Hybrid Deep Learning Model for Cyberbullying Detection on Social Multimodal Data

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By:

MUDITA SAXENA

2K17/CSE/11

Under the supervision of

Dr. AKSHI KUMAR

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

JUNE, 2019

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

DECLARATION

I, Mudita Saxena, Roll No. 2K17/CSE/11 student of M.Tech (Compter Science & Engineering), hereby declare that the Project Dissertation titled "**Hybrid Deep Learning Model for Cyberbullying Detection on Social Multimodal Data**" which is submitted by me to the Department of Computer Science & Engineering , Delhi Technological University, Delhi Report of the Major II which is being submitted to Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of degree of Master of Technology for the requirements of the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: DTU, Delhi Mudita Saxena
Date: (2K17/CSE/11)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled *"***Hybrid Deep Learning Model for Cyberbullying Detection on Social Multimodal Data***"* which is submitted by Mudita Saxena, Roll No. 2K17/CSE/11, Department of computer Science & Engineering, Delhi Technological University, Delhi in partial fulfillment for the requirement of the award of degree of Master of Technology (Computer Science and Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **(Dr. Akshi Kumar)**

Date: **SUPERVISOR**

Assistant Professor

Department of Computer Engineering

Delhi Technological University

ABSTRACT

Cyberbullying is the use of Information and Communication Technology (ICT) by individuals’ to humiliate, tease, embarrass, taunt, defame and disparage a target without any face-to-face contact. Social media is the “virtual playground” used by bullies with the upsurge of social networking sites such as Facebook, Instagram, YouTube, Twitter etc. It is critical to implement models and systems for automatic detection and resolution of bullying content available online as the ramifications can lead to a societal epidemic. This research proffers a novel hybrid model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image). The all-in-one architecture, CNN-BoVW-SVM, consists of a convolution neural network (CNN) for predicting the textual bullying content and a support vector machine (SVM) classifier trained using bag-of-visual-words (BoVW) 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. The processing of textual and visual components is carried out using the hybrid architecture and a Boolean system with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of text and image bullying truth value. The model achieves a prediction accuracy of 84% which is acquired after performing tuning of different hyper-parameters.

ACKNOWLEDGEMENT

I am most thankful to my family for constantly encouraging me and giving me unconditional support while pursuing this research.

I am extremely grateful to **Dr. Akshi Kumar** Asst. Professor, Department of Computer Science Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to her for the extensive support and encouragement she provided.

I also convey my heartfelt gratitude to all the research scholars of the web Research Group at Delhi Technological University, for their valuable suggestions and helpful discussions throughout the course of this research work.

Mudita Saxena

Roll No. 2K17/CSE/11

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

A	Accuracy
Adj	Adjective
Adv	Adverb
BoVW	Bag of Visual Words
CB	Cyberbullying
CNN	Covolution Neural Network
DNN	Deep Neural network
Emoti	Emoticon
F	F-measure
POS	Part-of-speech
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayesian
NLP	Natural Language Processing
P	Precision
R	Recall
RF	Random Forest
SM	Social Media
SVM	Support Vector Machine
URL	Uniform Resource Locator
Vb	Verb

CHAPTER 1 INTRODUCTION

Web 2.0 is extending and evolving in terms of the volume, velocity and variety of information accessible online across various social media portals which affirm that the social media (SM) has global reach and has become widespread [1]. The global and pervasive reach of social multimedia has in return given some unpremeditated consequences where people have discovered illegal & unethical ways to use the socially-connected virtual communities. There are various benefits of SM but few people use it in wrong way. One of its most severe upshots is known as *cyberbullying* where individuals find new means to bully one another over the Internet. The term 'Cyberbullying' was devised by anti-bullying activist Bill Belsey in the year 2003 [2]. Tokunaga defined cyberbullying as "*any behavior performed through electronic or digital media by individuals or groups that repeatedly communicates hostile or aggressive messages intended to inflict harm or discomfort on others*" [3]. Different characteristics are highlighted by this definition like, the technology part, the antagonistic behavior of the act, important reason for causing suffering, considered very crucial to the definition and repetitivenss by many of the scholars [1]. Cyberbullying over social networks has already been claimed as a major risk or threat. Cyberbullying can be possible through any of the media like mobile phones or using internet. Cyberbullying may be done through emails, instant messages, chat room, blogs, images, video clip, text messages etc. [4, 5]. It has grown as a social menace that puts a negative effect on the minds of both the victim and bully. It is more persistent way of bullying a person before an entire online network, that is, the social networking sites, which can ultimately result in emotional and psychological breakdown of the victim with developed feelings of depression, stress, lack of self-confidence, anger, sadness, loneliness, health degradation, and suicides etc.

1.1. OVERVIEW

Through Cyberbullying, an individual or victim can be humiliated or hurt before whole network on the web. It bothers the psychological and physical condition of an individual because of which an expansive number of suicides and discouragement cases happen. These days, Cyberbullying, through images or memes are extremely common. Pornographic images or pictures with oppressive, mean or defamatory remarks are being presented on one's profile in order to bully them. Protective measures must be taken so as to control this. Thus, in our work we have made an automated framework which will, in addition to detection of bullying for text, also recognizes the bullying in the pictures. There can be a straightforward picture with no content on it or there might be such sort of picture also where some content is embedded with that picture. More recently, as memes and GIFs dominate the social feeds; typo-graphic and info-graphic visual content has become a considerable element of social data. Thus, cyber bullying, through varied content modalities is very common. Researchers worldwide have been trying to develop new ways to detect cyber bullying, manage it and reduce its prevalence on social media. Advanced analytical methods and computational models for efficient processing, analysis and modeling for detecting such bitter, taunting, abusive or negative content in images, memes or text messages are imperative. Social media specificity, topic dependence and variety in hand-crafted features currently define the bottlenecks in detecting online bullying posts [6]. State-of-the-art results are achieved by deep learning methods on some specific language problems by using the capabilities of hierarchical learning and generalization [7]. Pertinent studies report the use of deep learning models like CNN, RNN and semantic image features for bullying content detection by analyzing textual, image based and user features [8].But most of the research on online cyber-aggression, harassment detection and toxicity has been limited to text-based analytics [9]. Visual analysis of images have also been reported by few related studies for determining bullying content [10-12] 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 research work puts forward a hybrid deep learning model for bullying content prediction, where the content, c ε {text, image, info-graphic}.The primary contribution of the work is that unlike previous models which are mono-modal dealing with a single type of content modality, we proffer a multi-modal model for cyberbullying detection. The all-in-one hybrid deep architecture, CNN-BoVW-SVM, consists of a convolution neural network (CNN) for predicting the textual bullying content and a support vector machine (SVM) classifier trained using bag-of-visual-words (BoVW) 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. The processing of textual and visual components is carried out using the hybrid architecture and Boolean system with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of text and image bullying truth value. This unifying model thus considers modalities of content and processes each modality type using a concord of deep learning and machine learning techniques for an efficient decision support for cyberbullying detection. The generic architectural workflow of the proposed model is given in fig.1.1.

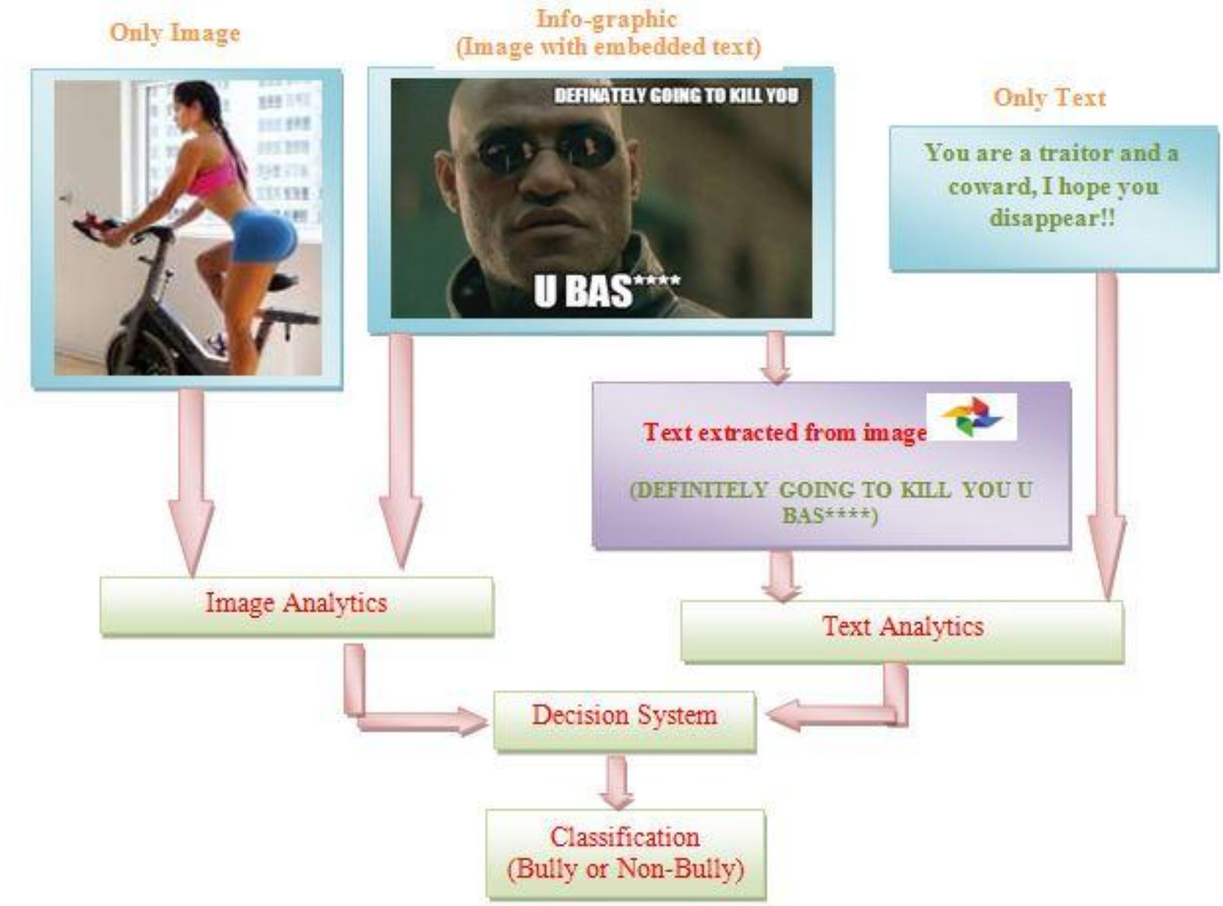

Only Image
Info-graphic
(Image with embedded text)
DEFINATELY GOING TO KILL YOU
U BAS****
Only Text
You are a traitor and a coward, I hope you disappear!!
Text extracted from image
(DEFINITELY GOING TO KILL YOU U BAS****)
Image Analytics
Text Analytics
Decision System
Classification
(Bully or Non-Bully)

Fig. 1.1 Generic architectural workflow of the proposed CNN-BoVW-SVM model

1.2. RESEARCH OBJECTIVES

This research proffers a novel hybrid model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image). The all-in-one architecture, CNN-BoVW-SVM, consists of a convolution neural network (CNN) for predicting the textual bullying content and a support vector machine (SVM) classifier trained using bag-of-visual-words (BoVW) 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. The processing of textual and visual components is carried out using the hybrid architecture and a Boolean system with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of text and image bullying truth value. The main target is to achieve a well accuracy for the detection of cyberbullying after performing tuning of different hyper-parameters.

1.3. ORGANIZATION OF THESIS

The project report has been divided into five chapters. Each chapter deals with one component related to this thesis. Chapter 1 being introduction to this thesis, gives us the brief introduction about the project, thereafter chapter 2 tells about the literature survey which further includes related work section. Following up is chapter 3 which tells about the proposed work. Chapter 4 provides us with the experiments and results followed by final chapter, chapter 5, which is the conclusion of the thesis.

