# CYBERBULLYING DETECTION ON SOCIAL MEDIA USING DEEP LEARNING MODELS

A THESIS

SUBMITTED TO THE DELHI TECHNOLOGICAL UNIVERSITY

FOR THE AWARD OF THE DEGREE OF

## DOCTOR OF PHILOSOPHY

IN

## COMPUTER SCIENCE & ENGINEERING

SUBMITTED BY

**NITIN SACHDEVA**


**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**
**DELHI TECHNOLOGICAL UNIVERSITY**
**(Formerly Delhi College of Engineering)**
**DELHI- 110042 (INDIA)**

**2021**

# CYBERBULLYING DETECTION ON SOCIAL MEDIA USING DEEP LEARNING MODELS

BY

**NITIN SACHDEVA**

SUBMITTED TO THE DELHI TECHNOLOGICAL UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

IN

COMPUTER SCIENCE & ENGINEERING


**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**DELHI TECHNOLOGICAL UNIVERSITY**
**(Formerly Delhi College of Engineering)**
**DELHI- 110042 (INDIA)**

**2021**

# CANDIDATE DECLARATION

I, Nitin Sachdeva, PhD scholar (2K17/PhD/CO/03), hereby certify that the research work which is being presented in this thesis entitled **"Cyberbullying Detection on Social Media using Deep Learning Models"** in fulfilment of requirements of the award of degree of Doctor of Philosophy, is an authentic record of my own research work carried out under the supervision of Dr. Akshi Kumar in Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. The matter presented in this thesis has not been submitted elsewhere in part or fully to any other University or Institute for award of any degree.

Date: 10/08/2021
Place: Delhi

**Nitin Sachdeva**
2K17/PHD/CO/03
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India-110042.

# DELHI TECHNOLOGICAL UNIVERSITY

(Govt. of National Capital Territory of Delhi)

BAWANA ROAD, DELHI – 110042

# CERTIFICATE

Date: 10/08/2021

This is to certify that the thesis entitled "**Cyberbullying Detection on Social Media using Deep Learning Models"** done by Nitin Sachdeva, Roll no. 2K17/PHD/CO/03 is an authentic work carried out by him under my guidance towards the fulfilment of requirements for the degree of Doctor of Philosophy in Department of Computer Science & Engineering, Delhi Technological University, Delhi, India. This work is based on original research and the matter embodied in this thesis has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

## Supervisor

**Dr. Akshi Kumar**
Assistant Professor
Dept. of Computer Science & Engineering,
Delhi Technological University,
Delhi, India

# ACKNOWLEDGEMENT

I feel pride in placing on record my deep gratitude to my Ph.D. supervisor, Dr. Akshi Kumar, Assistant Professor, Department of Computer Science & Engineering, Delhi Technological University, Delhi, who guided me throughout this long journey of four years of research work. Whether it was reviewing literature or clearing doubts, she spared quality time to guide me despite her busy schedule. She not only helped in solving all problems with comfortable ease but also encouraged and motivated me to sail through in difficult times. I have gained immensely from her comments and suggestions regarding the technical quality of the work. It is hard to imagine successful completion of such a research work without her guidance and care.

I am very grateful to Prof. Rajni Jindal, Head, Department of Computer Science & Engineering, Delhi Technological University, Delhi, for her constant encouragement and support to accomplish this task.

I am thankful to all the other faculty members of the Department of Computer Science & Engineering, Delhi Technological University, Delhi, for the motivation and inspiration. I am also thankful to all non-teaching staff of the Department of Computer Science & Engineering, Delhi Technological University who have helped me directly or indirectly in completion of this research work.

I wish to pay high regards to my mother and my late father for their invaluable support, best wishes, motivation and encouragement. I am thankful to my parents in law for their support as well. I am grateful to my wife and child for their love, patience and support for completing this project. Without them this research would not have been possible.

Nitin Sachdeva
2K17/PHD/CO/03 (PhD Scholar)
Department of Computer Science & Engineering
Delhi Technological University, Delhi
E-mail: nits.usit@gmail.com

# ABSTRACT

*"Unless and until our society recognizes cyberbullying for what it is, the sufferings of thousands of silent victims will continue."*— **Anna Maria Chavez**

Application of deep learning models for cyberbullying detection in social media is an upcoming area for both researchers and practitioners for finding, exploring and analysing the extensibility of human-based expressions. Automated cyberbullying detection is typically a classification problem in natural language processing where the intent is to classify each abusive or offensive comment or post or message or image as either bullying or non-bullying. It needs high-level semantic analysis as well. Most of the earlier attempts on cyberbullying detection rely on manual feature extraction methods. Such methods are not only time-consuming and cumbersome, but often fail to correctly capture the meaning of the sentence. This fosters the need to build an intelligent analytic paradigm for detecting cyberbullying in social media data to lower down its hazard with minimal human intervention. Motivated by it, this research utilizes deep learning models for cyberbullying detection in social media as they trivialize the need of explicit feature extraction and are highly skilful, fast and more efficient in retrieval of essential features and patterns by themselves. In our research, we have applied deep learning for cyberbullying detection on textual and non-textual social media content. With high volume and variety of user-generated content on complex social media platforms, the challenges to detect cyberbullying in real-time have amplified. The influx of content makes it challenging to timely regulate online expression. Moreover, the anonymity and context-independence of expressions in online posts can be ambiguous or misleading. Nowadays, cyberbullying, through varied content modalities is also very common. At the same time, cultural diversities, unconventional use of typographical resources and easy availability of native-language keyboards augment to the variety and volume of user-generated content compounding the linguistic challenges in detecting online bullying posts. In an effort to deal with this antagonistic online delinquency referred to as cyberbullying, this research computationally analysed the content, modality and language-use in social media using deep learning models. This research has shown that the use of embeddings with deep learning architectures show better representation learning capabilities and simplify the feature selection process with enhanced classification accuracy as compared to baseline machine learning methods. The goal of the research is to automatically detect cyberbullying on textual, multimodal and mash-up social media content using deep learning models. In our research, we build models for these using deep architectures including capsule network, convolution neural network, multi-layer perceptron, self-attention mechanism, bi-directional gated recurrent unit, long short-term memory & bi-directional long short-term memory using embeddings such as GloVe, fastText and ELMo on social media like Askfm.in, Formspring.me, MySpace, Twitter, YouTube, Instagram and Facebook. The results show superlative performance as compared to SOTA as well.

# Table of Contents

*Candidate declaration* *i*

*Certificate* *ii*

*Acknowledgement* *iii*

*Abstract* *iv*

*Table of contents* *v*

*List of abbreviations* *viii*

*List of figures* *x*

*List of tables* *xii*

*List of algorithms* *xiii*

**Chapter 1 Introduction.....1**

1.1 Introduction.....1

1.2 Statement of Research Question and Research Objectives.....3

1.3 Social Media.....4

1.4 Cyberbullying.....7

1.5 Deep Learning.....12

1.6 Organization of thesis.....15

1.7 Chapter Summary.....17

**Chapter 2 Literature Survey.....18**

2.1 Review Process.....18

2.2 Literature Survey.....23

2.3 Key Observations and Research Gaps.....34

2.4 Chapter Summary.....35

**Chapter 3 Cyberbullying Detection using Machine Learning.....36**

3.1 Methodology.....36
3.2 Dataset.....38
3.3 Findings.....39
3.4 Chapter Summary.....41

**Chapter 4 Cyberbullying Detection for Textual Data.....42**

4.1 Methodology.....42
4.1.1 First Phase.....44
4.1.2 Second Phase.....46
4.2 Dataset.....50
4.3 Findings.....51
4.3.1 Model Performance.....51
4.3.2 Ablation Study.....52
4.4 Chapter Summary.....54

**Chapter 5 Cyberbullying Detection for Multimodal Data.....55**

5.1 Methodology.....56
5.1.1 Textual Processing.....58
5.1.2 Visual Processing.....59
5.1.3 Prediction.....62
5.2 Dataset.....62
5.3 Findings.....63
5.3.1 Model Performance.....63
5.3.2 Ablation Study.....63
5.4 Chapter Summary.....65

**Chapter 6 Cyberbullying Detection for Mash-up Data.....66**

6.1Methodology.....68

6.1.1 Data Pre-processing & Feature Extraction……..69
6.1.2 English Sub-network……..72
6.1.3 Hindi Sub-network……..75
6.1.4 Typographic Sub-network……..77
6.1.5 Prediction……..78
6.2 Dataset……79
6.3 Findings……80
6.3.1 Model Performance……..81
6.3.2 Ablation Study……..81
6.4 Chapter Summary……82

**Chapter 7 Conclusion and Future Trends……..83**

7.1 Conclusion & Future Trends……83

**References ……91**

**Appendix-A List of publications ……..102**

# LIST OF ABBREVIATIONS

| | |
|---|---|
| A | Accuracy |
| ANN | Artificial Neural Network |
| Adb | Adaboost |
| Bos | Boosting |
| Bgg | Bagging |
| CB | Cyberbullying |
| CNN | Convolutional Neural Network |
| Cmc | C-Means Clustering |
| Cf | Confidence |
| CK | Cohen's Kappa Measure |
| Cr | Correlation |
| DL | Deep Learning |
| DT | Decision Tree |
| EC | Evolutionary Computing |
| FL | Fuzzy Logic |
| F | F Score/ F1 Score/ F1 Measure/ F Measure |
| Fp | False Positives |
| Fn | False Negatives |
| Hac | Hierarchical Agglomerative Clustering |
| KW | Kruskal–Wallis Test |
| Kmc | K-Means Clustering |
| Knn | K Nearest Neighbor |
| LSTM | Long Short Term Memory |
| LogR | Logistic Regression |
| LR | Linear Regression |
| M | Mean |
| MW | Mann–Whitney U-Test |
| Maxe | Maximum Entropy |
| MR | Multiple Regression |
| MLP | Multi-Layer Perceptron |

| | |
|---|---|
| ML | Machine Learning |
| NN | Neural Networks |
| NB | Naïve Bayesian |
| P | Precision |
| RF | Random Forests |
| RNN | Recurrent Neural Network |
| RQ | Research Question |
| R | Recall |
| RMSE | Root Mean Squared Error |
| SLR | Systematic Literature Review |
| SC | Soft Computing |
| SVM | Support Vector Machine |
| Sn | Sensitivity |
| Sd | Standard Deviation |
| Sp | Specificity |
| TFIDF | Term Frequency Inverse Document Frequency |
| Tp | True Positives |
| Tn | True Negatives |

# LIST OF FIGURES

Fig. 1.1. Multimedia support by popular social networking sites……5
Fig.1.2. Statistics on share of social media platforms where cyberbullying occurs……6
Fig.1.3. Example of cyberbullying……7
Fig.1.4. Classification of social media posts……9
Fig.1.5. Types of cyberbullying……9
Fig.1.6. Generic cyberbullying detection process……11
Fig.1.7. Categorization of soft computing techniques……13
Fig.1.8. Categorization of machine learning techniques……14
Fig.1.9. Relation between SC, ML and DL……15
Fig.2.1. Phases of review process……19
Fig. 2.2. Review procedure……23
Fig. 3.1. System architecture……37
Fig. 3.2. Results obtained for Formspring.me dataset……39
Fig. 3.3. Results obtained for MySpace dataset……40
Fig. 3.4. Results obtained for Ask.fm dataset……40
Fig.4.1. System architecture of proposed Bi-GAC model……43
Fig.4.2. ELMo-specific representation for “smart”……44
Fig.4.3. GRU Cell architecture……45
Fig.4.4. Self-Attention……47
Fig.4.5. CapsNet Architecture……48
Fig.4.6. Operations within a Capsule……49
Fig.4.7. Comparative analysis of deep models for MySpace & Formspring.me……53
Fig. 5.1. Types of visual content……55
Fig. 5.2. The proposed CapsNet-ConvNet model……57
Fig. 5.3. Visual Processing module……60
Fig. 5.4. Working of a typical ConvNet……61
Fig. 5.5. Modality distribution in dataset……62
Fig. 5.6. Performance of CapsNet-ConvNet Model……64
Fig. 5.7. Results for ablation study……64

Fig. 5.8. Comparative analysis of different classifiers used for image modality.....65
Fig.6.1. Example of translation ambiguity.....67
Fig.6.2. Proposed MIIL-DNN model.....69
Fig. 6.3. Feature extraction in MIIL-DNN.....71
Fig. 6.4. Operations within a capsule.....75
Fig. 6.5. MLP architecture.....77
Fig. 6.6(a). Early multi-lingual fusion.....78
Fig. 6.6(b). Model-level multi-lingual fusion.....78
Fig. 6.6(c). Late multi-lingual fusion.....79
Fig. 6.7. Performance of MIIL-DNN on DS-I and DS-II.....81
Fig.6.8. Comparative analysis of ablation architectures for Hindi using accuracy.....82
Fig. 7.1. Proposed vs. existing models for textual CB detection.....85
Fig. 7.2. Proposed vs. existing models for multimodal CB detection.....86
Fig. 7.3. Performance results of toxic comment categories.....88
Fig. 7.4. Proposed vs. existing models for code-mix CB detection.....89

# LIST OF TABLES

Table 1.1: Types of cyberbullying……10
Table 2.1. Quality Assessment……22
Table 2.2. Literature survey of the studies……24
Table 4.1. Sample distribution in datasets……50
Table 4.2. F-1 score for Bi-GAC model……51
Table 4.3. Comparison of Bi-GAC with existing works on MySpace dataset……51
Table 4.4. Comparison of Bi-GAC with existing works on formspring.me dataset……51
Table 4.5. Ablation architectures……53
Table 5.1. Data categorization ……63
Table 6.1. Tags and their counts for both the datasets……80
Table 6.2. Average post length in different class text for both the datasets……80
Table 6.3. Average word length in different class text for both the datasets……80

# LIST OF ALGORITHMS

Algorithm 4.1. Dynamic routing algorithm……49

# Chapter 1
# Introduction

This chapter briefs about cyberbullying, its harmful effects and impact on society. It also discusses its association with the proliferation of the Web and growing usage of social media. It talks about the challenges of the chosen research area and the need of deep learning for cyberbullying detection in social media. This chapter also briefly introduces the fundamental concepts related to the research area. It provides a brief description of the key terminologies namely, social media, cyberbullying, and deep learning, followed by a summary of the chapter.

## 1.1 Introduction

Social media has reshaped communication by facilitating healthy discussions and candid conversations in which people engage on the community-centric platform by sharing ideas, thoughts and information. As one of the most popular and modern means of communication, social networking sites provide a constructive platform for market research, decision-making process and government intelligence [1, 2]. Undoubtedly, its mass adoption, effortless availability and popularity can get users united in a very short time and allow gathering opinions from different people on an issue in just a click. But this virtual social world can also fuel and witness different anti-social activities such as scams, fake news, rumours and cyberbullying.

Cyberbullying (CB) is a form of manipulation, belittlement, and targeted abuse using mean-spirited messages and negative electronic postings [3]. It is the use of information technology networks by individuals to humiliate, mock, embarrass, insult, defame and criticize a target without any one to one contact. Cyberbullying can be as straightforward as sending mean, hurtful, rude texts or instant messages as devious as spreading secrets or rumours about people online. Though bullying in electronic form can have multiple-dimensions, such as exclusion, harassment, outing, trickery, cyber stalking, dissing, fraping, masquerading, trolling and flaming [4, 5], the obvious intention to hurt and harm is common. This inveterate nuisance creates mental, emotional and physical risks for the bullied. The targets (victims of cyberbullying) feel overwhelmed, powerless, vulnerable, unsafe, worthless, humiliated, isolated, depressed, embarrassed, vengeful and at times suicidal.

Technology (Web 2.0) allows the bullies to be anonymous, hard to trace and insulated from confrontation. To the targets of cyberbullying, it feels invasive and never-ending. An accurate detection can facilitate timely intervention by alarming the moderators to take countermeasures. But content moderation practices on these platforms by human moderators is often inconsistent and done in a non-transparent

manner. It also suffers from biasing and may apprehend freedom of expression online. Moreover, spotting bullying instances explicitly as well as spotting victims is tricky too. Blocking and reporting might augment the reality if the bully is within the same professional or personal community, for example, a classmate. Sadly, the scale and impact of cyberbullying can be seen across social media platforms even though its awareness is at an all-time high. Simultaneously, with huge volume and variety of user-generated content on complex social media platforms, the challenges to detect cyberbullying in real-time have amplified. The influx of content makes it challenging to timely regulate online expression. Moreover, the anonymity and context-independence of expressions in online posts can be ambiguous or misleading. Recently, as memes, online videos and other image-based, inter-textual content have become customary in social feeds; typo-graphic and info-graphic visual content has also become a considerable element of social data. Thus, cyber bullying, through varied content modalities is very common. At the same time, cultural diversities, country-specific trending topics hash-tags in social media, the unconventional use of typographical resources such as capitals, punctuation and emojis and easy availability of native language keyboards add to the variety and volume of user-generated content compounding the linguistic challenges in detecting online bullying posts.

Researchers worldwide have been trying to develop new ways to detect cyber bullying, manage it and reduce its prevalence in social media. Advanced analytical methods and computational models for efficient processing, analysis and modelling for detecting such bitter, taunting, abusive or negative content in images, memes or text messages are imperative. The automated cyberbullying detection has attracted growing interest over the past decade as it facilitates combating toxic online behaviour. A lot of research has been done on detecting cyberbullying in textual data using a myriad of features [5]. Many datasets have been made open-source to facilitate research enthusiasts.

As a classical problem in natural language processing (NLP), cyberbullying detection in real-time user generated content needs high-level semantic analysis. Most of the earlier attempts on cyberbullying detection rely on manual feature extraction methods [6]. Such methods are not only time-consuming and cumbersome, but often fail to correctly capture the meaning of the sentence. Few lexicon-based methods by maintaining a list of offensives, abusive and hateful words have also been used, but are quite limited in scope [7]. Recent research focuses on the application of deep learning models for various NLP tasks and has reported state-of-the-art results [8]. Basically, deep architectures are neural networks with multiple processing layers of neuron with each layer having a specific task [9]. Utilizing deep learning models trivializes the need of explicit feature extraction techniques as these models are highly skilful and fast in retrieval of essential features and patterns by themselves. With minimal human intervention these models report superior results than the conventional machine learning models. Various deep learning architectures have contributed significantly in

computational analytics of text [9]. Pre-trained word embedding's like Word2Vec, GloVe, ELMo, fastText that represent text in vector forms and deep neural networks such as CNN, RNN, GRU, LSTM, CapsNet & hierarchical networks that automate the task of feature extraction demonstrate best practices for solving text classification problems [10, 11]. Deep architectures with better representation learning capabilities have been substantiating its relevance in this field with improved results. Assessing the user-generated content in social media could be rewarding for automatic cyberbullying detection using deep neural architectures.

Thus, as an effort to deal with the antagonistic online delinquency referred to as cyberbullying, this research computationally analysed the content, modality and language-use in social media using deep learning models. The research demonstrates the feasibility, scope and relevance of using deep learning models for cyberbullying detection in social media portals.

## 1.2 Statement of Research Question and Research Objectives

The conventional methods used to analyse data are inadequate as unlike the traditional data, social media data is mainly unstructured and comprises multilingual text and in varied modalities such as audio, video, images, GIFs, Emojis' etc. Moreover, the linguistic complexities of user-generated content in social media makes it even more intricate to tap and analyse information using contemporary tools. Novel approaches to information discovery and decision making which use multiple intelligent technologies such as machine learning, deep learning, artificial intelligence, natural language processing and image recognition among others are required to understand data & then generate insights.

**Statement of Research Question:**

***"Can the linguistic complexities of user-generated content in social media be computationally analysed using deep learning models for automated cyberbullying detection?"***

Based on the statement of research question, the following research objectives (RO's) are identified:

**Research Objective I**– *To detect cyberbullying in textual social media content using deep learning model.*

**Research Objective II**– *To apply deep learning model for cyberbullying detection in content modalities other than text i.e., multimodal social media content*.

**Research Objective III**– *To computationally identify cyberbullying in mash-up social media content using deep learning model*.

## 1.3 Social Media

Information is power, but without a means to distribute information, people cannot harness this power. Social media come up as a key player that gives a platform for expression and content distribution in today's world [12]. The basic purpose of social media sites is to build interest, professional and interconnection-based virtual groups empowering better connections with other people all over the world. With the rapid growth of these sites (Instagram, Twitter, Snapchat, Tumblr, YouTube, Google+, Facebook, etc.), the netizen can share all type of social media data viz. text, audio, image, video utilizing the power of Internet without having ample information regarding the network topology and client-server architecture of Web. The social networking sites have given 'everyone a voice' but at the same time, we're drowning in abundance, complexity of choices and unfortunately, the misappropriation or misdirection of influence. Moreover, when lots of individuals come together and that too from different countries, communities, races, ethnicities, gender, and varied age-groups, there are bound to be conflicts, controversies, and intimidation vulnerabilities. But this virtual social world can also fuel and witness different anti-social activities such as scams, fake news, rumours and cyberbullying [13]. That is, although social networking sites proffer numerous benefits as these facilitate participation and collaboration but on the flip side hate speech, social distrust, cyberbullying, identity theft, cyber-stalking and cascading of rumours and fake stories are some antithetical concerns associated with it. The pervasive reach of these sites has irrefutably triggered, contributed and exacerbated bullying.

Social media may seem positive and safe, but it affects our daily lives more than we can think of. According to a study by Harvard University, "self-disclosure on social networking sites lights up the same part of the brain that also ignites when taking an addictive substance. The reward area in the brain and its chemical messenger pathways affect decisions and sensations" [14]. The overuse of social media can disrupt psychologically leading to social withdrawal, depression, anxiety and insomnia. Further, social media hacks and oversharing makes one's identity extremely vulnerable.

Social media is inherently an informal way of communication with all kinds of multimedia content. The following figure 1.1 [15] depicts the multimedia types supported by popular social networking sites.

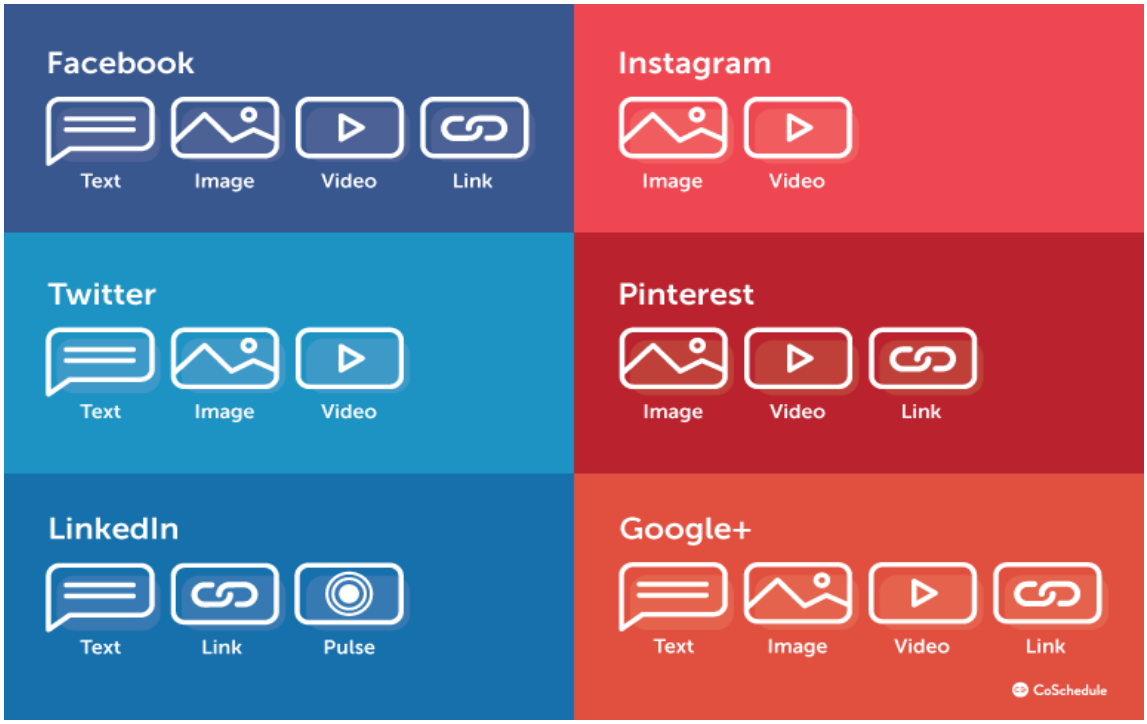


**Fig. 1.1.** Multimedia support by popular social networking sites

Social media dynamics keep changing with respect to increasing user base and user-activity which makes it a high dimensional, complex and ambiguous data space for analytical processing. Pertinent studies indicate that social media is one of the most favoured mediums by bullies and various factors such as socio-demography, physiological distress and time frames are related to cyberbullying. The massive volumes of human-centric, real-time, multimodal, heterogeneous and unstructured social media data makes manual detection intractable. Moreover, the social web applications/services are not restricted to the text-based data but extend to the partially unknown complex structures of image, audio and video. This fosters the need to develop intelligent tools and techniques for identifying, detecting and assessing cyberbullying from the available social media data to lower down its hazardous impact. Design and development of contemporary tools which tap and analyse online detrimental behaviour automatically from the high-dimensional social media are imperative. The substantial growth in the dimensionality, heterogeneity, subjectivity and multimodality of social media and the pressing need to timely curtail the damage instigated through cyberbullying, has fostered the need to devise automated mechanisms which detect such unfavourable activities. Social media has made cyberbullying a lot easier than it used to be due to it being much reckless in reach and virality that too with anonymity and without any restrictions. Social media cyberbullying is most prevalent in Instagram (42%), followed by Facebook (37%) and Snapchat (31%)[1]. Cyberbullying can be as straightforward as sending mean, hurtful,

[1]https://www.ditchthelabel.org/wp-content/uploads/2017/07/The-Annual-Bullying-Survey-2017-1.pdf

rude texts or instant messages as devious as spreading secrets or rumours about people online. Though bullying in electronic form can have multiple-dimensions, such as exclusion, harassment, outing, trickery, cyberstalking, dissing, fraping, masquerading, trolling and flaming [16, 17], the obvious intention to hurt and harm is common. This inveterate nuisance creates mental, emotional and physical risks for the bullied. The targets (victims of cyberbullying) feel overwhelmed, powerless, vulnerable, unsafe, worthless, humiliated, isolated, depressed, embarrassed, vengeful and at times suicidal. An accurate detection can facilitate timely intervention by alarming the moderators to take countermeasures. But content moderation practices on these platforms by human moderators is often inconsistent and done in a non-transparent manner. It also suffers from biasing and may apprehend freedom of expression online. Moreover, spotting bullying instances explicitly as well as spotting victims is tricky too. Blocking and reporting might augment the reality if the bully is within the same professional or personal community, for example, a classmate. Sadly, the scale and impact of cyberbullying can be seen across social media platforms even though its awareness is at an all-time high. Figure 1.2 shows how social media platforms are hotbeds of cyberbullying activities for young people.

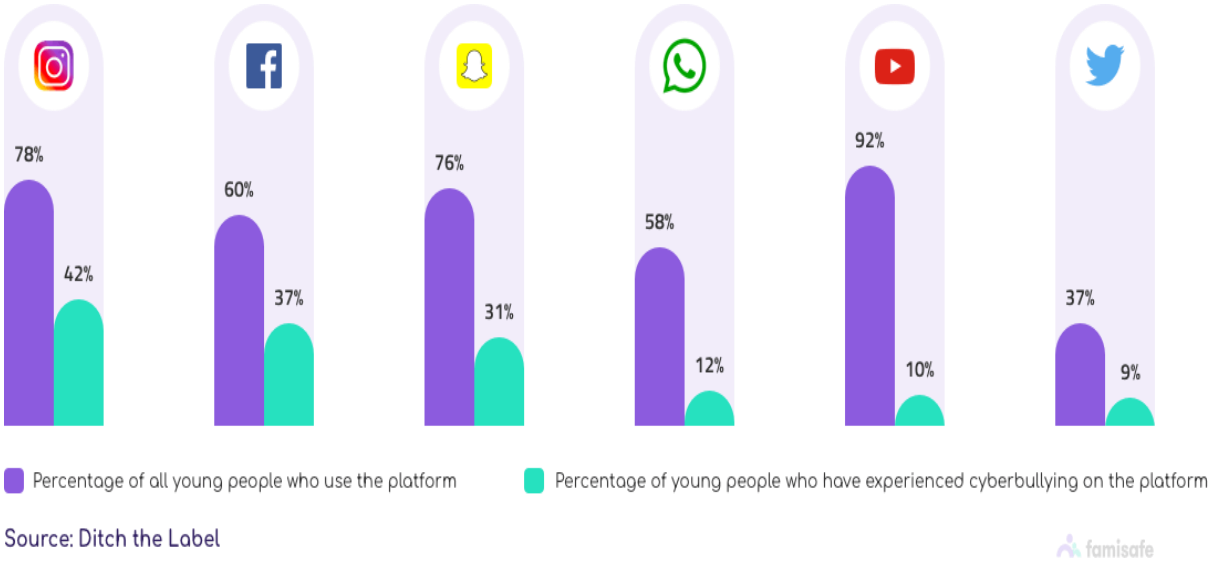


**Fig.1.2.** Statistics on share of social media platforms where cyberbullying occurs[2]

Typically, online bullying involves sending or posting harmful content or negative comments about a person. It intends to embarrass or humiliate a person in order to ruin his/her dignity, confidence and self-esteem [18]. The results of cyberbullying are dangerous and may affect the victim socially, mentally or psychologically. Hence, it is important to promptly detect cyberbullying in order to prevent it from becoming a global epidemic.

[2] https://www.ditchthelabel.org/

## 1.4 Cyberbullying

Cyberbullying is defined as bullying an individual or a group of individuals using the Internet, mobiles or any other electronic device by sending inappropriate textual or non-textual multimedia messages in order to hurt or cause embarrassment [19]. The one who bullies is called a 'bully' and the other is said to be 'victim'. The term 'Cyberbullying' was coined by Canadian educator and anti-bullying activist Bill Belsey in the year 2003 [20]. It is the repeated exposure of the negative actions on the part of one or more individuals in order to inflict humiliation, harassment, discomfort or injury upon another through the use of electronic medium [19] like emails, chat rooms, instant messaging, cell phones or by posting videos, audios, images etc. Bullying has been a part of human civilization history which involves hurting someone either by humiliating or harassing in any form, involving mental, verbal or physical damage. When this assault takes place in cyberspace, it is referred to as cyberbullying/ cyber-harassment/ cyber-victimization [21].

According to a study, nearly 43% of the teenagers in the United States are victims of cyberbullying [22]. It is more persistent way of bullying an individual in front of the entire online community especially within the social setting which can eventually lead to psychological, mental and emotional breakdown for the victim inculcating the sense of low self-esteem, low self-confidence, anger, depression, stress, loneliness, sadness, health degradation etc. [23]. Many of such intense cases have tragically ended in self-injury or suicides, underlining the grave nature of this critical issue [24]. The following figure 1.3 presents an example of cyberbullying from Twitter.

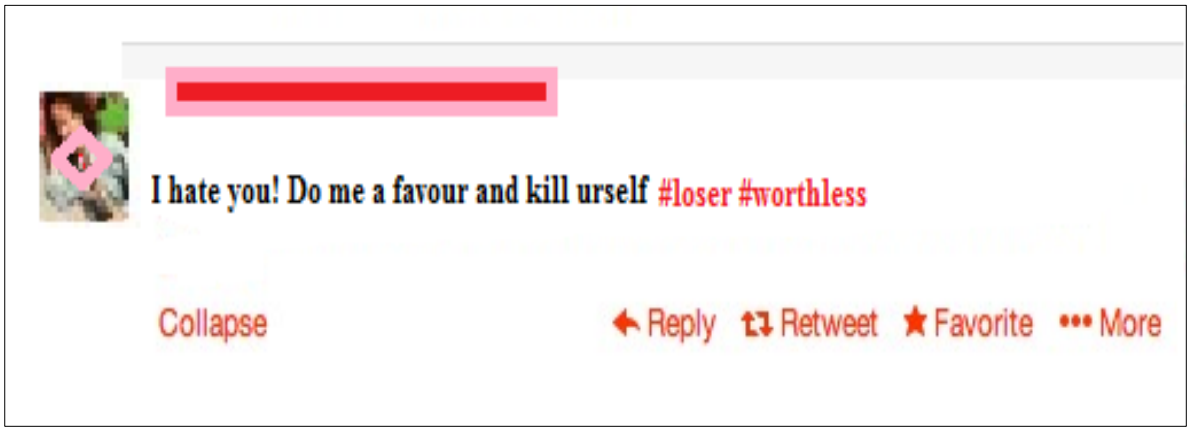


**Fig.1.3.** Example of cyberbullying

With technological advancement, the social freedoms that the networking sites give and larger audience, cyberbullying has spread manifolds affecting the individual not only limited to their workplace but also children and young adults in their daily lives. Anonymity further allows bullies to be more aggressive and offensive due to the reduced chance of being detected and punished, making it critical to efficiently detect cyberbullying behaviour in a real- time setting. This poses significant threat to the

physical and mental health of the victims making it a public health concern. Various studies have reported that victims of cyberbullying have lower self-esteem, higher levels of depression, suffer from behavioural issues and are addicted to substance abuse. Bullying victimization may trigger a sequence of events that results in suicidal behaviour. The first reported case of cyberbullying was of an American middle school student, Ryan Halligan of Vermont in 2003 [25]. Ryan was constantly bullied in person and online by his classmates and this bullying was attributed as his reason to commit suicide. As per the National Bullying Prevention Centre, 'Every child on Facebook likely has a bullying story, whether as the victim, bully or as a witness' [26].

Netiquette refers to good manners on the Internet and treating other people on the Internet as you would like to be treated yourself. Unfortunately, some people use the Internet and/or mobile phones to offend or harass others. This is referred to as cyberbullying. Automated cyberbullying detection is a proactive strategic technology-based mechanism. It is a typical inherent classification problem of natural language processing where the intent is to classify the social media messages as either bullying or non-bullying. Cyberbullying is a multi-step process (for predictive analysis) comprising various sub-tasks such as data collection and it's pre-processing; extracting and selecting relevant features and thereby classifying messages. The increasing use of social media at such a fast pace is adding both variety and volume to user-generated content, owing to which the manual classification (as either CB or non-CB) has become quite intractable. Simultaneously, it is generating an enormous number of features also. Choosing the most appropriate feature is a challenging task [27] and it influences the overall classification accuracy as well. This necessitates assessing & examining novel computational approaches that show better representation learning capabilities and simplify the feature selection process with enhanced classification accuracy and ensure result comprehensibility as well.

It is often characterized as a predictive learning model in the social setting which detects the presence of cyberbullying in an online post (textual/non-textual) so that it does not inflict seriously or damage the victim's emotional, psychological and social state. The posts classified as bullying can further be divided into two categories, namely, direct cyberbullying (DCB) and indirect cyberbullying (ICB) [28], as shown in figure 1.4. DCB involves direct sending of harmful content to a person either via email or SMS etc. ICB comprises of posting harmful contents about any person on social media or sharing it with others, for example posting an improper photograph of someone on Facebook is an example of ICB.

