International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Identifying Severity of Cyberbullying Using Scalable Labeled Multi-Platform Dataset

Madhura Vyawahare **Dr. Sharvari Govilkar**

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: Increasing invective posts on online social media platforms is of great concern considering the wellbeing of society and psychological health of youth. These invective posts many times take the form of cyberbullying if not tackled in an early stage. It is required to identify such posts which are harmful and may become even more dangerous for any netizens, to maintain a psychologically healthy society. Many machine learning and deep learning based systems were designed in the past for automated cyberbullying detection. Accurate and precise cyberbullying detection needs a large and correctly annotated dataset. The work is focused on resolving the issue of unavailability of appropriate dataset by designing an automated labeling system for creating and labeling the dataset to detect severity of cyberbullying. The meta-features apart from textual comments like semantic and syntactic features also contribute to learning of the machine. Principal components analysis is used for feature extraction and reduction. Rule based methodology is designed, developed and implemented which considers textual, semantic and syntactic features and results in a rich in features, multi-platform, multi-label dataset for severity of cyberbullying detection as well as cyberbullying prediction. Till now only two approaches have been used for Annotation of dataset: Manual labeling and filtration method. A new rule based automated approach is proposed and implemented in this work. Using this new approach the dataset of size 17 lakh entries with 5 labels is prepared and used for training the machines. To make the dataset standardized and usable for researchers in future, it is tested and verified with various methods. Evaluation of the proposed system based on accuracy, precision, recall and f-measure demonstrates that the performance of multiclass classification trained from the prepared dataset is highly improved.

Keywords: *Cyberbullying, Cybercrimes, Dataset Annotation, Social Media, Machine Learning*

1. Introduction

Increasing popularity of social media has also increased the social problems arising from the misuse of facilities provided by such platforms. The platform provides freedom of speech which many times results in cyber exacerbated crimes. Using abusive or profane words has become a trend in youth in recent years. Putting opinion in harsh words and blaming people is very commonly observed in all social media platforms. As can be seen from the figure 1, in general communication on twitter these are the most commonly used words. It can be observed that a lot of profane words are present in it. The hate content on these social media platforms is increasing and this also takes a direction towards targeting a single person for harassment [23] [34]. This injurious act is also known as cyberbullying. In cyberbullying the repetitive nastiness can be observed toward the victim. According to the Cyberbullying Research Centre, on average, 26.3% of teenagers from school in the United States have experienced cyberbullying at some point in their lives [26]. Cyberbullying creates deep impressions on the human mind which becomes difficult to remove [30]. Not only the victim but also the bully faces psychological problems if the bullying is taking place for a long time. Research shows the impact of severe bullying can also results in suicidal attempts [19] [32]. According to the survey done by Kowalski and Limber, almost 90% of teenagers do not disclose to their parents about the humiliation they are facing [18]. There are various reasons behind not sharing the fact with parents like: the victims are embarrassed that they are getting bullied, young victims are scared that the devices they are using to access social media platforms will be taken away etc. This leads to even difficult situations as detecting the cyberbullying act and taking some action to heal the mental state of adolescent is not possible for parents. This makes identifying this type of crime quite important for a healthy society and is also targeted by many researchers. Our research is not only limited to detecting traces of cyberbullying from social media platforms but also trying to predict the act before it becomes severe. Detection of such messages and posts from large social media platforms is quite a difficult task. Many times it results in a false alarm. This happens because of the incorrect interpretation of presence or absence of bullying traces. To detect cyberbullying the psychology behind the bully mind needs to be understood. For predicting such an act one needs to understand what type of sentences are heading towards bullying.