CHAPTER 2 LITERATURE SURVEY

The chapter explains various kinds of cyberbullying, the work done so far in this field and various background concepts.

2.1. TYPES OF CYBERBULLYING

Cyberbullying can be classified into direct Cyberbullying and indirect Cyberbullying [13]. Direct cyberbullying is to harass or humiliate a person directly either through email, SMS etc. Indirect Cyberbullying is done by posting harmful of humiliating content related to someone on SM. There are various types of cyberbullying shown in Fig.2.1.

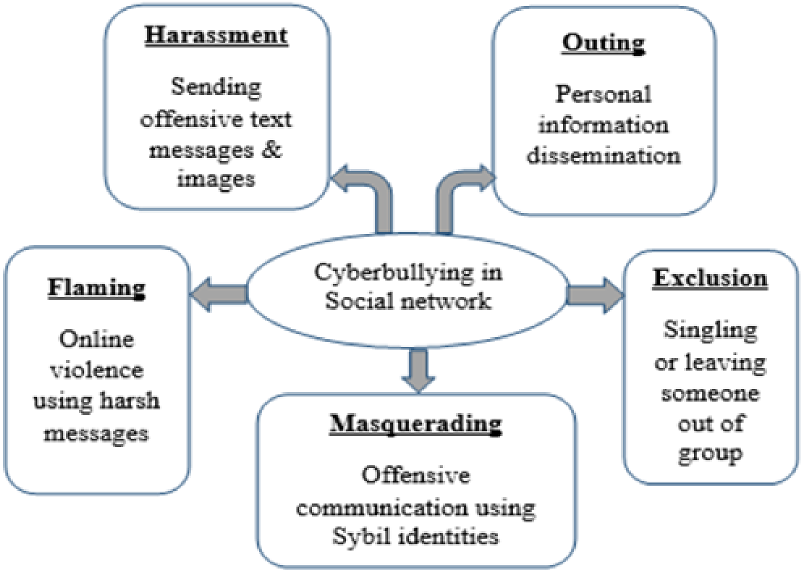

Harassment
Sending offensive text messages & images
Outing
Personal information dissemination
Cyberbullying in Social network
Flaming
Online violence using harsh messages
Exclusion
Singling or leaving someone out of group
Masquerading
Offensive communication using Sybil identities

Fig. 2.1 Various ways of Cyberbullying in Social Network

Fig.2 shows the various ways of cyberbullying generally occurs on the social network. As misinformation, insults and rumours can be immediately disseminated to a large audience, cyberbullying done in social networking websites is particularly painful and hurting for victims. Fig. 2.2 below illustrates the kinds of bullying in social web.

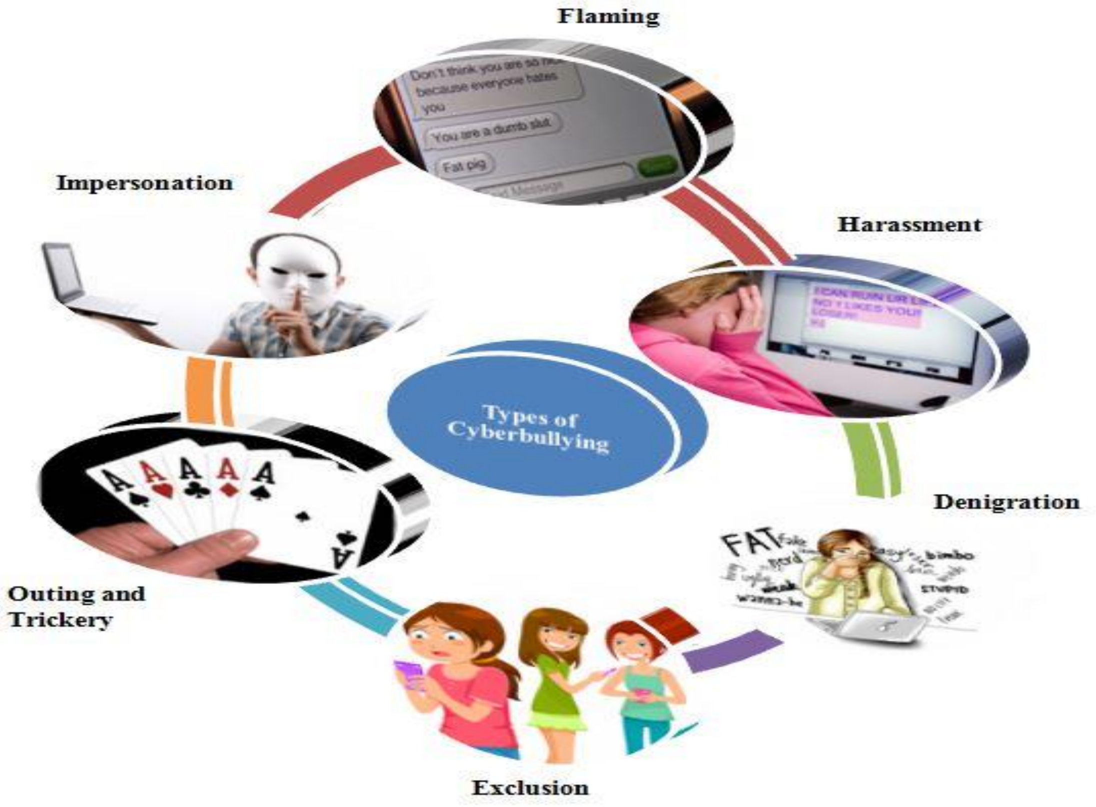


Fig. 2.2 Types of Cyberbullying

- ***Flaming*** : It is defined as online fight between people posting messages that contain offensive language.[13]
- ***Harassment*** : Harassment refers to posting messages in order to insult or threat the victim.[13]
- ***Denigration*** : Denigration is to spread rumors in order to harm the reputation of victim.[13]
- ***Impersonation*** : It is to impersonate the victim and using this fake identity in damaging way.[13]
- ***Outing and Trickery*** : It is the process of acquiring the trust of victim and the violating it by disclosing secrets of victim.[13]
- ***Exclusion*** : Excluding the victim from groups or other online activities is called Exclusion.[13]

2.2. RELATED WORK

Today, because of enormous development of web2.0, online presence of individuals is normal and permanent. Additionally the risk of cyberbullying and the pessimism brought about by cyberbullying is expanding. Hence a lot of research is currently being done around there, particularly for Cyberbullying Detection. Recent literature accounts the use of supervised machine learning and deep learning techniques for classifying hate speech, aggression, comment toxicity and bullying content on social forums [14, 15]. A portion of the work that offers sight to this issue is done by scientists in [16-18]. Theoretical aspects of cyberbullying and how it is prevailing among youths and youngsters have been examined in [19]. The characterstic profile of wrongdoers and victims and conceivable strategies for its preventions are introduced in [18]. Till now the majority of the work is devoted to text analysis for the most part. The work done in [9, 19, 20-24] basically is a research on Cyber-Aggression that has utilized text investigation approach on the comments. Work done in [22, 25-27] uses text based analysis for identification of cyberbullying utilizing the dataset from formspring.me and Myspace. Dinakar et al. in [19] construct a BullySpace, a common sense knowledge base that encodes particular knowledge about bullying situations and analyse the messages on Formspring (a social networking website) using AnalogySpace common sense reasoning technique. Hinduja et al. [17] explored the relationship between cyberbullying and suicide among adolescents. The characteristic profile of wrongdoers and victims and conceivable strategies for its preventions are introduced in [18]. Till now the majority of the work is devoted to text analysis for the most part. The work done in [19-24] basically is a research on Cyber-Aggression that has utilized text investigation approach on the comments. Work done in [22, 25-27] uses text-based analysis for identification of cyberbullying utilizing the dataset from Formspring.me and Myspace.

Yin, et al. in [28] use contextual and sentiment features for analyzing text and also noticed improved performance using these features for harassment detection achieving an accuracy of 50% on Kongregate, Myspace and Slashdot dataset. A very similar work determined an accuracy of 80% in the detection of bullying comments on YouTube using textual context features [9, 29]. Recently deep learning-based methods have been applied for cyberbullying detection. Agrawal et al. [14] have performed experimen with four Deep Neural Network (DNN) models for cyberbullying detection that include CNN, LSTM, BLSTM, and BLSTM with attention on various social networking platforms such as Formspring, Twitter, and Wikipedia. Huang et al. in [30] concatenated social network features with textual features for improved performance of cyberbullying detection. Authors analyzed attributes or features such as count of friends, network embeddedness, and centrality of relationship with the textual features and observed significant improvement in performnce. Some rule based classifications for the identification of bullies have also been done in [31] using FormSpring data set. A new sentence-level filtering model was established by Xu, et al. which semantically eliminates bullying words in texts by utilizing grammatical relations among words [32] on the YouTube dataset. Works have been done 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 [33].

Cyberbullying is prevailing and research is still carrying out in this area in a broad way, however, it is still rare to detect cyberbullying using mixed approach of image and text. Different aspect of cyberbullying detection is emphasized by survey of the literatures. Dinakar and Birago Jones et al. in [34] proposed the study of various aspects for social network and solutions to overcome the problem of cyberbullying by giving suggestions on the theoretical basis. None of the specific approach has been discussed Maral Dadvar et al. in [20] incorporated the users' characteristics information and post-harassing behavior. Users' behaviour is analysed by Cross-system i.e.; detection of cyberbullying can be done more accurately if the reactions of users in various social media platforms are monitored. This suggested technique for detection of cyberbullying performed by text analysis using few words occurrence in the comments or post in social network. Comment made by male or female is also one of the analysis factor included in this research. This research on detection of bullying content using text analysis with suggested techniques has been experimented with only few words. Thomas Vanhove et al. [35] proposed a pluggable architecture with components having features of reusability, mostly able to detect harmful content quickly. The platform uses text, image, and audio and video-based analysis modules to detect inappropriate content. Rybnicek M. et al. in 2013 proposed a research application providing the platform for detection of cyberbullying [36]. Research in this paper mentioned suggestions in the research platform is including the text, audio, video, images and addition to that social media analysis has been covered but for the implementation, the domain services aggregate this data and flag user profiles if necessary. Moderators on social network check the validity of the flagged profiles. The advance topic covered in this research that it provides the efficient knowledge upon key requirements of the platform, the essential components for architectural design and challenges to deal with cyberbullying using audio, video, and text and images but it only the survey of suggested techniques cannot guarantee for the efficient result as implementation strategies have not proved the result analysis.

The pertinent literature reports automated mono-modal models of cyberbullying detection in social media. The intent of this research is to build a three-in-one modality model which not only predicts cyberbullying in textual or visual content but also mix modal info-graphic content. A similar theoretical framework was proposed by Kansara et al. in [12] with Bag-of-Visual-Word (BoVW) model, local binary pattern (LBP) and SVM classification for image analysis and Bag-of-Word (BoW) model with Naïve Bayesian classification for text analysis. The framework lacked any implementation andresult details. This research offers a comprehensive model using a hybrid of deep learning for text analytics and BoVW with supervised machine learning for image analytics. In addition, we examine the image as well as utilize info-graphic property of the image (information which is the content/text embedded on that picture) to predict bullying content. However, as far as we know, apparently, there isn't any other work that is based on image features combined with text analysis in cyberbullying identification. Woks have been done 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 [37] however no such work has been done yet where all the potential outcomes of cyberbullying has been considered in one framework, so far either just image has been investigated or just content yet here with those two sorts of inputs, image with text embedded on it is additionally considered and after that choice is made where both image as well as text are contributing to the cyberbullying detection. A framework has been proposed by Kansara et al. in [12] where picture and comments both are mulled over yet no such practical implementation has been done yet.