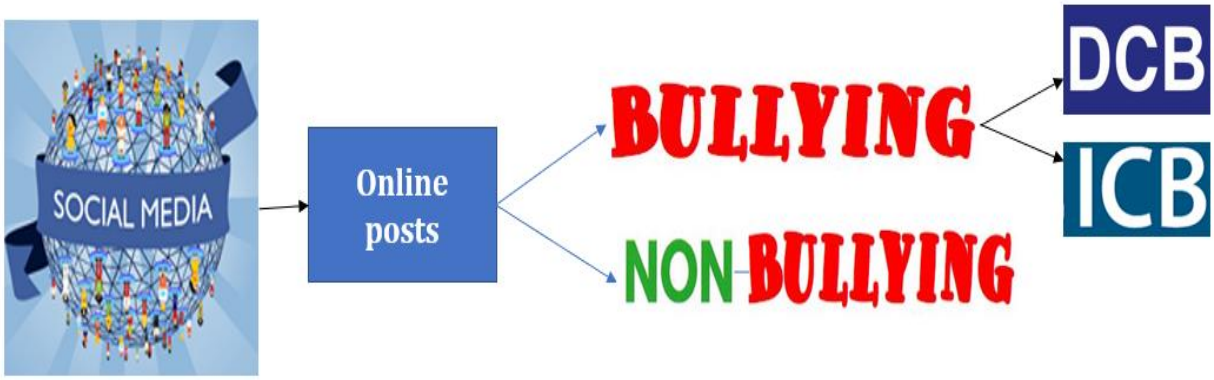


**Fig.1.4.** Classification of social media posts

The direct and indirect cyberbullying is categorized into the different types as depicted in figure 1.5.

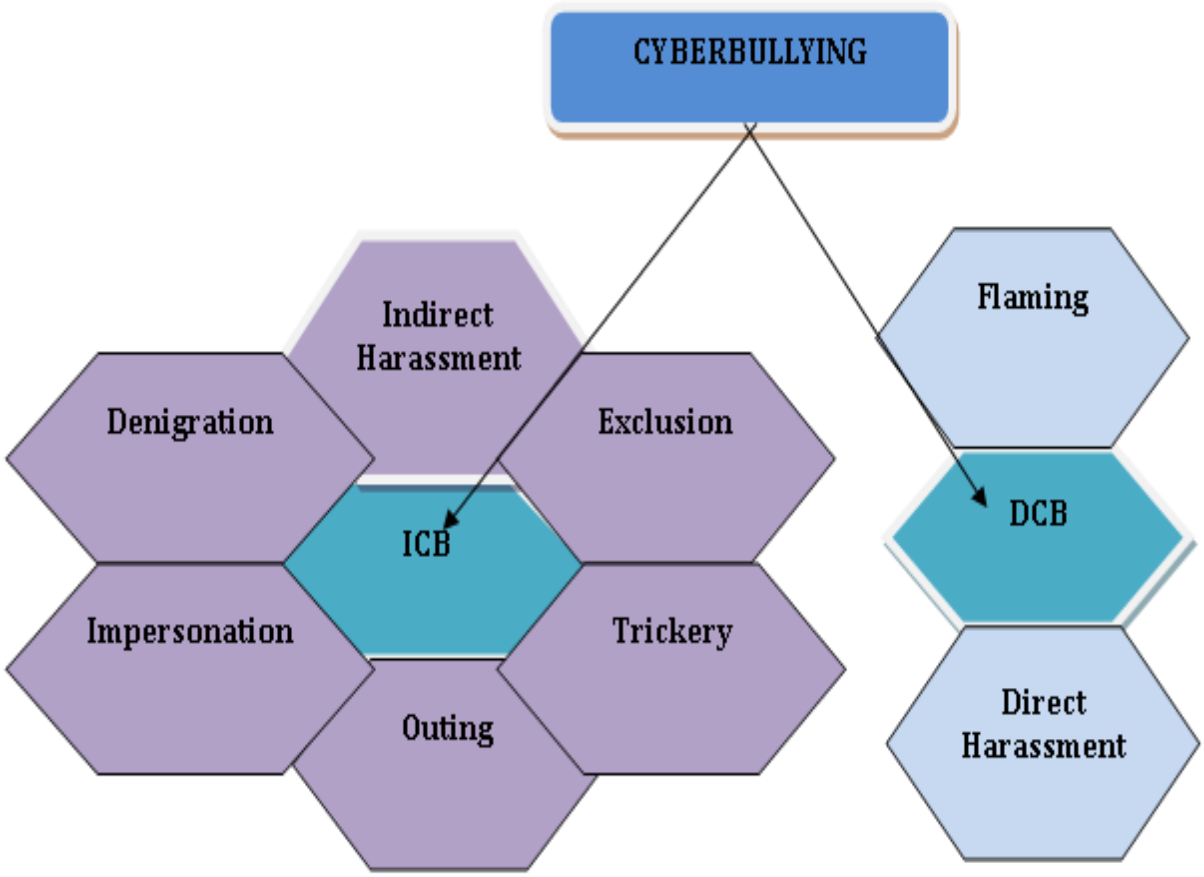


**Fig.1.5.** Types of cyberbullying

ICB messages are further categorized into six types [28]. These include Indirect Harassment, Denigration, Impersonation, Outing, Trickery and Exclusion, whereas DCB includes Flaming and Direct Harassment.

The following table 1.1 briefly explains these types of cyberbullying.

**Table 1.1.** Types of cyberbullying

| Types of cyberbullying | Details |
|---|---|
| **Flaming** | It is an online fight between individuals. They usually exchange vulgar electronic messages. People fighting on online forums exchanging obscene messages is an example of Flaming. |
| **Direct Harassment** | It is directly harassing a person either by insulting or threatening him or her via messages. It includes only two parties- the one who bullies and the other who is bullied. Threatening or harassing a person either by sending email or SMS directly is an example of direct harassment. |
| **Indirect Harassment** | It is indirectly harassing a person either by insulting or threatening him or her via messages posted online. It includes many parties. Posting embarrassing photos on social media in order to harass the other person indirectly is an example of indirect cyberbullying. |
| **Denigration** | It is spreading hearsay or rumours about others in order to ruin their reputation. It puts the status of the cyber-victim on stake. Posting skewed contents on forums or blogs etc. in order to turn down cyber-victim's reputation is an example of denigration. |
| **Impersonation** | It is acting or pretending as another person and then doing anti-social activities in order to embarrass or damage his or her reputation. Imitating cyber-victim either by creating a fake profile or through hacking and sending messages that may instigate other users to attack the victim is an example of impersonation. |
| **Outing** | It is sharing private information of a cyber-victim without his or her consent in order to hurt the victim. Posting a humiliating picture of someone in order to hurt the cyber victim is an example of outing. |
| **Trickery** | It is obtaining sensitive information about a user by faking the trust of cyber victims and then eventually violating that trust. Obtaining a personal video by faking as a close friend and then posting it online is an example of trickery. |
| **Exclusion** | It involves the exclusion of the cyber victim from online communities or groups etc. Excluding a person knowingly from a WhatsApp group is an example of exclusion. |

The elusive nature of cyberbullying undermines the self-esteem of the cyber victim, affecting him or her mentally, socially and psychologically. Automated detection model consists of multiple tasks which identify and classify posts as bullying or not. Considered as a generic classification problem, a typical cyberbullying detection process extracts the features from the pre-processed data and classifies the posts accordingly as shown in the figure 1.6 below.

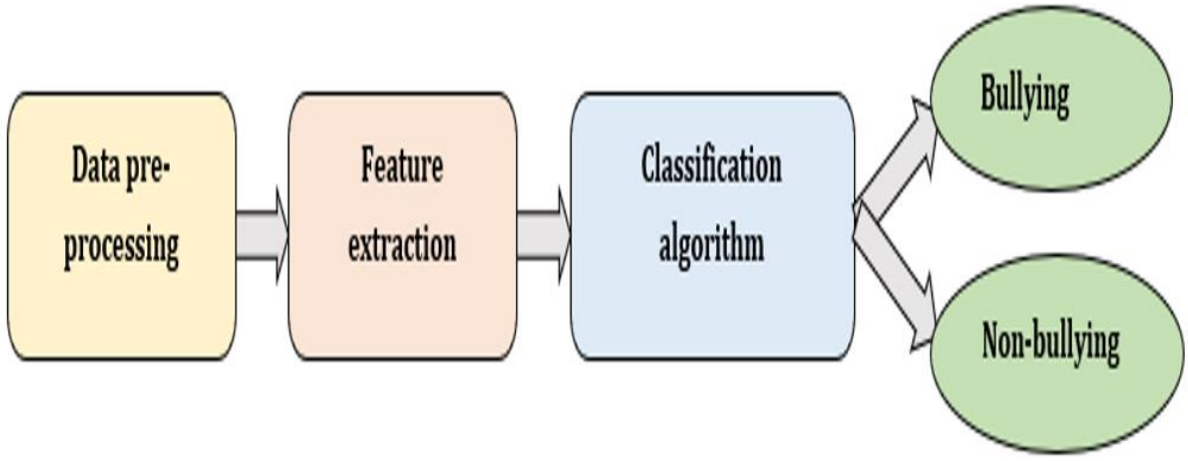


**Fig.1.6.** Generic cyberbullying detection process

The pre-processing phase includes cleaning the acquired data by removing unwanted URL's or strings etc., handling missing values, correcting words etc. and then transforming it into a representation suitable for feature extraction. After pre-processing, features such as keywords depicting bad/nasty/rude/abusive/hateful/attacking words, N-grams, pronouns, skip-grams are extracted. Next phase uses supervised learning techniques to classify the messages as either containing bullying content or not.

Cyberbullying is primarily associated with the utilization of digital media in order to bully someone. It has grown to a level where it is seriously affecting and damaging individual's lives where social media forums play a key role by providing a fecund means for bullies and the one's using such portals are more vulnerable to attacks. Nowadays, the Internet has drastically reformed the way people express their views-opinions-thoughts on social media. People rely more on the use of social forums like Twitter, Facebook, Formspring.me, MySpace, Ask.fm etc. for sharing their views & opinions which results in producing unprecedented volume of user generated online data that is available generally in the form of tweets, blog posts, reviews, question-answering forums etc. The heavy dependence of the mass on such multimedia content for appropriate opinions shows the increasing relevance of the utilization of Web 2.0 technologies and tools in our daily lives. Hence, we can say that social media has global reach and has become widespread. Its pervasive reach has in return given some unpremeditated consequences as well where people are discovering illegal & unethical ways of using such communities. One of its most severe upshots is known as cyberbullying where individuals are searching new means to bully one another over the Internet. It has grown as a social menace that puts a negative effect on the minds of both the bully and victim. It is more persistent way

of bullying a person before an entire online community especially when we talk in terms of social networking websites which can ultimately results in psychological and emotional breakdown for the cyberbullying victim developing the feeling of low self-confidence, depression, stress, anger, sadness, health degradation, loneliness, suicides etc. All this has gradually increased the linguistic challenges associated with the 'user-generated real time social media content' which further encourages the need to search for enhanced classification methods and paradigms which can cater well with cyberbullying detection in social media.

## 1.5 Deep Learning

The volume and variety of user-generated content on complex social media platforms have amplified the challenges to detect cyberbullying in real-time. The influx of content makes it challenging to timely regulate online expression. Moreover, the anonymity and context-independence of expressions in online posts can be ambiguous or misleading. Recently, as memes, online videos and other image-based, inter-textual content have become normal in social feeds; typo-graphic and info-graphic visual content has also become a substantial element of user-generated data. Thus, cyber bullying, through varied content modalities is very common. At the same time, cultural diversities, country-specific trending topics hash-tags in social media, the unconventional use of typographical resources such as capitals, punctuation and emojis and easy availability of native language keyboards add to the variety and volume of user-generated content compounding the linguistic challenges in detecting online bullying posts. Researchers worldwide have been trying to develop new ways to detect cyber bullying, manage it and reduce its prevalence in social media. Advanced analytical methods and computational models for efficient processing, analysis and modelling for detecting such bitter, taunting, abusive or negative content in images, memes or text messages are imperative. The automated cyberbullying detection has attracted growing interest over the past decade as it facilitates combating toxic online behaviour. A lot of research has been done on detecting cyberbullying in textual data using a myriad of features [29]. Many datasets have been made open-source to facilitate research enthusiasts.

As a classical problem in natural language processing (NLP), cyberbullying detection in real-time user generated content needs high-level semantic analysis. Most of the earlier attempts on cyberbullying detection rely on manual feature extraction methods [30]. Such methods are not only time-consuming and cumbersome, but often fail to correctly capture the meaning of the sentence. Few lexicon-based methods by maintaining a list of offensives, abusive and hateful words have also been used, but are quite limited in scope [31]. Recent research focuses on the application of deep learning models for various NLP tasks and has reported state-of-the-art results [32]. Basically, deep architectures are neural networks with multiple processing layers of neurons with each layer having a specific task [33]. Utilizing deep learning models trivializes the need

of explicit feature extraction techniques as these models are highly skilful and fast in retrieval of essential features and patterns by themselves. With minimal human intervention these models report superior results than the conventional machine learning (ML) models.

Deep learning (DL) is considered as a part of the broader family of ML based on learning data representations, in contrast to the task-specific algorithms and where learning can be supervised, semi-supervised or unsupervised. DL entails techniques such as deep neural network (DNN), recurrent NN, CNN, deep-belief networks etc., whereas NN is one of the sub-types of SC techniques [34] which includes feed forward; MLP; deep NN (DNN); radial-basis etc. (as shown in figure 1.7 and 1.8).

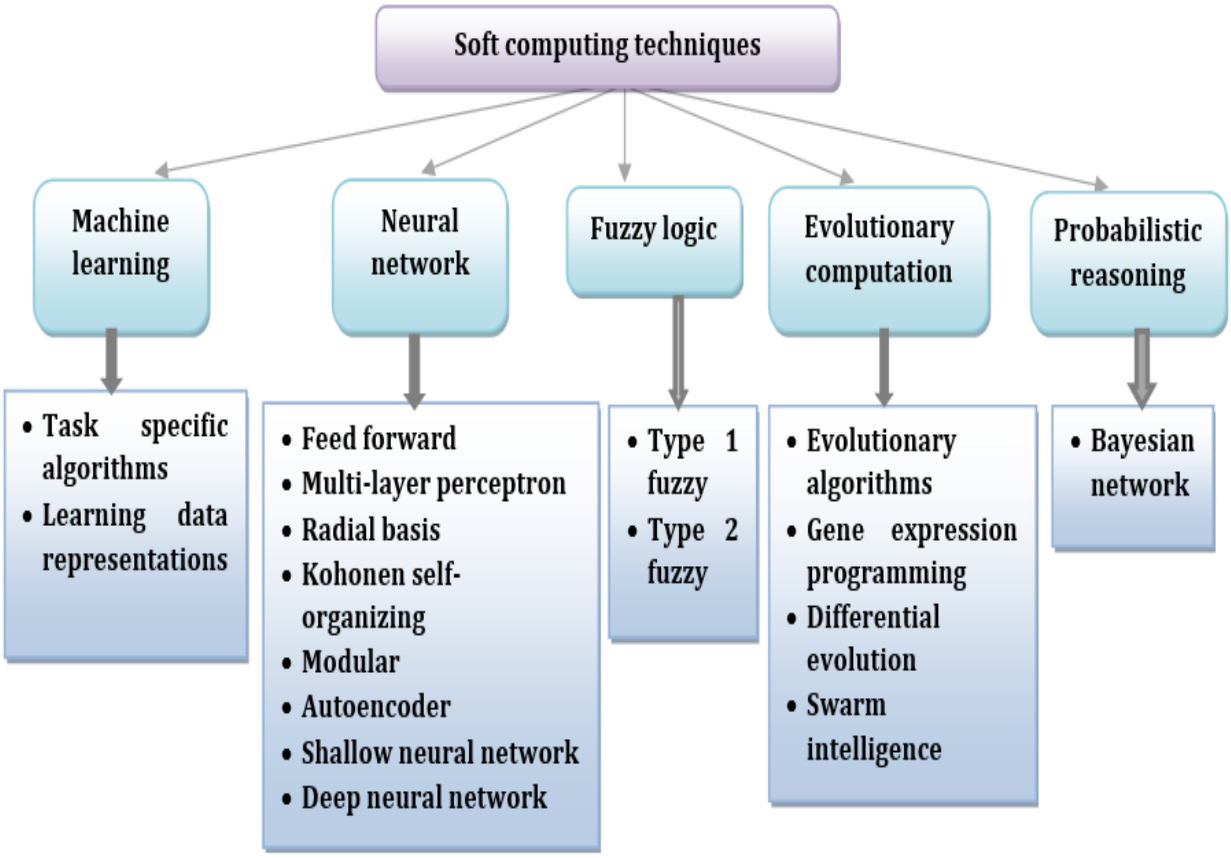


**Fig.1.7.** Categorization of soft computing techniques

From figure 1.7, it is apparent that soft computing is a 'blanket term' comprising several techniques which are themselves interrelated to one another. Also, referred to as computational intelligence techniques, soft computing (SC) techniques are categorized into machine learning (ML), neural networks (NN), evolutionary computation, fuzzy logic and probabilistic reasoning) [34].

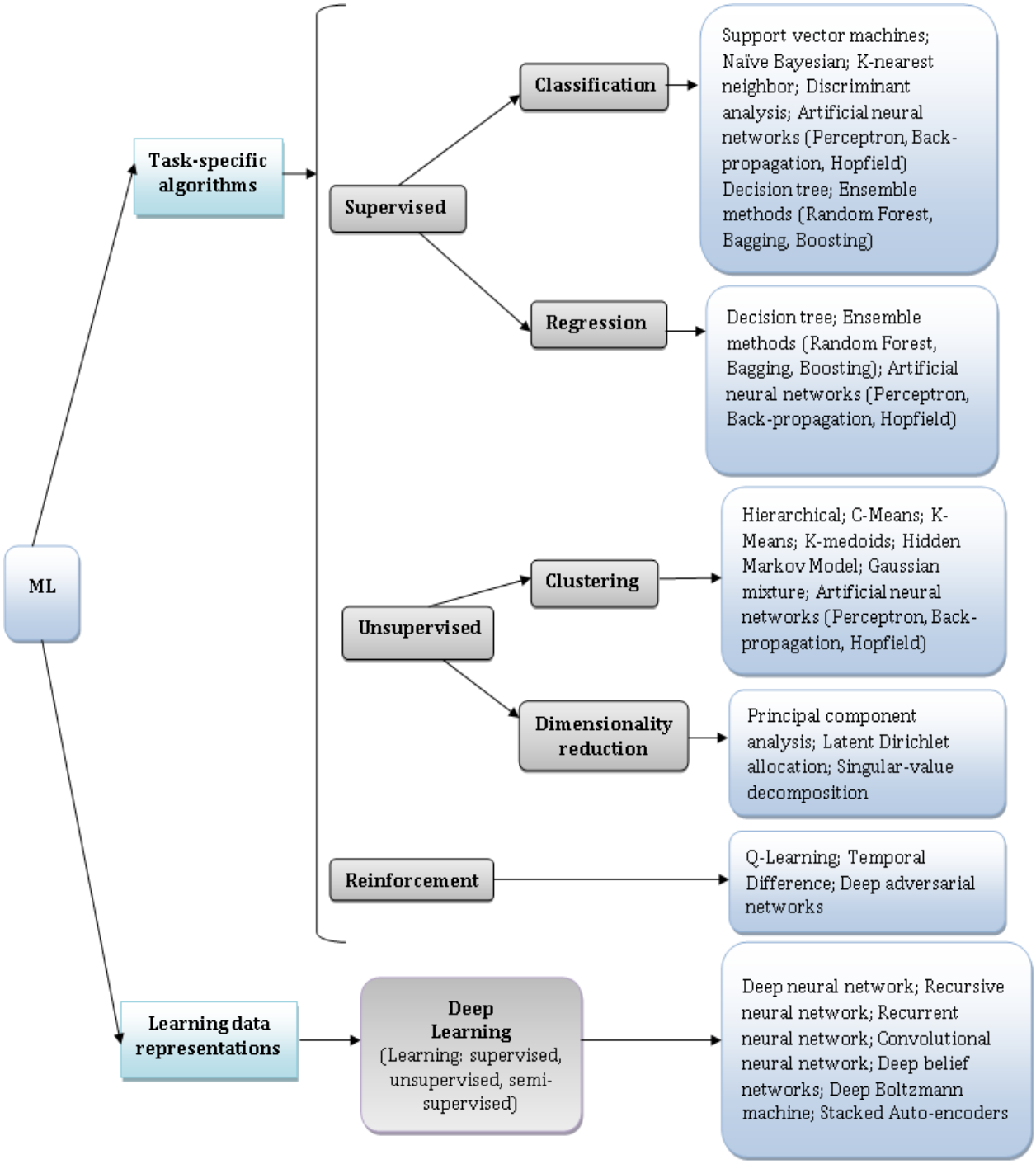


**Fig.1.8.** Categorization of machine learning techniques

From figures 1.7 and 1.8, we can infer that DL is considered as a sub-part of ML. Thus, it can be inferred that SC, ML and DL are inter-connected to each other (as shown in figure 1.9).

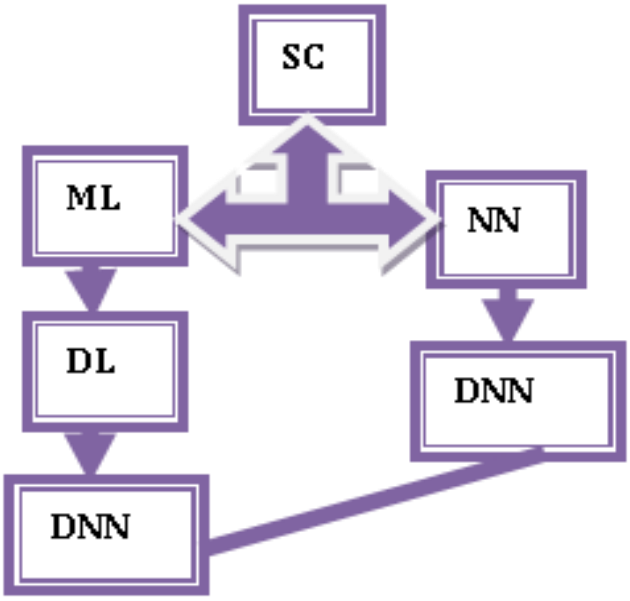


**Fig.1.9.** Relation between SC, ML and DL

These deep architectures have better representation learning capabilities. They perform automatic feature extraction for the desired outcomes and are also fast. Assessing the user-generated content in social media could be rewarding for automatic cyberbullying detection using the deep neural architectures. This research demonstrates the viability, scope & significance of using DL models for CB detection in social media portals. Application of deep learning models for cyberbullying detection in social media is an upcoming area of research for finding, exploring and analysing the extensibility of human-based expressions.

## 1.6 Organization of Thesis

This section presents the organization of the thesis.

**Chapter 1** will discuss the introduction of the chosen research area. It briefs about the harmful impacts of cyberbullying on the victim and the society. A brief description notifying the understanding and impact of cyberbullying on social media will be illustrated in this chapter. It talks about the challenges of the chosen research area and mentions the need of deep learning for cyberbullying detection in social media. The motivation and open scope of the chosen research area will be discussed which will be followed by the formation of statement of research question and research objectives. It also discusses the details about the fundamental concepts related to the chosen research area. A brief description of the key terminologies such as social media, cyberbullying and deep learning will be given. Further, this chapter will comprise of organization of the thesis with summary of the chapter at the end.

**Chapter 2** will comprise a literature review for cyberbullying detection using soft computing techniques in social media. The work will be represented in the form of a literature review within the promising area of cyberbullying detection using soft computing techniques in social media. Review process will describe the purpose of the stated phases to be performed during the conduct of a literature review. Review planning phase will contain the motivation and aim of the research, to gather and analyse the relevant primary studies of research. The next phase will elaborate searching strategy. Review reporting phase will document the results and discussion of the complete review. Thereafter, literature survey will be presented in a tabular format comprising the selected studies. Afterwards, key observations and the identified research gaps will be listed followed by the chapter summary.

**Chapter 3** will explicate about cyberbullying detection on textual social media content using baseline machine learning techniques on the datasets (from various social media such as Ask.fm, Formspring.me and MySpace). The methodology, dataset details and the findings of the work will be presented in this chapter. The details about the application of baseline machine learning techniques on social media for cyberbullying detection will be discussed. A brief summary of the above study will end the chapter.

**Chapter 4** will discuss cyberbullying detection for textual social media content using deep learning. It will brief about the proposed deep learning-based model for cyberbullying detection on textual social media such as Formspring.me and MySpace using attention-based mechanism. The methodology, dataset details and the findings of the work will be presented in this chapter. The details about the application of the proposed model on textual social media content for cyberbullying detection will be discussed. A brief summary of the above study will end the chapter.

**Chapter 5** will brief about multi-modal cyberbullying detection on social media content using deep learning. It will explain about the proposed model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image) on the data collected from social media namely YouTube, Instagram and Twitter. The methodology, dataset details and the findings of the work will be presented in this chapter. The details about the application of the proposed model on social media for multi-modal cyberbullying detection will be discussed. A brief summary of the above study will end the chapter.

**Chapter 6** will discuss cyberbullying detection on textual mash-up or code-mix social media content using deep learning on the dataset collected from social media such as Twitter and Facebook. It will focus on cyberbullying detection in the code-mix data, specifically the Hinglish, which refers to the juxtaposition of words from Hindi and English language. The methodology, dataset details and the findings of the work will be presented in this chapter. The details about the application of the proposed model on social media for cyberbullying detection will be discussed. A brief summary of the above study will end the chapter.

**Chapter 7** will discuss the conclusion. A thorough discussion of future scope and open areas of the research will be discussed, followed by discussions on the limitations of the work.

Chapters will be followed by a reference line-up which details out the citation sources used in the thesis.

## 1.7 Chapter Summary

This chapter has put forward the groundwork for this thesis. It briefs about the unison of social media, cyberbullying and deep learning. It also shows the relevance of using deep learning as a solution for overcoming the linguistic challenges related to social media for detecting cyberbullying. The need and motivation of the research area has been explained along with the organization of the thesis.

# Chapter 2

# Literature Survey

In order to comprehend the relevant work within the area of cyberbullying detection using deep learning in social media, a systematic literature review (SLR) for CB detection using all the techniques catered within the umbrella term, soft computing (SC) was conducted. SC is a 'blanket term' comprising several techniques which are themselves interrelated to one another. These techniques use approximate calculations to provide imprecise but usable solutions to complex computational problems. Also, referred to as computational intelligence techniques, SC techniques are generally divided into following categories (ML, NN, evolutionary computation, fuzzy logic and probabilistic reasoning) [34]. As already discussed in the previous chapter, deep learning is considered as a part of the broader family of ML based on learning data representations. DL consists of deep NN (DNN), recursive NN etc., whereas NN is also an established sub-type of SC techniques which includes feed forward; multi-layer perceptron; deep NN (DNN) and others. Thus, it can be inferred that SC, ML and DL are inter-connected to each other. So, in order to understand the recent trends within the area of cyberbullying detection using soft computing techniques in social media, a literature review was conducted. The SLR was conducted for a period from April 2003 till Sept 2018. Later, it was extended to include recent studies as well, primarily focusing on CB detection in social media using deep learning.

- This review was done based on the format given by Kitchenham & Charters [35]. A systematic literature review intends to identify, critically assess and combine the findings of all pertinent, high-quality primary studies addressing specific research questions (RQs).

- The prime focus was on understanding the viability, scope and significance of this alliance of using SC for cyberbullying detection in social media.

## 2.1 Review Process

The overall review-process was categorized in to six phases as depicted in figure 2.1. The first phase was denoted as formulation of research questions (RQs), followed by search strategy and study selection. Next phase was quality assessment and data extraction. Last phase was result reporting.

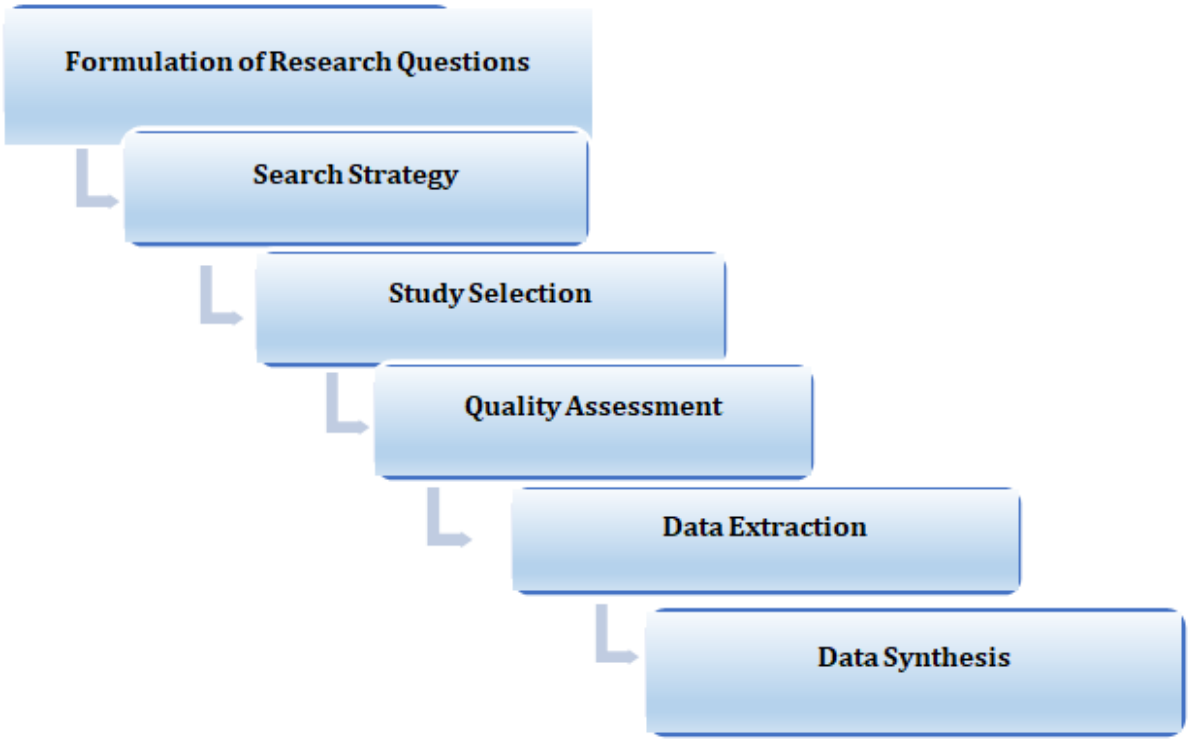


**Fig.2.1.** Phases of review process

The first phase ascertains and formulates the research questions (RQ's) as per the chosen research area. Second phase (i.e., search strategy) aids in recognizing and locating the appropriate studies in accordance to the defined RQs. Filtering of the studies is usually done using '*Inclusion-Exclusion'* criteria under the study selection phase. This phase yields the number of the relevant studies that can be included within the specified domain. Assessing the quality of the selected studies is also crucial. So, quality assessment is done for evaluating the actual relevance of the selected studies. Post this phase is the *'Data extraction'* phase which extracts the relevant and required data in order to answer the specified research questions. It produces a summarized critique to evaluate, extend and/or find research gaps that could actually aid in providing right directions for the future research as well. The last *'Data synthesis'* phase summarizes the results of these qualitative filtered studies using visualization tools such as graphs, charts and meta-summary tables. Following are the details related to each of the phases of the review process:

**Formulation of Research Questions**: This section deals about identifying and formulating the RQs as per the chosen research domain. Followed by mapping of the selected studies with those RQs. Following RQs were identified to conduct this SLR:

- RQ1: On which datasets and domains the studies using soft computing techniques for cyberbullying detection in social media have been conducted?
- RQ2: Which is the most frequently used soft computing techniques used for cyberbullying detection in social media?
- RQ3: What are the widely applied key performance indicators that are used for evaluating the applied techniques?

**Search Strategy:** An adept strategy of extensive search (starting the first reported study in April 2003-till Sept 2018), was set up in order to extract as many as possibly related research studies which expound the use of soft computing techniques in cyberbullying detection. This phase fragmented the selected RQ's into distinct concepts in order to generate the 'search-terms' which could further be searched in the identified databases/e-portals/digital libraries/digital portals. The search terms recognized for this review were: cyberbullying, SC techniques, ML, supervised, unsupervised and social media. These were then searched through the paper titles, the keywords and the abstracts of the selected studies belonging to high-repute journals. All the studies belonged to either Wiley, ACM, Elsevier, IEEE, Springer and Taylor and Francis. The grammatical-variations of these terms were also used to search exhaustively. During the search, Boolean expressions (like or/and) also aided in filtering out the non-related studies. Cross-citations were identified too by referring to the reference-section of the selected papers. Overall, this phase precisely helps in identifying, selecting and extracting the research-papers in order to conduct the review.

**Study Selection:** This phase (called as 'Inclusion-Exclusion criteria') deals with limiting and restricting the scope of the search. It works as a sort of filtration for selecting or rejecting the studies. The main criteria was that every selected study must match to at least one RQ. Studies were extracted utilizing the search-terms, publication-year, selected journal and citations. The subsequent Inclusion-Exclusion Criteria that was adopted is as follows:

***Inclusion criteria:***

- Studies published in the last fifteen years i.e., from April 2003-Sept 2018.
- Studies representative of cyberbullying detection in social media.
- Studies focusing on the application of unsupervised and supervised learning algorithms.
- Studies focusing on the application of supervised machine learning (ML) algorithms like Decision Tree (DT), Support Vector Machines (SVM), k Nearest Neighbour (kNN), Random Forests (RF), Linear-Regression (LR), Logistic-Regression (LogR), Boosting (Bos), Bagging (Bgg), Adaboost (Adb), Multiple Regression (MR), Maximum Entropy (MaxE) etc. for detecting cyberbullying in social media.
- Studies with unsupervised machine learning algorithms in soft computing such as K-Means Clustering (KMC), C-Means Clustering (CMC), Hierarchical Agglomerative Clustering (HAC) etc.

- Studies including soft computing techniques such as Probabilistic-Reasoning which includes Naïve Bayesian (NB) or Bayes Network (BN), Neural Networks (NN), Fuzzy logic (FL), Evolutionary Computing (EC) for cyberbullying detection in social media.
- Studies representing the application of deep learning (DL) techniques like Convolutional Neural Network (CNN) etc. for cyberbullying detection in social media.
- Studies with hybrids of soft computing techniques for detecting cyberbullying in social media.
- Studies involving the comparative analysis of aforesaid techniques.
- Studies involving detecting cyberbullying in social media in English language only.
- Studies involving cyberbullying detection in multimedia like images, texts, videos etc.

***Exclusion criteria:***

- Studies which are without appropriate empirical-analysis or benchmark comparisons.
- Studies that are purely reviews or surveys or theoretical concepts on cyberbullying detection without any experimentation or implementations.
- Studies on languages other than English (for example Dutch, Portuguese, Latin, Chinese, Arab, Spanish etc.) and multilingual cyberbullying detection (such as mash-up languages i.e., mixed usage of different languages.

