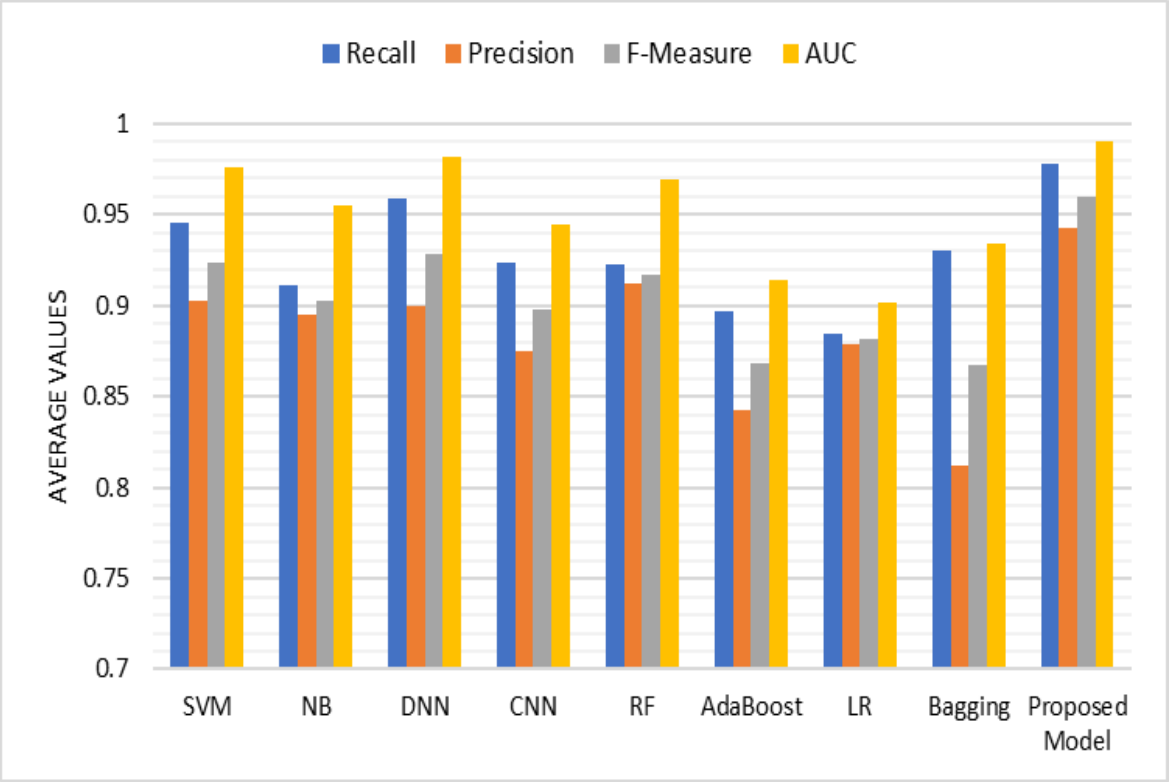
(i)


Recall
Precision
F-Measure
AUC
AVERAGE VALUES
0.9
0.85
0.8
0.75
0.7
0.65
0.6
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

(ii)


Recall
Precision
F-Measure
AUC
AVERAGE VALUES
1
0.95
0.9
0.85
0.8
0.75
0.7
SVM
NB
DNN
CNN
RF
AdaBoost
LR
Bagging
Proposed Model

(iii)



(iv)

Figure 5.8 (i, ii, iii, iv): Average Recall, Precision, F-Measure value of Models for Twitter#1, ASKfm, Formspring, and Twitter#2 datasets respectively

The boxplot reveals that even for some datasets the minimum recall, precision and f-measure value achieved by the proposed model is higher than the respective maximum values of other models. In the case of other models, DNN model performed relatively better than for Formspring and Twitter#2 Datasets. The evaluation results of recall, precision, and f-measure of all datasets are shown in Figure 5.8 (i, ii, iii, iv).