2.3. BACKGROUND CONCEPTS

Some of the concepts are explained below that forms the base of our work.

2.3.1. BoVW

Bag of visual words (BOVW) is generally utilized for classification of images. Its idea is adjusted from retrieval of images and Bag of words (BOW) approach of Natural language processing. In BOW, we check quantity of each word shows up in a file or document, then use the count of those words to get the knowledge about the keywords present in the document and finally a histogram is prepared using those keywords. Document is treated as Bag of words. A very similar idea is BOVW, yet rather than using just typical words, here image features are used as the "words". The particular pattern which we can obtain from an image is the image features which are unique for every image. The basic thought of BOVW is representing an image as a set of features. The Features comprises of key-points and descriptors. Key-points are the "emerge" focuses of a picture, so regardless of the picture is turned, contract, or grow, the key-points of the image will dependably be the equivalent. Also, descriptor is the depiction of the key-points. We utilize the key-points and descriptors to build vocabularies and show every image/picture as a frequency histogram of features present in the picture. From the frequency histogram, later, we can locate another comparative picture or foresee the classification of the image.

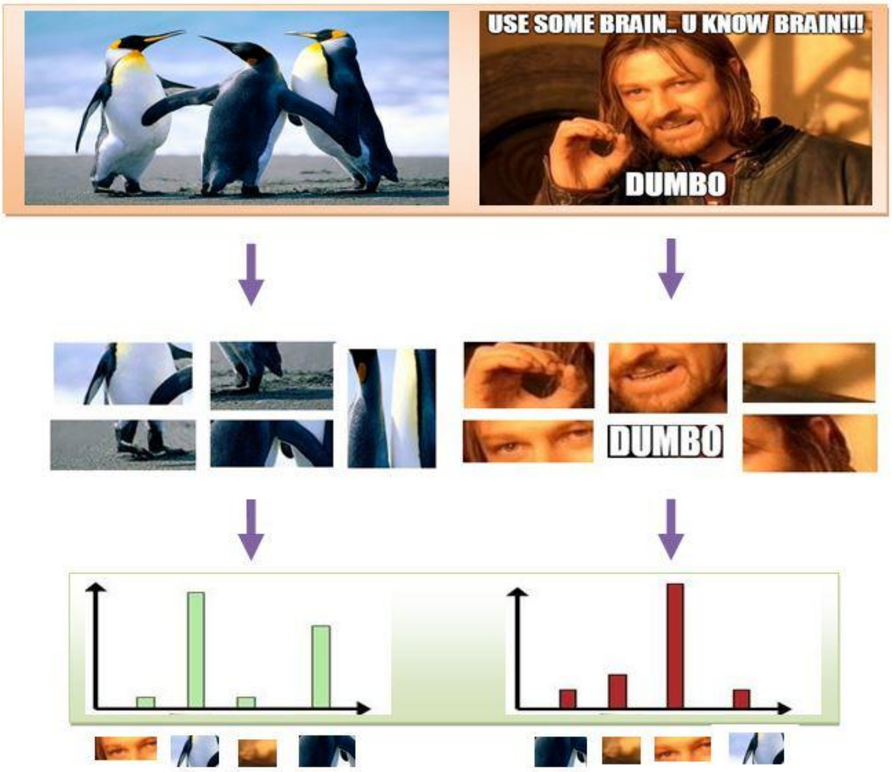


Fig. 2.3 Visual words Histogram

Features are detected, descriptors are extracted from every picture in the dataset, and a visual word references or a visual dictionary is constructed. Various algorithms of feature extraction can be used for the detection of features and extraction of descriptors. Those algorithms include SIFT, KAZE etc.

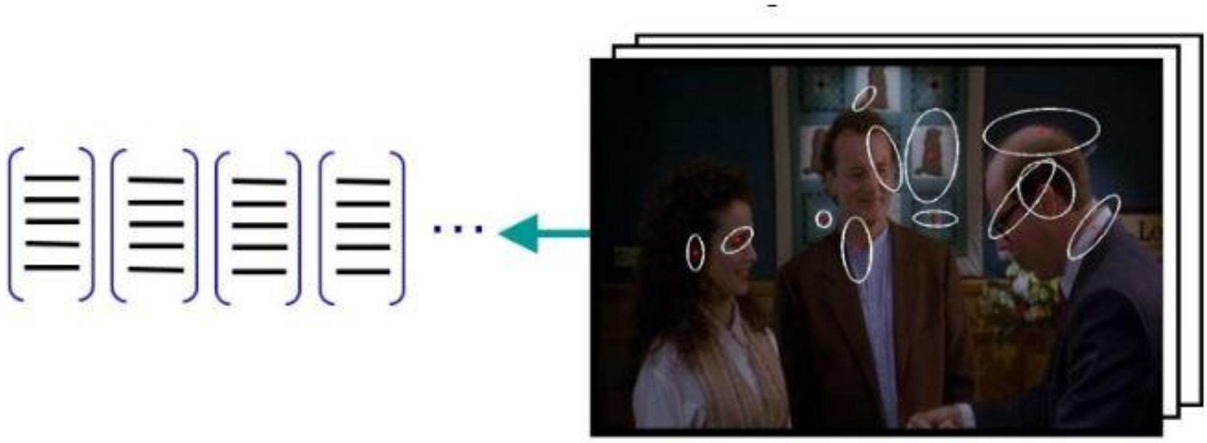


Fig. 2.4 Feature Detection and Extraction of Descriptors

Next step is to perform clustering using K-Means or any other clustering algorithm. Various clusters are made from the feature descriptors. The center of each cluster will be utilized as the visual word reference's vocabularies. Ultimately, a frequency histogram is created for every image using the vocabularies and the frequency of the vocabularies. The obtained histogram is our Bag of visual words.

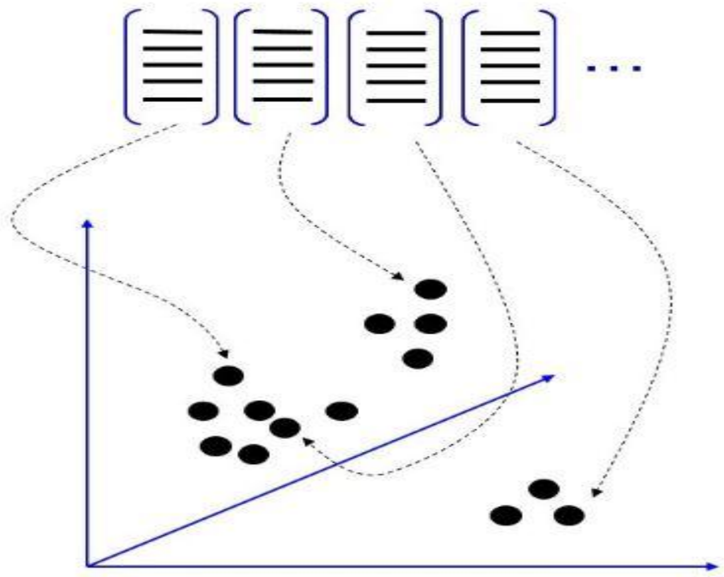


Fig. 2.5 Clustering of Descriptors

2.3.2. CNN

The CNN is just one of the feature-extraction architecture, alone itself is not useful, but is the first building block of a larger network. It is required to be trained together using classification layer for producing some useful results.

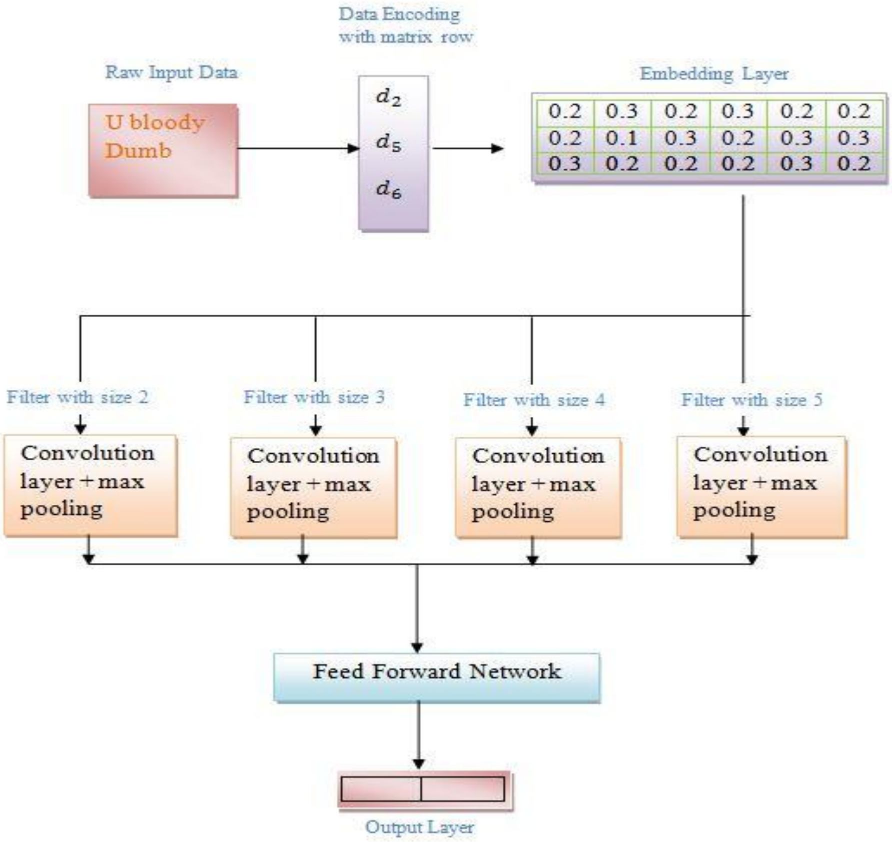
As summarized by Yoav Goldberg , “The CNN layer’s responsibility is to extract meaningful sub-structures that are useful for the overall prediction task at hand. A convolutional neural network is designed to identify indicative local predictors in a large structure, and to combine them to produce a fixed size vector representation of the structure, capturing the local aspects that are most informative for the prediction task at hand. In the NLP case the convolutional architecture will identify n-grams that are predictive for the task at hand, without the need to pre-specify an embedding vector for each possible n-gram.” The concepts used in Convolution Neural Network consist of various terminologies which are briefly defined as:

- **Convolution**: Applying filter to a fixed size window is the task of convolution operation.
- **Convolution Filter**: It is also known as convolution kernel. It is basically a matrix that is utilized for performing convolution operation.

- **Pooling**: It is the process of combining the vectors obtained as a result of various convolution windows into a vector single one dimension.
- **Feature_maps** : The significance of number of feature maps is that it directly controls capacity and is dependent on count of available examples and complexity of task.

CNN for Performing Text Analytics

The deep neural architecture for text analytics is shown in fig.2.6. The figure describes the process of applying CNN over text in order to perform the task of classification. This demonstrates a step by step explanation of every process involved during the applicarion of CMNN algorithm.



0.2	0.3	0.2	0.3	0.2	0.2
0.2	0.1	0.3	0.2	0.3	0.3
0.3	0.2	0.2	0.2	0.3	0.2

Fig. 2.6 CNN for text classification process

For the application of CNN over text, its components should be encoded before it is given as input to CNN. For this, a vocabulary is used which is formed as an index having words that appear in the posts/comments. Each word is mapped to an integer between 1 and vocabulary length. An example text encoding using vocabulary is shown in fig. 2.7.