**Quality Assessment**: In order to maintain the quality standard of the selected studies, novelty of the technique proposed and the technical content (data set and evaluation methods used) was also taken into consideration. The quality-check had been imposed in order to evaluate the worthiness, significance and strength of the selected studies based on various weighing parameters, as discussed next:

- **Novelty**: to judge whether the proposed technique is a novel one or just an enhancement or improvement over an existing one
- **Technical content**: to discover the real and clear motivation behind the proposed technique. Also, to find whether the scope and limitation of the proposed technique is evident and unambiguous.
- **Result and analysis**: to assess whether the proposed technique is tested on a standard benchmark data set or a random data set, with proper evaluation of efficacy measures and compared with existing techniques.
- **Publication**: to identify whether the selected study belongs to a conference or a high impact journal and the number of the citations that the study has. Although not much weightage has been given to this parameter as a recent study may not have many citations.

Thus, each selected study was evaluated out of 10 and scored on the following basis: 2 for novelty, 1.5 for publisher, 5 for results and analysis in which 2 was for data set, 2 for evaluation criteria used and 1 for the comparison with any of the existing techniques and the rest 1.5 for technical writing. This qualitative assessment is given in the following table 2.1.

**Table 2.1.** Quality Assessment

| Quality Level | Number of Studies | Percentage |
|---|---|---|
| Outstanding (9.5 < score <=1) | 4 | 7.69 |
| Excellent (8 < score <= 9.5) | 9 | 17.31 |
| Good (7< score <=8) | 10 | 19.23 |
| Average (5.5 < score <=7) | 18 | 34.61 |
| Below Average(4<score<=5.5) | 6 | 11.54 |
| Poor (score <=4) | 5 | 9.62 |

**Data Extraction**: In this phase, lastly the key information was extracted from the chosen studies and was summarized based on the mapping of the selected study to one or more RQ's. The information acquired from the extracted research studies included the detailing related to the author, publication-year, data sets used in those studies, techniques that were applied, specific domain that was targeted and the social media platforms that were used for analysis. Other useful details were cross-validation techniques that was employed, the efficacy measures that were used for validation or evaluation of the techniques, followed by remarks. All these acquired information were then represented in a tabular form for further data-synthesis.

**Data Synthesis**: The main aim of this phase is to sum up and expound the extracted information in order to get the answers for RQ's identified in first phase using discussions, analysis, visual and graphical representations like tables, charts and graphs etc. Searching the research papers in digital portals and selecting the relevant studies from them were done twice in order to get the most qualitative work. Meta-analysis then quantitatively combined results of the studies in the SLR. It collated data to generate statistically significant results and summaries from the pooled set of relevant studies. Identified search terms were input as a search query which yielded 320 papers. After eliminating duplicate studies, we got 255 studies on which the inclusion-exclusion criteria were applied. 52 potentially relevant studies were then filtered for further qualitative analysis out of which 47 high quality studies eventually formed the basis of this review. Figure 2.2 depicts the review process adopted in the SLR.

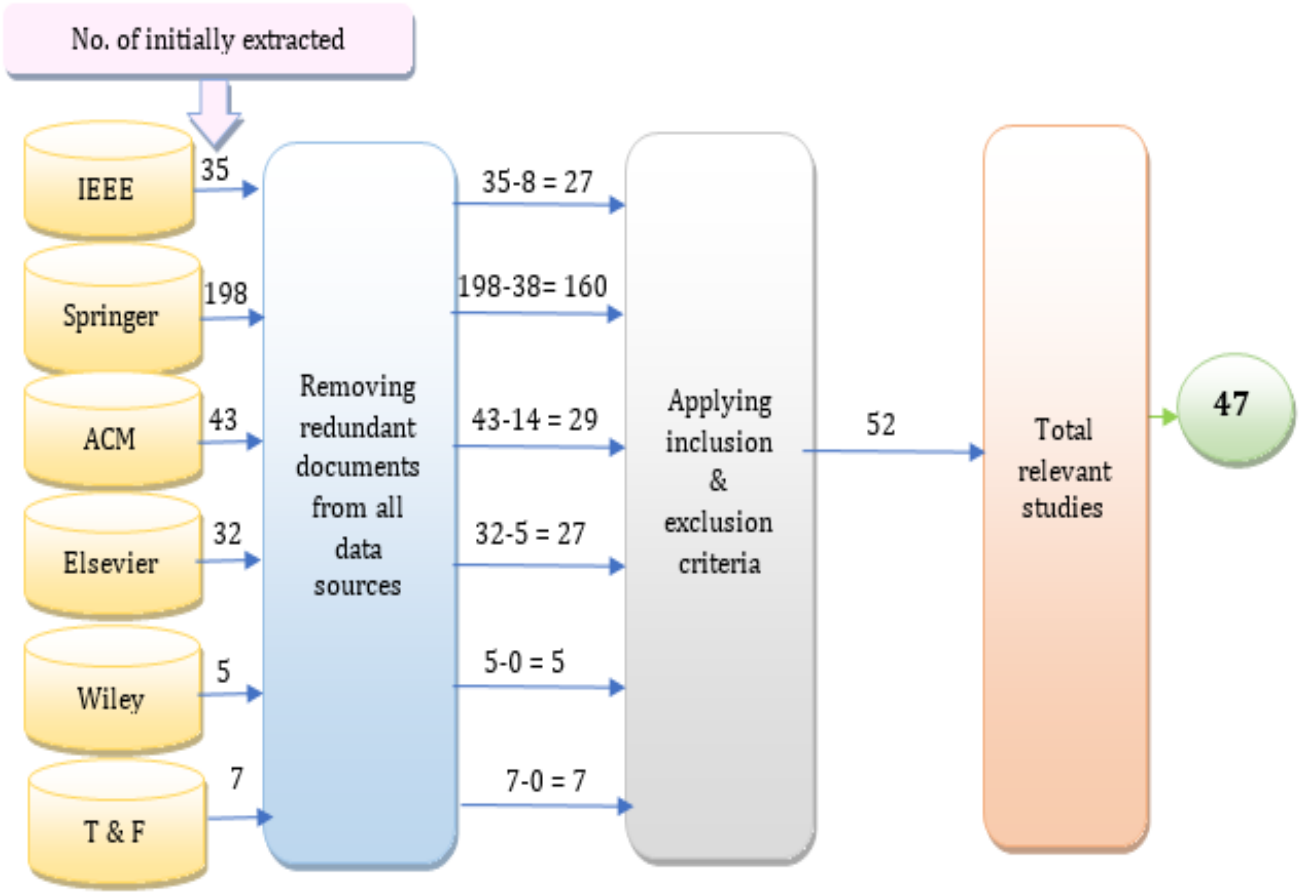


**Fig. 2.2.** Review procedure

## 2.2 Literature Survey

The survey of the short-listed studies identified for this review which demonstrate the use of SC techniques for cyberbullying detection on social media is illustrated in table 2.2. As discussed in the data extraction phase, the information extracted from the selected studies included details about the author, publication-year, data sets used in those studies, techniques that were applied, specific domain that was targeted, social media platforms that were used for analysis, domains targeted, tools which were used, cross-validation (CV) used, key-performance indicators (KPIs) [Accuracy (A), Precision (P), Recall (R), F-score/ F1-score/ F1-measure/ F-measure (F), Confidence (Cf), Sensitivity (Sn), Specificity (Sp), True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), ROC, AUC, Cohen's Kappa measure (CK), Root Mean Squared Error (RMSE), MSE, ACC, Mean (M), Standard Deviation (SD), Range (Re), Correlation (Cr), Mann–Whitney U-test (MW), Kruskal–Wallis test (KW)] and the remarks if any.

**Table 2.2.** Literature survey of the studies

| S.No. | Author & Publication Year | Techniques | Data set | Domain | CV | KPI | Remarks |
|---|---|---|---|---|---|---|---|
| 1. | Reynolds et al. [36] International Conference on Machine Learning and Applications, IEEE 2011 | DT, kNN, SVM | Author collected data from the Formspring.me and extracted the information from the sites of 18,554 users, containing 2696 posts for training and 1219 for testing purposes. | Random | 10 | A | DT outperformed the other classifier with A of around 79%. |
| 2. | Nahar et al. [37] Asia-Pacific Web Conference. Springer, Berlin, Heidelberg 2012 | BoW, PLSA, Bayes method, SVM | Data taken from the workshop on Content Analysis. | Games | 10 | A | Feature selection has shown improved accuracy. SVM performed best with 500 features. |
| 3. | Xu et al. [38] Conference of the North American chapter of the association for computational linguistics: Human language technologies, ACM 2012 | NB, L-SVM (linear), R-SVM(RBF) and LogR, CRF | Data taken from uniformly sampled 990 tweets for manual inspection by five experienced annotators. | Random | 5 | A, P, R, F | Key problems have been identified in using social media data and formulating them as NLP tasks, including text classification, role labelling, sentiment analysis, and topic modelling. |
| 4. | Kontostathis et al. [39] Proceedings of the 5th annual acm web science conference. ACM 2013 | BoW, EDLS, tf-idf | Author collected data from the Formspring.me and contained 13,652 posts. The data has been labelled using AMT. | Random | 2 | P, R, TP | Author proposed a model that resulted in better results for CD. |
| 5. | Dadvar et al. [40] European Conference on Information Retrieval, Springer, Berlin, Heidelberg 2013 | SVM | Author collected 4626 comments from 3858 distinct users | Movies | 10 | P, R, F | Author observed that incorporation of context in the form of User's activity history improves cyberbullying detection (CD) accuracy. |
| 6. | Sheeba et al. [41] International Conference on Computational Intelligence & Computing | FL, MaxE, CMC, Fuzzy C means, Fuzzy DT | Random | Meeting transcripts | - | A | Author obtained improved CD results using FL techniques. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Research, IEEE 2013 | | | | | | |
| 7. | Nahar et al. [42] In Australasian Database Conference Springer, Cham 2014 | Naive Bayes multinomial, Stochastic Gradient Descent, RF, LogR, Fuzzy SVM (FSVM, Kernel-based Fuzzy C-Means (K-FCM) clustering | Data provided by Fundacion Barcelona Media for the workshop on content analysis. | Random | 10 | P, R, F | Author proposed a semi-supervised method that had shown improved results as compared to traditional methods for CD. |
| 8. | Parime & Suri [43] International Conference on Circuit, Power & Computing Technologies [ICCPCT], IEEE 2014 | SVM | Myspace | Random | 10 | Cn | Author had taken into account the psychological factors related to cyberbullying for identifying the absence and presence of abusive content. |
| 9. | Dadvar et al. [44] Canadian Conference on Artificial Intelligence, Springer, Cham 2014 | NB, DT, SVM, MCES (Expert System) | Author collected 54,050 comments from 3,825 distinct users | Random | 10 | AUC | Author found that Naive Bayes outperformed the other two algorithms. |
| 10. | Michalopoulos et al. [45] Computers & security, Elsevier. 2014 | FL, NB, kNN, MaxE, SVM | Random | Romantic movies, chats | 10 | A, FN, FP | Author developed a 'Grooming Attack Recognition System' for real-time identification, assessment and control of cyber-attacks in favour of child protection. |
| 11. | Holt et al. [46] Journal of Criminal Justice, Elsevier 2014 | LogR | Author collected a self-administered questionnaire for 6th to 12th grade students in 14 middle and high schools in the Iredell- Statesville School System (ISS) in North Carolina. 1,972 students had completed the survey. | Adolescent problem behaviour | - | M, SD, Rg, Cr | Author discussed the implications of various demographic factors for policy responses to bullying victimization. |
| 12. | Byrne et al. [47] Journal of Computer-Mediated | LogR | C+R Research, a professional research firm in Chicago collected | Adolescent's problems, | - | A | Author obtained accuracy of more than 70% for the cases |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Communication, Wiley 2014 | | data from parents and children. The data was related to the survey involving questions on cyberbullying. | pornography or sexual imagery | | | where kids are cyberbullied by others. |
| 13. | Rafiq et al. [48] International Conference on Advances in Social Networks Analysis and Mining, ACM 2015 | Snowball sampling method, NB, Adb, DT, RF | Author collected 652K media sessions from Vine that contained information such as user id, profile information, videos posted by a user, post id's etc. | Well known celebrities | 10 | A, P, R | Amongst all, Adb had obtained the highest accuracy of around 76%. |
| 14. | Chavan and Shylaja [49] Advances in computing, communications and informatics (ICACCI), International Conference, IEEE 2015 | SVM, LogR | Author collected 2647 comments | Random | - | P, R, AUC,ACC | Author proposed that the suggested hypothesis increase the accuracy by 4%. |
| 15. | Balci and Salah [50] Computers in Human Behavior, Elsevier 2015 | DT, SVM, KMC, Bayes Point Machine (BPM) | Author gathered 800,000 Okey games along with the player interactions in the chat area over a period of six months. | Player demographics, statistics, game records, interactions and complaints | 10 | TP, TN, FP, FN, P, Sn, Sp. | Author proposed a model for assessing different types of features for detecting abuse automatically. |
| 16. | Nandhini and Sheeba [51] International Conference on Advanced Computing Technologies and Applications (special issue of Procedia Computer Science), Elsevier 2015 | NB, FL, FuzGen (hybrid of Fuzzy Logic and GA) | - | - | - | P, A, R, F | Author proposed a hybrid approach for CD where GA is used for optimizing the parameters and to obtain precise output and FL has been used to retrieve relevant data for classification from the input. |
| 17. | Balakrishnan V [52] Computers in Human Behavior, Elsevier | LogR | Author prepared a questionnaire of 393 participants consisting of questions related to cyberbullying | Sexting, self-esteem, family. | - | MW, KW | Author claimed that the proposed model for cyberbullying was significant where age and gender were found to be insignificant |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2015 | | involving victims and perpetrators, sexting (i.e., sharing sexually suggestive photos or messages through mobile phones and other mobile media), and personalities (i.e., questions related to their overall self-esteem and family). | | | | predictors for Cyber-victims and cyberbullies. |
| 18. | Zhang et al. [53] International Conference on Machine Learning and Applications, IEEE 2016 | PCNN, CNN | Author collected 1313 messages from Twitter and 13,000 messages were collected from Formspring.me and labelled by a web service called Amazon Mechanical Turk | Random | 5 and 10 | A, R, P, F, TP, FP, TN, FN | Author proposed a novel PCNN model for CD and the results show that the novel approach had outperformed the existing methods in terms of accuracy. |
| 19. | Zhao et al. [54] Proceedings of the 17th international conference on distributed computing and networking. ACM. 2016 | Continuous Bag of Words, Semantic-enhanced BoW Model, Embeddings-enhanced Bag-of-Words (EBoW), Latent Dirchilet Allocation (LDA), tf-idf, Latent Semantic Analysis (LSA), word embeddings, SVM | Author collected 1762 random tweets as of 6 Aug 2011. | Random | 5 | P, R, F | Author proposed a novel learning method called EBoW for CD that yielded enhanced results. |
| 20. | Hosseinmardi et al. [55] International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE | Snowball sampling method, backward feature selection approach, LogR, ridge regression classifier | Author collected data from around 41K Instagram user ids using a snowball sampling method via the Instagram API. | Elite users (Famous personality ies, like actors, singers etc | 5 | P, R, F, FP, ROC, AUC | Author achieved high performance in predicting cyberbullying using the proposed approach. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2016 | | | | | | |
| 21. | Gordeev [56] International Conference on Speech and Computer, Springer, Cham 2016 | RF, CNN-non-static, CNN (POS) | Author collected 1000 English messages | Random | 10 | F | Author observed that Random Forest classifier surpassed CNN for the task of detecting aggression for the English language |
| 22. | Hammer [57] International Conference on Industrial Networks and Intelligent Systems Springer, Cham 2016 | Logistic LASSO regression | Author collected a total of 24,840 sentences where 1,469 sentences were violent | Religious and political | - | MSE | Author proposed a method which can automatically detect threats of violence using machine learning. |
| 23. | Gordeev [58] Procedia-Social and Behavioral Sciences, Elsevier 2016 | word2vec, NN, RF | Author collected 1000 English messages | Random | 10 | A | Author proposed a method that detected automatic aggression with 88% accuracy. |
| 24. | Al-garadi et al. [59] Computers in Human Behavior, Elsevier 2016 | synthetic minority oversampling technique (SMOTE), NB, SVM RF, and kNN | Author collected data between January 2015 and February 2015 and contain 2.5 million geo-tagged tweets. Author randomly selected 10,606 tweets from collected data. | Random | 10 | P, R, AUC, F | The results exhibited that the Random Forest using SMOTE alone showed the best AUC (0.943) and f-measure (0.936). |
| 25. | Potha et al. [60] Knowledge-Based Systems, Elsevier 2016 | HAC, Bayesian hierarchical clustering, SVM | - | Sexual conversations | - | R | Author proposed a clustering-based method for extracting patterns in sexual cyberbullying data and had shown improved results. |
| 26. | Rafiq et al. [61] Social Network Analysis and Mining, Springer 2016 | LDA, Adb, DT, RF, Extra tree classifier, SVM (SVM Linear, SVM Polynomial, SVM rbf (radial basis function), SVM Sigmoid), kNN, NB, MLP, LogR | Author collected Vine information from 59,560 users about 652K media sessions. | Public or user profiles | 10 | A, P, R | RF yielded best A, P and R of more than 85% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 27. | Papegnies et al. [62] International Conference on Statistical Language and Speech Processing. Springer, Cham 2017 | Bag of words, TF-IDF, Probability of n gram emission, Context based (SVM) and graph-based classifier | Author accessed a database containing 4, 029, 343 messages where 779 messages were flagged by one or more users as abusive and 1558 as non-abusive. Total 2000 random messages were fetched. | Games | 10 | P, R, F | Author presented an approach based on graph features for automatically detecting online abuse. |
| 28. | Sedano et al. [63] International Conference on Artificial Intelligence and Soft Computing. Springer, Cham 2017 | SVM, FL | Author collected 18504 tweets from June to December 2016 | School students and staff members | - | A | Author presented a model where the output of SVM is fed as input to Fuzzy Logic for identifying the bullying severity. |
| 29. | Thu and New [64] International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distrib uted Computing (SNPD), IEEE 2017 | SVM | Author collected 674 real news-wire articles and 226 satire newswire articles. Also, they had gathered and pre-processed around 20K-30K tweets for each satire and non-satire corpus. Publicly available dataset was also used. | News-articles And Amazon Products Review s | 10 | P, R, F, A | Author proposed an approach for detecting satirical languages in both short text (tweets) and long text (Newswire articles and product reviews). Author had shown that the model with supervised weighting TFRF worked better in long text whereas the model with unsupervised weighting TFIDF worked better for short text. |
| 30. | Zhao and Mao [65] IEEE transactions on affective computing 2017 | Semantic-enhanced Marginalized Stacked Denoising Autoencoder - smSDA (deep learning method), mSDA, SVM, LSA, LDA, BoW, sBo W, BWM ( Bullying word matching), | Author collected 7321 tweets from Twitter from 6 Aug 2011 to 31 Aug 2011 and 1539 data samples from MySpace. | Random | 10 | A, F | Author proposed an approach called smSDA for text-based CD that showed improved results. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | word embeddings | | | | | |
| 31. | Raisi and Huang [66] International Conference on Advances in Social Networks Analysis and Mining, ACM 2017 | Participant-Vocabulary Consistency (PVC) using Alternating Least Squares, snowball sampling | Author collected 296,308 tweets from Twitter from 1 Nov 2015 to 14 Dec 2015. Author had also gathered 2,863,801 question-answer pairs from Ask.fm and 9,828,760 messages from Instagram. For labelling the data, a web service called Amazon Mechanical Turk (AMT) was used. | Random | 3 | P | Author proposed a weakly supervised PVC model for CD that had shown enhanced results. |
| 32. | Chatzakou et al. [67] Conference on Hypertext and social media, ACM 2017 | K means, Expectation-maximization algorithm, RF | Author collected 650K tweets about the Gamergate (GG) controversy. Author had also gathered a baseline dataset with 1M random tweets. | Gamergate Controversy [hate speech] and random | - | P, R, ROC | Author proposed an unsupervised machine learning analysis for better understanding the behaviours of abusive users in social media alike Twitter. |
| 33. | Chatzakou et al. [68] International Conference on World Wide Web Companion, ACM 2017 | RF | Author collected 650K tweets on hate related topics and around 1M baseline tweets random from Jun to Aug 2016. | Random and hate related (Gamergate controversy) [hate speech] | 10 | P, R, CK, ROC | The study depicted that the author's approach had produced promising results with high accuracy for detecting aggressive and bully users on Twitter. |
| 34. | Chatzakou et al. [69] Proceedings of the 2017 ACM on Web Science Conference. ACM 2017 | NB, DT, RF, NN, CBoW | Author collected 650K tweets on hate related topics and around 1M baseline tweets random from Jun to Aug 2016. | Random and hate related (Gamergate controversy) [hate speech] | 10 | P, R, ROC, CK, RMSE | Author proposed a robust approach for understanding the properties of bullies and aggressors on Twitter and improved results were achieved. |
| 35. | Garc´ıa-Recuero [70] International Conference on Advances in Social Networks Analysis and Mining, ACM 2017 | MinHashes, DT, RF , Extra Trees , Gradient Boosting, Adb and SVM, Voting | Author collected data from 163 trusted humans that provided 14193 annotations. | Elite users | 5 | P, R, F | MinHashes obtained better abuse detection rates with supervised learning and also minimized the amount of computation as well. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 36. | Ashktorab et al. [71] Web Science Conference, ACM 2017 | Snowball sampling, Latent Dirichlet Allocation topic modelling, NB | Author searched for the key terms that are generally associated with cyberbullying from Ask.fm and had then fetched random user profile information from it. | Random | - | CK, P, R, F | The proposed approach encouraged the performance of the classifier reasonably for accurate automatic detection of different discourse categories. |
| 37. | Bourgonje et al. [72] International Conference of the German Society for Computational Linguistics and Language Technology, Springer, Cham 2017 | Bayes, Bayes expectation maximizatio n, DT, Multivariate LogR, MaxE, Winnow2, BoW | Author gathered data from Twitter containing 15,979 tweets and from Wikipedia Talk pages containing 11,304 annotated Comments. | Random | 10 | A, P, R, F | Author observed that logistic regression implementation, using word unigrams, outperformed the best scoring feature set in Twitter dataset |
| 38. | Haidar et al. [73] Cyber Security in Networking Conference IEEE 2017 | NB, SVM | Author collected 91431 tweets | Random | - | P, R, F, TP, FP | Author presented a solution for detecting and stopping cyberbullying using SVM and Naive Bayes |
| 39. | Wint et al. [74] Digital Information Management (ICDIM), Twelfth International Conference, IEEE 2017 | CNN, NB | Author collected 1,578,627 tweets from Twitter and 18,554 users from Formspring.me | Random | - | A | Author measured accuracy on collected datasets using CNN and Naïve Bayes. |
| 40. | Sarna & Bhatia [75] International Journal of Machine Learning and Cybernetics, Springer 2017 | NB, kNN, DT, SVM | Author collected random tweets via a customized crawler written in Python. | Random | - | P, R, F | Author has shown that less users are involved in indirect cyberbullying than direct cyberbullying. |
| 41. | Rakib and Soon [76] Asian Conference on Intelligent Information and Database Systems, Springer, Cham 2018 | RF | Author collected 6,594 raw comments. | Random | - | AU C, P | Author depicted that the presented model had2% improvement of precision over the next best score. |
| 42. | Agrawal and Awekar [77] | CNN, LSTM, BLSTM, | Author collected 12k posts for FormSpring,16k | Random | 5 | P, R, F | This study analysed cyberbullying detection on various topics across |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | European Conference on Information Retrieval, ECIR, Springer, Cham 2018 | BLSTM with attention, logistic regression, SVM, RF, NB | for Twitter and 100k for Wikipedia | | | | multiple SMPs using deep learning-based models and transfer learning. |
| 43. | Chen et al. [78] Neural Computing and Applications, Springer 2018 | SVM, Logistic regression, LSTM+2D TF-IDF, CNN +2D TF-IDF, LSTM+ EMBEDDING, CNN +EMBEDDING | Author collected random aggressive comments from Twitter. | Random | - | A, Micro-AUC, Macro-AUC | Author achieved improvement in convolutional neural networks (CNN) using 2-dimensional tf-idf features. |
| 44. | Koban et al. [79] Computers in Human Behavior, Elsevier 2018 | MR | Author conducted an online survey of 256 participants concerning personality dispositions, participants' Internet and Facebook usage, as well as single-item measures of their interest and their level of expertise in four different news subjects (i.e., politics, sport, social issues, and terrorism). | Personality, politics, sport, social issues, and Terrorism. | - | M, SD, Cr | Author presented a study that examined participants' intention to comment in an uncivil manner that typically hinders a productive public discussion. |
| 45. | Coletto et al. [80] World Wide Web Conference 2018, ACM 2018 | LogR | Author used twitter datasets. The first one contained 977k English tweets from Dec 2008 to Jan 2009 and the other contained 1M English tweets collected in Dec 2015. | Random | - | P, R | The results showed that aggressive users smile less. Also, they appeared not happy in their profile pictures, |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 46. | Sharma et al. [81] International Conference on Advances in Computing & Communication Engineering (ICACCE) 2018, IEEE | LogR, SVM, RF, Gradient Boost | Author collected data that contained2235 samples. | Random | - | A, P, R AU C | Results depicted that Logistic Regression and Random Forest performed better than SVM and Gradient Boosting. |
| 47. | Bu et al. [82] International Conference on Hybrid Artificial Intelligence Systems, Springer 2018 | CNN, LSTM | Author collected data that contained 8815 comments. Amongst all, 2818 comments were labelled as cyberbullying. | Random | 10 | RO C, AU C, | Author proposed a hybrid architecture of character-level CNN and word-level LSTM that outperformed other machine learning methods. |

Few recent works also report the use of SC techniques for cyberbullying detection in social media. Ibrohim et al. [83] (2019) and Pratiwi et al. [84] (2019) studied hate speech and abusive language identification in Indonesian tweets. Haider et al. [85] proposed a multilingual cyberbullying detection system using machine learning and natural language processing techniques and validated their model on content written in the Arabic language from Facebook and Twitter data. In another study, Haider et al. [86] extended their previous work and provided a solution for detecting cyberbullying in Arabic content and stopping cyberbullying. Pawar et al. [87] (2019) proposed a Multilingual Cyberbullying Detection System for detection of cyberbullying in two Indian languages namely: Hindi and Marathi. Arreerard et al. [88] proposed a model for classification of defamatory Facebook comments in the Thai language using Machine learning classifiers. In 2019, Tarwani et al. [89] developed a system to detect cyberbullying in Hindi-English code-mixed Instagram and YouTube comments using eight machine learning techniques. In 2019, Gupta [90] utilized bi-directional sequence models to tackle a classification problem in categorizing social content written in Hindi-English into abusive, hate-inducing and not offensive categories. Various secondary studies on cyberbullying detection on multilingual content have also been reported [91, 92]. Works have been done (2019) where pictures are utilized for the discovery of cyberbullying utilizing deep learning models like CNN, RNN or where semantic image features are utilized for identifying bullying [83, 84, 86]. Meng et al. [93] (2020) applied Two-Branch Parallel Neural Network with Multi-Head Self-Attention Mechanism (MHSA), Capsule Network (CapsNet) and Independent Recurrent Neural Network (IndRNN). Paul et al. [94] (2020) did the comparative analysis with the slot-gated or attention-based DL models using BERT for CB identification. Liu et al. [95] (2020) proposed a model for multi-label text classification using ELMo and attention with GRU on Kaggle's toxic comment classification data. Krešnáková et al. [96] (2020) carried out experimentation on Kaggle Toxic Comment Classification dataset using different text pre-processing technique with DL models such as CNN, GRU, Bi-LSTM+CNN, Bi-GRU+CNN.

Muneer et al. [97] (2020) did comparative analysis of ML for CB detection on Twitter dataset.

## 2.3 Key Observations and Research Gaps

The literature review helped in identifying the following research gaps:

- Detecting cyberbullying behaviour is a non-trivial natural language processing task that suffers from anonymity, biasing and may apprehend freedom of expression online.
- Deep learning (DL) techniques, ensemble methods, evolutionary-computing and hybrids including neuro-fuzzy models have been least explored in order to substantiate their influence on cyberbullying detection.
- The existing models using SC techniques for cyberbullying detection on social media have majorly considered Twitter, MySpace, Formspring.me, Facebook as the database, making other technologies such as Reddit, Vine, Instagram, Flickr, Tumblr, Ask.fm, YouTube etc. open to further application and testing.
- Extracting, selecting and modelling computational features such as linguistic, visual, socio-demographic features (like person's economic status, age, gender, etc.), socio-ecological features (like parental monitoring, hours spent on Internet, racial/community differences etc.) and activity features (behavioural factors) needs further concurrence of soft computing techniques with natural language models and network analysis techniques.
- Most of the reported work done is to detect cyberbullying activity is using textual content, whereas other media types such as audio, video, images are open to research initiatives. Also, the use of animated GIF's, memes has recently been reported to embarrass or target people in social media, making it an open area of research.
- The informal, short, noisy and unstructured social media further add to the challenges. The use of slang, mal-formed or colloquial words, mash-up languages (mixed usage of different languages, for example, Hinglish is a mixture of English and Indian Hindi language) make detection of online bullying activities tricky and computationally hard.
- Cyberbullying is a rising public health concern that has multiple serious negative consequences including depression, anxiety, insomnia etc. Studies on cyberbullying mediated depression using computational analytics of social media text are important.

Thus, the need to exploit novel intelligent computation-based models to detect and predict cyberbullying in social media is abundant, making this research area more active and dynamic for social media researchers. It compels to look for models that combine the cognition, intelligence and optimally tuned hierarchical feature learning behaviour of deep learning with disciplines like natural language processing, psychology and artificial intelligence.

## 2.4 Chapter Summary

This chapter presented a systematic and comprehensive literature review on the research work done in different application areas of cyberbullying detection on social media using soft computing techniques. Some important conclusions have been drawn by answering identified research questions. The SLR helped us to identify the research gaps within the selected domain and aided in giving us various research directions to work upon.

**Publications:**

- Kumar, A.* & Sachdeva, N. (2019). "Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis.", *Multimedia Tools and Applications*, Vol. 78, pp. 23973–24010. https://doi.org/10.1007/s11042-019-7234-z **[SCIE- Impact Factor: 2.313],** ISSN: 1380-7501.

- Kumar, A., Sachdeva, N. (2021). "Cyberbullying Mediated Depression Detection in Social Media using Machine Learning", *In Second Doctoral Symposium on Computational Intelligence*, pp. 869-877. Springer.

# Chapter 3

# Cyberbullying Detection using Machine Learning

This research primarily focused on assessing the application of supervised baseline machine learning (ML) techniques for detection of cyberbullying on textual social media content, and to perform the comparative analysis which can further be applied to improve results of predictive analysis.

## 3.1 Methodology

Cyberbullying is primarily associated with the utilization of digital media in order to bully someone. It has grown to a level where it is seriously affecting and damaging individual's lives where social media plays a key role by providing a fecund means for bullies and the one's using such portals are more vulnerable to attacks. This online detrimental behaviour has instigated a need for devising an automated mechanism using data driven methodologies for critically analysing and detecting such unfavourable activities. This work presents the application of supervised baseline machine learning techniques namely Naïve Bayes, Support Vector Machines, K-Nearest Neighbour, Decision Tree, Logistic Regression and Artificial Neural Networks for identification and detection of textual cyberbullying content on Formspring.me, MySpace and Ask.fm datasets. The study was carried out using the Weka tool. The results were critically analysed using Accuracy (Ac), Precision (Pr), Recall (Re) and F-Measure (F) as an efficacy criterion.

Figure 3.1 illustrates the overall architecture and methodology of the process. Steps involved are data collection and data pre-processing, followed by feature extraction and application of baseline machine learning techniques for classifying the textual social media content as either bullying (CB) or non-bullying (N-CB). First phase involved the data collection from social media portals (Formspring.me, MySpace and Ask.fm). The collected data was then pre-processed for removing the language related irregularities. Then features were extracted. Thereafter, baseline machine learning techniques such as Naïve Bayes (NB), Logistic Regression (LogR), Support Vector Machines (SVM), K-Nearest Neighbour (kNN), Decision Tree (DT) and Artificial Neural Networks (ANN) were used for detection of cyberbullying in textual content.

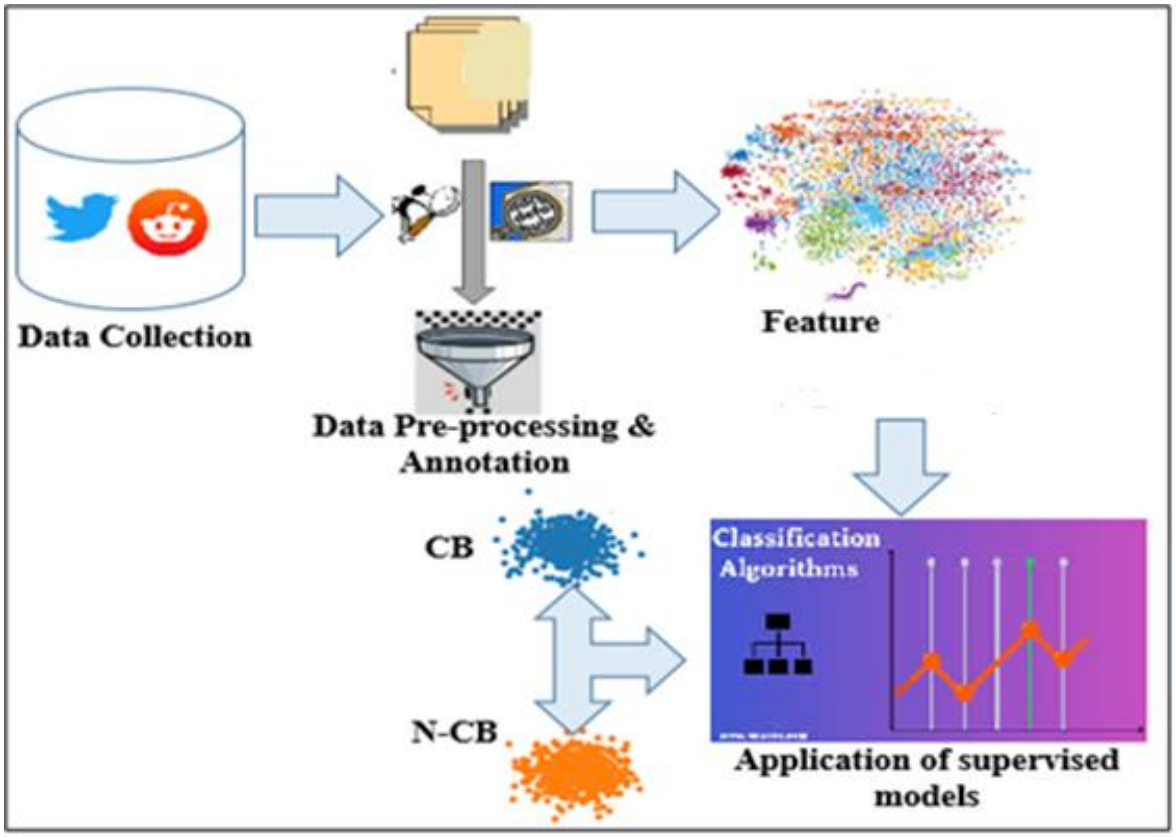


**Fig. 3.1.** System architecture

The training: test ratio was taken as 70:30 and 10-fold cross validation was used. Using 10 folds means that in each iteration of cross-validation the validation-set would be approximately 10% of the size of the total dataset. The results were assessed using performance measures (Pr, Ac, Re and F) [98] which are as follows:

**Pr:** It defines the exactness of any classifier. A higher precision value indicates fewer 'false positives' (FP) and vice versa. It is given as the ratio of true positives (TP) to all the predicted positives.