5.4 Discussion

It is well known that obtaining better predictive performance depends on different model characteristics. By integrating multiple individual learning algorithms, an ensemble learning method generally performs better than or comparable to single base-classifier. Therefore, it

becomes a popular and effective method in machine learning field. For the ensemble learning methods, there are two key issues to be solved. The first issue is how to combine "good and different" base-classifiers. The second is to assign reasonable tuning parameters to each base-classifier considered in the ensemble learning. To address the two issues, we propose a novel cuckoo inspired stacking ensemble framework to detect cyberbullying text in four different datasets from Twitter, ASKfm, and Formspring. For achieving optimal performance in different datasets, the CS algorithm dynamically adjusts the base-classifier combination and their tuning parameters. By comparing with the other eight state-of-art classification methods, the proposed framework performs the best or comparable in four different datasets in terms of recall, precision, f-measure and AUC.

The following chapter (Chapter 6) concludes the thesis work along with its future directions.

Chapter 6

Conclusion and Future Scope

Content-based cybercrime is a significant problem with several dire effects on victims. It takes place in online social networks such as Twitter, Formspring, ASKfm, Facebook etc. This thesis work provides comprehensive discussions on content-based cybercrime and investigated the possibilities of building a framework capable of efficiently identifying cyberbullying in online social networks. Section 6.1 concludes the thesis work and section 6.2 discusses the contribution of proposed techniques. Finally, section 6.3 provides the future scope of extension in the proposed solutions.

6.1 Conclusion

The focus of this thesis is to propose an efficient framework for content-based cybercrime detection in online social networks.

A novel hybrid CS-SVM model has been proposed for optimal selection of tuning parameters and features for cyberbullying detection from Formspring, Twitter and ASKfm datasets. To increase the accuracy of SVM classifier, the CS approach has been used as an optimizer, that select an appropriate feature subset and kernel parameters. In the second approach, a multiconfiguration detection framework has been proposed for the selection of right choice of preprocessing steps, feature selection techniques and classification models, that efficiently detect cyberbullying text in

four different online social media datasets i.e. from Twitter, ASKfm, and Formspring. An ensemble learning method generally performs better than or comparable to single base-classifier. In the third approach, a cuckoo inspired stacking ensemble framework has been proposed that integrates multiple individual learning algorithms and dynamically adjusts the base-classifier and meta-classifier combination and their tuning parameters. These approaches outperformed other recent techniques on all the datasets, giving high predictive recall value via 10-fold cross-validation.

6.2 Future Scope

The state of the art in cyberbullying classification involves training a machine learning model using supervised learning. The research conducted is mostly focusing on feature engineering, i.e. finding features that can separate bullying comments from non-bullying comments. Finding good features is difficult and problematic. Features that work well for Youtube comments may not work well for comments on ASKfm, due to different social media platforms being likely to have varying vocabulary and expressions in part caused by restrictions on communication (e.g. Twitter 140-character cap), different age groups and users' interests. The future work will focus on developing a responsive, scalable, and diverse model with better recall and precision to detect cybercrime related to cyberbullying like expressions of hate, threats, racism, and curses. For another future work, the parallel computing architecture may be exploited to reduce the computation time of the proposed approach. The work can be further extended by incorporating other cybercrime topics like expressions of racism, threats, hate and curses that would also aid in developing a responsive, scalable and diverse model. Another future work direction will be focusing on identifying cyber-predator specifically indulged in cyberbullying.

References

Aboujaoude, E., Savage, M.W., Starcevic, V. and Salame, W.O., 2015. Cyberbullying: Review of an old problem gone viral. Journal of adolescent health, 57(1), pp.10-18.

Agrawal, S. and Awekar, A., 2018, March. Deep learning for detecting cyberbullying across multiple social media platforms. In European Conference on Information Retrieval (pp. 141-153). Springer, Cham.

Al-garadi, M.A., Varathan, K.D. and Ravana, S.D., 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, pp.433-443.

Balakrishnan, V., Khan, S., Fernandez, T. and Arabnia, H.R., 2019. Cyberbullying detection on twitter using Big Five and Dark Triad features. Personality and individual differences, 141, pp.252-257.

Bayzick, J., Kontostathis, A. and Edwards, L., 2011. Detecting the presence of cyberbullying using computer software.

Ben-Hur, A. and Weston, J., 2010. A user's guide to support vector machines. In Data mining techniques for the life sciences (pp. 223-239). Humana Press.