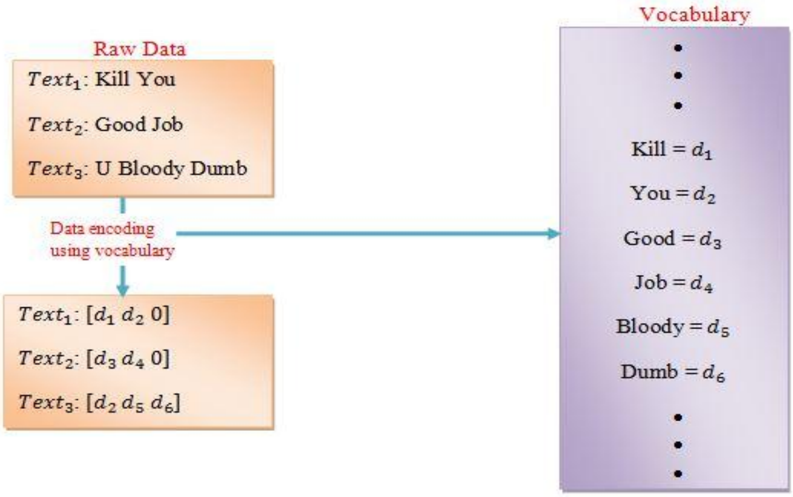


Fig. 2.7 Text Encoding using Vocabulary

Padding is used to maintain the fixed input dimensionality feature of CNN, in which zeros are filled in the matrix to get the maximum length amongst all comments in dimensionality.

In the next step the encoded texts are transformed into matrices where each row represents one word. The constructed matrices pass through the embedding layer where each word (row) is converted into a lower-dimensional representation by a dense vector. Next step is to perform convolution, one way to think of convolution is that we're sliding the filter over the input text. For each position of the filter, we are multiplying the overlapping values of the filter and text together, and add up the results. This sum of products will be the value of output at the point in the input text where the filter is centered.

Let us assume we have a post or comment of length m denoted as $X_{1:m} = X_1, X_2, \ldots., X_m$ where $X_1, X_2 \ldots, X_m$ are the words of sentence represented as a k dimensional vector. Concatenation of those vectors is a matrix represented by $X_{1:m}$. Using a filter $W \in h\ x\ k$ of height h or a window of h words, a convolution operation on h consecutive word vectors starts from t^{th} word outputs the scalar feature (equation 2)

$$c_t = f\ (W^T \cdot X_{t:t+h-1} + b_f) \qquad (2)$$

where, $X_{t:t+h-1} \in R^{h\,X\,k}$ is the matrix whose i^{th} row is $X_i \in R^{\,k}$ and $b_f \in$ R is a bias. The symbol ·refers to the dot product and f is linear unit function used.

We perform convolution operations with n different filters, and denote the resulting features as $c_t \in R^n$, each of whose dimensions comes from a distinct filter. Repeating the convolution operations for each window of h consecutive words in the text, we obtain $c_{1:\,m-h+1}$. Next, Pooling is done which is generally max-pooling where the most important activation is captured from the obtained convolution output. A short-text representation $s \in R^n$ is computed in the max pooling layer, as the element-wise maximum of$c_{1:\,m-h+1}$. An n-dimensional representation of text is finally obtained after this operation (Fig. 2.8).

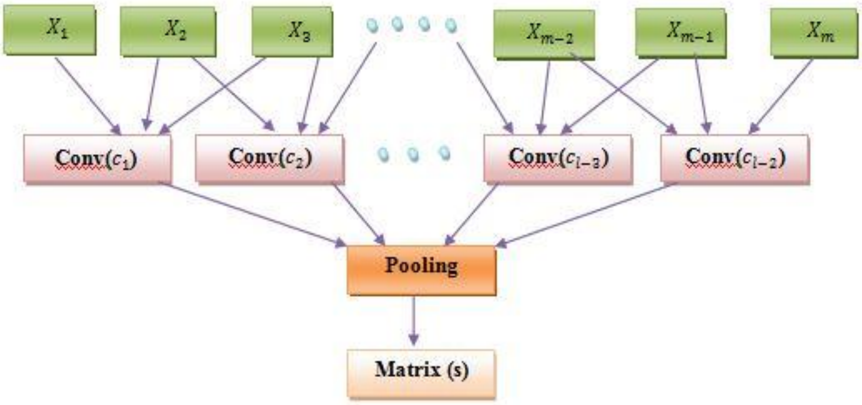


Fig. 2.8 Convolution and pooling

The process then continues following the generic CNN model like passing this obtained n-dimension matrix to the feed forward network and finally result is obtained by the output layer.

2.3.3. SVM

A Support Vector Machine (SVM) works by finding a hyper-plane that can efficiently divide the set of objects in different classes.SVM takes a labelled training data, and outputs an optimal hyper-plane which can then be used to categorize new examples. A decision plane separates set of objects having memberships of different classes. Generalization error of the classifier is lowered if the distance between the separating planes to the closest training data point belonging to any class is largest. The query points or test points are then mapped to that particular space and predicted to belong to a category depending on the side of the gap they then fall on as shown in Figure 2.9.

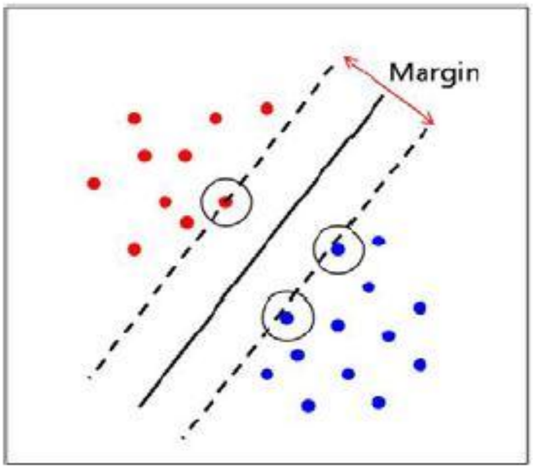


Fig. 2.9 SVM Classification

For the case of multidimensional space, SVM finds the hyperplane that maximizes the margin between two different classes. A few samples control the decision boundary. There are only a few training samples that touch the decision boundary. These are the ones that actually control the decision boundary and are known as support vectors. Here the support vectors are those dots that have been circled. One of the important advantages of using SVM classification is that it performs very well on datasets having many attributes, even when there are just a few cases that are available for training process. Speed limitation and size limitation during the phases of algorithm training and testing are some of the disadvantages of SVM.

CHAPTER 3 PROPOSED METHOD

The proposed deep classification model reinforces the strengths of deep learning nets in combination to machine learning to deal with different modalities of data in online social media content. The proposed CNN-BoVW-SVMmodel consists of four modules, namely, text analytics module, image analytics module, discretion module and decision module. The basic architecture of the work done has been clarified beneath (Fig 3.1). The steps included can be comprehended as

1. Analyzing the sort of information.
2. Passing it to the separate module for processing.
3. Decision module is utilized to analyze the outcome.

Analyzing the sort of information includes checking whether the input is just text or it is a picture or it is a picture with text embedded on it. This is vital in light of the fact that once we have investigated this then we can perform further handling in the respective modules.

Text only: On the off chance that the input is as just text, at that point we will perform pre-processing of text, extract the features, create the feature vector and after that utilization CNN is used for performing the task of classification.

Image only: If the input is image only, at that point we will perform the pre-processing of image like converting it to gray scale or resizing, then extricate features utilizing BoVW approach, produce histogram and after that uses SVM for doing the classification of the image.

Image with Text: If the input is the image with text embedded on it on it, at that point an additional step will be included to separate that text from the picture, which we are doing utilizing Google Photos as a tool. When we have the text separated from the image, we can utilize the means utilized for performing text analysis and for the picture we will utilize the image handling steps. The result of those two modules will be encouraged as contributions to a Boolean framework that will at that point shows the outcome whether it is a bully or not.

When one of the input to Boolean framework isn't accessible like we have only text or in the event that we have picture just, at that point that input to Boolean framework will be unfilled or false since we are utilizing an OR operation in the Boolean framework, if the text or image is a sort of bully, it will get identified.

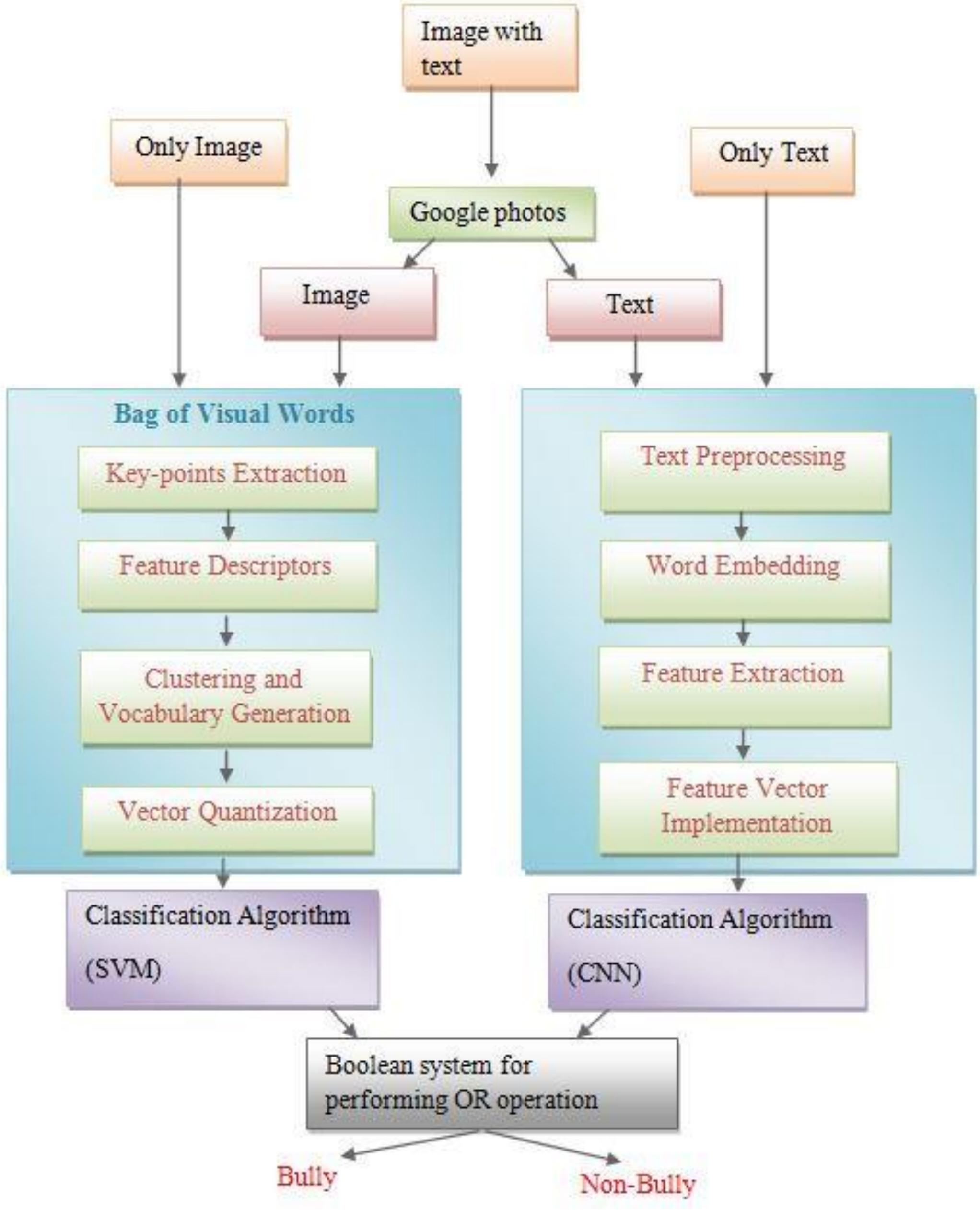


Fig. 3.1 Proper flow of proposed model

The whole architecture can be understood in three sub-modules

1. Text Analytics Module
2. Image Analytics Module

3. Discretization Module
4. Decision Module

3.1. TEXT ANALYTICS MODULE

Deep learning architectures have proven capabilities for extrapolating new features from a limited set of features contained within a training set, without human intervention and without the need to label everything. These have given superlative results in comparison to conventional machine learning techniques for various natural language processing task [38].To analyze textual bullying content the model uses a convolution neural network (CNN). A CNN is a deep neural architecture which works using multiple copies of the same neuron in different places. It has the power of self-tuning & learning skills by generalizing from the training data. CNN model enhances feature extraction in online posts/comments, which improves the generic classification task. The text analytics process is shown is fig.3.2.