**Re**: It defines the sensitivity or the completeness of any classifier. A higher recall value indicates less 'FPs' and vice-versa. Re and Pr are bounded by inverse relation with each other. It is given as the ratio of TP to all the actual positives.

**Ac**: It is defined as closeness of a measurement to its true-value. It is calculated as a proportion of TP to true negatives (TN) among total inspected cases.

**F:** It is defined as the 'weighted' harmonic-mean of Recall & Precision. It is mere a combination of Re-Pr.

## 3.2 Dataset

The data was collected from the social media platforms like Formspring.me, Myspace.com and Ask.fm. Formspring.me [99] was peculiarly chosen as it is a very famous website amongst all the teens and college going students. It is heavily populated by the young minds and generates quality level answers to their queries or questions. Since its inception, it is gaining popularity based on social Q&A format (Question and Answering) allowing anyone to ask you the questions and providing you a platform for answering them thus simulating an interview-based layout. The most distinguishing feature of this website is that you invite others to question yourself over any topic that they want & the other person has two choices in this situation. Either he can leave the question anonymously or leave his user information. It is principally this option of anonymity that makes this website highly prone to online bullying. It is expected to contain a huge percentage of cyberbullying content that could be fruitful to be used for investigation in our study. To obtain data from Formspring.me, we crawled the website and mined information from the sites of 19,000 randomly selected users. The opted range for selection of questions per user consisted of 1 post to 1000 posts. MySpace is another social networking [100] website that offers an interactive & user submitted network of known friends. Data was crawled from MySpace groups. The dataset had 2800 posts from more than 10 separate chats. The third selected social media was Ask.fm [101] which is also one of the famous social networking websites that allows users to create their own profiles and later they may send each other questions as well. It was earlier considered to be a form of anonymous social-media as it allowed users to submit their questions anonymously. Data was crawled from the website during summer of 2014 for 72 normal users and 38 common users with over 1lakh posts per ID. We selected one public ID at random and used it for our work. After data collection, data pre-processing was done. Data pre-processing involves cleaning of all the three datasets to deal with the language related nuances. Pre-processing included removal of words that were not in English language. Thereafter, we removed highly repeated words like question and answer from all the posts. Some data points also consisted of html elements which were downloaded as text during crawling. Such data points were not relevant for our study and were hence removed. Post pre-processing, data labelling was carried out. Labelling of data instances was done to facilitate supervised learning. The Formspring.me dataset was pre-labelled using Amazon's mechanical Turk (AMT). However, the other two datasets required labelling. As cyberbullying detection of textual messages is a type of subjective task, for this, three workers were employed that labelled each post as either "yes" for having cyberbullying content in it and "no" for not having any such content. The final labelling was selected with majority voting. 10-fold cross-validation technique was used for choosing the randomly selected files for training and testing purposes.

## 3.3 Findings

This section describes the experimentation results involved in this study. Weka tool (version-3.8.1) was used for performing the empirical assessment of the aforesaid ML techniques when applied to the chosen datasets. Above mentioned datasets were analysed for textual CB detection (as bullying and non-bullying) using above specified ML techniques. The efficacy measures, namely Ac, Pr, F & Re are used to evaluate the overall performance of CB classification tasks. Results are presented in the following figures (expressed in percentages). Figures 3.2 to 3.4 illustrate the results obtained by application of supervised ML techniques on Formspring.me, MySpace and Ask.fm datasets respectively based on Ac, Pr, Re and F.

- **Findings for Formspring.me Dataset**

Figure 3.2 depicts the results obtained by application of baseline ML techniques on Formspring.me dataset.

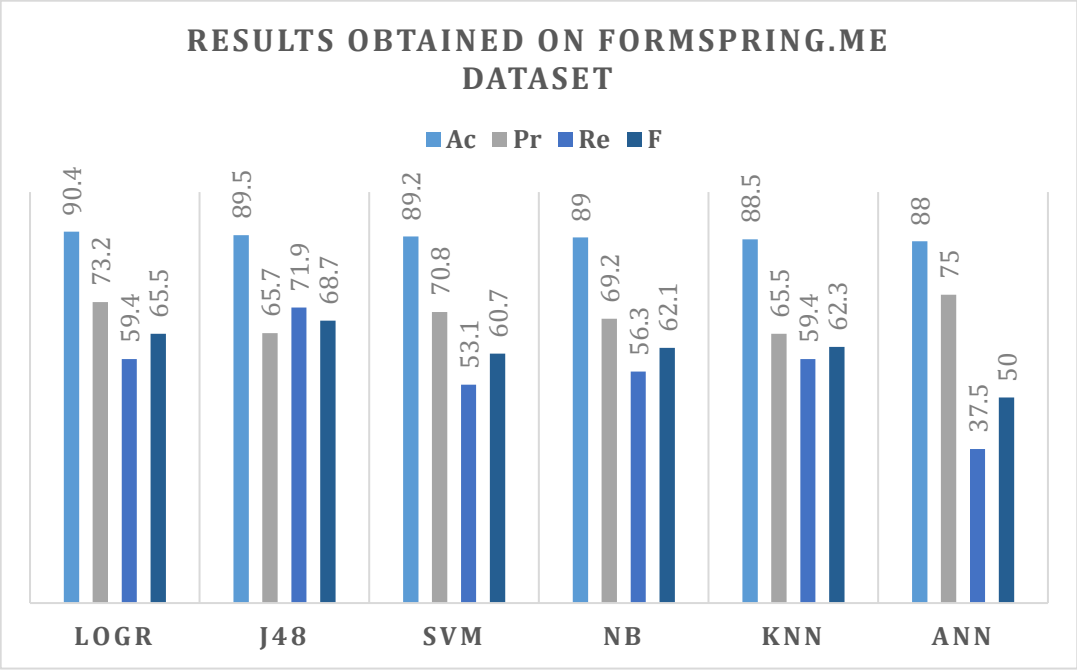


**Fig. 3.2.** Results obtained for Formspring.me dataset

- **Findings for MySpace Dataset**

Figure 3.3 depicts the results obtained by application of baseline ML techniques on MySpace dataset.

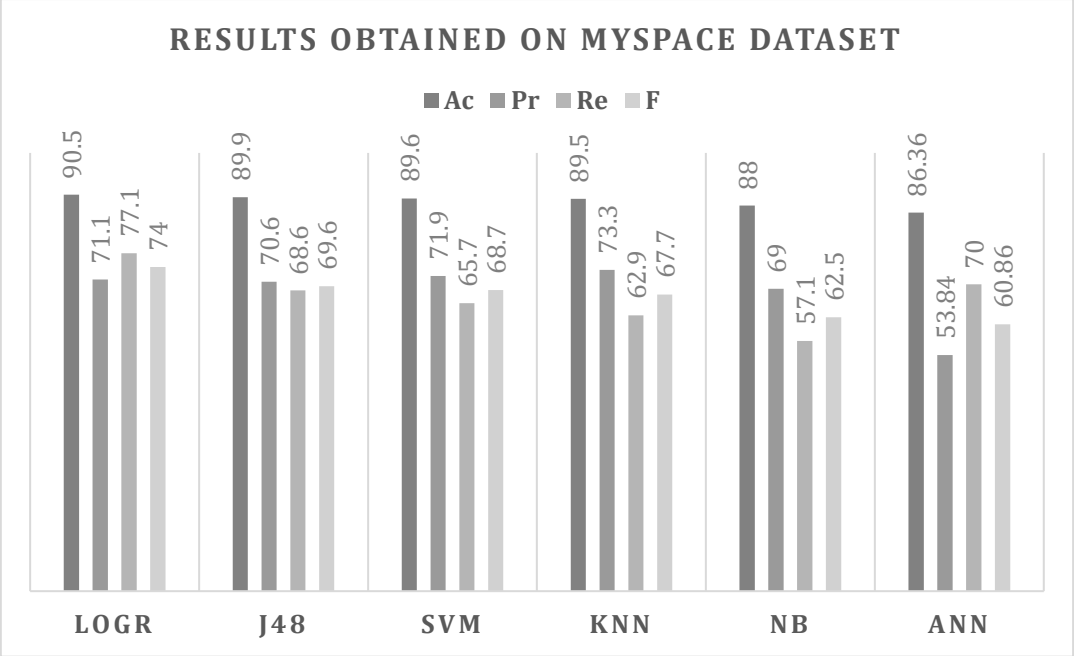


**Fig. 3.3.** Results obtained for MySpace dataset

- **Findings for Ask.fm Dataset**

Figure 3.4 depicts the results obtained by application of baseline ML techniques on Ask.fm dataset.

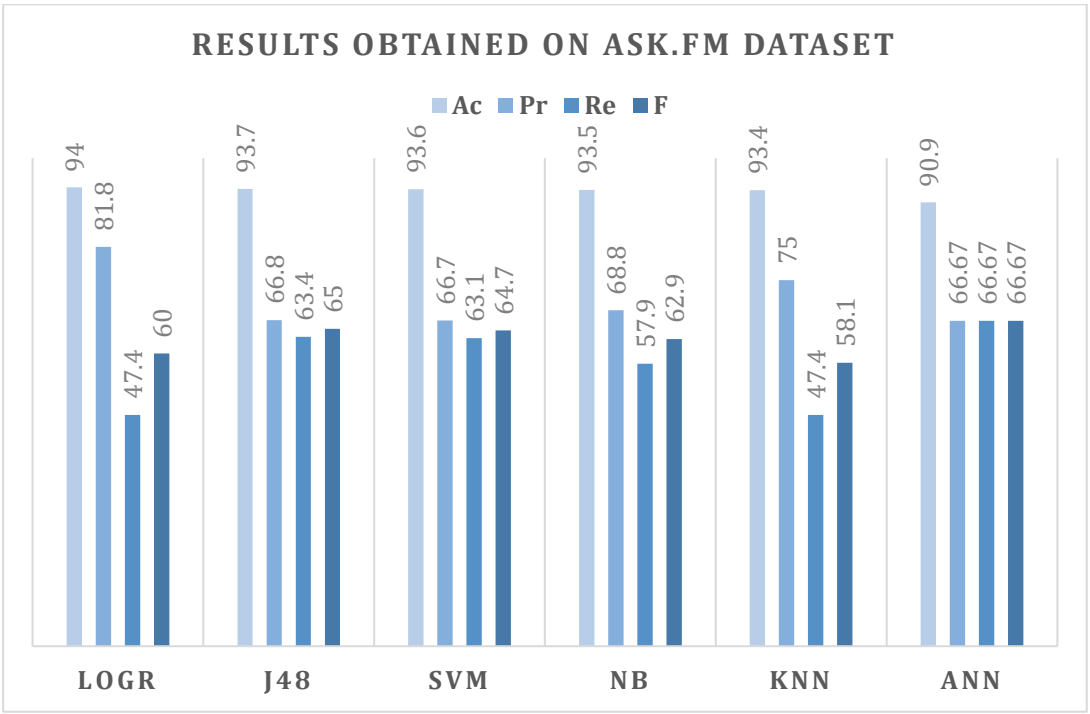


**Fig. 3.4.** Results obtained for Ask.fm dataset

The results showed that the best accuracy was achieved by logistic regression followed by J48, support vector machines, naïve Bayesian and k-nearest neighbour for all the three datasets. NB had comparable accuracy quite akin to kNN for all the datasets. Whereas artificial neural network reported the lowest accuracy.

## 3.4 Chapter Summary

This chapter assessed and empirically contrasted the three social media portals, Formspring.me, Myspace.com and Ask.fm. for cyberbullying detection using baseline machine learning techniques, namely, Naïve Bayes (NB), Logistic Regression (LogR), Support Vector Machines (SVM), K-Nearest Neighbour (kNN), Decision Tree (DT) amd Artificial Neural Networks (ANN). The results were evaluated for the classifier performance, based on precision (Pr), recall (Re), accuracy (Ac) & F score (F). Amongst the aforesaid techniques, Logistic Regression outperformed the other techniques and gave accuracy within the range of 90% -94% for all the three datasets, followed by Decision Tree, Support Vector Machines, Naïve Bayesian, K-Nearest Neighbour and Artificial Neural Networks which demonstrated the lowest accuracy.

**Publication:**

- Kumar, A., Sachdeva, N. (2020). "Cyberbullying Checker: Online Bully Content Detection using Hybrid Supervised Learning", *In International Conference on Intelligent Computing and Smart Communication 2019*, pp. 371-382. Springer.

# Chapter 4
# Cyberbullying Detection for Textual Data

As a constructive mode of information sharing, collaboration and communication, social media platforms offer users with limitless opportunities. The same hypermedia can be transposed into a synthetic and toxic milieu that provides an anonymous, destructive pedestal for online bullying and harassment. Automatic cyberbullying detection on social media using synthetic or real-world datasets is one of the proverbial natural language processing problems. Analysing a given text requires capturing the existent semantics, syntactic and spatial relationships. Learning representative features automatically using deep learning models efficiently captures the contextual semantics and word order arrangement to build robust and superlative predictive models. This research puts forward a hybrid model utilizing deep architectures, ***Bi-GRU-Attention-CapsNet (Bi-GAC)***, that benefits by learning sequential semantic representations and spatial location information using a Bi-GRU with self-attention followed by Capsule networks (CapsNet) for cyberbullying detection in the textual content of social media. The improved text representation and feature learning offers a robust model which can avoid the vanishing gradient problem in comparison to baseline neural models. The model is validated on two benchmark datasets, Formspring.me and MySpace. The proposed Bi-GAC model is evaluated for performance using F1-score. This chapter primarily focuses on binary classification of mono-lingual textual social media content for cyberbullying detection on benchmark datasets taken from social media like Formspring.me and MySpace. An ablation study was also done to ratify the results. The methodology and findings related to this research objective is presented in this chapter. A brief summary of the above study will end the chapter.

Thus, the primary work undertaken in this research includes:

- Building a Bi-GRU-Attention-CapsNet (Bi-GAC) model for cyberbullying content classification.
- Validating improved classifier performance in small sequences like social media posts by capturing semantic information, context and dependencies.
- Evaluation of Bi-GAC on two benchmark datasets using F1 score.

## 4.1 Methodology

The Bi-GAC model which gets its nomenclature from its constituent core components, realizes the complexities of textual data in user-generated content where data representation is learned as real-valued vectors. The model has been divided into two phases. Embedding, encoding and self-attention layer constitutes the first phase whereas

Capsule Network and fully connected layer with sigmoid comprises the classification or the prediction (second) phase. Here, the pre-trained ELMo word embedding was used to create the input embedding matrix. A Bi-GRU encoder is trained using ELMo embedding to generate a sequence context feature vector. This feature vector is flawed due to the existence of redundant and irrelevant features. Consequently, a self-attention mechanism is added to capture significant information. Next the CapsNet generates semantic representation using a dynamic routing algorithm which is finally used for classification of the posts. The feature encoding ability of capsules and dynamic routing allow aggregating the important information. Simultaneously, bi-directional gated recurrent units (Bi-GRU) are known for their sequential modelling capabilities which learns relationships between words from both directions. GRUs are easy and fast to train on smaller sequence data like social media posts, making them appropriate for various text classification tasks on social media. Further, self-attention mechanism helps to model dependencies between different parts of the sequence and captures important information using the mutual input interaction. Bi-GAC uses the fully connected output layer with sigmoid activation to finally classify the positive as bullying or non-bullying. Thus, apprehending the dexterity of Bi-GRU, self-attention and CapsNet, in this research, we present a hybrid model, Bi-GAC to classify online textual posts into bullying categories. Figure 4.1 depicts the architecture of Bi-GAC model.

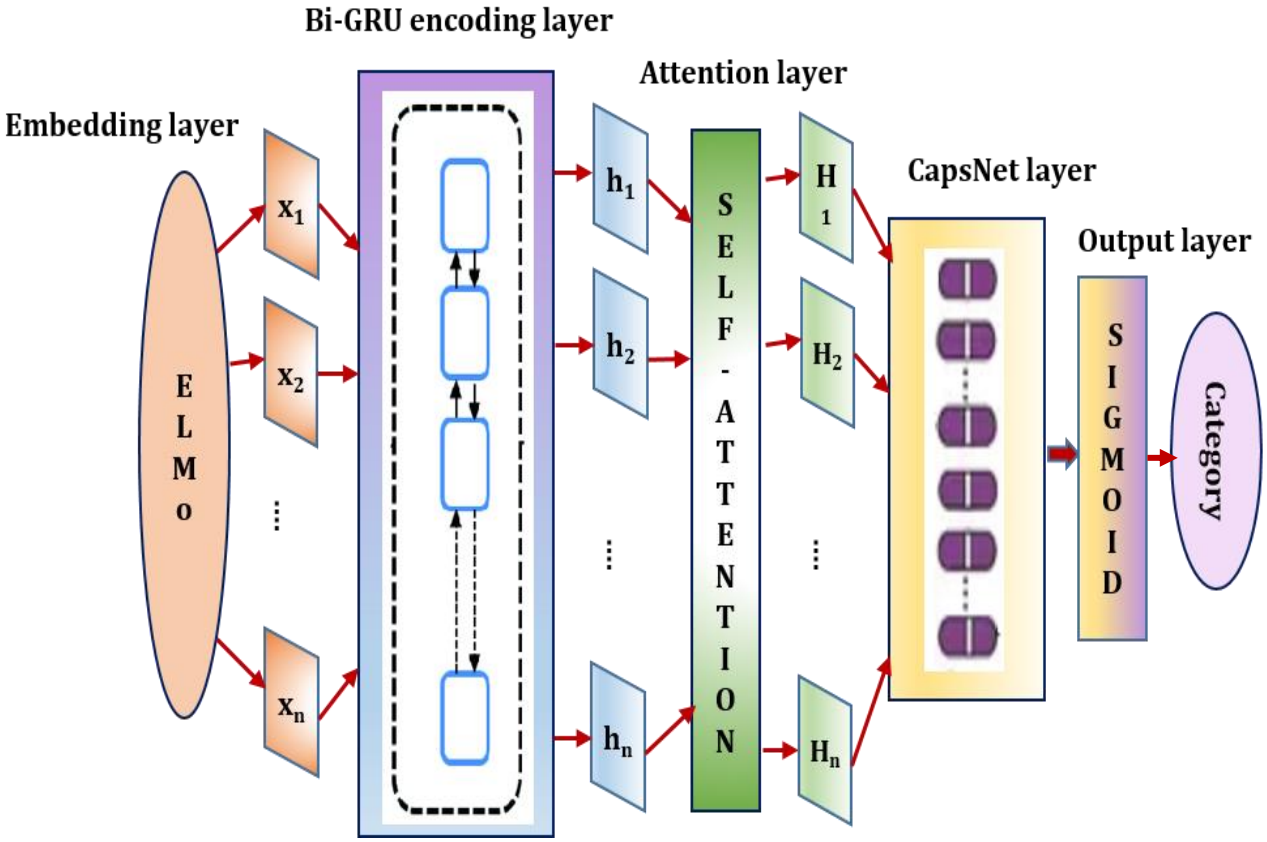


**Fig.4.1.** System architecture of proposed Bi-GAC model

### 4.1.1 First Phase

Embedding layer, encoding layer and self-attention layer constitutes the first learning phase.

**Embedding layer:** Word embeddings are based on the idea of distributional meaning: the fact that semantically (or morphologically) related words tend to appear in similar contexts. These represent each word using a continuously valued, lower-dimensional vector so that it reserves semantic information of the word. Once word embeddings have been trained, we can use them to derive similarities between words, as well as other relations. The pre-trained ELMo (Embedding's from Language Models) embedding [102] are used in this work. ELMo has an advantage above other conventional embedding's such as GloVe and word2vec as it encapsulates context in the word feature representations. These are high-dimensional representations of words, based on the contexts that different words appear in. ELMo uses a 2-layer bidirectional LSTM for learning words and their context. This design allows ELMo to learn more context-dependent aspects of word meanings in the higher layers along with syntax aspects in lower layers. Figure 4.2 shows an example of how an ELMo specific representation is generated by combining the bidirectional hidden representations.

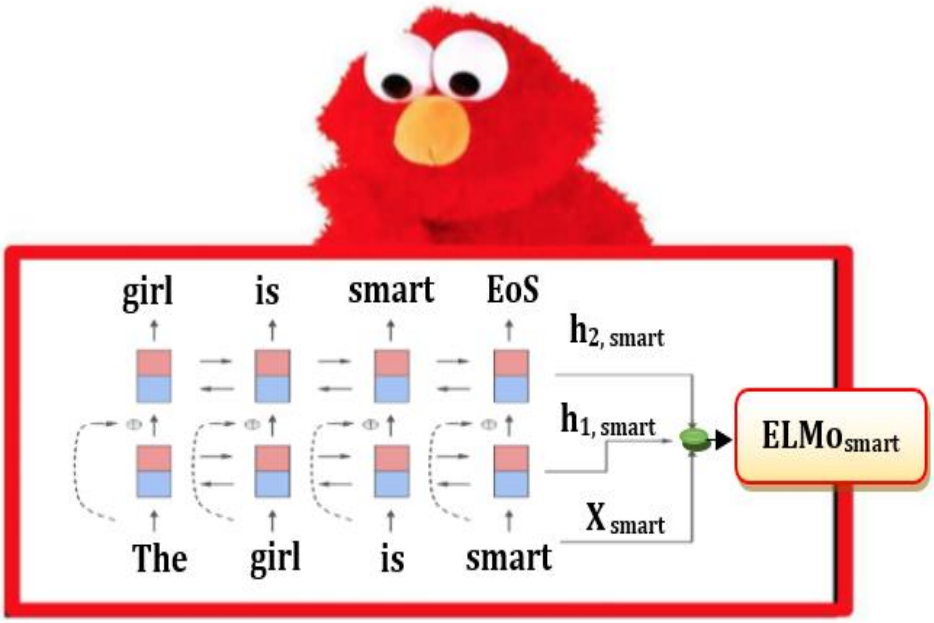


**Fig.4.2.** ELMo-specific representation for "smart"

**Encoding layer:** Recurrent neural network (RNN) models [103] have been popularly used in the field of natural language processing. But this seq2seq encoder/decoder architecture suffers from the vanishing gradients problem that is the inability of the RNN unit to reminisce values that showed initially in the sequence [104]. But this is vital within a typical NLP task like text classification, as few words rely on words that appear very early in the sentence. As a solution, both gated recurrent unit (GRU) & long-short-

term memory (LSTM) solve the problem of vanishing gradients in RNN, by replacing the neurons of the hidden layer with the memory-units to store early sequence data [105]. GRU uses few training parameters, less memory, needs fewer data to generalize, execute faster and train faster than LSTM whereas LSTM is more accurate on dataset using longer sequence. As social media posts are shorter, GRU is an apt choice for encoding.

The two gates of a GRU are as follows: -

- Update Gate: It determines how much of the past knowledge needs to be passed along into the future.
- Reset Gate: It determines how much of the past knowledge to forget.

The reset gate sits between the previous activation and the next candidate activation to forget the previous state, and the update gate decides how much of the candidate activation to use in updating the cell state. These gates aid in dealing with the long-term-dependencies. They forward and backward pass the information from the previous state to the next state. The update gate keeps a track of retaining all the important features and also aids in solving the long-term temporal dependency issues. A vector having values from 0 to 1 is received via update gate with pointwise-multiplication-operation (PMO). It uses sigmoid activation function for squashing values b/w 0 and 1. It basically helps in updating or forgetting (or disappearing) the data as the PMO would result in 0 for any multiplication with a vector of zeros. In such a scenario, the resulting values would be considered to have 'disappeared or be forgotten'. Contrastingly, for any multiplication with a vector of one's would result in the 'same value or be kept'. This mechanism would eventually help in retaining the relevant data and forgetting the non-relevant ones. The other gate which is the reset gate is normally used to ensure the amount of past information that can be forgotten. Figure 4.3 depicts the architecture of a basic GRU cell.

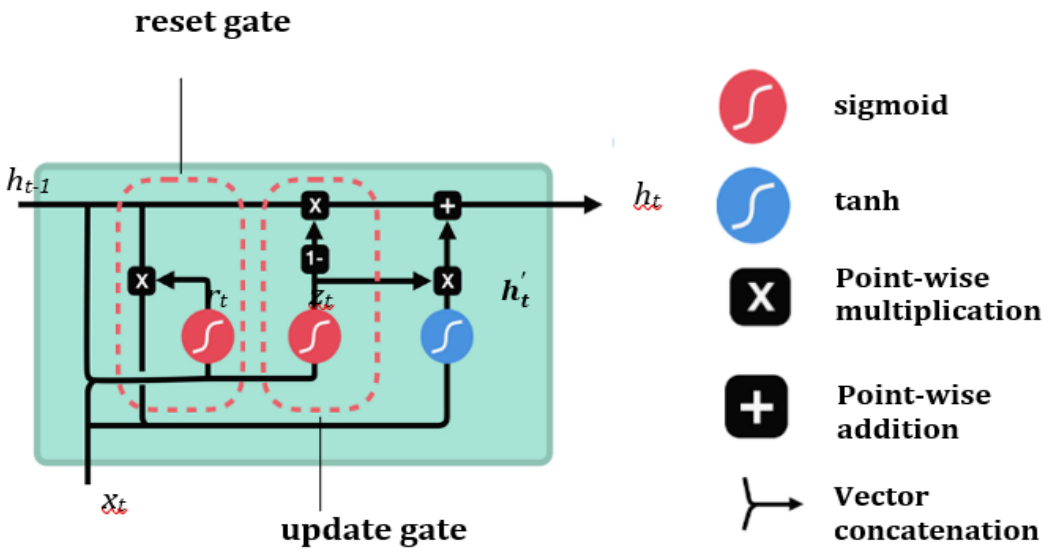


**Fig.4.3.** GRU Cell architecture

In this research, we use a bidirectional GRU's (Bi-GRU) [106, 107] encoder which takes the input sequence and encapsulates the information as the internal state vectors. A Bi-GRU allows capturing information from both previous time steps and later time steps to make predictions about the current state. Bi-GRU enables apprehending meaning and context for the sentences than a simple GRU.

The forward GRU $\vec{f}$ reads the sentence $s_i$ from $w_{i1}$ to $w_{iT}$ as given in equation 4.1.

$$\vec{h}_{it} = \overrightarrow{GRU}(x_{it}),\ t=1, 2, 3, ....T\text{-}1, T \qquad (4.1)$$

The backward GRU $f^{\leftarrow}$ reads sentences from $w_{iT}$ to $W_{i1}$, as given in equation 4.2.

$$h^{\leftarrow}_{it} = GRU^{\leftarrow}(x_{it}),\ t=T, T\text{-}1, T\text{-}2, ....2, 1 \qquad (4.2)$$

The annotation of the word $w_{it}$ is calculated by combining the forward and backward hidden states i.e., $h_{it}=[\vec{h}_{it}, h^{\leftarrow}_{it}]$.

**Attention layer:** Attention mechanism allows output to focus attention on input while producing output [108, 109] whereas a self-attention model allows inputs to interact with each other, that is, calculate attention of all other inputs with respect to one input. Self-attention is good at modelling dependencies between different parts of the sequence. Self- attention includes both location and observation value information and replaces conditioning on the entire sequence with pairwise comparisons (the importance of one word and its position conditional on some other word and its location) given as vector representations of both.

Consider the sentence "*The dog didn't eat the food because it was too full".* The word *"it"* refers to the dog. If we replace *"full"* with *"much"*, the word *"it"* now refers to the food. Attention mechanism helps to understand this, that is, in the former case there's high attention linking *"it"* and *"dog"* but in the latter case high attention shifts to *"food".* In this research, a self-attention mechanism [109] as shown in figure 4.4 is applied on the outputs of GRU layer.

## 4.1.2 Second Phase

Capsule Network and fully connected layer with sigmoid comprises the second phase i.e., prediction phase.

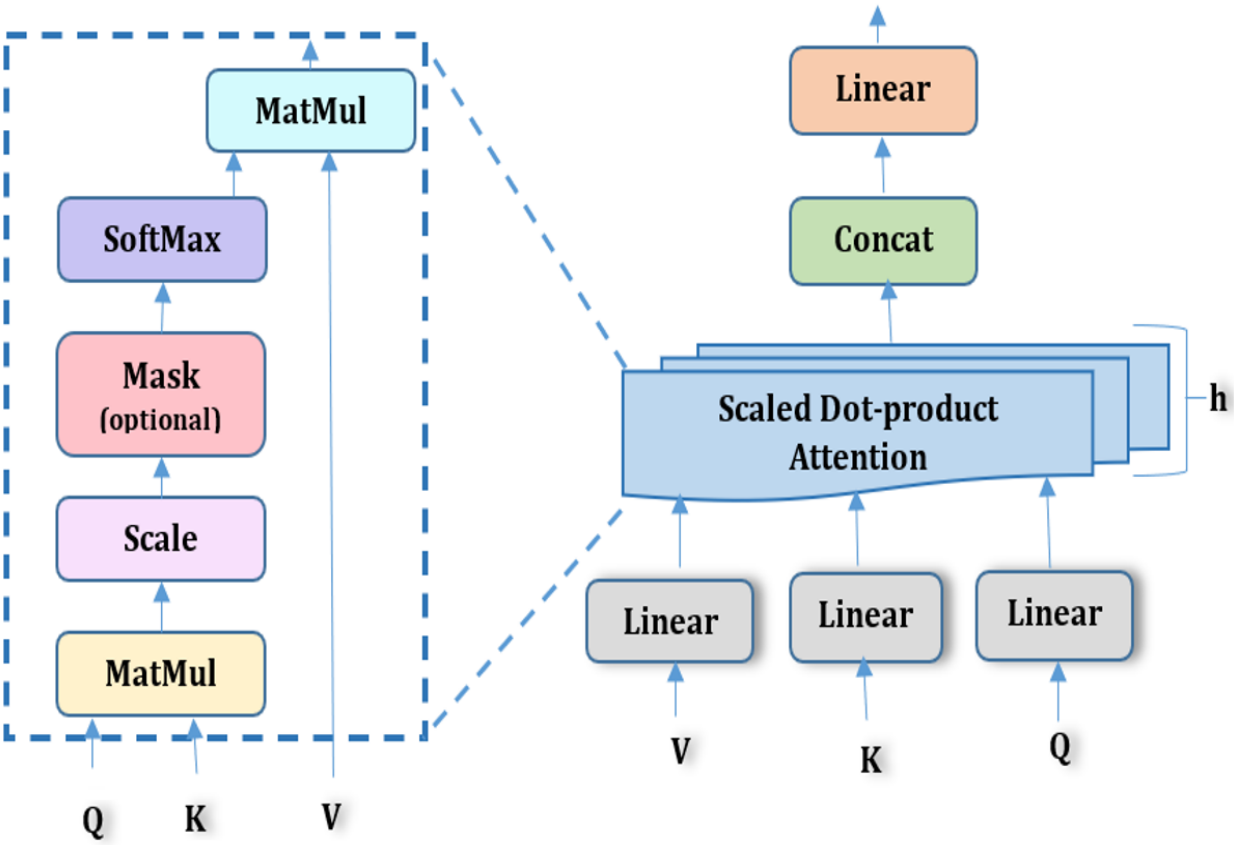


**Fig.4.4.** Self-Attention

**CapsNet Layer:** Conventionally, artificial neurons output a scalar, real-valued activation that loosely represents the probability of an observation. CapsNets [110, 111] replace the scalar output feature detectors with vector output capsules and max-pooling with routing-by-agreement. A capsule is a collection of neurons which represent distinctive properties of the same entity as outputs. Capsule's activity vector represents the instantiation parameters of a specific type of entity which adequately captures the important and relevant information (features). The CapsNet is divided into two layers: primary capsule layer and digit capsule layer. As the preceding Bi-GRU attention layer outputs high-level feature representation, these feature maps are input to the primary caps and subsequent digit caps layer as shown in figure 4.5. Capsules with vector outputs are generated by the primary caps layers as given in equation 4.3.

$$u_{j|i} = W_{ij} u_i \qquad (4.3)$$

where, 'i' is in the current lower-level primary-caps and 'j' is in the next-level layer. Connection-weight updating (as shown in equation 4.4) will occur during network training using dynamic routing algorithms.

$$s_j = \sum_i c_{ij}\, u_{j|i} \qquad (4.4)$$

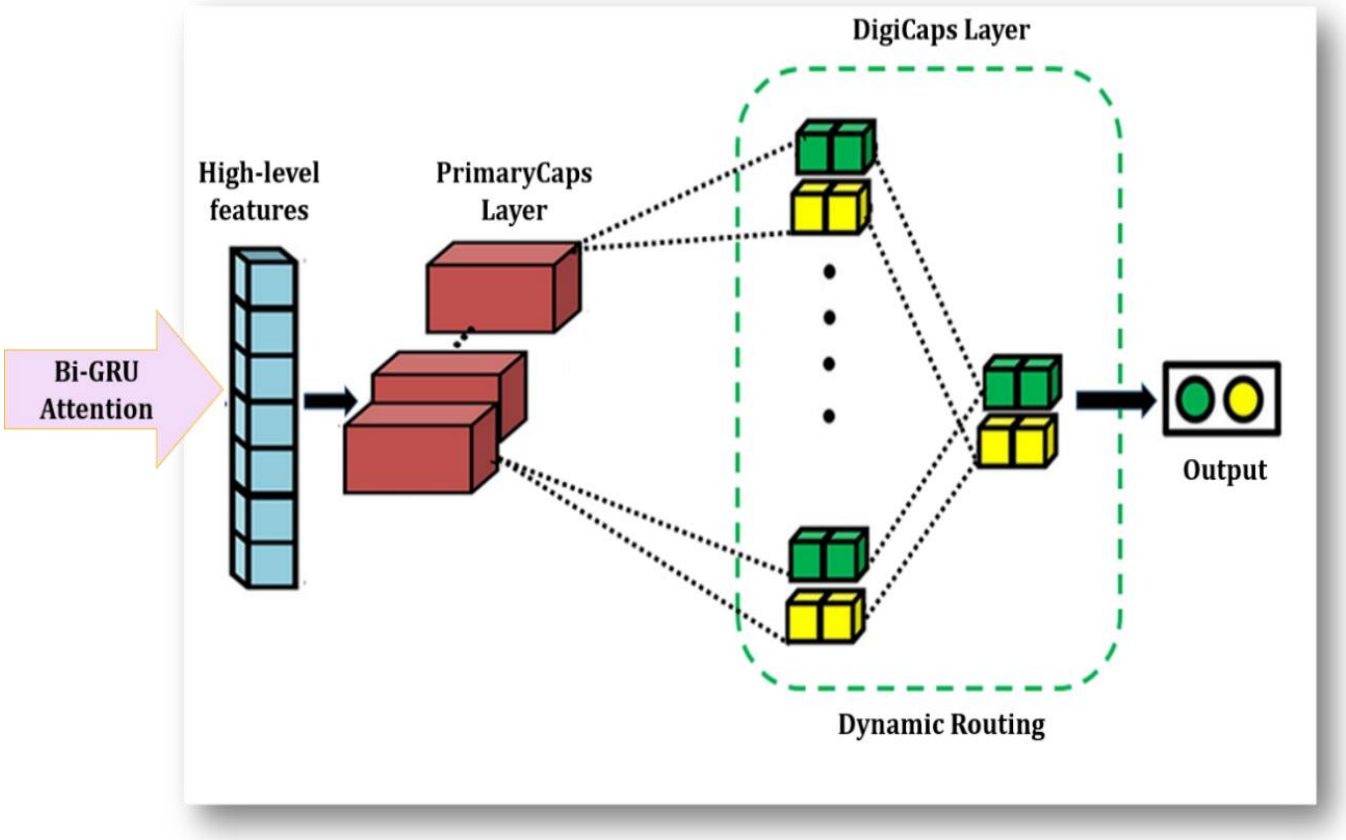


**Fig.4.5.** CapsNet Architecture

The key operations within the capsule are as follows [112]:

- Multiplying input vector matrix and weights to encode significant features and their relationships within the text.
- Dynamic routing for sending output from one low-level capsule to a higher-level capsule.
- Weighted input vectors summation.
- "Squash" function to add non-linearity. This function takes a vector and "squashes" it to have a maximum length of 1, and a minimum length of 0 while retaining its direction.