Beran, T. and Li, Q., 2008. The relationship between cyberbullying and school bullying. The Journal of Student Wellbeing, 1(2), pp.16-33.

Bhattacharya, M. and Das, A., 2010. Genetic algorithm based feature selection in a recognition scheme using adaptive neuro fuzzy techniques. International Journal of Computers Communications & Control, 5(4), pp.458-468.

Bonanno, R.A. and Hymel, S., 2013. Cyber bullying and internalizing difficulties: Above and beyond the impact of traditional forms of bullying. Journal of youth and adolescence, 42(5), pp.685-697.

Bottino, S.M.B., Bottino, C., Regina, C.G., Correia, A.V.L. and Ribeiro, W.S., 2015. Cyberbullying and adolescent mental health: systematic review. Cadernos de saude publica, 31, pp.463-475.

Chang, C.C. and Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), pp.1-27.

Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. Machine learning, 46(1-3), pp.131-159.

Chawla, K., Bansal, V., Arya, K.V. and Shukla, A., 2010. Face Recognition using Principal Component Analysis and multi-class Support Vector Machine. In IPCV 2010: proceedings of the 2010 international conference on image processing, computer vision, & pattern recognition (Las Vegas NV, July 12-15, 2010) (pp. 331-336).

Chen, H., Mckeever, S. and Delany, S.J., 2017. Harnessing the power of text mining for the detection of abusive content in social media. In Advances in Computational Intelligence Systems (pp. 187-205). Springer, Cham.

Chen, J., Huang, H., Tian, S. and Qu, Y., 2009. Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36(3), pp.5432-5435.

Cheng, L., Li, J., Silva, Y.N., Hall, D.L. and Liu, H., 2019, January. Xbully: Cyberbullying detection within a multi-modal context. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (pp. 339-347).

Civicioglu, P. and Besdok, E., 2013. A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. Artificial intelligence review, 39(4), pp.315-346.

Cooke, M. and Buckley, N., 2008. Web 2.0, social networks and the future of market research. International Journal of Market Research, 50(2), pp.267-292.

Dadvar, M., Trieschnigg, D., Ordelman, R. and de Jong, F., 2013, March. Improving cyberbullying detection with user context. In European Conference on Information Retrieval (pp. 693-696). Springer, Berlin, Heidelberg.

Dadvar, M. and Eckert, K., 2018. Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study. arXiv preprint arXiv:1812.08046.

Dani, H., Li, J. and Liu, H., 2017, September. Sentiment informed cyberbullying detection in social media. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 52-67). Springer, Cham.

Dehue, F., Bolman, C. and Völlink, T., 2008. Cyberbullying: Youngsters' experiences and parental perception. CyberPsychology & Behavior, 11(2), pp.217-223.

Dewang, R.K. and Singh, A.K., 2018. State-of-art approaches for review spammer detection: a survey. Journal of Intelligent Information Systems, 50(2), pp.231-264.

Dinakar, K., Reichart, R. and Lieberman, H., 2011, July. Modeling the detection of textual cyberbullying. In fifth international AAAI conference on weblogs and social media.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R., 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3), pp.1-30.

Dooley, J.J., Pyżalski, J. and Cross, D., 2009. Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. Zeitschrift für Psychologie/Journal of Psychology, 217(4), pp.182-188.

Finklea, K.M. and Theohary, C.A., 2015, January. Cybercrime: Conceptual issues for congress and US law enforcement. Congressional Research Service, Library of Congress.

Fister, I., Yang, X.S. and Fister, D., 2014. Cuckoo search: a brief literature review. In Cuckoo search and firefly algorithm (pp. 49-62). Springer, Cham.

Galán-García, P., Puerta, J.G.D.L., Gómez, C.L., Santos, I. and Bringas, P.G., 2016. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. Logic Journal of the IGPL, 24(1), pp.42-53.

Gandomi, A.H., Yang, X.S. and Alavi, A.H., 2013. Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. Engineering with computers, 29(1), pp.17-35.

Görzig, A. and Frumkin, L.A., 2013. Cyberbullying experiences on-the-go: When social media can become distressing. Cyberpsychology, 7(1), p.4.