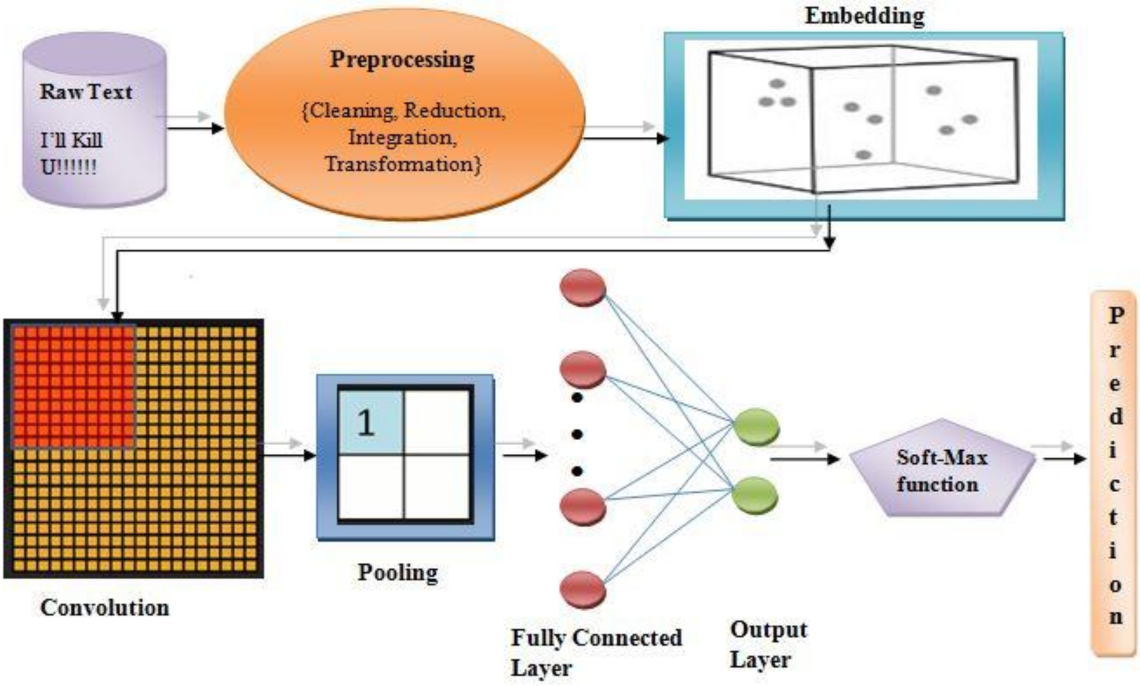


Fig. 3.2 Text analytics module

3.1.1. Preprocessing: The textual data can be of any length and may contain misspelled words, emoticons, special symbols etc. All these words are trivial and exemplify noise. Pre-processing is an important step in text classification [39]. It includes removal/replacement of emojis, replacing urls and hashtags with keywords, tokenization, stop words removal, lemmatization, lowercasing and stemming.

3.1.2. Word Embedding: The pre-processed posts are input into the embedding layer. The feature representation and extraction in the CNN is learned in a hierarchical way using word embeddings making it distinctive and better than the lexical or syntactic feature extraction. The embedding layer thus uses GloVe [40] to build word embeddings and the model learn geometrical encodings (vectors) of words in each post. We run our model on top of GloVe word embedding using 100 dimension representation of word. We train the system to learn the vectors for each word (which would be represented as one hot vector initially), thus we convert each word to a vector of integers of 100 dimensions and thereforewe have a comment matrix of size equals to number of words in the vocabulary multiplied by 100. Now our text data is in the form of numerical data that can further be used for performing convolutions.

3.1.3. Convolution and pooling: For textual data, we need convolution for one dimension only unlike image where 2d convolutions work well so convolution in 1d can generally be defined as (equation 1)

$$(g * h)[n] = \sum_{i=-m}^{m} g[n-i]h[i] \qquad (1)$$

where, g is the input vector which we have obtained after applying word embedding and length of input vector g is k, h is the filter or kernel used whose length is m. we usually multiply the terms of g[n] by the terms of a time-shifted h[i] and add them up [41].

3.1.4. 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 bully (+1) or non-bully (-1).

3.2. IMAGE ANALYTICS MODULE

Local Binary Pattern which is a kind of visual descriptor is used for detecting any image. Using the clustering algorithm like k-means, visual vocabulary is generated, where similar kind of features forms the center of cluster and becomes one visual word. LBP is used to extract the features and then those features are mapped to the existing visual word in vocabulary or codebook. Eventually vector of visual word frequencies is generated. The occurrence of few visual words provides certain hints for the presence of offensive material in image. Finally, the SVM classifier is used for bullying classification. The image analytics process is shown is fig.3.3.

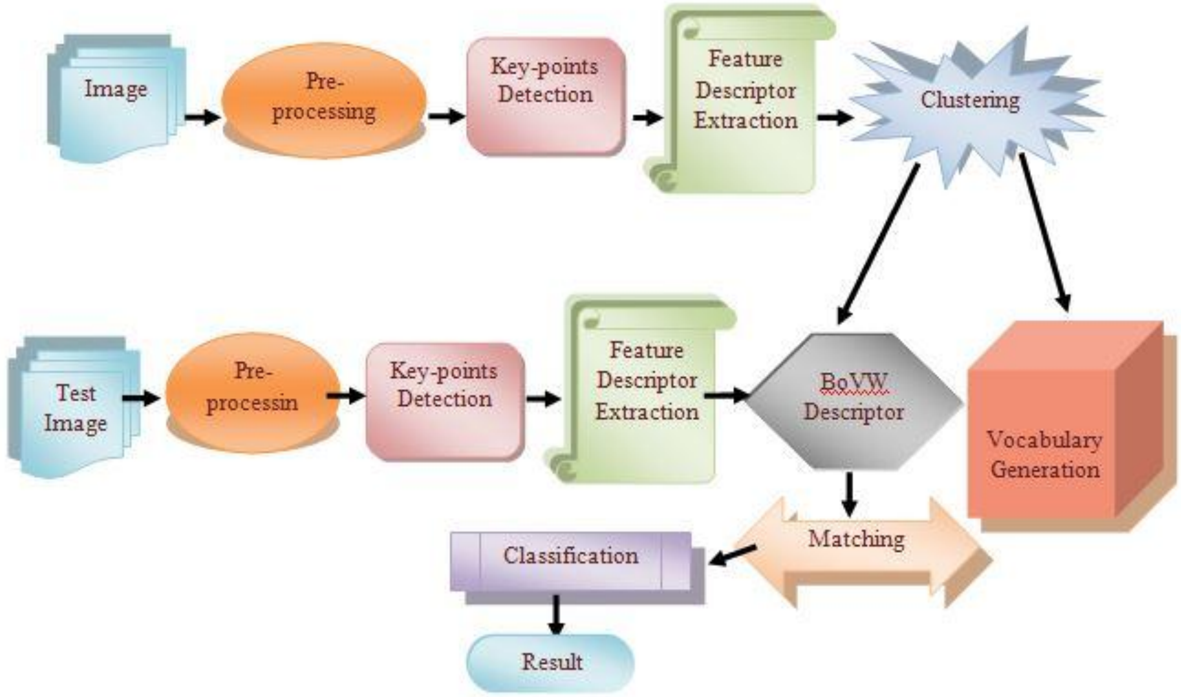


Fig. 3.3 Image Analytics Module

3.2.1. Pre-processing: Similar to text pre-processing, prior to performing any analysis on the image, pre-processing is done. Pre-processing generally transforms the image in a form that is suitable to work on. Here the noise from the images is removed, images are resized and conversion of images to gray scale is done.

3.2.2. **Feature extraction:** Images exhibit some local points of interest generally around edges and corners. Local descriptors are used for describing the local points. The detected points are described by local descriptors. Feature extraction is done using local binary pattern which itself uses texture analysis. Local binary pattern is basically to threshold the window with the center pixel value. This encodes local contrast and pattern making it highly discriminative. Also, it is easy to compute. For constructing feature vector, image is divided into blocks of say 16 x 16 pixels or 32 x 32 pixels for each cell. Every pixel in a cell is compared with all its eight neighbors (that is pixels in top-left, top, top-right, left etc.) The value is assigned following a rule that wherever center pixel's value is more than the neighbor pixel's value, the value 1 is written else 0 is written in its 8 x 8 neighborhood. An 8-digit binary number is obtained this way which is then converted to decimal for ease of understanding. And then that number is assigned to the center pixel. Then histogram is computed, for each block and it is concatenated to get the feature vector for the image. Normalization of histogram can be performed before performing the concatenation. Since Bag-of-Visual-Words (BoVW) is considered as an order less collection of features therefore information regarding the spatial layout of features is discarded and a limited description is provided by this approach. To be more precise, it can be said that from object’s background, BoVW can’t break that object. To remove this drawback of the basic BoVW model, spatial pyramid matching can be used which sub-divides the image repeatedly and form histograms at finer resolutions. This way the spatial information can be encoded in the BoVW model. This results in feature extraction with great accuracy.

3.2.3. Visual vocabulary - (BoVW) model

BoVW model can be defined as an unordered collection of image features. It is similar to the Bag of Words representation used in information retrieval for textual data. It is a representation of histogram made from independent features.

This model works in two steps coding and pooling. The process of hard assigning each local descriptor to the closest visual word is coding. Pooling is the process of performing the average of the local descriptor projections. After these steps, finally a histogram is generated counting the occurrence of each visual word in the image. Each image is represented by various local patches and in order to represent those patches, vectors are generated using feature representation methods like sift or surf [42]. Those vectors are known as feature descriptors. After obtaining the feature descriptors, now we have vectors of same dimension for every image. Now these vectors or feature descriptors are mapped to the visual words that belong to visual vocabulary. Vocabulary size can be defined as the number of visual words existing in the dictionary. There are various visual words that occur in the vocabulary and those visual words can be very similar to each other thus we need to find a visual word so that it can act as a representative for the several same kind of image patches. This is done using the k-means clustering process and then each cluster represents a visual word. And then the histogram of the image is generated based on the visual word and their frequencies. Finally a classifier, which is SVM here, is used to classify a image as bully or not.

So the overall BoVW model can be summarized in these three steps

- Detect feature descriptors from training images and cluster them with k-means.
- Test the training and testing images feature descriptors in the k-means classifier and make a histogram of classification results.
- Use these histograms as feature vectors for SVM classification.

3.2.4. Classification: The image analytics module uses a supervised learning Support vector machine (SVM) trained using the BoVW features to predict bully or non-bully image. A Support Vector Machine (SVM) works by finding a hyper-plane that can efficiently divide the set of objects in different classes.SVM takes a labelled training data, and outputs an optimal hyper-plane which can then be used to categorize new examples. A decision plane separates set of objects having memberships of different classes. For a 2d space, this hyper-plane or decision boundary is a straight line. In this image analytic module, SVM analyzes data and recognizes image patterns. A set of training examples is provided to the algorithm and it generates a boundary in order to differentiate between the classes learning from training examples. Whenever a new image is given as input, SVM applies the learned rules and then if particular image contains some bully or abusive content, that image gets placed in the bully class else in the non-bully class.