Figure 4.6 summarizes the operations within a capsule [113].

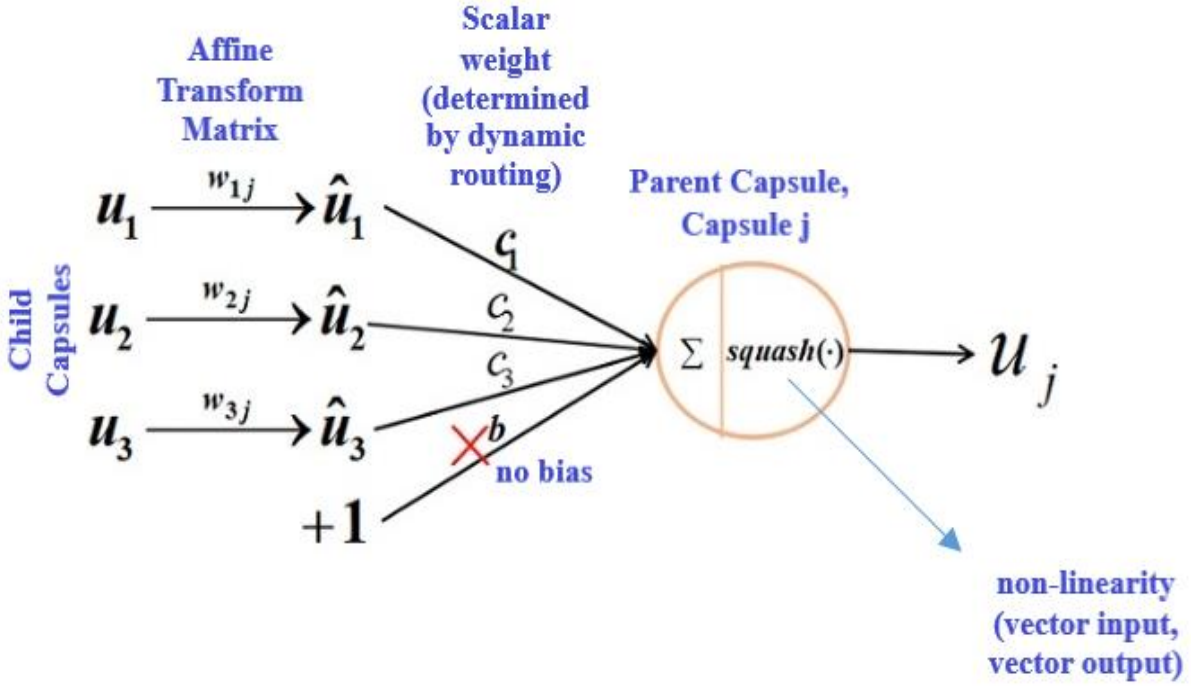


**Fig.4.6.** Operations within a Capsule

The dynamic routing algorithm [114] is given below (algorithm 4.1).

**Algorithm 4.1.** Dynamic routing algorithm

**procedure** ROUTING ($\hat{u}_{j|i}$, r, l)
for all capsule i in layer l and capsule j in layer (l + 1): $b_{ij} \leftarrow 0$.
**for** r iterations **do**
for all capsule i in layer l: $c_i \leftarrow \text{softmax}(b_i)$
for all capsule j in layer (l + 1): $s_j \leftarrow \sum_i c_{ij} \hat{u}_{j|i}$
for all capsule j in layer (l + 1): $v_j \leftarrow \text{squash}(s_j)$
for all capsule i in layer l and capsule j in layer (l + 1): $b_{ij} \leftarrow b_{ij} + \hat{u}_{j|i} . v_j$
**return** $v_j$

**Output Layer:** The final prediction layer of the proposed model is the fully-connected layer with sigmoid activation that eventually helps in obtaining the probabilities for the binary classification (target classes: bullying and non-bullying). It basically maps a real valued number using a threshold to a probability (i.e., to a no. b/w 0 to 1).

## 4.2 Dataset

This section briefs about the datasets used. Various social media datasets from Twitter, YouTube, MySpace, Kongregate, Formspring and Slashdot have been created for automatic cyberbullying detection. Two datasets from varied social media have been taken under consideration for analysing the effectiveness of our proposed model. In this research, two benchmark datasets, namely, question-answering Formspring.me and thread- style MySpace social network cyberbullying classification datasets were used for experimentation. The datasets [99, 100] are labelled for cyberbullying and non-bullying type and the sample distribution is as given in table 4.1.

**Table 4.1.** Sample distribution in datasets

| Dataset | Bullying samples | Non-Bullying samples | Total samples |
|---|---|---|---|
| Formspring.me [99] | 836 | 12288 | 13124 |
| MySpace [100] | 357 | 1396 | 1753 |

From above we could infer that both the datasets are quite imbalanced in nature. It is always evident from our pertinent literature to use a balanced dataset for any classification [115]. Getting the proper insights from the classification remains a challenging task. In totality, the dataset distribution, the performance criteria and the execution of the right model together counts for the efficiency of any CB detection model. Evidently, the datasets suffer from class imbalance and we over-sampled the data from the bullying class (sporadic class) thrice to deal with this skewness. This technique is usually used for balancing the corpus in similar studies [115].

Thereafter, the proposed model was trained and tested with the data using F1-score as the performance measure. Here, 70:30 ratio was used for training & testing purposes. We performed 10-fold cross-validation. Using 10 folds means that in each iteration of cross-validation the validation-set would be approximately 10% of the size of the total dataset. The results were then averaged across the folds, using suitable performance measures.

We used the Scikit-learn library and Keras deep learning library with Theano backend. The hyperparameters in experiments were set as follows: The Bi-GRU layer had 50 units and a dropout value of 0.2. The CapsNet layer also had a drop out of 0.2 and had 3 iterations for dynamic routing. Adam optimizer was used with the learning rate of 0.0001.

## 4.3 Findings

In this section, we discuss the findings of this research.

### 4.3.1 Model Performance

The Bi-GAC model was evaluated using the F-1 score. The results for both datasets are shown in table 4.2.

**Table 4.2.** F-1 score for Bi-GAC model

| Dataset | F1-Score |
|---|---|
| Formspring.me | 94.03 |
| MySpace | 93.89 |

The results are also compared with the existing techniques used on both datasets. The performance comparison of the proposed Bi-GAC model with the existing techniques is shown in table 4.3 and table 4.4 for MySpace and formspring.me datasets respectively. A superior performance is observed for the proposed Bi-GAC model.

**Table 4.3.** Comparison of Bi-GAC with existing works on MySpace dataset

| Reference Study | Techniques | F1-score |
|---|---|---|
| Zhang et al. [116] | Logistic Regression | 78 |
| Zhang et al. [116] | Support Vector Machine | 79 |
| Zhang et al. [116] | CNN | 85 |
| **Proposed Bi-GAC Model** | **Bi-GAC** | **93.89** |

**Table 4.4.** Comparison of Bi-GAC with existing works on formspring.me dataset

| Reference Study | Techniques | F1-score |
|---|---|---|
| Agrawal & Awekar [115] | Random Forest | 29.8 |
| Agrawal & Awekar [115] | Naive Bayesian | 35.9 |

| | | |
|---|---|---|
| Agrawal & Awekar [115] | Support Vector Machine | 42.2 |
| Agrawal & Awekar [115] | Logistic Regression | 44.8 |
| Agrawal & Awekar [115] | Bi-LSTM | 86 |
| Paul & Saha [94] | RNN+LSTM | 88 |
| Paul & Saha [94] | Bi-LSTM with Attention | 91 |
| Paul & Saha [94], Agrawal & Awekar [115] | CNN | 91 |
| **Proposed Bi-GAC Model** | **Bi-GAC** | **94.03** |

### 4.3.2 Ablation Study

We also performed the ablation study [113] which is very useful for doing the comparative study of varied architectures that will eventually help in better analysis. An ablation study is done to learn about the network by removing and/or replacing parts of the complex neural network architecture and study the model performance. The two variations studied are: using Bi-LSTM instead of Bi-GRU (Bi-LSTM+Attention+CapsNet) and using CNN instead of CapsNet (Bi-GRU + Attention+ CNN). Table 4.5 presents the F1-score of these variations where the proposed Bi-GAC (Bi-GRU+Attention+CapsNet) outperformed the other two. In this research, we did it in the following ways:

- **Demonstrating the effect of exchanging Bi-GRU with Bi-LSTM**

We replaced Bi-GRU with Bi-LSTM in our model and performed the experimentation with this ablation architecture by training and testing both the datasets with this ensemble. Here, Bi-directional LSTM is an extension of conventional LSTM. It comprises two LSTMs which tends to improve the model efficiency. These are considered better than traditional RNNs as they cater well with the gradient and long-term dependency problems of RNN. These are widely used for text classification. These are often considered as siblings of Bi-GRU in terms of efficiency but are more complex in nature as compared to Bi-GRU. Thereafter, the highly correlated features obtained from Bi-LSTM with attention are fed to the capsule network and to output layer with sigmoid for final classification. The results are shown in figure 4.7.

- **Demonstrating the effect of exchanging CapsNet with CNN**

Similarly, another variant of ablation that we demonstrated in our study includes the usage of CNN instead of capsule network. Again, we trained and tested the methods on both the datasets with this ensemble. Pertinent literature reports capsule networks to have a slightly complex architecture in comparison to other deep learning methods such as CNNs. The results are shown in figure 4.7.

Table 4.5 shows the comparative analysis with ablation architectures.

**Table 4.5.** Ablation architectures

| Datasets | Bi-LSTM+ Attention +CapsNet | Bi-GRU+ Attention +CNN | Bi-GRU+ Attention + CapsNet |
|---|---|---|---|
| **Formspring.me** | 92.67 | 91.83 | **94.03** |
| **MySpace** | 93.10 | 92.35 | **93.89** |

Figure 4.7 shows the comparative analysis of various deep architectures with the proposed model on the basis of their performance when applied to MySpace and Formspring.me datasets.

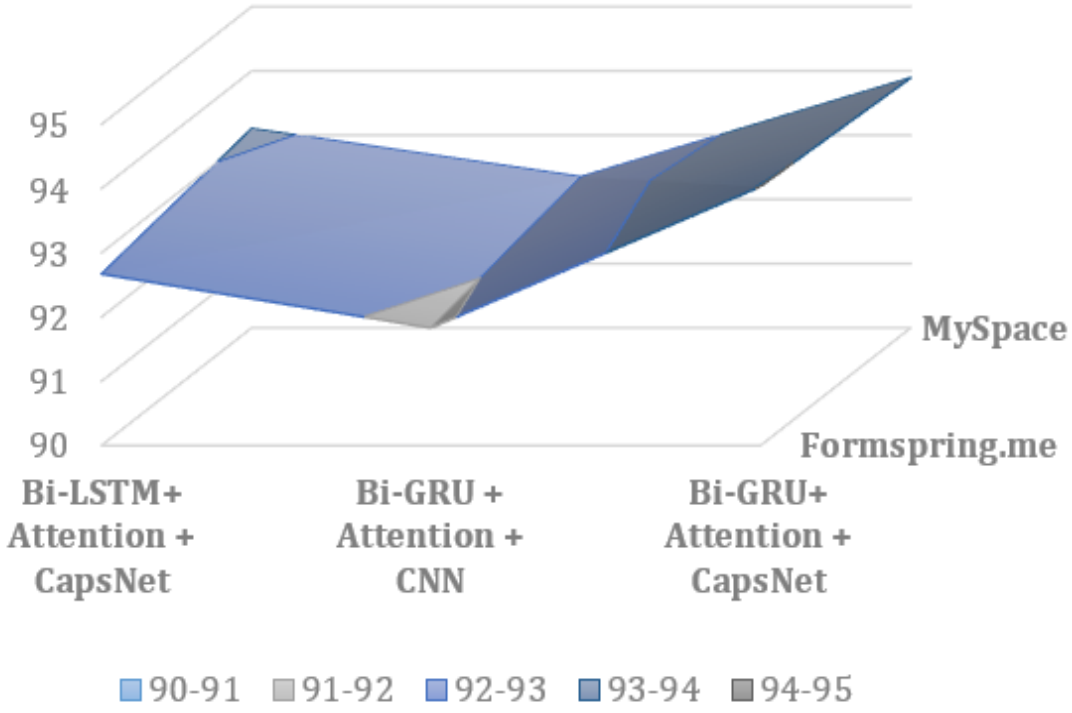


**Fig.4.7.** Comparative analysis of deep models for MySpace & Formspring.me

Having performed the ablation study, we observed that the highest F1-score was obtained with our proposed deep learning-based hybrid model. While, we could notice that there is slight variation in the accuracies obtained by Bi-LSTM and Bi-GRU, however Bi-GRU outperformed Bi-LSTM. Consequently, we could deduce that having used Bi-GRU with attention enhances the F1-score as well as reduces the complexity of the model. Similarly, in the case of CNN, we noticed that although it reduces feature dimensionality considerably, the correlation b/w the words and the final classification were not same for all the input words. Subsequently, this hampers the understanding of the contextual information. Also, CNN ensemble reports lower F1-score as compared to Bi-GAC.

## 4.4 Chapter Summary

This chapter discusses the proposed hybrid deep learning model for cyberbullying classification task which combines the advantages of self-attention-based Bi-GRU encoder and capsule network for cyberbullying detection of mono-lingual textual data in social media. ELMo contextual embeddings are used as input. Automatic detection of cyberbullying on social media has become essential & this chapter extends an attention-based deep model for dealing with real-time textual messages or posts on social media using deep neural architectures. The uniqueness of this proposed model is that it competently produces enhanced predictive results. The results are validated on the benchmark datasets taken from social media namely Formspring.me and MySpace. The proposed model achieved a superior F1-score of 94.03 and 93.89 for formspring.me and MySpace benchmark cyberbullying datasets respectively.

**Publication:**

- Kumar, A., Sachdeva, N. (2021). "A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media". World Wide Web, *Springer*. https://doi.org/10.1007/s11280-021-00920-4. **[SCIE JOURNAL, IMPACT FACTOR: 2.892]**

# Chapter 5

# Cyberbullying Detection for Multimodal Data

The misuse of social networks by netizens to embarrass, mock, defame and disparage a victim without any direct contact is known as cyberbullying. Social networks emerged as "virtual playground" which is utilized by bullies for doing antisocial activities. It is important to develop models for automatic detection and prevention of bullying content as it can turn out to a societal outbreak. Researchers worldwide have been trying to develop new ways to detect cyberbullying, manage it and reduce its prevalence on social media. Advanced analytical methods and computational models for efficient processing, analysis and modelling for detecting such bitter, taunting, abusive or negative content in images, memes or text messages are imperative. More recently, as memes, online videos and other image-based, inter-textual content have become customary in social feeds; typo-graphic and info-graphic visual content (figure 5.1) have become a considerable element of social data [117, 118].


**Fig. 5.1.** Types of visual content

Cyberbullying through varied content modalities is very common. Social media specificity, topic dependence and variety in hand-crafted features currently define the bottlenecks in detecting online bullying posts [119]. Deep learning methods are proving useful and obtaining state-of-the-art results for various natural language tasks with end-to-end training and representation learning capabilities [120, 121, 122]. Pertinent studies report the use of deep learning models like CNN, RNN and semantic image features for bullying content detection by analysing textual, image based and user features [119, 123]. But most of the research on online cyber-aggression, harassment detection and toxicity has been limited to text-based analytics. Few related studies have

also re-counted analysis of images to determine bullying content but the domain of visual text which combines both text and image has been least explored in literature. The combination can be observed in two variants: typo-graphic (artistic way of text representation) or info-graphic (text embedded along with an image). This chapter presents a deep neural model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image).

The primary contribution of this research is:

- Building all-in-one hybrid deep architecture, ***CapsNet-ConvNet***, consists of CapsNet with dynamic routing for predicting the textual bullying content and ConvNet for predicting the visual bullying content.
- The info-graphic content is discretized by separating text from the image using Google Lens of Google Photos App[3].
- The processing of textual and visual components is carried out using the hybrid architecture and the late-fusion decision layer is then used to output the final prediction.
- The performance of CapsNet-ConvNet is validated on 10000 comments and posts (text, image, and info-graphic) prepared using three social media sites YouTube, Instagram and Twitter.

This unifying model thus considers modalities of content and processes each modality type using deep neural learning techniques for efficient decision support for cyberbullying detection. Experimental evaluation was done on a mix-modal dataset which contains 10000 comments and posts scrapped from YouTube, Instagram and Twitter. The modalities within the dataset were 60% textual, 20% visual and 20% info-graphic. We performed 10-fold cross validation and calculated the AUC-ROC curve. The methodology and findings related to this research objective is presented in this chapter. A brief summary of the above study will end the chapter.

## 5.1 Methodology

The methodology adopted is shown by the proposed CapsNet-ConvNet model. In this research, we propose a deep neural model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image). The proposed deep neural model comprehends the complexities of natural language and deals with different modalities of data in online social media content where the representations of data in different forms, such as text and image, is learned as real-valued vectors. In addition to text, we examined the image as well as utilized the info-graphic property of the image (information which is the content/text embedded on that picture) to predict bullying content. The proposed CapsNet-ConvNet model consists of four modules, namely, modality discretization module, textual

[3] https://photos.google.com/

processing module, visual processing module, and prediction module (as shown in figure 5.2).

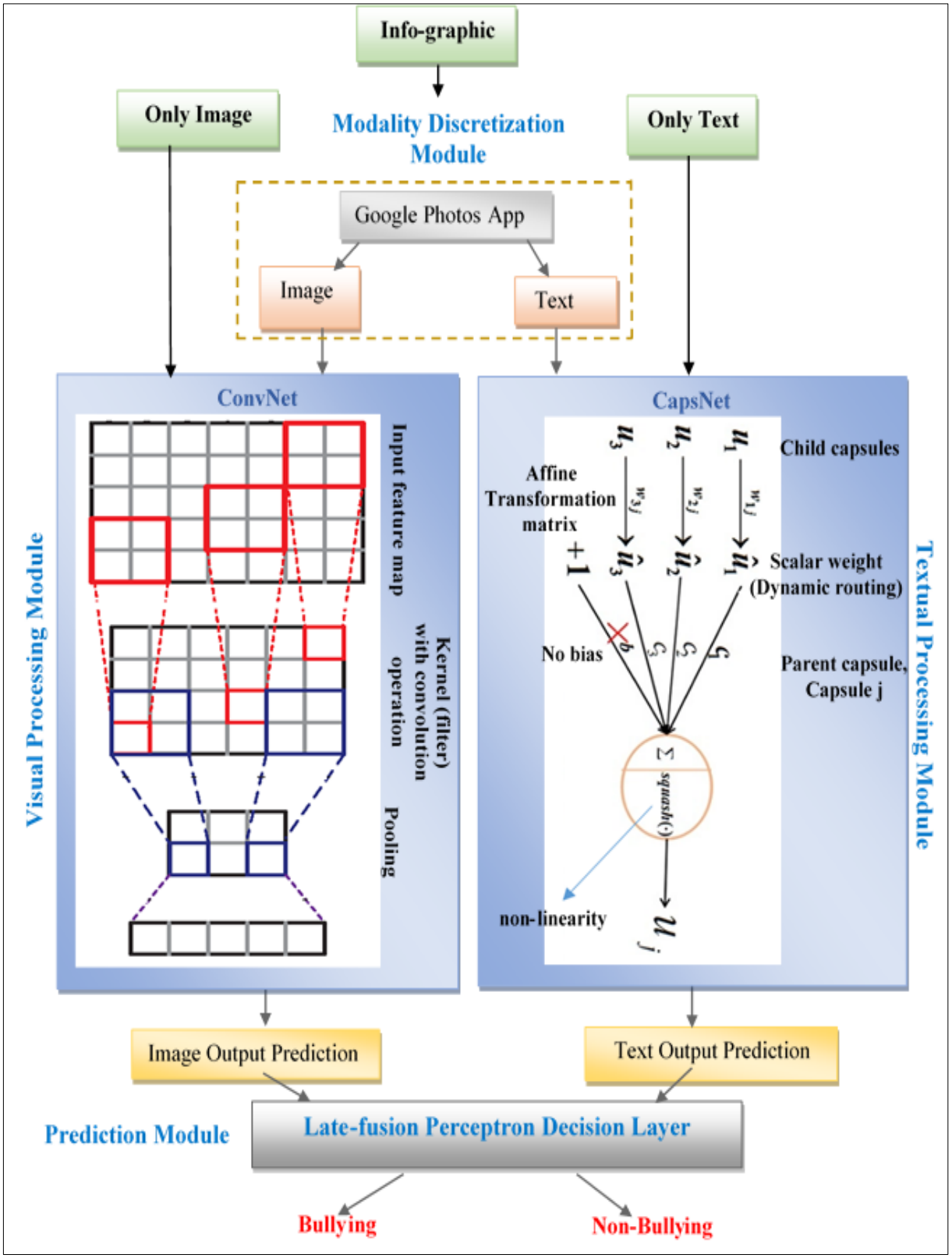


**Fig. 5.2.** The proposed CapsNet-ConvNet model

Depending on the input modality, that is text only or image only, the content is forwarded to the respective processing modules. If the input is an info-graphic post/ comment, which is the image with text embedded on it, the CapsNet-ConvNet model utilizes a Google Photos App to extract text from an image. This visual analysis tool separates the text from the image and sends it to the respective textual processing and visual processing modules for analysis. The Google Lens feature has the ability to recognize texts in the images recorded utilizing the Optical Character Recognition (OCR). The Google Cloud's Vision API offers powerful pre-trained machine learning models which can detect and extract text from images. There are *two* annotation features that support OCR, namely TEXT_DETECTION that detects and extracts text from any image and DOCUMENT_TEXT_DETECTION which extracts text from an image, but the response is optimized for dense text and documents.

### 5.1.1 Textual Processing

Textual processing module is implemented using CapsNet with dynamic routing. CapsNet belongs to the class of deep neural networks consisting of a set of capsules [124]. These are further composed of groups of neurons arranged in a layer and do the actual internal computations in order to predict instantiation parameters of any feature, such as orientation, color etc. at any given location. Pertinent literature reports the use of many routing techniques for text classification such as dynamic, attention based, clustering, static, where dynamic routing reported major applicability.

The embedding layer of a neural network converts an input from a sparse representation into a distributed or dense representation. In this research, we use the state-of-the-art pre-trained ELMo 5.5B word embeddings [125] to generate the word vectors. We preferred ELMo over the conventional embedding models such as Word2Vec or GloVe, as ELMo offers contextualized word representations, which essentially means that the representation for each word depends on the entire context in which it is used. The same word can have two different vector representations based on different contexts. ELMo creates vectors on-the-go by passing words through the deep learning model rather than having a dictionary of words and their corresponding vectors, as is the case with traditional word embedding models. Also, ELMo representations are purely character-based, which allows the network to form representations for words that are not seen in training. All this motivated us to use the ELMo 5.5B model for implementing the embedding layer.

Encoding layer, thereafter, reshapes the word vector matrix into feature vectors of single dimension, where this encoding layer is executed as a capsule network. This network comprises convolution, primary caps and class caps layers. Here, the scalar outputs of each convolution layer are fed as input to the primary caps layer that generates capsules. It must be noted that the output of a capsule is a vector that exhibits the object's

existence whereas, the vector's orientation represents the object's properties. The vector is an input to all the possible parents in the network.

These capsules work towards detecting the parts of the object under consideration in order to associate the random parts of the object to the whole. To accomplish this, CapsNet uses a nonlinear-dynamic routing algorithm in order to capture the capsules part-whole relationship dynamics. Thus, ensuring that the output of the capsule is sent to the possible and relevant parent. Lower-level capsule vectors are multiplied with weight matrices in order to encode spatial and other relationships between features of lower and higher-level using equation 5.1.

$$u_{j|i} = W_{ij} u_i \qquad (5.1)$$

Where 'i' is low level capsule, 'j' is high level capsule and $W_{ij}$ is the translation matrix

Lower-level capsule knows which upper-level capsule accommodates its results in an efficient way and therefore adjusts its coupling coefficient. Thus, previous step output is multiplied with coupling coefficients using equation 5.2.

$$s_j = \sum_i c_{ij} \, u_{j|i} \qquad (5.2)$$

Where $c_{ij}$ is coupling coefficient and $u_{j|i}$ is the output vector from equation 5.1.

Post this, squashing is applied for normalizing the length of each capsule's output vector in the range of [0, 1] using equation 5.3.

$$v_j = \frac{||s_j||^2}{1+||s_j||^2} \frac{s_j}{||s_j||} \qquad (5.3)$$

## 5.1.2 Visual Processing

Next module is visual processing which is used to analyse visual bullying content via ConvNet. A ConvNet is a deep neural architecture [126] which works using multiple copies of the same neuron in different places. It has the power of self-tuning and learning skills by generalizing from the training data. The visual processing is shown in figure 5.3.

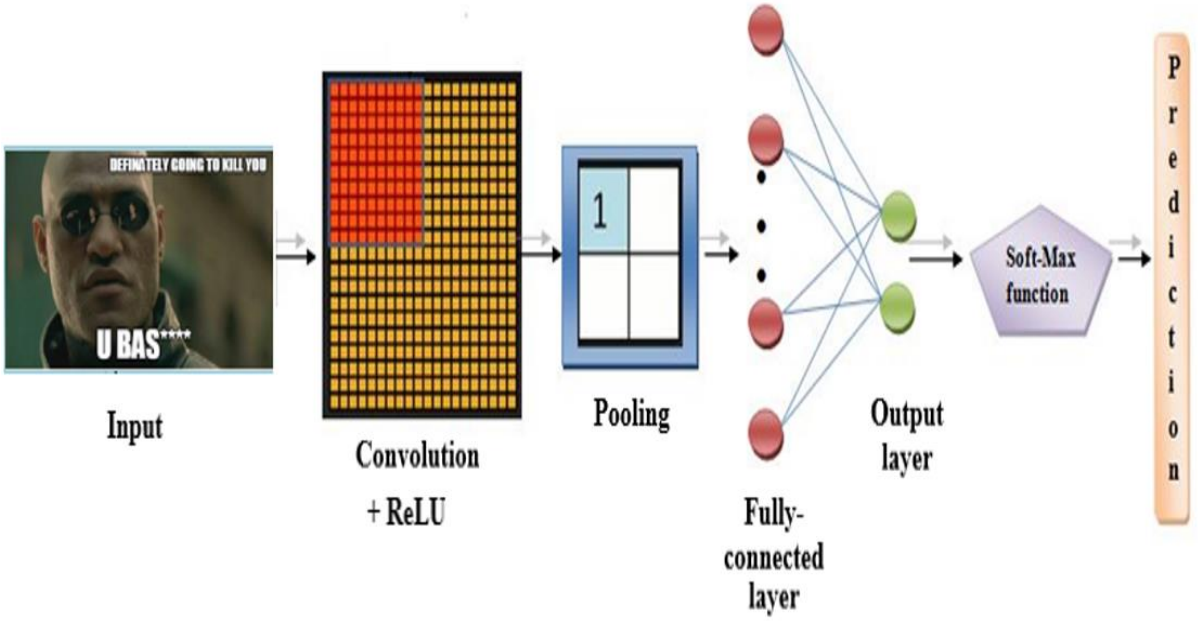


**Fig. 5.3.** Visual Processing module

A ConvNet convolves learned features with input data and uses 2D convolutional layers. It usually consists of several convolutional networks with filters (kernels) in combination with non-linear and pooling layers [127]. The image is passed through convolution layers such that the output of the primary layer becomes the input for the subsequent layer. Convolution is a linear operation but images are non-linear. Therefore, non-linearity is added post every convolution operation using an activation function such as ReLU, Leaky_ReLU, tanh or sigmoid. Each non-linear layer is followed by a pooling layer which reduces the spatial size of the image and performs a down sampling operation. Pooling operation thus helps to progressively reduce the size of the input representation and control overfitting too. We can either use max, average or sum pooling. A fully connected layer is then attached to this series of convolution, non-linear and pooling layers which outputs the information from the convolutional networks. The working of a typical ConvNet is shown in figure 5.4.

In this research, the visual processing module has three convolutional layers followed by three max-pooling layers to extract the features of images, a flatten layer which takes the output from the previous max-pooling layer and convert it to a 1D array such that it can be feed into the dense layers and to the output layer for prediction.

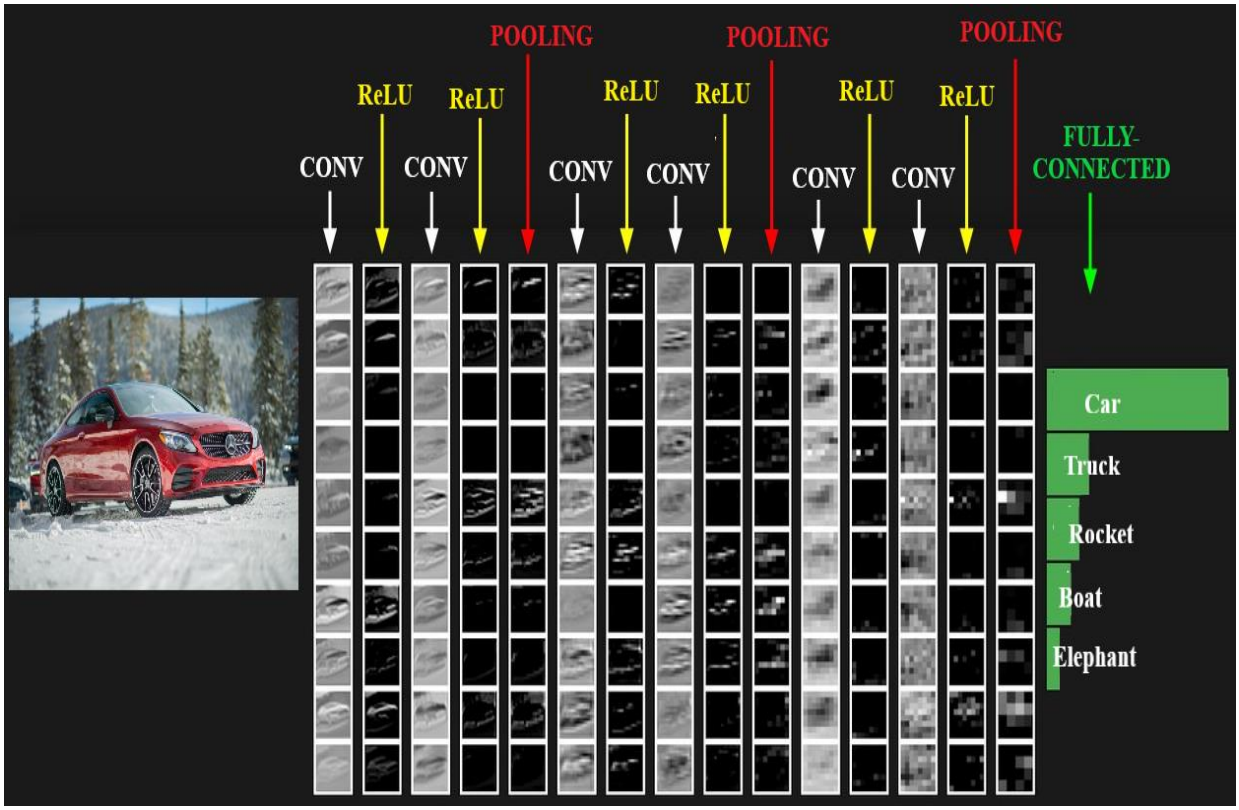


**Fig. 5.4.** Working of a typical ConvNet

The details of the layers are as follows:

**Convolution Layer:** The convolution layer transforms the input image to extract the features. This is done by convolving the image with a filter (kernel) which is specialized to extract certain features. Mathematically, the convolution operation (a.k.a. scalar product) is the summation of the element-wise product of two matrices (filter-sized patch of the input and filter) which results in a single value.

**Activation layer and Pooling layer:** The activation (ReLU) layer is intended to introduce non-linearity to the system and produces a rectified feature map which is inserted into the pooling layer where a max-pooling operation is applied to each convolution $c_{max} = max\ (c)$. The max-pooling operation extracts the 'k' most important features for each convolution. The output of the final convolution layer, that is, the pooled feature map is a representation of the original image.

**Fully Connected layer:** A fully connected neural network is a feed forward network that will have the feature vector of n dimension obtained after concatenating every $c_i$ obtained by the application of n filters. Now we train the network using back-propagation algorithm. Gradients are back propagated and when we reach at the convergence, we finally stop the algorithm. A softmax function is used to classify the post as bullying (+1) or non-bullying (-1).

### 5.1.3 Prediction

The perceptron-based decision-level late fusion strategy for multimodal learning is used to dynamically combine the predictions of discrete modalities and output the final category as bullying or non-bullying type. Thus, the final prediction is usually done using an additional decision layer implementing multimodal classification fusion. Typically, there are two strategies to multimodal fusion: model-free and model-level. Model-free fusion can be further classified into early fusion (feature-level) and late fusion (decision-level). In early fusion, the different types of input features are firstly concatenated and then fed into a classifier, whereas in late fusion, the predictions of different classifiers trained for distinct input types are combined to provide us with the final output. Model-level fusion combines the advantages of both of these strategies by concatenating high-level feature representations from different classifiers. In this work, late fusion strategy for multimodal learning is used, that is, the bullying content prediction of mono-modalities (text and image separately) is done by the respective classification models. Late fusion allows the use of different models on different modalities, thus allowing more flexibility. It is easier to handle a missing modality as the predictions are made separately. The class probabilities are thus fused together to join information from the two modalities to perform a final prediction task.

## 5.2 Dataset

The dataset prepared for the experiment contained 10000 comments and posts (text, image, and info-graphic) prepared using three social media platforms namely YouTube, Instagram and Twitter. The modalities within the dataset were 60% textual, 20% visual and 20% info-graphic (figure 5.5).