Görzig, A. and Ólafsson, K., 2013. What makes a bully a cyberbully? Unravelling the characteristics of cyberbullies across twenty-five European countries. Journal of Children and Media, 7(1), pp.9-27.

Griffin, R.S. and Gross, A.M., 2004. Childhood bullying: Current empirical findings and future directions for research. Aggression and violent behavior, 9(4), pp.379-400.

Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS technical report, 14(1), pp.5-16.

Harridge-March, S., Dunne, Á., Lawlor, M.A. and Rowley, J., 2010. Young people's use of online social networking sites–a uses and gratifications perspective. Journal of Research in interactive Marketing.

Hinduja, S. and Patchin, J.W., 2010. Bullying, cyberbullying, and suicide. Archives of suicide research, 14(3), pp.206-221.

Hinduja, S. and Patchin, J.W., 2014. Bullying beyond the schoolyard: Preventing and responding to cyberbullying. Corwin Press.

Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q. and Mishra, S., 2015. Detection of cyberbullying incidents on the instagram social network. arXiv preprint arXiv:1503.03909.

Hsu, C.W., Chang, C.C. and Lin, C.J., 2003. A practical guide to support vector classification.

Huang, J., Li, Y.F. and Xie, M., 2015. An empirical analysis of data preprocessing for machine learning-based software cost estimation. Information and software Technology, 67, pp.108-127.

Huang, Q., Singh, V.K. and Atrey, P.K., 2014, November. Cyber bullying detection using social and textual analysis. In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia (pp. 3-6).

Jahankhani, H., Al-Nemrat, A. and Hosseinian-Far, A., 2014. Cybercrime classification and characteristics. In Cyber Crime and Cyber Terrorism Investigator's Handbook (pp. 149-164). Syngress.

Joachims, T., 1998, April. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer, Berlin, Heidelberg.

Juvonen, J. and Gross, E.F., 2008. Extending the school grounds?—Bullying experiences in cyberspace. Journal of School health, 78(9), pp.496-505.

Kim, W., Jeong, O.R., Kim, C. and So, J., 2011. The dark side of the Internet: Attacks, costs and responses. Information systems, 36(3), pp.675-705.

Kowalski, R.M., Limber, S.P. and Agatston, P.W., 2012. Cyberbullying: Bullying in the digital age. John Wiley & Sons.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T. and Scherlis, W., 1998. Internet paradox: A social technology that reduces social involvement and psychological well-being?. American psychologist, 53(9), p.1017.

Li, Q., 2006. Cyberbullying in schools: A research of gender differences. School psychology international, 27(2), pp.157-170.

Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X. and Wang, S., 2011. An improved particle swarm optimization for feature selection. Journal of Bionic Engineering, 8(2), pp.191-200.

Loia, V., Parente, D., Pedrycz, W. and Tomasiello, S., 2018. A granular functional network with delay: some dynamical properties and application to the sign prediction in social networks. Neurocomputing, 321, pp.61-71.

Maldonado, S., López, J., Jimenez-Molina, A. and Lira, H., 2020. Simultaneous feature selection and heterogeneity control for SVM classification: An application to mental workload assessment. Expert Systems with Applications, 143, p.112988.

Mangaonkar, A., Hayrapetian, A. and Raje, R., 2015, May. Collaborative detection of cyberbullying behavior in Twitter data. In 2015 IEEE international conference on electro/information technology (EIT) (pp. 611-616). IEEE.

Nahar, V., Unankard, S., Li, X. and Pang, C., 2012, April. Sentiment analysis for effective detection of cyber bullying. In Asia-Pacific Web Conference (pp. 767-774). Springer, Berlin, Heidelberg.

Nahar, V., Li, X., Pang, C. and Zhang, Y., 2013, November. Cyberbullying detection based on text-stream classification. In The 11th Australasian Data Mining Conference (AusDM 2013).

Nahar, V., Al-Maskari, S., Li, X. and Pang, C., 2014a, July. Semi-supervised learning for cyberbullying detection in social networks. In Australasian Database Conference (pp. 160-171). Springer, Cham.