3.3. DISCRETIZATION MODULE

If the input is an info-graphic post/ comment, which is the image with text embedded on it, the CNN-BoVW-SVM model utilizes a Google Photos App to extract text from an image. This visual analysis tool separates the text from the image which is then processed as discrete entities sent to the respective text analytics and image analytics modules.

3.4. DECISION MODULE

The bullying content prediction of mono-modalities (text and image separately) is done by the respective classification models. An additional decision system is augmented to cater the multi-modal info-graphic content. This decision system is a Boolean system with an OR operation that resolves the output as bullying or not. The hypothesis for the same is based on the fact that if either of text or image is classified as bullying, any shift in class (bullying to non-bullying or vice versa) between the textual and visual modality types should not alter the classification. The logical operator requires at least one of two inputs to be present and if we have only text or only image as input then the respective second input for the system is by default 0. Table 3.1 depicts the logical OR.

Table 3.1: Boolean Decision System

Modality	Text Classifier	Image Classifier	Classification
Text Only	+	Null	Bullying
	-	Null	Non-bullying
Image Only	Null	+	Bullying

	Null	-	Non-bullying
Info-graphic	+	+	Bullying
	+	-	Bullying
	-	-	Non-bullying
	-	+	Bullying

+: Positive of bullying content; **-:** Negative of bullying content

CHAPTER 4 IMPLEMENTATION AND RESULTS

The data used here is in generally three forms with percentages shown below. The dataset prepared for experiments contains 10000 data out of which 55% data is in the form of text, 21% is in image form and there are 24% of data that contains image embedded with text as shown in Fig. 4.1. Table 4.1 below shows actual distribution of data in numbers. The dataset is prepared using the social media sites like Youtube, Instagram and Twitter. Comments and posts from these sites have been gathered and used for experiments.

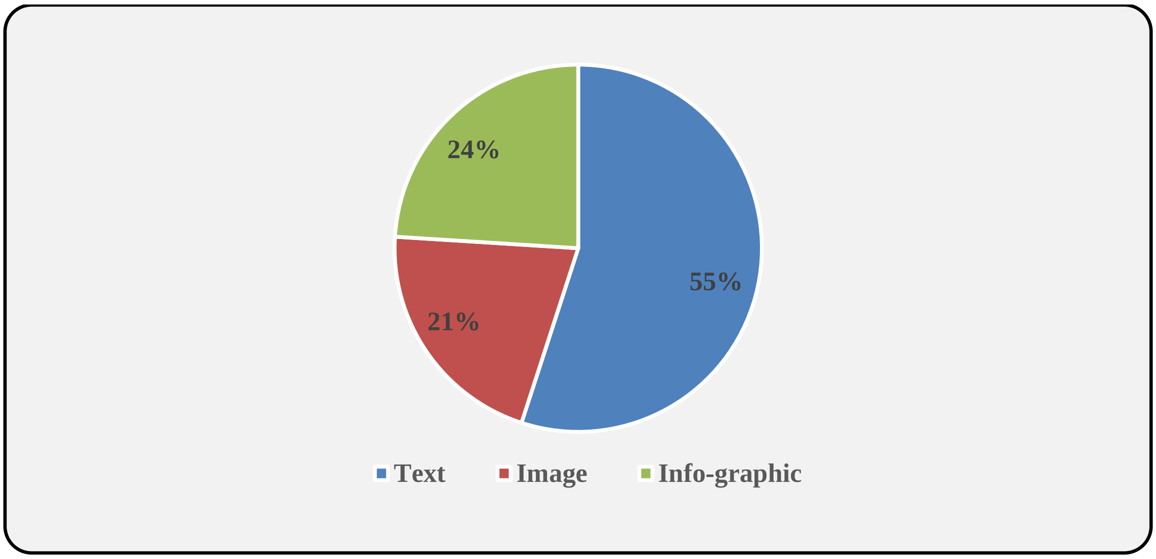


Fig. 4.1 Distribution of Data

Type of modality	Number of instances	
	Bully	**Non bully**
Image only	1260	840
Text only	3300	2200
Info-graphic	1440	960

Table 4.1: Categorization of data used in input

After passing the image to the image module defined above, a histogram is generated that is shown below in Fig.4.2. We have used the local binary pattern SIFT for extracting the features of the images and clustering is done using k-means algorithm.

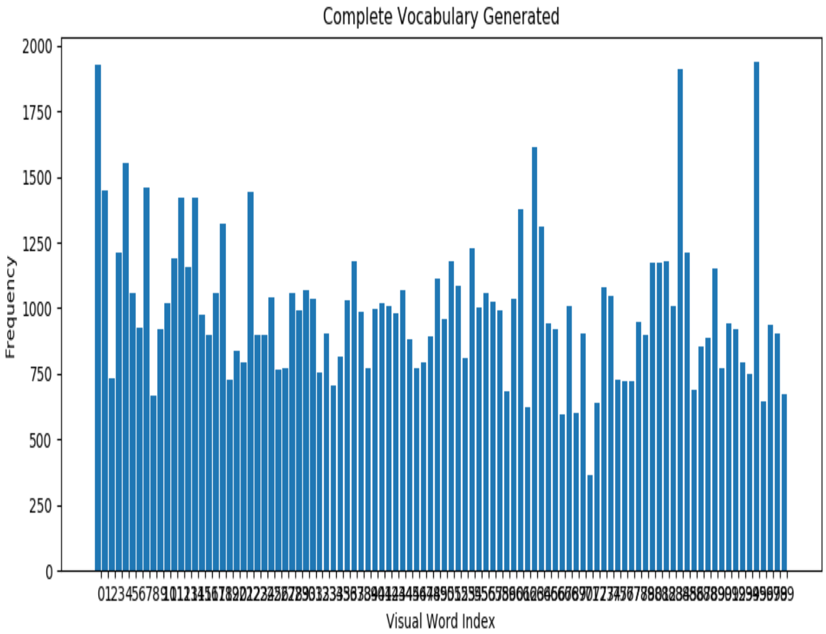
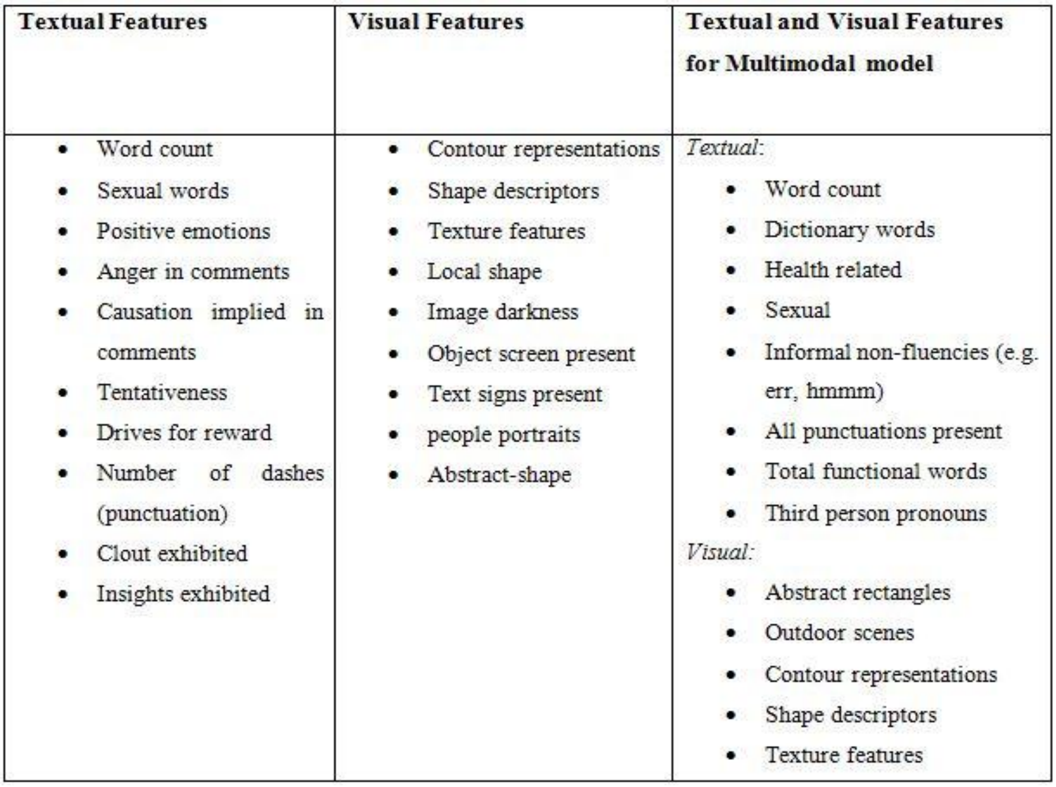


Fig. 4.2 Vocabulary Histogram

Features used by the algorithm for the analysis of text, image and text+ image type of modality are represented below in Figure 4.3.



Textual Features	Visual Features	Textual and Visual Features for Multimodal model
• Word count • Sexual words • Positive emotions • Anger in comments • Causation implied in comments • Tentativeness • Drives for reward • Number of dashes (punctuation) • Clout exhibited • Insights exhibited	• Contour representations • Shape descriptors • Texture features • Local shape • Image darkness • Object screen present • Text signs present • people portraits • Abstract-shape	*Textual:* • Word count • Dictionary words • Health related • Sexual • Informal non-fluencies (e.g. err, hmmm) • All punctuations present • Total functional words • Third person pronouns *Visual:* • Abstract rectangles • Outdoor scenes • Contour representations • Shape descriptors • Texture features

Fig. 4.3 Features Selected for Cyberbullying Detection

4.1. PERFORMANCE MEASURES

We have used three evaluation metrics: Precision, Recall, Accuracy.

Precision: It is the ratio of data elements that are correctly classified (for both the minority and majority class) to total number of classified instances.

$$P = TP/\ (TP + FP)$$

Recall: The ratio of the minority class instances that are correctly classified to the total number of actual minority class instances.

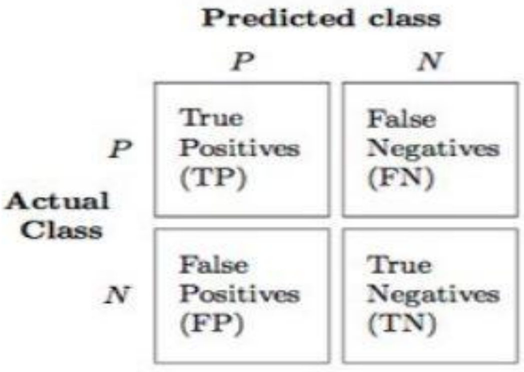
$$R = TP/\ (TP + FN)$$

F-Measure: Precision and Recall are used for performing the calculation of F-measure**.** It is calculated by taking the harmonic mean of Precision & Recall. We can say that it is essentially an average between the two percentages. It really simplifies the comparison between the classifiers.

$$F\text{-}measure = 2\ /\ (1/R + 1/P)$$

CONFUSION MATRIX

It maps the relation between what the model has predicted and what the actual result should be as shown in fig 4.4. If the predicted class is positive and actual class is positive as well, then we get the true positive section. If the predicted class is positive but actual class is negative then we get false positive section, on similar bases if the actual class is positive but the predicted class in negative then it is false negative and if the actual class is negative and predicted class is also negative we get true negative section.



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Fig. 4.4 Confusion Matrix

CNN is used for the input which is in the form of text. Experiments have been performed several times by using set of different parameters in order to get better results. The Table 4.2 below represents some of those set of parameters, it also shows the parameter setting for which CNN obtained best results.