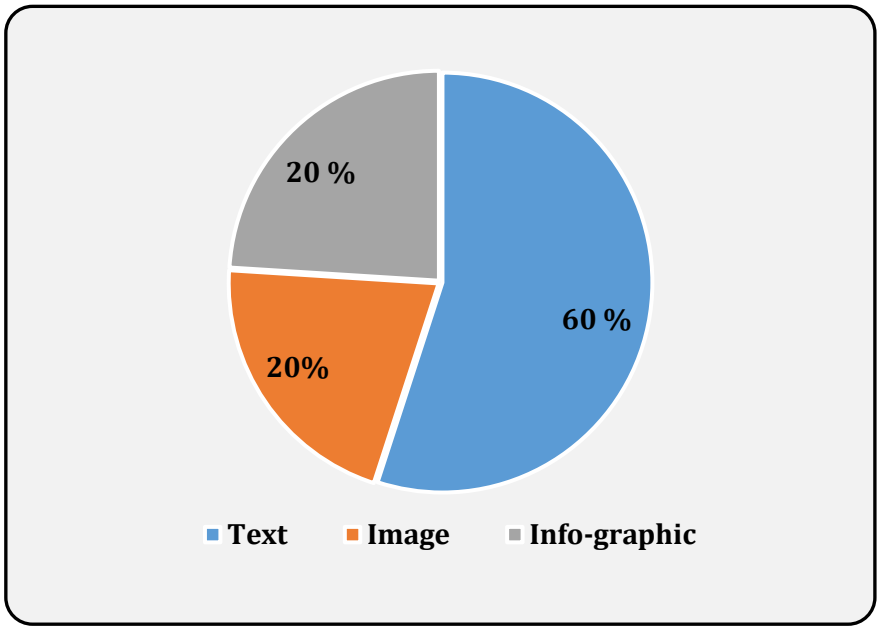


**Fig. 5.5.** Modality distribution in dataset

Table 5.1 below shows the actual distribution of data in numbers.

**Table 5.1.** Data Categorization

| Type of modality | Number of instances | |
|---|---|---|
| | **Bullying** | **Non-bullying** |
| Image only | 1260 | 740 |
| Text only | 3000 | 3000 |
| Info-graphic | 1440 | 560 |

We performed 10-fold cross validation and calculated the AUC-ROC curve. We used the Scikit-learn library and Keras deep learning library with Theano backend.

## 5.3 Findings

The findings are as follows.

### 5.3.1 Model Performance

In this proposed model, the perceptron-based decision-level late fusion strategy for multimodal learning was used to dynamically combine the predictions of discrete modalities and output the final category as bullying or non-bullying type. The performance of CapsNet-ConvNet was validated on 10000 comments and posts (text, image, and info-graphic) scrapped from YouTube, Instagram and Twitter. The proposed model achieved a superlative performance with the AUC-ROC of 0.98 (as shown in figure 5.6). This unifying model considers modalities of content and processes each modality type using a deep neural learning techniques for an efficient decision support for cyberbullying detection. The uniqueness of the proposed hybrid deep learning model, CapsNet-ConvNet is that it deals with different modalities of content, namely, textual, visual (image) and info-graphic (text with image).

### 5.3.2 Ablation Study

In order to assess the performance of the proposed model, we performed the ablation study (figure 5.7) as well where we interchanged the roles of a few of the deep architectures and then tested the performance of the model. We reversed the hybrid by using a ConvNet for the textual processing module and the CapsNet for the visual processing module and it was observed that the original set-up achieved superlative results.

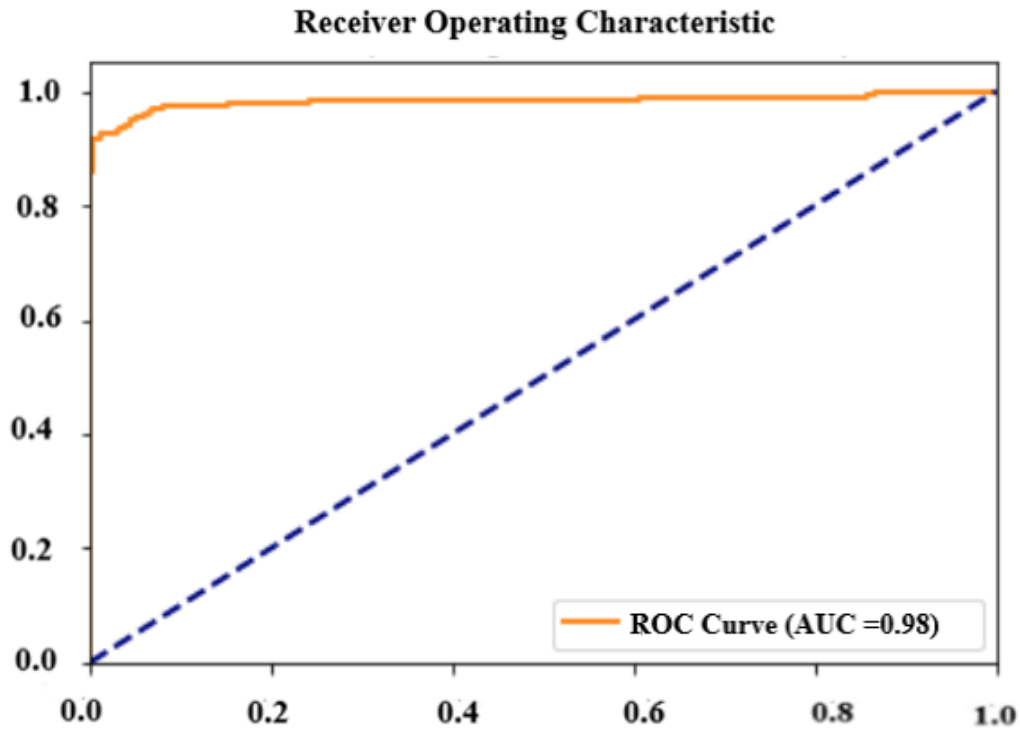


**Fig. 5.6.** Performance of CapsNet-ConvNet Model

The ROC-AUC for this variation is shown in figure 5.7.

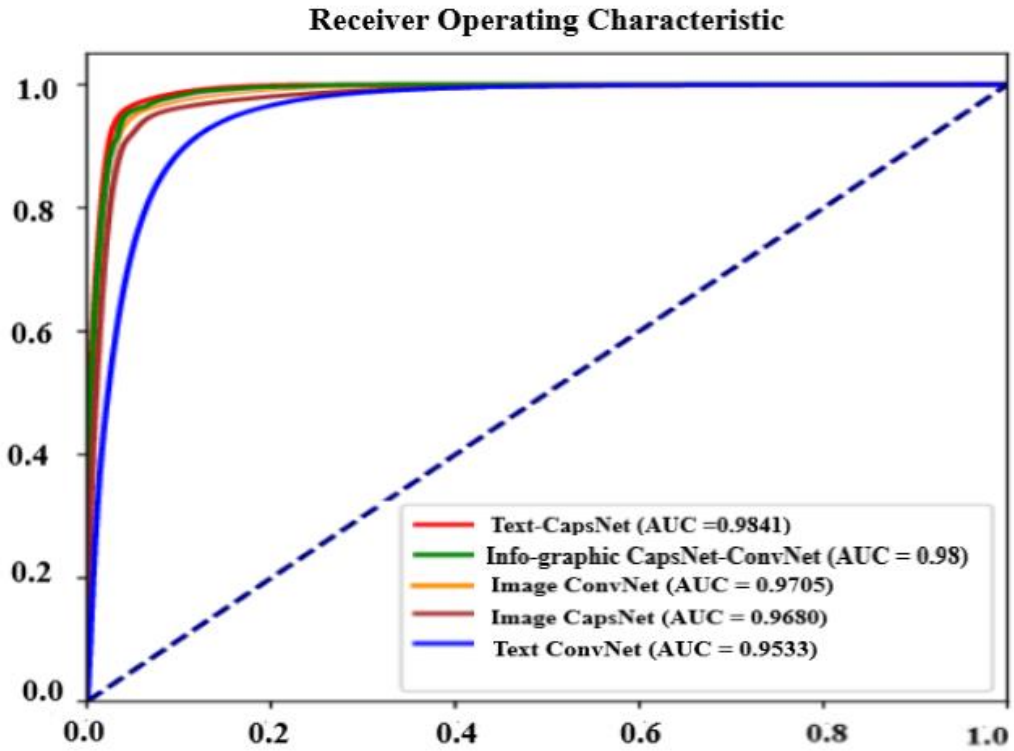


**Fig. 5.7.** Results for ablation study

Furthermore, in order to evaluate the performance of the model, comparative analysis of classifiers was also done using supervised machine learning techniques for image modality. Three machine learning classifiers, namely, K- nearest neighbour (K-NN) and Naïve Bayesian (NB) and support vector machine (SVM) were compared with deep neural ConvNet image classifier. The Bag-of-Visual words (BoVW) approach was used to extract the features and train the three machine learning classifiers. It was observed that the ConvNet outperformed the other classifiers. Comparative analysis of the image classification algorithms is shown via figure 5.8.

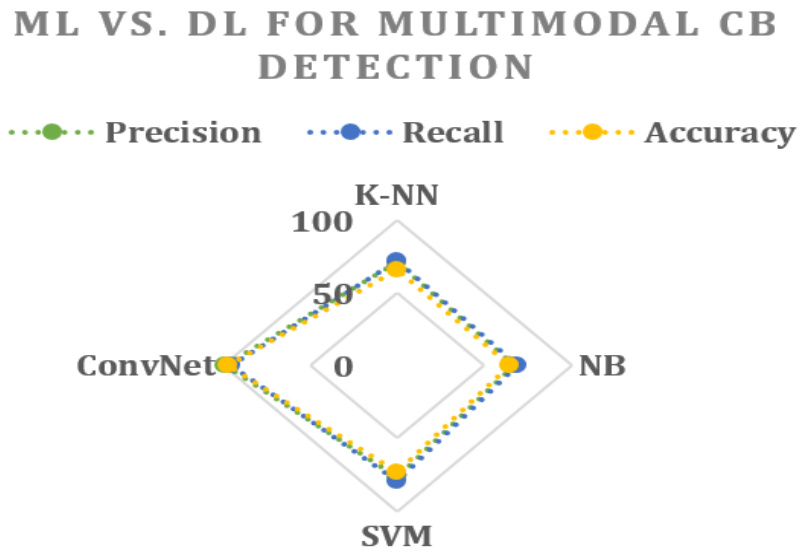


**Fig. 5.8.** Comparative analysis of different classifiers used for image modality

## 5.4 Chapter Summary

This chapter discusses the proposed model based on deep learning, CapsNet-ConvNet for CB detection in social media that deals with different modalities of content, namely, textual, visual (image) and info-graphic (text with image). The proposed deep neural model comprehends the complexities of natural language and deals with different modalities of data in online social media content where the representations of data in different forms, such as text and image, is learned as real-valued vectors. In addition to text, we examine the image as well as utilize the info-graphic property of the image (information which is the content/text embedded on that picture) to predict bullying content. It was observed that the CapsNet-ConvNet model achieved superlative performance with the AUC-ROC of 0.98.

**Publication:**

- Kumar, A.* & Sachdeva, N. (2021). "Multimodal Cyberbullying Detection Using Capsule Network with Dynamic Routing and Deep Convolutional Neural Network", *Multimedia Systems, Springer.* https://doi.org/10.1007/s00530-020-00747-5 **[SCIE-Impact Factor: 1.563]** ISSN: 1432-1882

# Chapter 6
# Cyberbullying Detection for Mash-up Data

Automatic cyberbullying detection in social networks data is a generic text classification and natural language understanding task. The traditional diversities, country-specific topics, hash-tags, the unorthodox use of typographical styles such as punctuation, capitals, emojis and easy accessibility of local language keyboards add up to the volume and variety of user-generated content aggravating the linguistic challenges [128]. This research focused on code-mix cyberbullying detection, especially the Hinglish, which points to the juxtaposition of Hindi and English language words. We explored cyberbullying detection problem and proposed ***MIIL-DNN***, a multi-input integrative learning deep neural network model. MIIL-DNN merge learning's from three sub-networks to identify and classify cyberbullying in code-mix data. Three inputs viz. English language features, Hindi language features and typographic features are learned individually utilizing sub-networks (CapsNet for English language, Bi-LSTM for Hindi language and MLP for typographic features). Subsequently, these are merged into one unified representation which is used as an input to a fully connected network for final prediction. Model-level fusion for multi-lingual data has an advantage that, for each input type, it works with the unique distribution without increasing the input space dimensionality. Validation of the model is done on two datasets which are created by scraping data from social media sites viz. Facebook and Twitter.

As a typical natural language text classification task, automatic detection of bully content depends on feature engineering and learning model. Social media [129] has created a new 'text-speak' genre of language which is more direct or casual or polemical. Shorthand English has become a social norm and is full of abbreviations, hashtags, emojis and new-fangled uses of punctuation. It consists of some novel words (such as selfie), wordplay (greaaatttttttt for great), neologisms (l8r for later), and Internet slangs (TTLY for talk to you later). While English dominates this shortened text-speak, a vast amount of static and dynamic web content is continuously generated by non-native writers. Multi-linguality is a commonly observed phenomenon [130]. All these multiple inputs make the task of automated text analytics computationally intensive.

A critical challenge is to find techniques for multilingual input-type fusion which can either be done at an early or a later stage (early fusion, late fusion) or at a model-level. While early fusion takes a combined representation to train the network, in late fusion, the features of each language are examined and classified independently and the results are fused as a decision vector to obtain the final decision. Early fusion suffers because it increases the dimensionality of the input data without considering the unique distribution of each input type and further demands normalization to avoid giving added weight to the input type with more dimensions. Though late or decision level fusion is

easy as compared to early feature fusion and facilitates the use of the best suitable classifier or model to learn its features, it significantly isolates interactions among different features. Moreover, as different classifiers are used for the analysis task, the learning process of all these classifiers at the decision-level fusion stage becomes tedious and time-consuming. We propose a medial fusion strategy, that is, the model-level fusion which resolves the cons of both early and late fusion. It exploits correlation in data as different sub-networks are used to operate over features which are learned separately for each input type and then combined into one unified representation.

Text classification in a multilingual code-mix input can either be done by translating the input into a mono-lingual dataset (English only) or by using a language-dependent method (English and Hinglish) without translation. The translation method has a serious shortcoming as it may cause an ambiguity or failure of the translation resulting in wrong semantics and feature vector generation used to train the model. For example, the English translation of the Hindi transliterated text "yeh ladki ekdum chaalu hai'' is wrongly translated to "This girl is on the move" (figure 6.1).

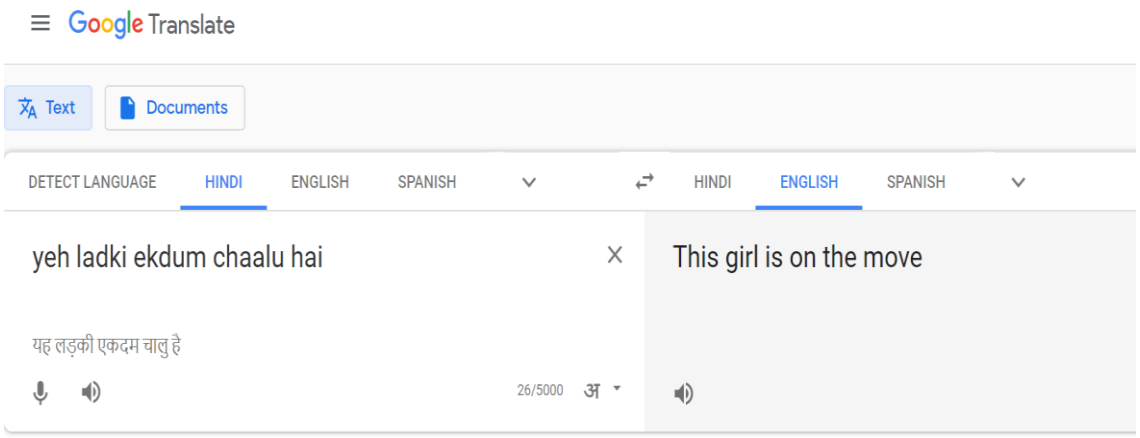


**Fig.6.1**. Example of translation ambiguity

On the other hand, the language-dependent model requires a large labelled training dataset for every new language, which is a computationally expensive job. Therefore, to bridge the limitation between translated method and language-dependent method, we use the Google Transliteration – Translator toolkit, such that word-level transliteration is done to convert Hinglish text to Hindi. This also enables to capture the right textual interpretation, for example, the correct transliteration to Hindi for "yeh ladki ekdum chaalu hai'' is "यह लड़की एकदम चालु है".

The hierarchical learning capabilities and generalization offered by deep learning architectures have made them a popular choice within natural language text processing [131]. The most sophisticated bullying classification methods are trained on general corpora with vast amounts of labelled data which are not suitable to a code-mix data (English and a low-resource language like Hindi). Transfer learning methods look like a

promising solution to this challenge of the scarcity of labelled data. In transfer learning, we first train a base network on a base dataset and task, and then we repurpose the learned features, or transfer them, to a second target network to be trained on a target dataset and task. The core idea behind these models is that by training language models on very large corpora and then initializing down-stream models with the weights learned from the language modelling task, a much better performance can be achieved. The initialized layers can range from the single word embedding layer to the whole model.

The methodology and findings related to this research is presented in this chapter. A brief summary of the above study will end the chapter.

## 6.1 Methodology

The proposed integrative learning network, ***MIIL-DNN*** combines information from *three sub-networks trained using three-inputs, namely English, Hindi and typographic* respectively. We used transfer learning by fine-tuning the pre-trained word embeddings (*GloVe for English and fastText for Hindi*) for the domain-specific words to increase the size of the training dataset. These three sub-networks include *Capsule Network with dynamic routing* [132] sub-network to generate English semantic context vectors using pre-trained GloVe embeddings. Hindi *Bi-directional LSTM* sub-network used to generate the feature vector using pre-trained word embedding for the Hindi language provided by fastText and Typographic feature sub-network where *MLP* is used to operate over typographic input data.

Subsequently, a model-level feature fusion of sub-network outputs is done to generate the output class. That is, these sub-networks are then concatenated together to form the final multi-input integrative learning model which generates a output with linear activation. Figure 6.2 depicts the architecture of the proposed MIIL-DNN network. Characteristically, MIIL-DNN is the foremost model-level feature fusion deep neural architecture for code-mix data which also uses transfer learning to increase the size of the training dataset. The performance of the model is validated on two datasets taken from the popular social networking sites namely, Twitter and Facebook.

.

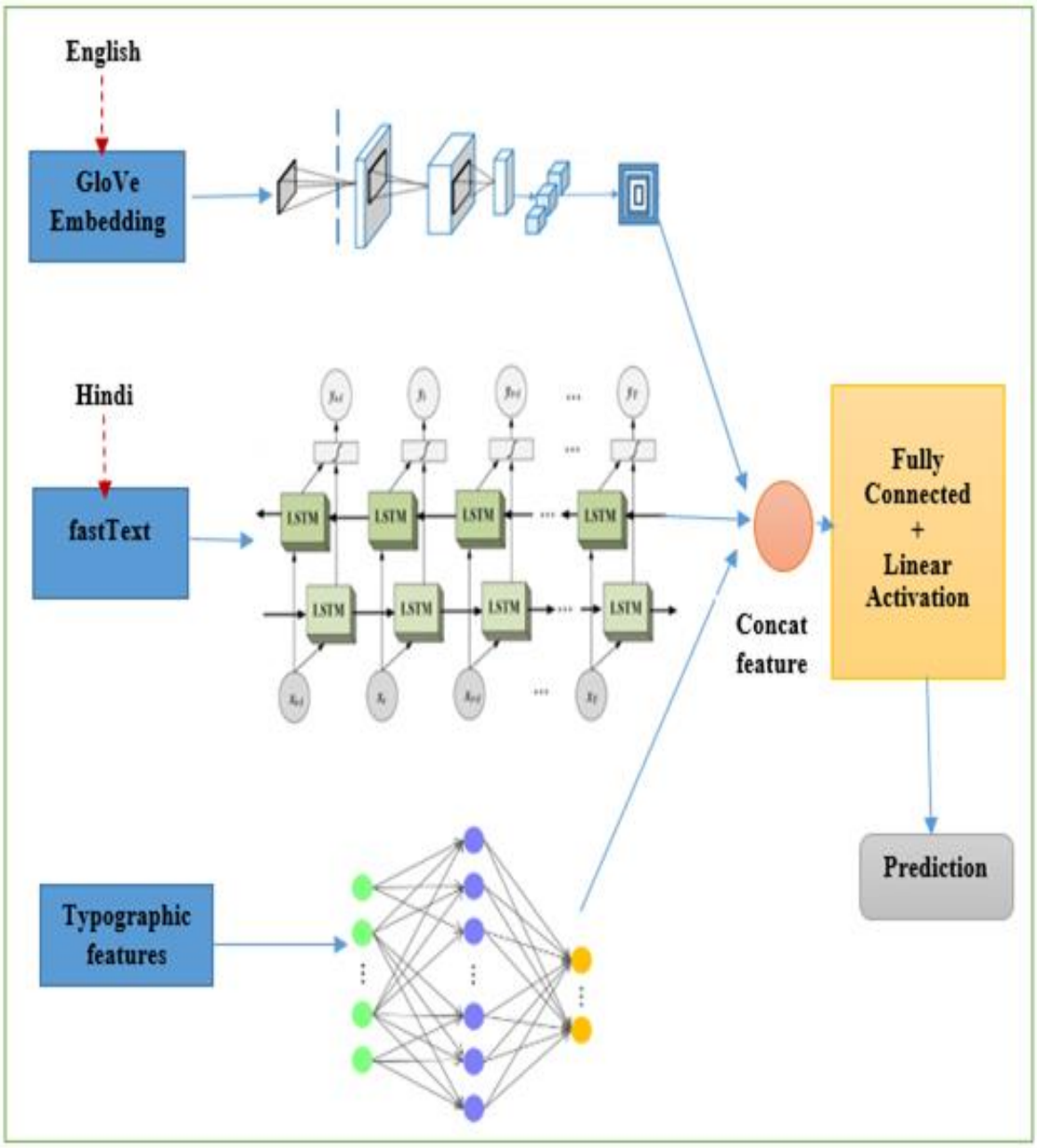


**Fig.6.2**. Proposed MIIL-DNN model

### 6.1.1 Data Pre-processing & Feature Extraction

The primary intent of pre-processing was to clean, prepare and transform the data for the extraction of features [133]. The process included:

- Removing tags, numbers, URLs and stop words.
- Spell check, lemmatization and stemming
- The tokens are converted to lowercase.

- Substituting slangs and emojis using the SMS Dictionary[4] and emojipedia[5] respectively.
- Punctuations are usually discarded during data pre-processing phase but in casual or informal writing such as text message or online posts, these are used as a technique to add emphasis to written text. Therefore, the count of each punctuation mark (!, ?, ., capitalization, 'x', "x") is extracted as a typographic feature set to train the model [134].

Tokenization [135] of Facebook posts and tweets were then done using the TreebankWordTokenizer of Python Natural Language Toolkit (NLTK)[6]. Subsequently, language transformation is done to decode the Hinglish language tokens using transliteration into Hindi. Transliteration is the conversion of text written in one script (language) into text written in another script (language), while maintaining the pronunciation to the greatest possible extent [136]. There is no change in grammar or meaning. Unlike translation which tells the meaning in the target language, transliteration is based on the pronunciation in the target language, and not on the meaning. For example, for the Hindi phrase 'मुझे उसका तरीका बिलकुल अच्छा नहीं लगता', its translation in English would be 'I don't like her way' and 'Mujhe uska tareeka bilkul achha nahi lagta' is the transliterated Hindi. We use the Google Transliteration – Translator toolkit to implement this language transformation module where the transliterated Hindi text is converted to the Hindi language.

Manual feature extraction is computationally expensive [137] whereas feature learning techniques such as word embeddings enable vector representations of a word in a vector space where words sharing certain semantic or syntactic relationships exist in close vicinity of each other. Such knowledge allows us to do away with manual feature engineering required to gain semantic and local contextual insight. Subsequently in transfer learning, the embedding layer is initialized using third-party embeddings such as GloVe, Word2Vec or fastText and the semantic information between words that was learned during the embedding process is transferred.

In this work, we used the GloVe pre-trained embedding and the fastText pre-trained embedding to initialize the English and Hindi sub-networks respectively. The count-based GloVe embedding is used to seed the sub-network for the English language feature vector generation [138]. This feature vector is given as input to the CapsNet. A capsule is trained to specify the features of the object and its likelihood. Thus, the objective of the capsule is not just feature detection but also to train the model to capture the context features. Similarly, we use a pre-trained word embedding for the

[4] SMS Dictionary. Vodacom Messaging. Retrieved 16 March 2012.

[5] https://emojipedia.org/

[6] https://www.nltk.org/

Hindi language provided by fastText to train a bidirectional LSTM sub-network such that word features H = ($h_1$, $h_2$, ..., $h_n$) are concatenated from both directions.

Additionally, punctuations such as exclamation mark, quotation marks, capitalization add emphasis in written informal text and are significant signs which assist to comprehend the context inconsistency or intensity within the text [134]. Similarly, target curse words[7] also act as textual indicators and therefore the presence of offensive/ profane words must be included as an important typographic feature.

Thus, the typographic feature vector *t* with six tuples is, <*r, e, p, u, q, c*>, where, *r* is the frequency of recurring alphabetic character, (that is, if recurrence > 2 set *r* =1, else 0) and *e*, *p*, *u* and *q* defines the count of exclamation marks, periods, uppercase letters, single quotes(‘’) or double quotes (“”) respectively and *c* defines the presence of curse word within the text.

The conceptual flow of feature extraction is shown in figure 6.3.

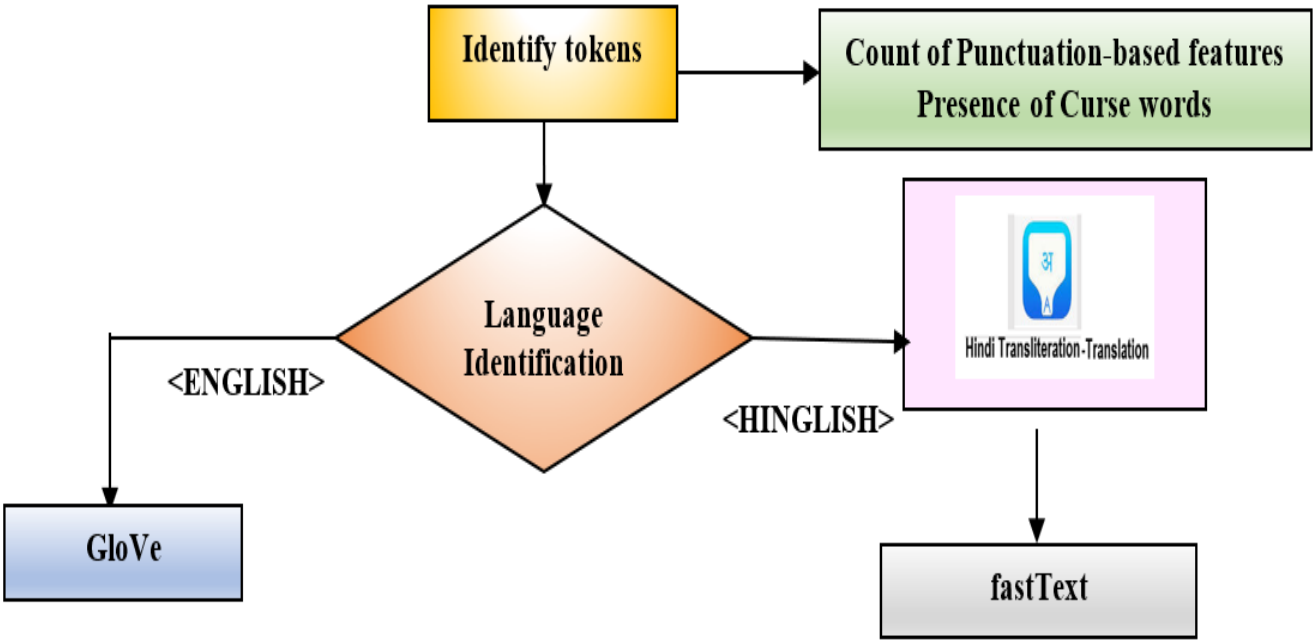


**Fig. 6.3.** Feature extraction in MIIL-DNN

Different deep learning models are then applied, as sub-networks for this multi-input data that is English and Hindi language input mapped to real-valued vectors using pre-trained word embeddings GloVe and fastText respectively and numeric/ categorical pragmatic data. These inputs are fed into the respective sub-networks namely, CapsNet for English, Bi-LSTM for Hindi and Multi-layer perceptron (MLP) for pragmatic to model an integrative learning network which combines information from the sub-networks.

[7] https://www.cs.cmu.edu/~biglou/resources/bad-words.txt

### 6.1.2 English Sub-network

English sub-network comprised of Capsule Network (CapsNet) with dynamic routing for the English input. A capsule network is composed of many capsules [139]. A capsule can be a neuron or set of neurons which output a vector rather than a single value scalar. This vector usually carries additional information that would otherwise be lost by the summation process (max-polling). The key concepts of capsule network include substituting the scalar-output feature detectors of CNNs with vector-output capsules and replacing the max-pooling with "routing-by-agreement." The purpose of the capsule is not only to detect a feature but, also to train the model to learn the variant. Various routing algorithms such as static, dynamic, clustering and attention based have been proposed in the relevant literature on text classification. Most of the work relies on the customary dynamic routing algorithm where basically the capsules 'vote' which capsule to output to [132]. In contrast to CNNs which require training on large datasets, the generalization capabilities of CapsNets on smaller datasets make them competent and conducive for use in various real-life applications. The following subsections explain its details.

**Embedding Layer:** The embedding layer of a neural network converts an input from a sparse representation into a distributed or dense representation. Word Embedding facilitates natural language understanding by means of semantic parsing such that the meaning from text is extracted preserving the contextual similarity of words. In this research, we pre-trained the model on a general dataset using GloVe word embeddings and used transfer learning to train it on the domain-specific problem. The GloVe embedding was pre-trained on Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 200d vectors) data.

**Encoding Layer (Capsule Network):** The matrix of word vectors produced by GloVe is converted into a feature vector of one dimension by the encoding layer. The encoding layer is implemented as a capsule network. The network consists of the convolution layer, the primary caps layer, and the class caps layer such that the outputs from one capsule (child) are routed to capsules in the next layer (parent). The detailed functionality of each layer is explained in the following subsections.

- **Convolution Layer:** Extraction of features from the given input is done by the convolution layer. This layer performs the convolution operation and generates a feature $f_i$ for the given filter as given in the following equation (6.1).

$$f_i = \boldsymbol{\varphi}\left(\sum_{\mathbf{h=1}}^{\mathbf{h}} \sum_{\mathbf{d=1}}^{\mathbf{d}} K_{h,d} X_{i+h,d} + b_i\right) \qquad (6.1)$$

where,
$f_i$ $(i = 1,2,3, \ldots . , n)$ represents the feature produced,
$\varphi$ is ReLU function used or activation,
$K_{h,d} \in R^{h \times d}$ is the filter,
$X_i$ represents the input word vector,
$b_i$ is a bias term

- **Primary Caps Layer:** This low- level layer takes the previous convolution layer scalar output to generate vector outputs called capsules. Capsule networks can be visualized as tree-like representations that learn transformations to associate the parts of an object to the whole. Capsules provide a way to detect parts of objects identifying the child and parent capsules such that the output of the capsule gets sent to an appropriate parent in the layer above. The key question that needs to be answered is which parts belong to which parents. A powerful non-linear dynamic routing captures the part-whole relationship dynamics of the capsules and ensures the output of a capsule is sent to a suitable parent. This ensures that if after applying a transformation to the part, we have the same or a similar feature vector to that of the parent, then we update a parameter for the likelihood that the two capsules are linked as parent/child. Another key issue is that of whether a part actually exists or not. This is determined by the length of the feature vector of a capsule.

  The output of a single capsule $u_i$ is multiplied by a translation matrix $W_{ij}$ to produce a vector $u_{j|i}$.

$$u_{j|i} = W_{ij} u_i \qquad (6.2)$$

  Where capsule 'i' is in the current lower-level primary caps layer whereas capsule 'j' is in the next level layer.

  Using the iterative routing-by-agreement mechanism, a lower-level capsule sends its output to higher-level capsules whose activity vectors have a big scalar product with the prediction coming from the lower-level capsule (equation 6.3).

$$s_j = \sum_i c_{ij} \, u_{j|i} \qquad (6.3)$$

  Where $c_{ij}$ is the coupling coefficient that is calculated using a softmax function during the dynamic routing process (equation 6.4).

$$c_{ij} = \frac{\exp (b_{ij})}{\sum_k \exp (b_{ik})} \qquad (6.4)$$

Where, $i \in [1, a]$ and $j \in [1, k]$ , and k represents the number of classes

$b_{ij} = b_{ij} + u_{j|i}.\, v_j$

$b_{ij}$ is the initial logit (prior probabilities that capsule i should be coupled to capsule j).

That is, till now we have multiplied the output of the previous capsule by weight matrices to encode the spatial relationships, then multiplied them with coupling coefficients to just receive the relevant information from the previous capsules. A non-linear "squashing" function is used to normalize the length of the output vector of each capsule to [0, 1]. Thus, on applying the squashing function $s_{j,}$ the output vector $v_j$ is given as in equation (6.5).

$$v_j = \frac{||s_j||^2}{1+||s_j||^2} \frac{s_j}{||s_j||} \qquad (6.5)$$

- **Class Caps layer**: The output of the lower-level capsule which is a linear combination of different predictions is sent to an appropriate parent in the layer above. Moreover, according to the degree of agreement, as measured by the dot product, between the prediction and the final output of the higher-level capsule, i.e., after the squashing, the routing-by-agreement algorithm increases or decreases the coupling to adjust different contributions of different capsules. After the prediction vectors are calculated, they are linearly summed as in (Equation 6.3) to get the total input of the capsule, which is then squashed as (Equation 6.5) to calculate the output of this capsule. The prediction made by the network after convergence is of course the class with the largest output vector norm. The final class caps layer outputs a vector to represent the existence of the entity. That is, the length of the activation vector characterizes the probability of the existence of the entity. We refer to the normalized outputs from the class caps layer and use them as features for our bully detection classifier. Figure 6.4 summarizes the operations within a capsule.