Nahar, V., Li, X., Zhang, H.L. and Pang, C., 2014b. Detecting cyberbullying in social networks using multi-agent system. Web Intelligence and Agent Systems: An International Journal, 12(4), pp.375-388.

Nixon, C.L., 2014. Current perspectives: the impact of cyberbullying on adolescent health. Adolescent health, medicine and therapeutics, 5, p.143.

Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R. and Menesini, E., 2010. Cyberbullying: Labels, behaviours and definition in three European countries. Journal of Psychologists and Counsellors in Schools, 20(2), pp.129-142.

Park, A., Phillips, M. and Johnson, M., 2004. Young people in Britain: the attitudes and experiences of 12 to 19 year olds. Nottingham: Department for Education and Skills.

Park, J.H., Sung, Y., Sharma, P.K., Jeong, Y.S. and Yi, G., 2017. Novel assessment method for accessing private data in social network security services. The journal of supercomputing, 73(7), pp.3307-3325.

Patchin, J.W. and Hinduja, S., 2006. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. Youth violence and juvenile justice, 4(2), pp.148-169.

Pham, T. and Adesman, A., 2015. Teen victimization: Prevalence and consequences of traditional and cyberbullying. Current opinion in pediatrics, 27(6), pp.748-756.

Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S. and Mattson, S.A., 2015, August. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 617-622). IEEE.

Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q. and Mishra, S., 2018, April. Scalable and timely detection of cyberbullying in online social networks. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (pp. 1738-1747).

Raisi, E. and Huang, B., 2017, July. Cyberbullying detection with weakly supervised machine learning. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 409-416).

Reynolds, K., Kontostathis, A. and Edwards, L., 2011, December. Using machine learning to detect cyberbullying. In 2011 10th International Conference on Machine learning and applications and workshops (Vol. 2, pp. 241-244). IEEE.

Sabella, R.A., Patchin, J.W. and Hinduja, S., 2013. Cyberbullying myths and realities. Computers in Human behavior, 29(6), pp.2703-2711.

Sampasa-Kanyinga, H. and Hamilton, H.A., 2015. Use of social networking sites and risk of cyberbullying victimization: A population-level study of adolescents. Cyberpsychology, Behavior, and Social Networking, 18(12), pp.704-710.

Schenk, A.M. and Fremouw, W.J., 2012. Prevalence, psychological impact, and coping of cyberbully victims among college students. Journal of school violence, 11(1), pp.21-37.

Sharma, H., Jadon, S.S., Bansal, J.C. and Arya, K.V., 2013, December. Lèvy flight based local search in differential evolution. In International Conference on Swarm, Evolutionary, and Memetic Computing (pp. 248-259). Springer, Cham.

Shih, Y.E., 2007. Setting the new standard with mobile computing in online learning. The International Review of Research in Open and Distributed Learning, 8(2).

Singh, V.K., Huang, Q. and Atrey, P.K., 2016a, August. Cyberbullying detection using probabilistic socio-textual information fusion. In 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 884-887). IEEE.

Singh, V.P., Srivastava, A., Kulshreshtha, D., Chaudhary, A. and Srivastava, R., 2016b. Mammogram classification using selected GLCM features and random forest classifier. International Journal of Computer Science and Information Security, 14(6), p.82.

Singh, V.K., Ghosh, S. and Jose, C., 2017, May. Toward multimodal cyberbullying detection. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2090-2099).

Slonje, R. and Smith, P.K., 2008. Cyberbullying: Another main type of bullying?. Scandinavian journal of psychology, 49(2), pp.147-154.

Smith, P.K., Catalano, R., Slee, P., Morita, Y., Junger-Tas, J. and Olweus, D. eds., 1999. The nature of school bullying: A cross-national perspective. Psychology Press.

Sourander, A., Klomek, A.B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen, M., Ristkari, T. and Helenius, H., 2010. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. Archives of general psychiatry, 67(7), pp.720-728.

Sticca, F. and Perren, S., 2013. Is cyberbullying worse than traditional bullying? Examining the differential roles of medium, publicity, and anonymity for the perceived severity of bullying. Journal of youth and adolescence, 42(5), pp.739-750.