Embedding Dimensions	Filters	Hidden Dimensions	Batch Size	Epochs	Speed	Accuracy
100	150	350	64	5	26 µs/step	93.00 %
50	150	350	64	3	40 µs/step	92.36%
75	75	200	64	3	25 µs/step	92.33%
80	350	300	64	5	59 µs/step	92.31%
50	100	250	64	5	25 µs/step	91.29%

Table 4.2: Hyper-parameter tuning

Thus, various parameters have been used for both the modules of Image analysis and Text analysis during the experiment. The values of those parameters and kind of functions used can be summarized in the Table 4.3 below.

Parameter	Value
Number of Filters	150 for each size
Filter sizes	2,3,4,and 5
Drop out	0.5
Local Binary pattern	SIFT
Non-linearity function	ReLU
Word embedding	GloVe

Table 4.3: Parameters used in model

The classification results are evaluated on the basis of classification accuracy, precision and recall. It was noticed then when emoticons used in the comments or posts are taken into consideration [43-44], it improves the results by nearly 5% for the model. The result is shown graphically below in fig.4.5.

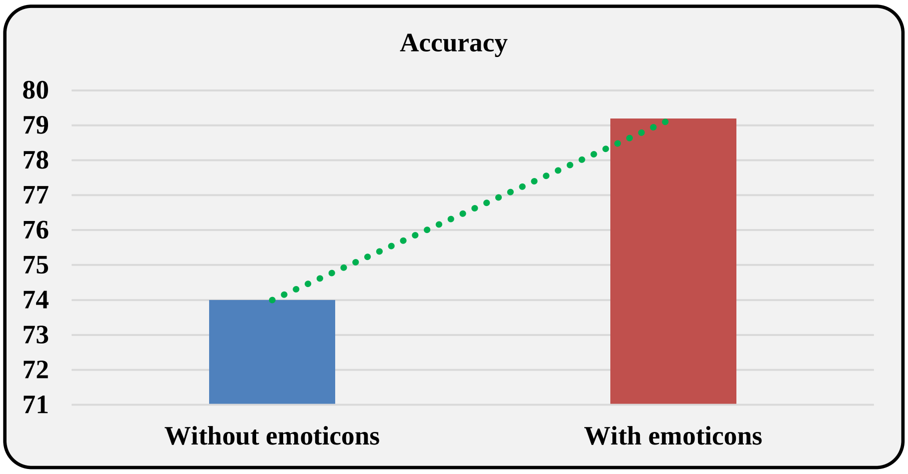


Fig. 4.5 Accuracy with or without emoticons

4.2. EXPERIMENTAL RESULTS

In our overall model the text used is the text without removal of emoticons during pre-processing. Confusion matrices for all type of modalities are shown in Table 4.4, 4.5 and 4.6.

Actual classification	**Predicted classification**			
TEXT	**Bully**	**Non-bully**	**Precision**	**Recall**
Bullying	5100	900	79.94%	85%
Non-bullying	1280	2720	75.14%	68.75%

Table 4.4: Confusion matrix for textual modality

Actual classification	**Predicted classification**			
IMAGE	**Bully**	**Non-bully**	**Precision**	**Recall**
Bullying	4750	1250	76.86%	79.17%
Non-bullying	1430	2570	67.28%	64.25%

Table 4.5: Confusion matrix for visual (image) modality

Actual classification	**Predicted classification**			
INFO-GRAPHIC	**Bully**	**Non-bully**	**Precision**	**Recall**
Bullying	5510	490	84.25%	91.83%
Non-bullying	1030	2970	85.84%	74.25%

Table 4.6: Confusion matrix for info-graphic (image + text) modality

For the textual module, five classifiers, namely, Naïve Bayesian (NB), Random forest (RF), Support vector machine (SVM), K-nearest neighbor (KNN) and Sequential Minimal Optimization (SMO) were compared with CNN. CNN achieved the highest accuracy of 78.2%. Two classifiers, namely, KNN and NB were compared with SVM using the BoVW features for the image analytics module and it was observed that SVM outperformed both the other classifiers. Comparative analysis of the classification algorithms used for discrete textual and visual modalities is given in Table 4.7 and Table 4.8.

Brief Introduction of the classifiers:-

Naive Bayes :- This classifier belongs to the probabilistic group of classifiers in the domain of machine learning. The base of this classifier is the Bayes Theorem where the features are considered to be independent of each other. It is a very popular when it comes to classification. It is a simple model where the test (unknown) instances are assigned class tags based on the trained model.

K-NN :- K-nearest neighbor model can be used as classification model or regression model. For an unclassified instance as the input we consider the k classified instances in a constraint region and accordingly the unclassified instance is given a class whose instances are most in that region. In case K=1, the unclassified instance is given the class whose neighbour is nearest to it, there is no need for count as the value of k is 1.

SVM :- A Support Vector Machine (SVM) works by finding a hyper-plane that can efficiently divide the set of objects in different classes.SVM takes a labelled training data, and outputs an optimal hyper-plane which can then be used to categorize new examples. A decision plane separates set of objects having memberships of different classes. For a 2d space, this hyper-plane or decision boundary is a straight line. In this image analytic module, SVM analyzes data and recognizes image patterns. A set of training examples is provided to the algorithm and it generates a boundary in order to differentiate between the classes learning from training examples.

SMO :- Sequential minimal optimization helped the support vector machine (SVM) with the problem of quadratic programming. It was developed at the Microsoft Research in 1988 by John Platt. SMO is used in the training phase of the SVM so as to get rid of the problem. It was quiet an important development as in early days if was very expensive to get rid of the quadratic programming problem of SVM using 3-party software.

Random Forest :- It is also known as - Random decision forests, It is an ensemble learning technique used for both regression and classification. It works by generating large number of decision trees in the training phase and in the test phase gives the result according to whether it is for classification or regression. It is better than decision tress as it removes its limitation of getting too precise depending on the training dataset. Its first creation was done by Tin Ham Ho in the year 1995.

Classifier	P	R	A
NB	79.36%	66.66%	69.6%
RF	77.50%	68.33%	69.1%
SVM	76.62%	66.66%	67.8%
KNN	76.79%	71.66%	70%
SMO	74.75%	74%	69.4%
CNN	**79.94%**	**85%**	**78.2%**

Table 4.7: Comparative Analysis of different classifiers used for text modality

Classifier	P	R	A
KNN	71.3%	71.8%	65.8%
NB	67%	69.2%	64.4%
SVM	**76.86%**	**79.17%**	**73.2%**

Table 4.8: Comparative Analysis of different classifiers used for Image Modality

The accuracy achieved for the multi-modal model is nearly 85% which is an improvement over the accuracies obtained after validating text and image modules individually as shown in Table 10.

Modality	P	R	A
Text	79.94	85	78.2
Image	76.86	79.17	73.2
Info-graphic	84.25	91.83	84.8

Table 4.9: Classification Results for Textual, Visual and Info-graphic modalities

Fig.4.6 depicts the results graphically.

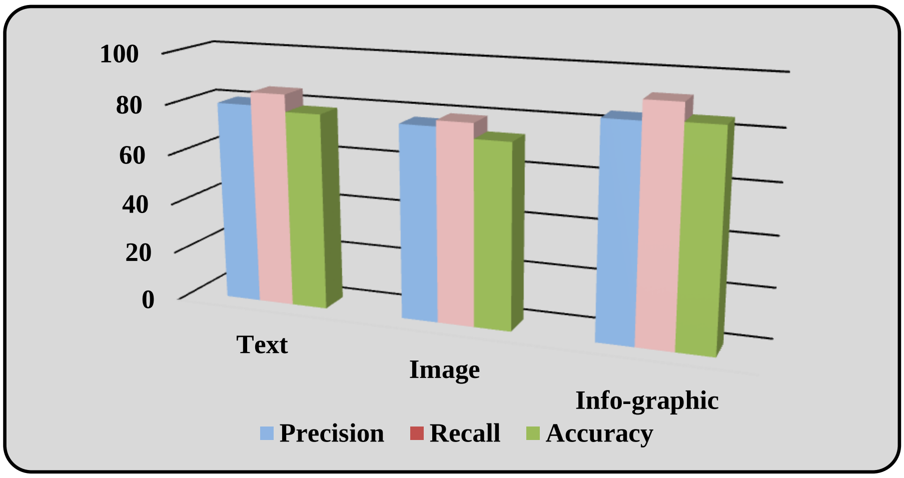


Fig. 4.6 Recall, Precision and Accuracy of proposed model

CHAPTER 5 CONCLUSION AND FUTURE WORK

In this chapter, we first briefly summarize the main work in the thesis. And then gather the findings and make some comments on them. At last, we suggest possible future work in order to better tackle the problem.

5.1 CONCLUSION

Social media and the internet have opened up new forms of both empowerment and oppression. Meaningful engagement has transformed into a detrimental avenue where individuals are often vulnerable targets to online ridiculing. Predictive models to detect this cyberbullying in online content is imperative and this research proffered a prototype model for the same. The uniqueness of the proposed hybrid deep learning model, CNN-BoVW-SVM is that it deals with different modalities of content, namely, textual, visual (image) and info-graphic (text with image). The results have been evaluated and compared with various baselines and it is observed the proposed model gives superlative performance accuracy. The limitations of the model arise from the characteristics of real-time social data which are inherently ‘high-dimensional’, ‘imbalanced or skewed’, ‘heterogeneous’, and ‘cross-lingual’. The growing use of micro-text (wordplay, creative spellings, slangs) and emblematic markers (punctuations and emoticons) further increase the complexity of real-time cyberbullying detection.

5.2 SUMMARIZATION

Our aim in this thesis is to detect whether a comment or post posted on social media would be a bully or not. The data set involved is crawled from the new and uprising social media sites like Instagram, Facebook, Youtube. We have gathered the data from all these sites and build our own data set. We had the data instances labeled according to its comment contents as cyberbullied or non-cyberbullied. Then, with the help of descriptive caption of the instance and the user information, we try to build a model to accurately classify the multimodal posts.

In chapter 2 we review the non-technical and technical studies dedicated to cyberbullying. Cyberbullying is an epidemic phenomenon and is generating severe harm to people, especially teenagers. The chapter deals with the kinds of cyberbullying, the work done in this field and also the background studies that are important for performing the analysis. In this thesis, our target media object consists of multimodal data that contains text, images and photos and their attached text information like caption and user information. Although there is barely any work trying to detect cyberbullying taking into consideration all these features, we have tried to implement such an all in one model here that will take care of all these aspects of bullying.

Chapter 3 illustrates the methodology proposed by us. We have gathered the data from different social media sites and perform the process of data acquisition and feature extraction. After the preprocessing of the data, we have them labeled as bully or non-bully. The proposed CNN-BoVW-SVM model consists of four modules, namely, text analytics module, image analytics module, discretization module and decision module. We further explained every module with the sub-modules involved within those. The explanation consists of techniques like CNN, BoVW, SVM etc. The chapter also introduces the involve features like color histogram, local binary pattern, bag of visual words and so on.

Chapter 4 is where we show the implementation details, Experimental setup and classification results. Here we have explained the setting of various parameters that has been used for performing the experiments. We have defined the proper distribution of the data as in what proportion the modalities are used in our model. Further we have analyzed our model individually for each type of modality and analyzed the results. The results are also compared by using different classification algorithms like naïve bayes, SVM, KNN etc. and observed CNN gave the best results for text modality among all the methods and for image modality SVM proved to be the best classifier. The best result obtained after setting all the hyper-parameters is 84.8 % for our model. Whereas the accuracy obtained for only text modality is 78.2% and that of only image modality is 73.2%.