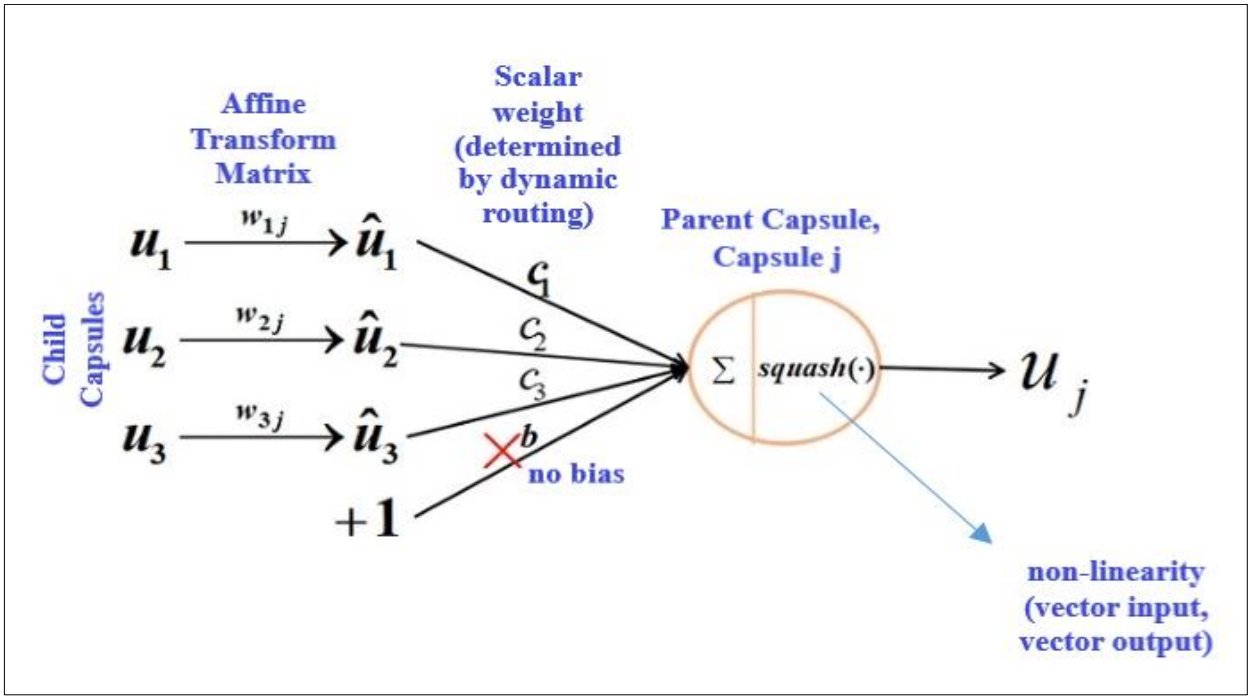


**Fig.6.4.** Operations within a capsule

### 6.1.3 Hindi Sub-network

Hindi sub-network comprises bi-directional Long Short-Term Memory Sub-network for Hindi Input. Training a model on a huge dataset and then re-using the pre-trained model for a target task (transfer learning), can be valuable to low-resource languages such as Hindi, where the amount of labelled data is limited. Here, we used a pre-trained word embedding for Hindi language provided by fastText to train a bidirectional LSTM sub-network such that word features H = ($h_1$, $h_2$, ..., $h_n$) are concatenated from both directions. This model was trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives.

A Long Short-Term Memory (LSTM) [140] is a type of recurrent neural network that addresses the vanishing/exploding gradient problem with RNNs. LSTMs introduce the concept of cell states, which provide "highways" for the gradient to flow backward through time freely, thereby making it more resistant to the vanishing gradient problem. The cell state can be thought of almost like data stored in a computer's memory. LSTMs can "remember" or "forget" information in the cell state by using specialized neurons called "gates".

This way, LSTMs can retain long-term dependencies and connect information from the past to the present. There are three major gates, namely, the forget gate, input gate and the output gate (equation 6.6 to 6.11).

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \quad (6.6)$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (6.7)$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \quad (6.8)$$

$$s_t = tanh\ \left(W_s \cdot [h_{t-1}, x_t] + b_s\right) \quad (6.9)$$

$$c_t = f_t * c_{t-1} + i_t * s_t \quad (6.10)$$

$$h_t = tanh\left(c_t\right) *\ o_t \quad (6.11)$$

where,

- $x_t$ is the *t*-th word vector that it denotes the word representation of $w_t$
- $W_i, W_f, W_o, W_s$ are model parameters
- $b_i, b_f, b_o, b_s$ represents the bias vectors
- $\sigma$ is the sigmoid function used as the gate activation function
- $*$ represents the product (element-wise)
- *tanh* is the hyperbolic tangent function

Bidirectional learning in LSTMs trains two LSTMs to allow the propagation of input in both backward (previous time steps) as well as forward (later time steps) direction in time to make predictions about the current state. This adds past and future context as a bonus to the network and improves the results. We used bidirectional LSTM [140] to obtain word features H = ($h_1$, $h_2$, ..., $h_n$) concatenated from both directions. A forward LSTM processes the sentence (tweet/post) from $x_1$ to $x_n$, while a backward LSTM process from $x_n$ to $x_1$. For word $x_t$, a forward LSTM obtains a word feature as $\overrightarrow{h_t}$ and a backward LSTM obtains the feature as $h_t^{\leftarrow}$ [134]. Then, $h_i$ is calculated using (equation 6.12).

$$h_i = \left(\overrightarrow{h_i} \odot \overleftarrow{h_i}\right) \quad (6.12)$$

Where, $h_i$ is the output of the *i*-th word,

$\odot$ function is a concatenation function. Generally, different merge modes can be used to combine the outcomes of the Bi-LSTM layers. These are concatenation (default), multiplication, average, and sum.

$\vec{h}$ is the forward hidden sequence and $h^{\leftarrow}$ is the backward hidden sequence calculated iteratively for time step from t = T to 1 for the backward layer and t =1 to T for the forward layer.

### 6.1.4 Typographic Sub-network

Typographic sub-network comprised of Multi-layer Perceptron for Pragmatic features. The pragmatic feature vector is trained using a multilayer feed forward neural network which is a special type of *fully-connected* network with multiple single neurons. MLP can be viewed as a logistic regression classifier where the input is first transformed using a learnt non-linear transformation $\Phi$. This transformation projects the input data into a space where it becomes linearly separable. This intermediate layer is referred to as a hidden layer. A single hidden layer is sufficient to make MLPs a universal approximator.

The types of layers in a typical MLP are as follows as shown in figure 6.5.

- ***Input Layer****:* Input variables, sometimes called the visible layer.
- ***Hidden Layers****:* Layers of nodes between the input and output layers. There may be one or more of these layers.
- ***Output Layer****:* A layer of nodes that produce the output variables.

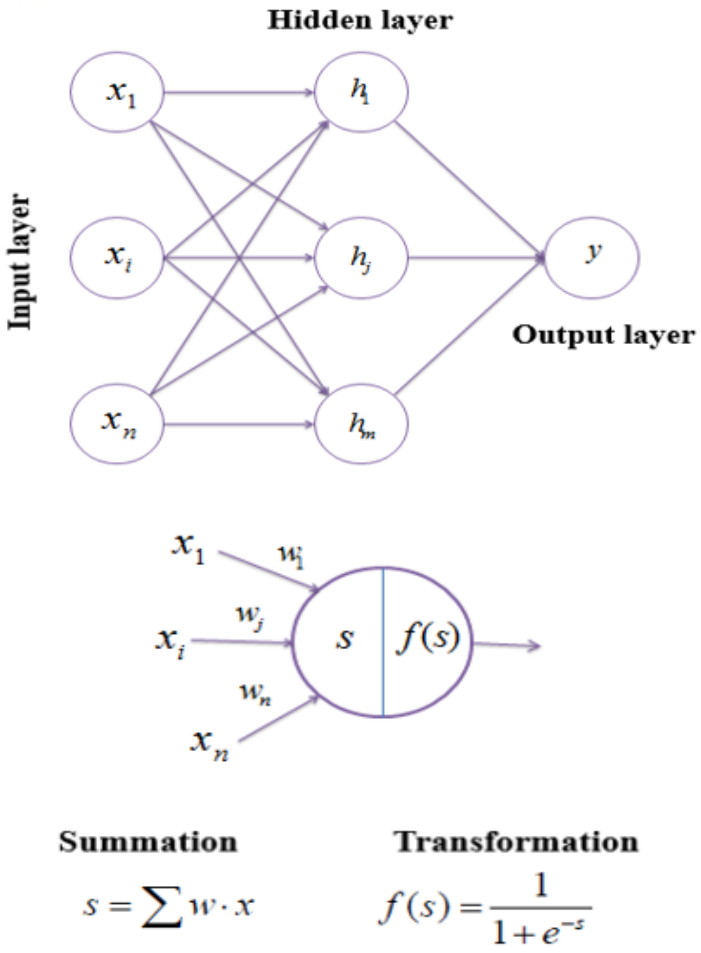


**Fig. 6.5.** MLP architecture

In this work, the MLP consists of a single hidden layer which is fully-connected to the input layer as well as the output layer. The standard logistic sigmoid is used as the activation function in the MLP.

### 6.1.5 Prediction

This phase consisted of the concatenation of output features from all the sub-networks in order to generate the final *concat feature* using model-level multi-lingual fusion strategy. Typically, multi-lingual fusion strategies can be categorized into early, model-level and late fusion. The early multilingual fusion strategy involves concatenation of features from different languages, the model-level multi-lingual fusion involves concatenation of high-level feature representations from different languages and the late multi-lingual fusion involves fusion of predictions from different languages as shown in figure 6.6 (a, b and c).

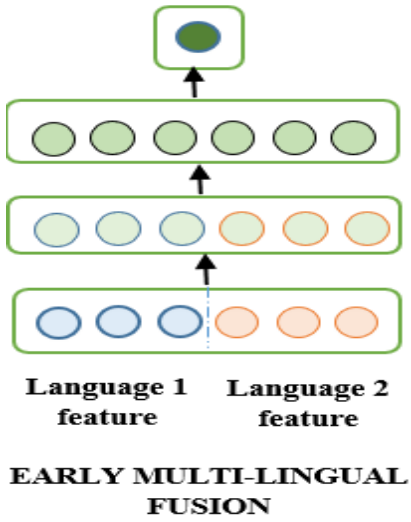


**Fig. 6.6(a).** Early multi-lingual fusion

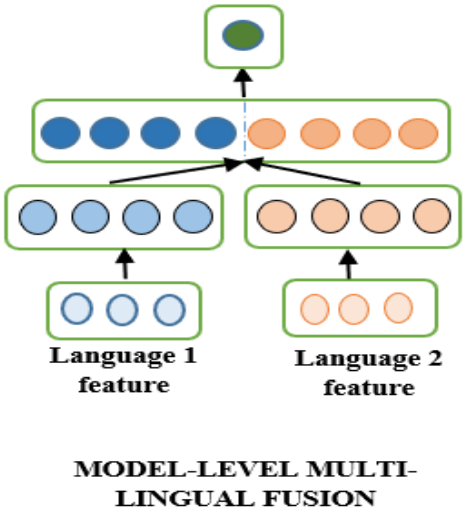


**Fig. 6.6(b).** Model-level multi-lingual fusion

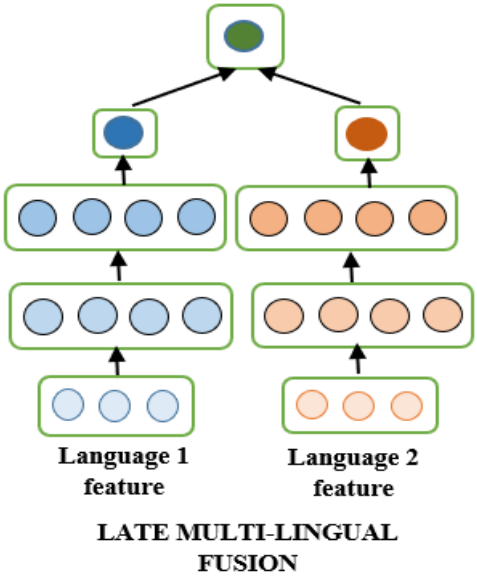


**Fig. 6.6(c).** Late multi-lingual fusion

In this research, the output features from the sub-networks are concatenated to generate the final *concat feature* using model-level multi-lingual fusion strategy. This *concat feature* is the shared representation which combines the high-level representation features of each input type. This fusion strategy helps to complete the essence of multi-input integrative learning proposed in this work. Unlike early fusion, this strategy helps to circumvent the curse of dimensionality and synchronization between different features and at the same time does not isolate interactions among different languages as in late fusion. Finally, the shared representation is given to the fully-connected layer which generates an output with linear activation to detect cyberbullying for code-mixed social media textual content.

## 6.2 Dataset

Two datasets were created by scraping data from the popular social networking sites namely, Twitter and Facebook. Data was based on selection of certain hashtags and keywords from the domain of politics, public figures, and entertainment etc. and was restricted to code-mixed 'Hinglish (Hindi+English)' language. The posts fetched from Facebook were "profile-based". The most popularly searched profiles of Sh. Narendra Modi ji (Prime Minister of India), Mr. Shahrukh Khan (actor), public profiles of NDTV (news channel) and Jawaharlal Nehru University (university in India) were observed for the data analysis. GraphAPI was used for the extraction of Facebook comments. For Twitter, "topic-based" tweets were scraped that belonged to the most trending topics such as "#Ind VS Pak, #Beef Ban, #movies". Tweepy tool was used for the extraction of tweets from Twitter. Also, the posts that were solely written in English or Hindi were removed using manual filtering. Finally, two datasets with 6500 (English-Hindi) code-

mixed posts each, for both Facebook, and Twitter, were created. The datasets were annotated for two categories, namely, cyberbullying (B) and non-bullying (NB).

The details for the tag-categorization are given in table 6.1.

**Table 6.1**. Tags and their counts for both the datasets

| | **Facebook** | **Twitter** |
|---|---|---|
| **Bullying (B)** | 3275 | 3350 |
| **Non-Bullying (NB)** | 3225 | 3150 |
| **Total** | 6500 | 6500 |

Table 6.2 and 6.3 gives the details about the average post and word length in different class text respectively.

**Table 6.2**. Average post length in different class text for both the datasets

| | **Facebook** | **Twitter** |
|---|---|---|
| **Bullying (B)** | 27.75 | 27.035 |
| **Non-Bullying (NB)** | 27.63 | 26.75 |

**Table 6.3.** Average word length in different class text for both the datasets

| | **Facebook** | **Twitter** |
|---|---|---|
| **Bullying (B)** | 4.505 | 4.76 |
| **Non-Bullying (NB)** | 4.24 | 4.10 |

## 6.3 Findings

Above specified Facebook and Twitter (mash-up or code-mix) datasets were analysed for cyberbullying detection on code-mix social media data using deep learning models. The findings are given below.

### 6.3.1 Model Performance

The performance of the model was assessed using the aforesaid datasets comprising around 6500 tweets and posts each. The Facebook dataset consisted of 3275 posts as bullying and 3225 as non-bullying posts and the Twitter dataset consisted of 3350 tweets as bullying and 3150 tweets as non-bullying. We performed 10-fold cross-validation and calculated the AUC curve.

The proposed model reports a performance of AUC-ROC of 0.97 for both the datasets as shown in figure 6.7. This is primarily because it combines an automatic feature extraction mechanism with the robustness, dynamism and flexibility of the deeper neural architectures such as CapsNet and Bi-LSTM.

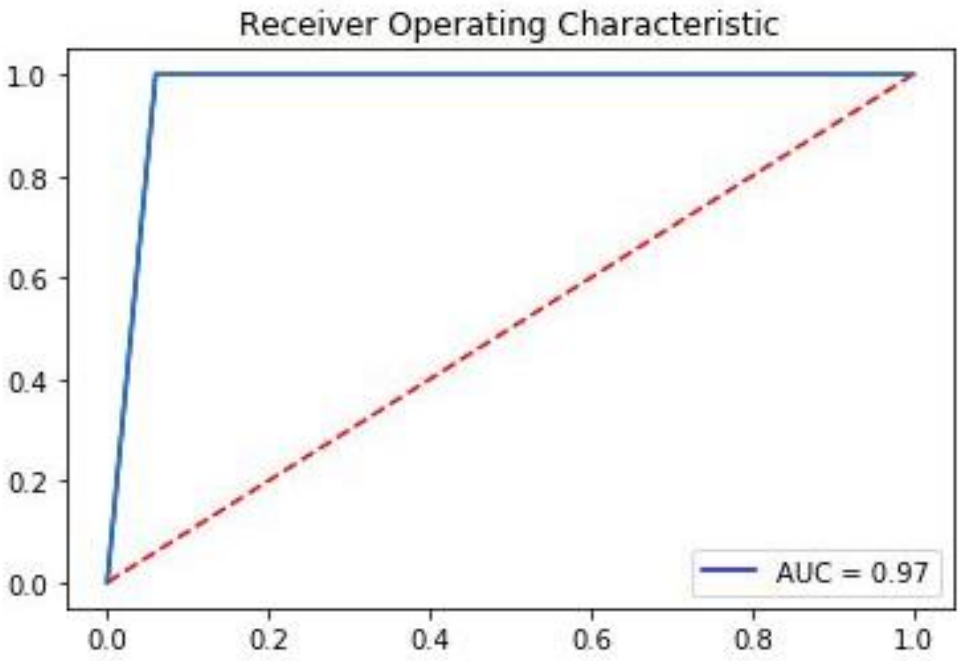


**Fig. 6.7.** Performance of MIIL-DNN on DS-I and DS-II

### 6.3.2 Ablation Study

The proposed model was evaluated using ablation architectures as well. The Hindi Bi-LSTM model was compared with other deep neural architectures, namely Convolution neural network (CNN) and LSTM. The accuracy results are shown in figure 6.8.

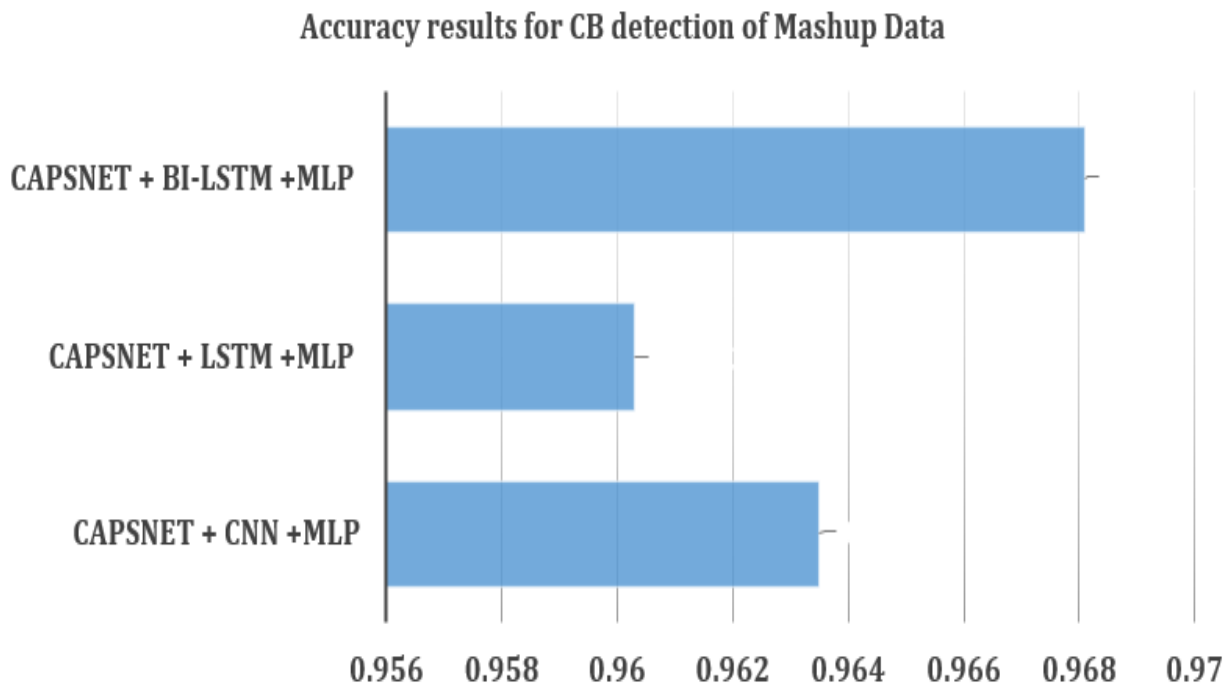


**Fig.6.8.** Comparative analysis of ablation architectures for Hindi using accuracy

## 6.4 Chapter Summary

This chapter discusses the proposed model which entails focusing cyberbullying detection using deep learning for mash-up or code-mix data. It also explains about the model-level feature fusion using deep neural networks to classify the incoming real-time post into bullying or non-bullying categories. It explains the problem of cyberbullying as a generic case where classification is done into two broad categories that is bullying and non-bullying as compared to earlier works on either toxic comment classification or hate-speech detection. This comprehensibility and generalization of the proposed model makes it easily scalable and applicable to high dimensional cross-platform, cross-lingual real-time datasets as well. Also, the chapter explains about the proposed integrative learning network, MIIL-DNN which combines the information from three sub-networks to generate the final output. The model yielded appreciable ROC-AUC of 0.97 on both the datasets.

**Publication:**

- Kumar, A.* & Sachdeva, N. (2020). "Multi-input Integrative Learning using Deep Neural Networks and Transfer Learning for Cyberbullying Detection in Real-time Code-Mix Data" *Multimedia Systems, Springer.* https://doi.org/10.1007/s00530-020-00672-7 **[SCIE-Impact Factor: 1.563]** ISSN:1432-1882

# Chapter 7
# Conclusion and Future Trends

This chapter discusses the conclusion of the research and the future trends.

## 7.1 Conclusion & Future Trends

Online bullying is an adverse societal issue which is rising at an alarming rate. In general, a bullying behaviour can be categorized on the basis of type of behaviour (verbal, social and physical), the environment (in person and online), its mode (direct and indirect), the visibility (overt and covert), the damage caused (physical and psychological) and the context in terms of place of occurrence (home, workplace and school). Cyberbullying is typically a social behaviour bullying within the online setting done covertly by direct or indirect means causing short-term and long-term psychological harm. The increasing availability of reasonable data services and social media presence has given some uninhibited effects where online users have discovered wrong & unlawful ways to harm and humiliate individuals through hateful comments on online platforms or apps. The persistence, audience size and damage speed make cyberbullying even more damaging than face-to-face bullying causing serious mental health and wellbeing issues to victims and making them feel totally overwhelmed.

Cyberbullying can result in increased distress for the victims along with low self-esteem, increased anger, frustration, depression, social withdrawal and in some cases, developing violent or suicidal traits. Technology allows the bullies to be anonymous, hard to trace and insulated from confrontation. To the targets of cyberbullying, it feels invasive and never-ending. With the amount of emotional and psychological distress caused to victims it is urgently required to find appropriate provisions which can detect and prevent it. Effective prevention relies on the timely and satisfactory detection of potentially toxic posts. The information overload on the chaotic and complex social media portals necessitates advanced automatic systems to identify potential risks proactively. Social media is one of the most favoured mediums by bullies and the huge amount of real-time, multi-modal and mash-up social media data makes manual bullying detection intractable. Additionally, the social media is more of social multimedia comprising of text, image, audio and video. Simultaneously, as the users are usually more comfortable conversing in their native language, the native-language keyboards add up to the volume and variety of user-created content aggravating the linguistic challenges. This fosters the need to design and develop contemporary models which tap and analyse online detrimental behaviour automatically from user-generated content in social media. Researchers worldwide have been trying to develop new ways to detect cyberbullying automatically, manage it and reduce its prevalence on social media. For detecting CB

posts in social-media, an intelligent data-driven model is required that can effectively identify and categorize the bullying messages. Assessing the user-generated content in social media could be rewarding for automatic cyberbullying detection using deep neural architectures as they are proving useful and obtaining state-of-the-art results for various natural language tasks with end-to-end training and representation learning capabilities. So, this research gives the overall holistic view of CB detection where the aim is to perform automatic bully detection using deep learning models for computationally analysing the content, modality and language-use in social media.

In this research, we applied supervised baseline machine learning techniques for cyberbullying detection on (mono-lingual) textual social media content taken from social media namely Formspring.me, MySpace & Ask.fm. This work presented the application of baseline machine learning techniques namely Naïve Bayes, Support Vector Machines, K-Nearest Neighbour, Decision Tree, Logistic Regression and Artificial Neural Networks for identification and detection of textual CB content in Formspring.me, MySpace & Ask.fm datasets. The study was carried out using the Weka tool. The results were critically analysed using Accuracy (Ac), Precision (Pr), Recall (Re) and F-Measure (F) as an efficacy criterion. It was observed that logistic regression outperformed all other supervised classification algorithms in terms of accuracy, followed by J48, support vector machines, naïve Bayesian and k-nearest neighbour for all the three datasets. Whereas artificial neural networks reported the lowest accuracy. Naïve Bayesian had comparable accuracy quite akin to k-nearest neighbour for all the datasets.

Furthermore, we also did detection of online bullying (binary classification: bullying & non-bullying) on the textual social media content using deep learning models. This research presented a hybrid model utilizing deep architectures, Bi-GRU-Attention-CapsNet (Bi-GAC), that benefits by learning sequential semantic representations and spatial location information using a Bi-GRU with self-attention followed by Capsule networks (CapsNet) for cyberbullying detection in the textual content of social media. The model was divided into two phases. Embedding, encoding and self-attention layer constituted the first phase whereas Capsule Network [150] and fully connected layer with sigmoid comprised the classification or the prediction (second) phase. Here, the pre-trained ELMo word embedding was used to create the input embedding matrix. A Bi-GRU encoder was trained using ELMo embedding to generate a sequence context feature vector. Consequently, a self-attention mechanism was added to capture significant information. Next the CapsNet generates semantic representation using a dynamic routing algorithm which was finally used for classification of the posts. Bi-GAC used the fully connected output layer with sigmoid activation to finally classify the positive as bullying or non-bullying. The improved text representation and feature learning offered a robust model which can avoid the vanishing gradient problem in comparison to baseline neural models. The model was validated on two benchmark and standard datasets, Formspring.me and MySpace. The proposed Bi-GAC model was evaluated for performance using F1-score. This study primarily focused on binary classification of

mono-lingual textual social media content for cyberbullying detection. Bi-GAC observed F1 score of 0.9403 and 0.9389 for Formspring.me and MySpace respectively.

The results obtained for the proposed Bi-GAC model were compared with the existing techniques for textual (mono-lingual) cyberbullying detection. Amongst all, it is observed that the Bi-GAC model showed a superior performance on both the benchmark datasets (Formspring.me and MySpace). We also performed the comparative analysis between the Bi-GAC model and the SOTA [51, 53, 94, 115, 116, 141] (as shown in figure 7.1). The darker lines depicted the performance obtained via the proposed model.

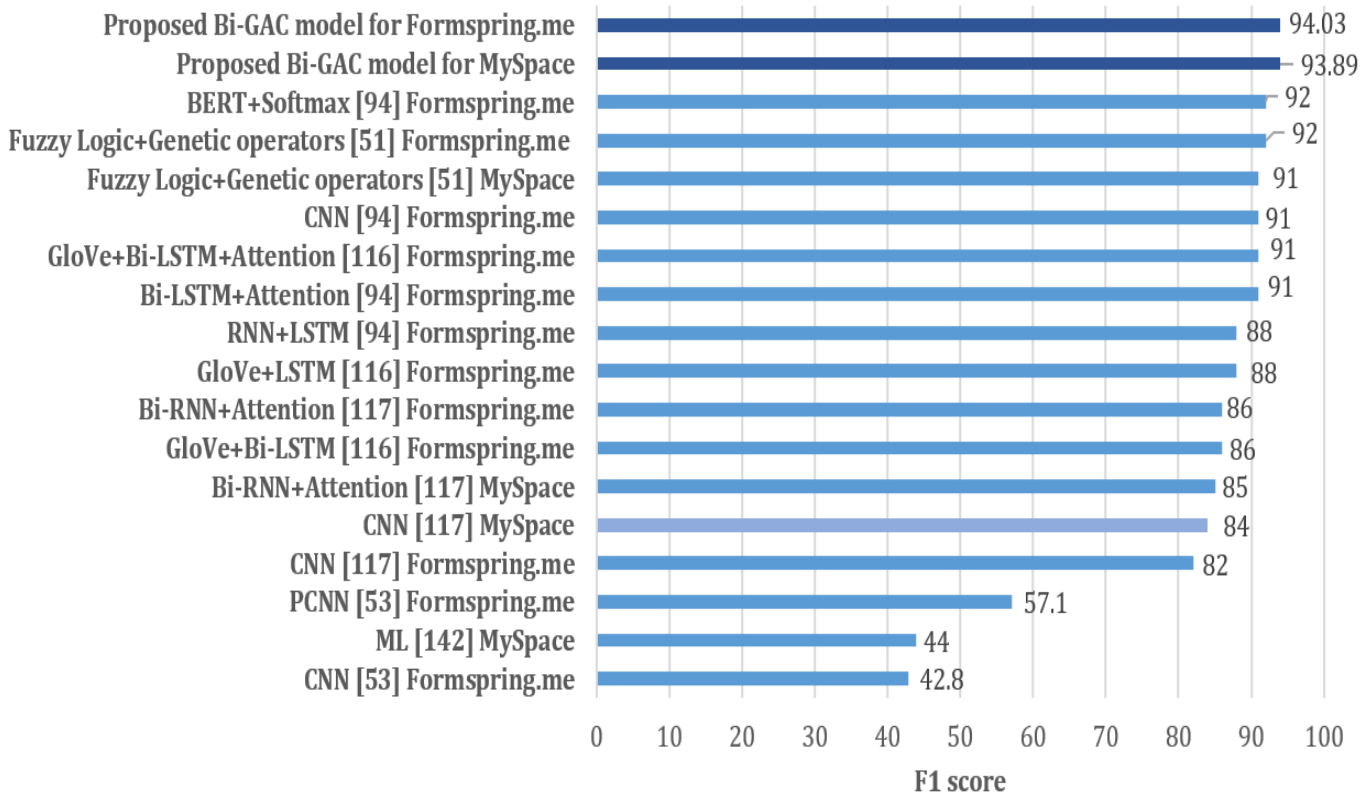


**Fig. 7.1.** Proposed vs. existing models for textual CB detection

Also, when we performed multimodal cyberbullying detection on social media using deep learning, it was observed that deep architectures efficiently dealt with three different modalities of social media data. In this work, we proposed a deep learning-based model for CB detection catering to textual, visual & info-graphic contents of social media. The proposed CapsNet- ConvNet model consisted of four different modules, namely, modality discretization module, textual processing module, visual processing module, and prediction module. The proposed CapsNet-ConvNet model, comprised of a Capsule network with dynamic-routing for detecting CB of textual content and a ConvNet for detecting CB of visual content. Discretization of the info-graphic content (i.e., separating text from image) was done using Google-Lens of Google Photos App. Perceptron-based-decision-level-late-fusion approach was used for multimodal learning in order to dynamically merge the predictions of distinct modalities into bullying or non-bullying type. Performance of the model was validated using mix-modal dataset containing

around 10000 comments & posts from YouTube, Instagram and Twitter. The modalities within the dataset were 60% textual, 20% visual and 20% info-graphic. We performed 10-fold cross validation and calculated the AUC-ROC curve. The proposed model produced best performance with the AUC-ROC of 0.98. The results obtained using the proposed CapsNet-ConvNet model were also compared with the existing techniques [99, 100, 104, 143, 144, 145] (SOTA) for multimodal cyberbullying detection. Amongst all, it was observed (from figure 7.2) that the CapsNet-ConvNet model outperformed the current best model [99] with an accuracy gain of around 7%. The darker line depicted the performance obtained via the proposed model.

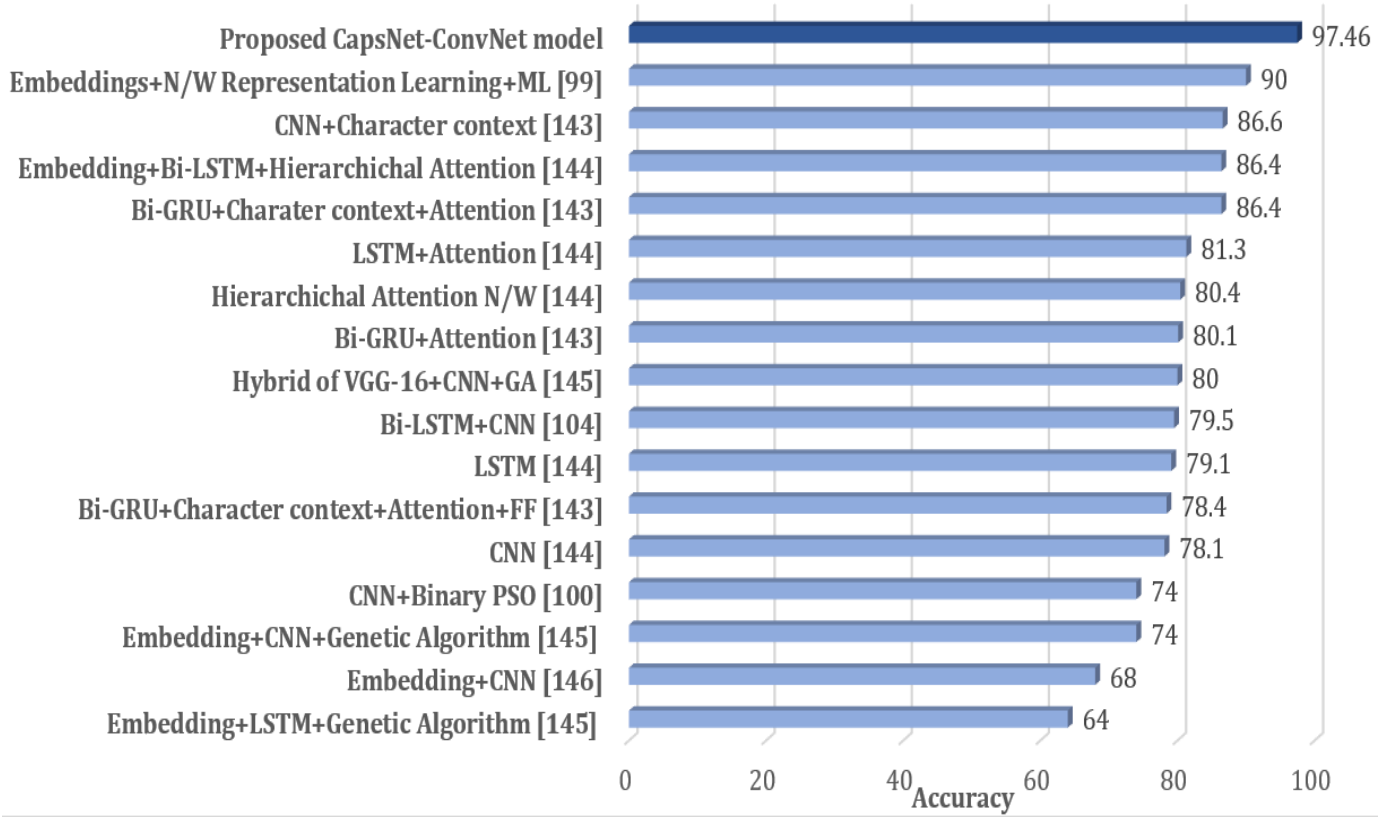


**Fig. 7.2.** Proposed vs. existing models for multimodal CB detection

Catering to the linguistic challenges pertaining to the use of mash-up or code-mix languages in social media, we proposed a MIIL-DNN model for cyberbullying detection using deep learning on code-mix or mash-up data. This research discussed CB detection of the code-mix data, precisely the Hinglish, which denotes about the association of words from Hindi & English language. This work proffered a model-level feature fusion model using deep neural networks to classify the incoming real-time post into bullying or non-bullying categories. MIIL-DNN merge learning's from three sub-networks to identify and classify cyberbullying in code-mix data. Three inputs viz. English language features, Hindi language features and typographic features are learned individually utilizing sub-networks (CapsNet for English language, Bi-LSTM for Hindi language and MLP for

typographic features). Subsequently, these are merged into one unified representation which is used as an input to a fully connected network for final prediction. Model-level fusion for multi-lingual data has an advantage that, for each input type, it works with the unique distribution without increasing the input space dimensionality. Validation of the model is done on two datasets which are created by scraping data from social media sites viz. Facebook and Twitter. The contribution of the research is two-fold: firstly, the problem of cyberbullying was taken as a generic case such that classification was done into two broad categories that is bullying and non-bullying as compared to earlier works on either toxic comment classification or hate-speech detection. This comprehensibility and generalization of the proposed model made it easily scalable and applicable to high dimensional cross-platform, cross-lingual real-time datasets as well. Secondly, the proposed integrative learning network, MIIL-DNN used a model-level multi-lingual fusion to combine information from three sub-networks to generate the final output. The proposed model reported a performance of approximately 0.97 (AUC-ROC) for both the datasets. This was primarily because it combines an automatic feature extraction mechanism with the robustness, dynamism and flexibility of the deeper neural architectures such as CapsNet and Bi-LSTM.