Thangiah, M., Basri, S. and Sulaiman, S., 2012, June. A framework to detect cybercrime in the virtual environment. In 2012 International Conference on Computer & Information Science (ICCIS) (Vol. 1, pp. 553-557). IEEE.

Tokunaga, R.S., 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in human behavior, 26(3), pp.277-287.

Valentini, G. and Dietterich, T.G., 2004. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. Journal of Machine Learning Research, 5(Jul), pp.725-775.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V., 2018. Automatic detection of cyberbullying in social media text. PloS one, 13(10).

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V., 2015. Automatic detection and prevention of cyberbullying. In International Conference on Human and Social Analytics (HUSO 2015) (pp. 13-18). IARIA.

Vandebosch, H. and Van Cleemput, K., 2008. Defining cyberbullying: A qualitative research into the perceptions of youngsters. CyberPsychology & Behavior, 11(4), pp.499-503.

Willard, N.E., 2007. Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Research press.

Williamson, D.F., Parker, R.A. and Kendrick, J.S., 1989. The box plot: a simple visual method to interpret data. Annals of internal medicine, 110(11), pp.916-921.

Wold, S., Esbensen, K. and Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), pp.37-52.

Wolpert, D.H., 1992. Stacked generalization. Neural networks, 5(2), pp.241-259.

Xu, J.M., Jun, K.S., Zhu, X. and Bellmore, A., 2012, June. Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 656-666). Association for Computational Linguistics.

Yadav, S., Ekbal, A. and Saha, S., 2018. Feature selection for entity extraction from multiple biomedical corpora: A PSO-based approach. Soft Computing, 22(20), pp.6881-6904.

Yang, X.S., 2010. Nature-inspired metaheuristic algorithms. Luniver press.

Yang, X.S. and Deb, S., 2009, December. Cuckoo search via Lévy flights. In 2009 World congress on nature & biologically inspired computing (NaBIC) (pp. 210-214). IEEE.

Yang, X.S. and Deb, S., 2014. Cuckoo search: recent advances and applications. Neural Computing and Applications, 24(1), pp.169-174.

Yao, M., Chelmis, C. and Zois, D.S., 2019, May. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In The World Wide Web Conference (pp. 3427-3433).

Yar, M. and Steinmetz, K.F., 2019. Cybercrime and society. SAGE Publications Limited.

Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A. and Edwards, L., 2009. Detection of harassment on web 2.0. Proceedings of the Content Analysis in the WEB, 2, pp.1-7.

Yu, H. and Kim, S., 2012. SVM Tutorial-Classification, Regression and Ranking. Handbook of Natural computing, 1, pp.479-506.

Zhang, J., Otomo, T., Li, L. and Nakajima, S., 2019, October. Cyberbullying Detection on Twitter using Multiple Textual Features. In 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST) (pp. 1-6). IEEE.

Zhao, R. and Mao, K., 2016. Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Transactions on Affective Computing, 8(3), pp.328-339.

Zhao, R., Zhou, A. and Mao, K., 2016, January. Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking (pp. 1-6).

List of Publications

• Published

1) Amanpreet Singh, and Maninder Kaur. "Detection framework for content-based cybercrime in online social networks using metaheuristic approach." *Arabian Journal for Science and Engineering* (2020*)*: 2705-2719. doi: *https://doi.org/10.1007/s13369-019-04125-w*. (SCIE Indexed - IF: 1.711)
2) Amanpreet Singh, and Maninder Kaur. "Intelligent content-based cybercrime detection in online social networks using cuckoo search metaheuristic approach." *The Journal of Supercomputing* (2019): 1-23. doi: *https://doi.org/10.1007/s11227-019-03113-z*. (SCIE Indexed - IF: 2.469)
3) Amanpreet Singh, and Maninder Kaur, "Cuckoo Inspired Stacking Ensemble for Content based Cybercrime Detection in Social Media Platforms." *Transactions on Emerging Telecommunications Technologies* (2020): e4074. doi: https://doi.org/10.1002/ett.4074 (SCIE Indexed - IF: 1.594)
4) Amanpreet Singh, and Maninder Kaur, "Content-based Cybercrime Detection: A Concise Review." *International Journal of Innovative Technology and Exploring Engineering* SCOPUS Indexed, 2019.