5.3 DISCUSSION

According to the results, the contribution of general image features to the classification work is limited. One of the possible explanations is that the patterns of bullying are beyond merely image characteristics. On social media sites, a viewer comment under a post knowing information a lot more than just the post itself which is hard to simulate in a classification model. While trying different techniques to classify the data set and enhance the performance, we discovered a big diversity of the image content and patterns of bullying. Due to the popularity of Instagram, Facebook or Youtube users post photos or comments with different purposes. Some post to promote products, some post to report news, some are organizations and post to gain popularity among viewers and some are common individuals who post to share experiences of their life. The bully might be intrigued by the identity of a user posting the comment in form of text, photo, or the content of images that the people who comment have strong sentiment about. These factors might require specific common sense knowledge to be recognized, which is sometimes hard for others without it to see and increase the difficulty of this classification problem.

Obviously, images are much more expressive as compared with text and if the image is embedded with text image or text itself represented as image then it further describes this function of expressiveness. We presented a model for cyberbullying detection that work for both, typo-graphic or info-graphic contents as well as simple text or image in order to capture this expressiveness. The proposed model for cyberbullying detection offers uniqueness in a way that it is able to handle various dimensions in the comments like: text, image and text + image to analyze the bullying. Further, we have explored the use of deep learning technique and word embeddings for performing context-aware analysis of text. The performance results of the proposed model are motivating and improve the generic cyberbullying detection task. We have seen that if emoticons are taken into consideration then it improves the accuracy of entire model by 2.5 to 3 percent. The model works as a visual listening tool for brand management for enhanced social media monitoring and analytics. The main limitation from which the model suffers is that the text recognition for bullying is defined to only English language. As social media is a non-formal way of having communication, a prominent use of mash-up languages, like, a mix of English and Hindi is widely seen, but such content be it text or text within image could not be processed.

5.4 FUTURE SCOPE

One way to alleviate the diversity of bullying is to find other ways in labeling. In this thesis, we just label bullied/non-bullied posts by going through the comments, pictures or memes which would be affected by commenters’ preferences. To give a simple example, in sports, fans tend to verbally attack their team’s rivals whenever they get a chance. In our situation, a photo in which a football player is visiting a children’s hospital might receive some offensive comments just because there are supporters of his rival teams among the viewers. Yet the image probably looks innocent to other viewers if they are not aware of who the person is. If we can increase the number of labelers for each image and ask them to label the instances only after viewing the photos themselves based on their own subjective opinions, the diversity of bullying could be mitigated. Besides that, we should design or look for features that can more efficiently recognize the factors that cause viewers to bully. Emotional information from the posts would be a possible choice to implement since the bully is a product of extreme emotions. Also, strong object detection techniques could be considered. For image modality, possible object targets are human body parts, sports scenes (football field, basketball court), celebrity detection etc. Another idea is to better simulate the context under which the data instances are exposed to viewers in virtual world. For example, the photos can be categorized so that we can tell if it’s a sharing of a life moment of an exhibition of products being promoted. Thus for each category we can look for different features to detect bullying. Moreover, we have only considered communicative media of the text and image type whereas other categories such as animated GIFs, videos etc. define an open problem within the research domain.

APPENDICES

APPENDIX 1: LIST OF PUBLICATIONS (COMMUNICATED)

Hybrid Deep Learning Model for Cyberbullying Detection in Social Multi-modal Data

Akshi Kumar[1], Mudita Saxena[2]
[1,2]Department of Computer Science & Engineering, Delhi Technological University, Delhi, India
akshikumar@dce.ac.in, saxena.mudita987@gmail.com

Abstract: Cyberbullying is the use of Information and Communication Technology (ICT) by individuals' to humiliate, tease, embarrass, taunt, defame and disparage a target without any face-to-face contact. Social media is the "virtual playground" used by bullies with the upsurge of social networking sites such as Facebook, Instagram, YouTube, Twitter etc. It is critical to implement models and systems for automatic detection and resolution of bullying content available online as the ramifications can lead to a societal epidemic. This research proffers a novel hybrid model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image). The all-in-one architecture, CNN-$_{\text{BoVW}}$SVM, consists of a convolution neural network (CNN) for predicting the textual bullying content and a support vector machine (SVM) classifier trained using bag-of-visual-words (BoVW) 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[1]. The processing of textual and visual components is carried out using the hybrid architecture and a Boolean system with a logical OR operation is augmented to the architecture which validates and categorizes the output on the basis of text and image bullying truth value. The model achieves a prediction accuracy of 84% which is acquired after performing tuning of different hyper-parameters.

Keywords: Social media, cyberbullying, BoVW, deep learning, modality

1. Introduction

The global and pervasive reach of social multimedia has in return given some unpremeditated consequences where people have discovered illegal & unethical ways to use the socially-connected virtual communities. One of its most severe upshots is known as *cyberbullying* where individuals find new means to bully one another over the Internet. Formally, cyberbullying is any behavior performed through electronic or digital media by individuals or groups that repeatedly communicates hostile or aggressive messages intended to inflict harm, embarrassment or discomfort on others [1]. It has grown as a social menace that puts a negative effect on the minds of both the bully and victim.

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

REFERENCES

[1]. Aboujaoude E, Savage MW, Starcevic V, Salame WO. Cyberbullying: review of an old problem gone viral. J Adolesc Health. 2015;57(1):10–18. doi: 10.1016/j.jadohealth.2015.04.011.

[2]. Campbell MA (2005) Cyber bullying: An old problem in a new guise? Journal of Psychologists and Counsellors in Schools 15(1):68-76.

[3]. Tokunaga Following you home from school: a critical review and synthesis of research on cyberbullying victimization. Comput Hum Behav. 2010; 26:277–287. doi: 10.1016/j.chb.2009.11.014.

[4]. Centers for Disease Control and Prevention. Youth violence: technology and youth protecting your child from electronic aggression; 2014. http://www.cdc.gov/violenceprevention/pdf/ea-tipsheet-a.pdf. Accessed 11 September 2017.

[5]. Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N. Cyberbullying: its nature and impact in secondary school pupils. J Child Psychol Psychiatry. 2008;49(4):376–385. doi: 10.1111/j.1469-7610.2007.01846.

[6]. Hinduja, S. & Patchin, J. W. (2014). Cyberbullying Identification, Prevention, and Response. Cyberbullying Research Center (www.cyberbullying.us)

[7]. https://machinelearningmastery.com/best-practices-document-classification-deep-learning/

[8]. Dadvar, Maral, and Kai Eckert. "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study." *arXiv preprint arXiv:1812.08046*(2018).

[9]. K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In The Social Mobile Web, 2011(21, 30).

[10]. Hosseinmardi, Homa, et al. "Prediction of cyberbullying incidents in a media-based social network." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016.

[11]. Alkhawlani, Mohammed, Mohammed Elmogy, and Hazem Elbakry. "Content-based image retrieval using local features descriptors and bag-of-visual words." *Int J Adv Comput Sci Appl* 6.9 (2015): 212-219

[12]. K. B. Kansara and N. M. Shekokar. A framework for cyberbullying detection in social network. 2015.

[13]. Kumar, A. & Sachdeva, N. Multimed Tools Appl (2019). https://doi.org/10.1007/s11042-019-7234-z.

[14]. Agrawal, Sweta, and Amit Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms." *European Conference on Information Retrieval*. Springer, Cham, 2018.

[15]. https://github.com/jowoojun/kaggle-Toxic-Comment-Classification-Challenge.

[16]. K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans. Interact. Intell. Syst., 2(3):18:1-18:30, Sept. 2012}.

[17]. Hinduja, Sameer, and Justin W. Patchin. "Bullying, cyberbullying, and suicide."Archives of suicide research 14.3 (2010): 206-221.

[18]. Kokkinos, Constantinos M., Nafsika Antoniadou, and Angelos Markos. "Cyberbullying: an investigation of the psychological profile of university student participants." Journal of Applied Developmental Psychology 35.3 (2014): 204-214.

[19]. K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans. Interact. Intell. Syst., 2(3):18:1-18:30, Sept. 2012.

[20]. M. Dadvar, F. M. G. de Jong, R. J. F. Ordelman, and R. B. Trieschnigg. Improved cyberbullying detection using gender information. In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, pages 23-25, Ghent, February 2012. University of Ghent.

[21]. V. Nahar, S. Unankard, X. Li, and C. Pang. Sentiment analysis for effective detection of cyber bullying. In Web Technologies and Applications, pages 767-774. Springer, 2012.

[22]. V. Nahar, S. Unankard, X. Li, and C. Pang. Semi-supervised learning for cyberbullying detection in social networks. In Databases Theory and Applications, LNCS'12, pages 160-171, 2014.

[23]. A. K. K. Reynolds and L. Edwards. Using machine learning to detect cyberbullying. Machine Learning and Applications, Fourth International Conference on, 2:241-244, 2011.

[24]. M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. In the service of online order tackling cyberbullying with machine learning and affect analysis. 2010.

[25]. B. Nandhini and J. Sheeba. Cyberbullying detection and classifcation using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015.

[26]. A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards. Detecting cyberbullying: query terms and techniques. In Proceedings of the 5th Annual ACM Web Science Conference, pages 195-204. ACM, 2013.

[27]. B. S. Nandhini and J. Sheeba. Online social network bullying detection using intelligence techniques. Procedia Computer Science, 45:485-492, 2015.

[28]. Yin, Dawei, et al. "Detection of harassment on web 2.0." Proceedings of the Content Analysis in the WEB 2 (2009): 1-7.

[29]. Marathe, Shivraj Sunil, and Kavita P. Shirsat. "Approaches for Mining YouTube Videos Metadata in Cyberbullying detection." *International Journal of Engineering Research & Technology, International Journal of Engineering Research & Technology (IJERT)* 4.05 (2015): 680-684.

[30]. Q. Huang, V. K. Singh, and P. K. Atrey. Cyber bullying detection using social and textual analysis. In Proceedings of the 3rd International Workshop on Socially-Aware Multimedia, pages 3-6. ACM, 2014.

[31]. Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using machine learning to detect cyberbullying." Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on. Vol. 2. IEEE, 2011.

[32]. Xu, Zhi, and Sencun Zhu. "Filtering offensive language in online communities using grammatical relations." Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. 2010.

[33]. Zerr, Sergej, et al. "Privacy-aware image classification and search."Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[34]. Lieberman, Henry, Karthik Dinakar, and Birago Jones. "Let's gang up on cyberbullying." *Computer* 44.9 (2011): 93-96.

[35]. Vanhove, Thomas, et al. "Towards the design of a platform for abuse detection in OSNs using multimedial data analysis." 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013). IEEE, 2013.

[36]. Rybnicek, Marlies, Rainer Poisel, and Simon Tjoa. "Facebook watchdog: a research agenda for detecting online grooming and bullying activities." 2013 IEEE International Conference on Systems, Man, and Cybernetics. IEEE, 2013.

[37]. Zerr, Sergej, et al. "Privacy-aware image classification and search."Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[38]. Zhao, Rui, et al. "Deep learning and its applications to machine health monitoring." *Mechanical Systems and Signal Processing* 115 (2019): 213-237.

[39]. Uysal, Alper Kursat, and Serkan Gunal. "The impact of preprocessing on text classification." *Information Processing & Management* 50.1 (2014): 104-112.

[40]. Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

[41]. http://www.cs.cornell.edu/courses/cs1114/2013sp/sections/S06_convolution.pdf

[42]. https://kushalvyas.github.io/BOV.html

[43]. P. Kralj Novak, J. Smailovic, B. Sluban, I. Mozetic, Sentiment of Emojis, PLoS ONE 10(12): e0144296, doi:10.1371/journal.pone.0144296, 2015.

[44]. Kumar, A., Dogra, P. and Dabas, V., 2015, August. Emotion analysis of Twitter using opinion mining. In 2015 Eighth International Conference on Contemporary Computing (IC3)(pp. 285-290). IEEE