As we proposed training a CapsNet model for English tweets/posts, it was imperative to evaluate the robustness of this sub-network as well. We compared its performance with the existing state-of-the-art Toxic Comment Classification Challenge dataset[8] from a Kaggle competition. The dataset contains 159571 Wikipedia manually labelled comments categorized as: toxic; severe toxic; obscene; threat; insult and identity hate. All these categories accounted for cyberbullying whereas any comment with value = 0 in all fields indicated non-cyberbullying i.e., non-toxic comments. As per www.kaggle.com, the first-place solution reported a performance of 0.9885 using a Bi-GRU with the pseudo-labelling technique. The performance of the best single model of the competition was around 0.9869 and a single layer RNN-Capsule Network with GRU cell performed at 0.9857. One of the other works [147] used a capsule network with focal loss and achieved a ROC-AUC of 0.9846 on the Kaggle toxic comment dataset. The performance of the proposed CapsNet was thus comparable at 0.9841. The figure 7.3 showed the ROC curves for all the toxic comment categories.

[8] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

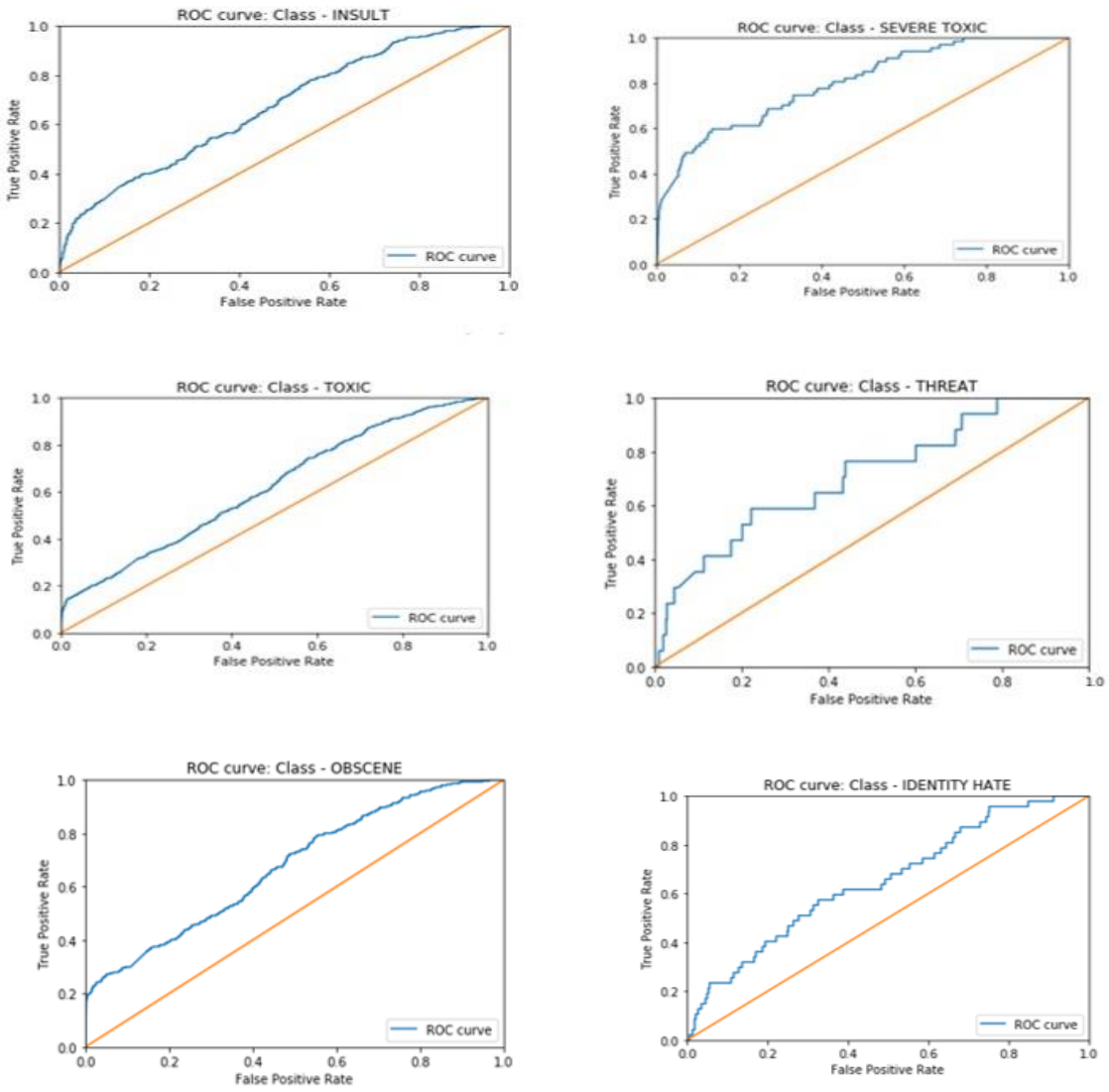


**Fig. 7.3.** Performance results of toxic comment categories

The results obtained using the proposed MIIL-DNN model were also compared with the existing techniques (SOTA) [89, 90, 148, 149, 150] of code-mix (Hinglish) cyberbullying detection. Amongst all, it was observed (from figure 7.4) that MIIL-DNN showed superlative performance with 96.81% accuracy, which was approximately 5% better than the current best GloVE + GRU model [90]. The darker line depicted the performance obtained via the proposed model.

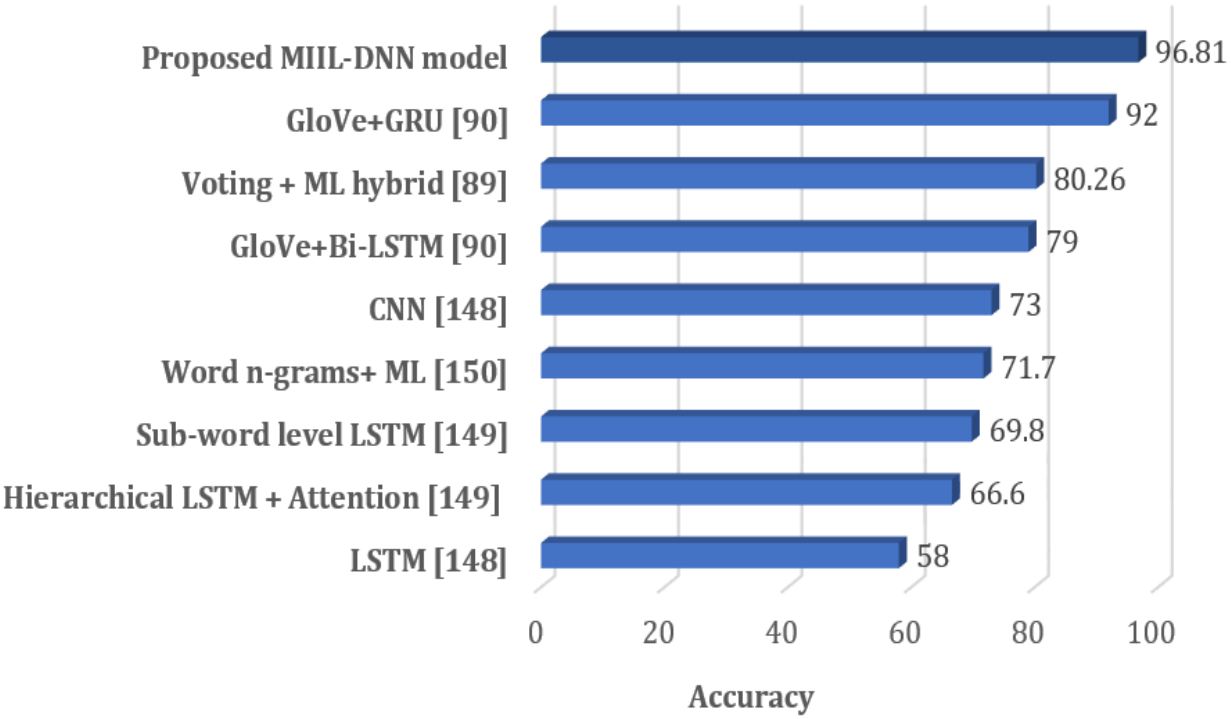


**Fig. 7.4.** Proposed vs. existing models for code-mix CB detection

Based upon the observations and findings, we can conclude that:

- Cyberbullying can take many forms; however, it typically refers to repeated and hostile behaviour online to intentionally and repeatedly harass or harm individuals.
- With the pervasive use of social media, cyberbullying is becoming rampant.
- Conventional methods to combat cyberbullying included guidelines on cyber-ethics, human moderators, and blacklisting based on the use of profane words.
- This research was to understand the alliance of cyberbullying on social media using deep learning.
- We developed, assessed and analysed models for mono-lingual textual CB detection, multi-modality-based CB detection and code-mix or mash-up based CB detection.
- The result of this analysis and comparison with SOTA validated the work in terms of improvement in the performance of the proposed models.
- MIIL-DNN model (for code-mix CB detection) showed superlative performance with 96.81% accuracy, which is approximately 5% better than the current best model.
- CapsNet-ConvNet model (for multi-modal CB detection) outperformed the current best model with an accuracy gain of around 7%.
- Bi-GAC model (for mono-lingual textual CB detection) showed superior performance in F-score for the MySpace and Formspring.me dataset respectively.
- This research showed the use of different types of embeddings that simplified the feature selection process effectively.

- ❖ This research also illustrated the comparative analysis of the proposed models for the various ablation architectures. The proposed models produced superlative results as compared to those ablation models.

As promising future direction, models and benchmark datasets for multi-lingual, multi-modal cyberbullying detection research tasks are ardently desired. At times, the available datasets have smaller size of training sets for classifier learning models. Most of the real-time datasets suffer from imbalance class distribution (skewed data), it encourages the use of learning techniques that could handle this. Computational approaches to deal with linguistic code switching is yet another potential area for future work. Our research primarily focused on content-based CB detection, whereas handling user profile-based features, socio-demographic features etc. is another dimension for further research. Another promising future direction may include the use of explainable artificial intelligence models for such predictive analysis.

# REFERENCES


1. Bounegru L, Gray J, Venturini T, Mauri M. A Field Guide to 'Fake News' and Other Information Disorders. A Field Guide to" Fake News" and Other Information Disorders: A Collection of Recipes for Those Who Love to Cook with Digital Methods, Public Data Lab, Amsterdam (2018). 2018.
2. Shrivastava, G., Kumar, P., Ojha, R. P., Srivastava, P. K., Mohan, S., & Srivastava, G. (2020). Defensive modeling of fake news through online social networks. IEEE Transactions on Computational Social Systems, 7(5), 1159-1167.
3. Kumar, A., Nayak, S., & Chandra, N.: Empirical Analysis of Supervised Machine Learning Techniques for Cyberbullying Detection. In: International Conference on Innovative Computing and Communications, pp. 223-230. Springer, Singapore (2019).
4. Mladenović, M., Ošmjanski, V., & Stanković, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. ACM Computing Surveys (CSUR), 54(1), 1-42.
5. Kumar, A., & Jaiswal, A. (2019). Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. Multimedia Tools and Applications, 78(20), 29529-29553.
6. Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. PloS one, 13(10), e0203794.
7. Hang, O. C., & Dahlan, H. M. (2019, December). Cyberbullying lexicon for social media. In 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS) (pp. 1-6). IEEE.
8. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75.
9. Ballal N., Saritha S.K. (2020) A Study of Deep Learning in Text Analytics. In: Shukla R., Agrawal J., Sharma S., Chaudhari N., Shukla K. (eds) Social Networking and Computational Intelligence. Lecture Notes in Networks and Systems, vol 100. Springer, Singapore. https://doi.org/10.1007/978-981-15-2071-6_16
10. Kumar, A., Srinivasan, K., Cheng, W. H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Information Processing & Management, 57(1), 102141.
11. Kumar, A., & Jaiswal, A. (2020). A Deep Swarm-Optimized Model for Leveraging Industrial Data Analytics in Cognitive Manufacturing. IEEE Transactions on Industrial Informatics, 17(4), 2938-2946.
12. Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. Concurrency and Computation: Practice and Experience, 32(1), e5107.

13. Sangwan, S. R., & Bhatia, M. P. S. (2020). D-BullyRumbler: a safety rumble strip to resolve online denigration bullying using a hybrid filter-wrapper approach. Multimedia Systems, 1-17.
14. Brown, L. (2012). New Harvard Study Shows Why Social Media Is So Addictive for Many. [online] WTWH Marketing Lab. Available at: http://marketing.wtwhmedia.com/new-harvard-study-shows-why-social-media-is-so-addictive-for-many/ [Accessed 27 Jan. 2020].
15. https://coschedule.com/ Accessed on 14 July 2018.
16. Salawu S, He Y, Lumsden J (2017) Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Transactions on Affective Computing (1):1-20.
17. Sarna, G., & Bhatia, M. P. S. (2020). Structure-Based Analysis of Different Categories of Cyberbullying in Dynamic Social Network. International Journal of Information Security and Privacy (IJISP), 14(3), 1-17.
18. Bounegru L, Gray J, Venturini T, Mauri M. A Field Guide to 'Fake News' and Other Information Disorders. A Field Guide to" Fake News" and Other Information Disorders: A Collection of Recipes for Those Who Love to Cook with Digital Methods, Public Data Lab, Amsterdam (2018). 2018.
19. Foody, M., Samara, M., & Carlbring, P.: A review of cyberbullying and suggestions for online psychological therapy. Internet Interventions 2(3), 235-242 (2015).
20. Campbell MA (2005) Cyber bullying: An old problem in a new guise?. Journal of Psychologists and Counsellors in Schools 15(1):68-76.
21. Barlett, C. P., Simmers, M. M., Roth, B., & Gentile, D. (2021). Comparing cyberbullying prevalence and process before and during the COVID-19 pandemic. The Journal of Social Psychology, 1-11.
22. Ybarra M (2010) Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression. National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda.
23. Cyberbullying, The National Crime Prevention. http://www.ncpc.org/cyberbullying. Accessed 10 March 2018.
24. Ngo, A. T., Tran, A. Q., Tran, B. X., Nguyen, L. H., Hoang, M. T., Nguyen, T. H. T., ... & Ho, C. S. (2021). Cyberbullying Among School Adolescents in an Urban Setting of a Developing Country: Experience, Coping Strategies, and Mediating Effects of Different Support on Psychological Well-Being. Frontiers in Psychology, 12, 930.
25. http://www.ryanpatrickhalligan.org/ Accessed on 14 July 2018.
26. National Bullying Prevention Center. https://www.pacer.org/bullying/. Accessed 26 July 2018.
27. Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., & Rajan, A. R. (2021). Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. Computers & Electrical Engineering, 92, 107186.
28. Farley, S., Coyne, I., & D'Cruz, P. (2021). Cyberbullying at work: Understanding the influence of technology. Concepts, Approaches and Methods, 233-263.

29. Chia, Z. L., Ptaszynski, M., Masui, F., Leliwa, G., & Wroczynski, M. (2021). Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. Information Processing & Management, 58(4), 102600.
30. Fang, Y., Yang, S., Zhao, B., & Huang, C. (2021). Cyberbullying detection in social networks using Bi-gru with self-attention mechanism. Information, 12(4), 171.
31. Kumar, R. (2021). Detection of Cyberbullying using Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(9), 656-661.
32. Bharti, S., Yadav, A. K., Kumar, M., & Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. Kybernetes.
33. Mahat, M. (2021, March). Detecting Cyberbullying Across Multiple Social Media Platforms Using Deep Learning. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 299-301). IEEE.
34. Aggarwal, C. C. Neural Networks and Deep Learning: A Textbook. Springer 2018.
35. B. Kitchenham, S. Charters, (2007) Guidelines for performing Systematic Literature Reviews in Software Engineering, Tech. Rep. EBSE. 1 1–57.
36. Reynolds K, Kontostathis A, Edwards L (2011) Using machine learning to detect cyberbullying. InMachine learning and applications and workshops (ICMLA), 2011 10th International Conference, IEEE 2: 241-244.
37. Nahar V, Unankard S, Li X, Pang C (2012) Sentiment analysis for effective detection of cyber bullying. Asia-Pacific Web Conference, Springer, Berlin, Heidelberg: 767-774.
38. Xu JM, Jun KS, Zhu X, Bellmore A (2012) 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, Association for Computational Linguistics:656-666.
39. Kontostathis A, Reynolds K, Garron A, Edwards L (2013) Detecting cyberbullying: query terms and techniques. InProceedings of the 5th annual acm web science conference: 195-204.
40. Dadvar M, Trieschnigg D, Ordelman R, de Jong F (2013) Improving cyberbullying detection with user context. In European Conference on Information Retrieval, Springer, Berlin, Heidelberg: 693-696.
41. Sheeba JI, Vivekanandan K (2013) Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique. In IEEE International Conference on Computational Intelligence and Computing Research (ICCIC): 1-5.
42. Nahar V, Al-Maskari S, Li X, Pang C (2014) Semi-supervised learning for cyberbullying detection in social networks. In Australasian Database Conference, Springer, Cham: 160-171.
43. Parime S, Suri V (2014) Cyberbullying detection and prevention: Data mining and psychological perspective. In Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference IEEE: 1541-1547.

44. Dadvar M, Trieschnigg D, de Jong F (2014) Experts and machines against bullies: A hybrid approach to detect cyberbullies. In Canadian Conference on Artificial Intelligence, Springer, Cham: 275-281.
45. Michalopoulos D, Mavridis I, Jankovic M (2014) GARS: Real-time system for identification, assessment and control of cyber grooming attacks. Computers & security 42:177-90.
46. Holt TJ, Turner MG, Exum ML (2014) The impact of self control and neighborhood disorder on bullying victimization. Journal of Criminal Justice 42(4):347-55.
47. Byrne S, Katz SJ, Lee T, Linz D, McIlrath M (2014) Peers, predators, and porn: Predicting parental underestimation of children's risky online experiences. Journal of Computer-Mediated Communication. 19(2):215-31.
48. Rafiq RI, Hosseinmardi H, Han R, Lv Q, Mishra S, Mattson SA (2015) Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM: 617-622.
49. Chavan VS, Shylaja SS (2015) Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Advances in computing, communications and informatics (ICACCI), 2015 International Conference on IEEE: 2354-2358.
50. Balci K, Salah AA (2015) Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. Computers in Human Behavior 53:517-26.
51. Nandhini BS, Sheeba JI (2015) Online social network bullying detection using intelligence techniques. Procedia Computer Science 45:485-92.
52. Balakrishnan V (2015) Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency. Computers in Human Behavior 46:149-57.
53. Zhang X, Tong J, Vishwamitra N, Whittaker E, Mazer JP, Kowalski R, Hu H, Luo F, Macbeth J, Dillon E (2016) Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA): 740-745.
54. Zhao R, Zhou A, Mao K (2016) Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking: 43-48.
55. Hosseinmardi H, Rafiq RI, Han R, Lv Q, Mishra S (2016) Prediction of cyberbullying incidents in a media-based social network. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: 186-192.
56. Gordeev D (2016) Detecting state of aggression in sentences using CNN. In International Conference on Speech and Computer, Springer, Cham: 240-245.
57. Hammer HL (2016) Automatic detection of hateful comments in online discussion. In International Conference on Industrial Networks and Intelligent Systems, Springer, Cham: 164-173.

58. Gordeev D (2016) Automatic detection of verbal aggression for Russian and American image boards. Procedia-Social and Behavioral Sciences 236:71-5.
59. Al-garadi MA, Varathan KD, Ravana SD (2016) Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior 63:433-43.
60. Potha N, Maragoudakis M, Lyras D (2016) A biology-inspired, data mining framework for extracting patterns in sexual cyberbullying data. Knowledge-Based Systems 96:134-55.
61. Rafiq RI, Hosseinmardi H, Mattson SA, Han R, Lv Q, Mishra S (2016) Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. Social Network Analysis and Mining 6(1):88.
62. Papegnies E, Labatut V, Dufour R, Linares G (2017) Graph-based Features for Automatic Online Abuse Detection. In International Conference on Statistical Language and Speech Processing, Springer, Cham: 70-81.
63. Sedano CR, Ursini EL, Martins PS (2017) A Bullying-Severity Identifier Framework Based on Machine Learning and Fuzzy Logic. In International Conference on Artificial Intelligence and Soft Computing, Springer, Cham: 315-324.
64. Thu PP, New N (2017) Implementation of emotional features on satire detection. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017 18th IEEE/ACIS International Conference, IEEE: 149-154.
65. Zhao R, Mao K (2017) Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. IEEE Transactions on Affective Computing 8(3):328-39.
66. Raisi E, Huang B (2017) Cyberbullying detection with weakly supervised machine learning. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM: 409-416.
67. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Hate is not binary: Studying abusive behavior of# gamergate on twitter. In Proceedings of the 28th ACM conference on hypertext and social media: 65-74.
68. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Detecting aggressors and bullies on Twitter. In Proceedings of the 26th International Conference on World Wide Web Companion 767-768.
69. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Mean birds: Detecting aggression and bullying on twitter. In Proceedings of the 2017 ACM on web science conference: 13-22.
70. García-Recuero Á (2017) Efficient Privacy-preserving Adversarial Learning in Decentralized Online Social Networks. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM: 1132-1135.
71. Ashktorab Z, Haber E, Golbeck J, Vitak J (2017) Beyond Cyberbullying: Self-Disclosure, Harm and Social Support on ASKfm. In Proceedings of the 2017 ACM on Web Science Conference: 3-12.

72. Bourgonje P, Moreno-Schneider J, Srivastava A, Rehm G (2017) Automatic classification of abusive language and personal attacks in various forms of online communication. In International Conference of the German Society for Computational Linguistics and Language Technology, Springer, Cham: 180-191.
73. Haidar B, Chamoun M, Serhrouchni A (2017) Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In Cyber Security in Networking Conference (CSNet), IEEE: 1-8.
74. Wint ZZ, Ducros T, Aritsugi M (2017) Spell corrector to social media datasets in message filtering systems. In Digital Information Management (ICDIM), 2017 Twelfth International Conference, IEEE: 209-215.
75. Sarna, G., & Bhatia, M. P. S. (2017). Content based approach to find the credibility of user in social networks: an application of cyberbullying. International Journal Of Machine Learning and Cybernetics, 8(2), 677-689.
76. Rakib TB, Soon LK (2018) Using the Reddit Corpus for Cyberbully Detection. In Asian Conference on Intelligent Information and Database Systems, Springer, Cham: 180-189.
77. Agrawal S, Awekar A (2018) Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms. In European Conference on Information Retrieval, Springer, Cham: 141-153.
78. Chen J, Yan S, Wong KC (2018) Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis. Neural Computing and Applications:1-10.
79. Koban K, Stein JP, Eckhardt V, Ohler P (2018) Quid pro quo in Web 2.0. Connecting personality traits and Facebook usage intensity to uncivil commenting intentions in public online discussions. Computers in Human Behavior 79:9-18.
80. Coletto M, Lucchese C, Orlando S (2018) Do Violent People Smile: Social Media Analysis of their Profile Pictures. In Companion of the Web Conference 2018. International World Wide Web Conferences Steering Committee, ACM: 1465-1468.
81. Sharma HK, Kshitiz K, Shailendra (2018) NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms. International Conference on Advances in Computing and Communication Engineering (ICACCE) 2018, IEEE: 265-272.
82. Bu SJ, Cho SB (2018) A Hybrid Deep Learning System of CNN and LRCN to Detect Cyberbullying from SNS Comments. International Conference on Hybrid Artificial Intelligence Systems 2018, Springer: 561-572.
83. Ibrohim, M. O., & Budi, I. (2019, August). Multi-label hate speech and abusive language detection in Indonesian twitter. In Proceedings of the Third Workshop on Abusive Language Online, 46-57.
84. Pratiwi, N. I., Budi, I., & Jiwanggi, M. A. (2019, July). Hate Speech Identification using the Hate Codes for Indonesian Tweets. In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, 128-133.

85. Haidar, B., Chamoun, M. and Serhrouchni, A., 2017, October. Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In 2017 1st Cyber Security in Networking Conference (CSNet), 1-8. IEEE.
86. Haidar, B., Chamoun, M. and Serhrouchni, A., 2017. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Advances in Science, Technology and Engineering Systems Journal, 2(6), 275-284.
87. Pawar, R. and Raje, R.R., 2019, May. Multilingual Cyberbullying Detection System. In 2019 IEEE International Conference on Electro Information Technology (EIT), 040-044. IEEE.
88. Arreerard, R. and Senivongse, T., 2018, July. Thai Defamatory Text Classification on Social Media. In 2018 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD) 73-78. IEEE.
89. Tarwani, S., Jethanandani, M. and Kant, V., 2019, April. Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification. In International Conference on Advances in Computing and Data Sciences 543-551. Springer, Singapore.
90. Gupta VK. (2019). "Hinglish" Language--Modeling a Messy Code-Mixed Language. arXiv preprint arXiv:1912.13109. 2019 Dec 30.
91. Haidar, B., Chamoun, M. and Yamout, F., 2016, November. Cyberbullying detection: a survey on multilingual techniques. In 2016 European Modelling Symposium (EMS) 165-171. IEEE.
92. Al-Hassan, A. and Al-Dossari, H., 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In 6th International Conference on Computer Science and Information Technology.
93. Meng, Z., Tian, S., & Yu, L. (2020). Regional Bullying Text Recognition Based on Two-Branch Parallel Neural Networks. Automatic Control and Computer Sciences, 54(4), 323-334.
94. Paul, S., & Saha, S. (2020). CyberBERT: BERT for cyberbullying identification. Multimedia Systems, 1-8.
95. Liu, W., Wen, B., Gao, S., Zheng, J., & Zheng, Y. (2020). A multi-label text classification model based on ELMo and attention. In MATEC Web of Conferences (Vol. 309, p. 03015). EDP Sciences.
96. Maslej-Krešňáková, V., Sarnovský, M., Butka, P., & Machová, K. (2020). Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification. Applied Sciences, 10(23), 8631.
97. Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. Future Internet, 12(11), 187.
98. Kumar, A., & Jaiswal, A. (2017). Empirical study of twitter and tumblr for sentiment analysis using soft computing techniques. In Proceedings of the world congress on engineering and computer science (Vol. 1, pp. 1-5).

99. Özel, S. A., & Sarac, E. (2016). Effects of Feature Extraction and Classification Methods on Cyberbully Detection. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 21(1), 190-200.
100. Çiğdem, A. C. I., Çürük, E., & Eşsiz, E. S. (2019). Automatic Detection of Cyberbullying in FORMSPRING. Me, Myspace and Youtube Social Networks. Turkish Journal of Engineering, 3(4), 168-178
101. Hosseinmardi, H., Li, S., Yang, Z., Lv, Q., Rafiq, R. I., Han, R., & Mishra, S. (2014, December). A comparison of common users across instagram and ask. fm to better understand cyberbullying. In 2014 IEEE Fourth International Conference on Big Data and Cloud Computing (pp. 355-362). IEEE.
102. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
103. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
104. Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6(02), 107-116.
105. Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.
106. Banerjee, V., Telavane, J., Gaikwad, P., & Vartak, P. (2019, March). Detection of cyberbullying using deep neural network. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (pp. 604-607). IEEE.
107. Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306.
108. Kumar, A., Sangwan, S. R., Arora, A., Nayyar, A., & Abdel-Basset, M. (2019). Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE access, 7, 23319-23328.
109. Jain, D., Kumar, A., & Garg, G. (2020). Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. Applied Soft Computing, 91, 106198.
110. Shao, Y., Lin, J. C. W., Srivastava, G., Jolfaei, A., Guo, D., & Hu, Y. (2021). Self-attention-based conditional random fields latent variables model for sequence labeling. Pattern Recognition Letters, 145, 157-164.
111. Kim, J., Jang, S., Park, E. and Choi, S., 2019. Text classification using capsules. Neurocomputing.
112. Srivastava, S., & Khurana, P. (2019, August). Detecting aggression and toxicity using a multi dimension capsule network. In Proceedings of the Third Workshop on Abusive Language Online (pp. 157-162).

113. Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S. and Zhao, Z., 2018. Investigating capsule networks with dynamic routing for text classification. arXiv preprint arXiv:1804.00538.
114. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. arXiv preprint arXiv:1710.09829.
115. Agrawal, S., & 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.
116. Zhang, A., Li, B., Wan, S., & Wang, K. (2019). Cyberbullying detection with birnn and attention mechanism. In International Conference on Machine Learning and Intelligent Communications (pp. 623-635). Springer, Cham.
117. Kumar, A. Srinivasan, K., Cheng, W.H & Zomaya, A.Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Information Processing & Management, Elsevier, vol. 57, no.1, 102141.
118. Kumar, A. (2020). Using cognition to resolve duplicacy issues in socially connected healthcare for smart cities. Computer Communications 152 (2020): 272-281. https://doi.org/10.1016/j.comcom.2020.01.041.
119. Arif, M. (2021). A Systematic Review of Machine Learning Algorithms in Cyberbullying Detection: Future Directions and Challenges. Journal of Information Security and Cybercrimes Research, 4(1), 01-26.
120. Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep learning algorithm for cyberbullying detection. International Journal of Advanced Computer Science and Applications, 9(9), 199-205.
121. Nimmi, K., Menon, V. G., Janet, B., & Kumar, A. (2020). Deep Learning for Next-Generation Inventive Wireless Networks: Issues, Challenges, and Future Directions. Handbook of Research on Emerging Trends and Applications of Machine Learning, IGI Global, 183-199.
122. Young T., Hazarika D., Poria S. & Cambria, E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence magazine, vol. 13, no. 3, pp. 55-75.
123. Dadvar, M., & Eckert, K. (2018). Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study. arXiv preprint arXiv:1812.08046.
124. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In Advances in neural information processing systems (pp. 3856-3866).
125. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv preprint arXiv:1802.05365. 2018 Feb 15.
126. Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352-2449.
127. Caldeira, M., Martins, P., Costa, R. L. C., & Furtado, P. (2020). Image Classification Benchmark (ICB). Expert Systems with Applications, 142, 112998.

128. Ali, W. N. H. W., Mohd, M., & Fauzi, F. (2018, November). Cyberbullying detection: an overview. In 2018 Cyber Resilience Conference (CRC) (pp. 1-3). IEEE.
129. Hani, J., Nashaat, M., Ahmed, M., Emad, Z., Amer, E., & Mohammed, A. (2019). Social media cyberbullying detection using machine learning. International Journal of Advanced Computer Science and Applications, 10(5), 703-707.
130. Haidar, B., Chamoun, M., & Yamout, F. (2016, November). Cyberbullying detection: A survey on multilingual techniques. In 2016 European Modelling Symposium (EMS) (pp. 165-171). IEEE.
131. Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2020). Cyberbullying detection solutions based on deep learning architectures. Multimedia Systems, 1-14.
132. Srivastava, S., & Khurana, P. (2019, August). Detecting aggression and toxicity using a multi dimension capsule network. In Proceedings of the Third Workshop on Abusive Language Online (pp. 157-162).
133. Tarwani, S., Jethanandani, M., & Kant, V. (2019, April). Cyberbullying Detection in Hindi-English Code-Mixed Language Using Sentiment Classification. In International Conference on Advances in Computing and Data Sciences (pp. 543-551). Springer, Singapore.
134. Loper E., Bird S. (2002). "NLTK: The natural language toolkit", Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, Association for Computational Linguistics vol. 1, 63-70.
135. Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on twitter. International Journal of Computer Science Issues (IJCSI), 9(4), 372.
136. Knight, K. and Graehl, J., 1998. Machine transliteration. Computational linguistics, 24(4), 599-612.
137. Kumar, A., & Jaiswal, A. (2019). Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on twitter. Multimedia Tools and Applications, 78(20), 29529-29553.
138. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.
139. Yang, M., Zhao, W., Ye, J., Lei, Z., Zhao, Z., & Zhang, S. (2018). Investigating capsule networks with dynamic routing for text classification. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 3110-3119).
140. Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding, 273-278. IEEE.
141. Dadvar, M., Jong, F. D., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). University of Ghent.
142. Vijayaraghavan, P., Larochelle, H., & Roy, D. (2021). Interpretable multi-modal hate speech detection. arXiv preprint arXiv:2103.01616.

143. Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. (2020, July). Multi-modal cyberbullying detection on social networks. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
144. Kumari, K., & Singh, J. P. (2021). Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Transactions on Emerging Telecommunications Technologies, 32(2), e3907.
145. Kumari, K., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2020). Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Computing, 24(15), 11059-11070.
146. Srivastava, S., Khurana, P., & Tewari, V. (2018, August). Identifying aggression and toxicity in comments using capsule network. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 98-105.
147. Singh, V., Varshney, A., Akhtar, S. S., Vijay, D., & Shrivastava, M. (2018, October). Aggression detection on social media text using deep neural networks. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) 43-50.
148. Santosh, T. Y. S. S., & Aravind, K. V. S. (2019, January). Hate Speech Detection in Hindi-English Code-Mixed Social Media Text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data 310-313.
149. Bohra, A., Vijay, D., Singh, V., Akhtar, S.S. and Shrivastava, M., 2018, June. A dataset of hindi-english code-mixed social media text for hate speech detection. In Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media 36-41.
150. Gangwar, A. K., & Ravi, V. (2020). A Novel BGCapsule Network for Text Classification. arXiv preprint arXiv:2007.04302.

# APPENDIX-A
# LIST OF PUBLICATIONS

**Journal(s)**

1. Kumar, A.* & Sachdeva, N. (2019). “Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis.”, *Multimedia Tools and Applications*, Vol. 78, pp. 23973–24010. https://doi.org/10.1007/s11042-019-7234-z [SCIE-Impact Factor: 2.313], ISSN: 1380-7501.

2. Kumar, A.* & Sachdeva, N. (2020). “Multi-input Integrative Learning using Deep Neural Networks and Transfer Learning for Cyberbullying Detection in Real-time Code-Mix Data” *Multimedia Systems, Springer,* https://doi.org/10.1007/s00530-020-00672-7 [SCIE-Impact Factor: 1.563] ISSN:1432-1882

3. Kumar, A.* & Sachdeva, N. (2021). “Multimodal Cyberbullying Detection Using Capsule Network with Dynamic Routing and Deep Convolutional Neural Network”, *Multimedia Systems, Springer* https://doi.org/10.1007/s00530-020-00747-5 [SCIE-Impact Factor: 1.563] ISSN: 1432-1882

4. Kumar, A., & Sachdeva, N. (2021). "A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media”, *World Wide Web, Springer*. https://doi.org/10.1007/s11280-021-00920-4. [SCIE JOURNAL, IMPACT FACTOR: 2.892]

**Conference(s)**

1. Kumar, A., Sachdeva, N. (2020). “Cyberbullying Checker: Online Bully Content Detection using Hybrid Supervised Learning”, *In International Conference on Intelligent Computing and Smart Communication 2019*, pp. 371-382. Springer.

2. Kumar, A., Sachdeva, N. (2021). “Cyberbullying Mediated Depression Detection in Social Media using Machine Learning”, *In Second Doctoral Symposium on Computational Intelligence*, pp. 869-877. Springer.