[1] *Pillai College of Engineering, New Panvel, Maharashtra, India*
ORCID ID: 0000-0002-8981-7636
[2] *Pillai College of Engineering, New Panvel, Maharashtra, India*
** Corresponding Author Email:madhura.vyawahare@email.com*

Fig 1. Most Commonly Used Words on Twitter

In the recent past mostly machine learning and deep learning models are used for the detection of cyberbullying. Researchers in recent studies have proved with results that the state of the art results are achieved by using deep learning algorithms for detecting cyberbullying [37][6][39][2]. Training such models correctly plays a crucial role in achieving good results; and which needs an appropriate dataset [8]. Especially for deep learning models, large size and appropriately labeled dataset is the primary requirement. Unavailability of such datasets is the major problem faced by researchers working in this area. The features apart from textual comments like semantic and syntactic features also contribute to learning of the machine. In dataset preparation, data scraping is not a problem but providing appropriate labels is a tedious and time consuming task. If labeling is done manually, it is costly and not scalable. Instead of manual labeling, automated systems for labeling will give more freedom and will reduce human efforts. Hence in this work a rule based methodology is designed, developed and tested; which is suitable for detecting the severity of bullying and providing labels accordingly. The work includes preparation of a rich in features, multi-platform and multi-labelled dataset which is auto annotated using the designed algorithm, with multiple labels in the respect of identifying severity of cyberbullying. PCA is used for the feature extraction and reduction purpose. PCA being an unsupervised learning algorithm, made it possible to reduce the features before labeling. To make the dataset standardized and usable for researchers, it is tested and verified with various methods. The work presents all the results of dataset validation and verification. For the testing purpose of the proposed model, Random forest, Decision Tree and XGBoost algorithms are used. Results also show the machine learning models for multiclass classification trained from this dataset are providing improved results.

2. Related Work

In recent years a large number of researchers have got attracted towards solving the problem of cyberbullying which is taking place on all different social media platforms. Many researchers have identified the most popular OSN or the platform where maximum bullying is taking place. Few researchers have also presented work for multi-platform cyberbullying detection [31]. Most of the researchers in recent years have used the machine learning and deep learning methods and very few have actually provided the details of the dataset used for the work [35] [37]. Dataset plays a crucial role in learning of any machine. Particularly for deep learning, large size datasets are required. Authors Akshita A. et al., 2020 have shown with the experimental results that for smaller dataset deep learning algorithms perform poorly [1]. To avoid overfitting and underfitting problems, a precisely annotated large and balanced dataset is very important. For the classifier to extract features and learn the information dataset plays an important role. Availability of the proper dataset is always a major issue faced by researchers [21] [20]. Especially in the area of cyberbullying detection and prediction, no standard datasets are publically available and most of the researchers have prepared their own dataset by manual labeling [24].

Authors Hugo Rosa et al, in the year 2019 have done a detailed survey of 22 recent models and stated their observation that there is no uniformity regarding the methodology used for this annotation. As per the survey done in [35] very few researchers in recent years have worked towards preprocessing and dataset annotation for cyberbullying detection and the size of datasets of Twitter is not more than 3lakh entries. Recently authors Daniyar Sultan et al. have worked on creating a dataset in Kazakh language for cyberbullying detection. They have represented the complete process in detail [42]. Authors Tabassum Gull Jan et al. have proposed a semi-supervised algorithm for labeling a twitter dataset for identifying legitimacy of the user [6]. They have collected the most recent tweets and used self-learning techniques for labeling the tweets. Till now no work is presenting the creation process of dataset for English language to identify cyberbullying severity. Multi-Platform dataset is released by few researchers but most of such datasets do not include other parameters than textual comments from social media. As per the literature many other parameters and meta-information features do have a positive impact on cyberbullying detection and should also be considered in the dataset [8]. Authors Yeungjeom Lee et al. have considered gender as a meta-information and identified the impact of gender analysis on cyberbullying detection [26]. Major work done is on binary classification where the post is classified either as: bullying or non-bullying; hence the datasets are labeled accordingly. The Available datasets are not as per the requirements of cyberbullying prediction and severity detection. To identify the severity multiclass classification is required and binary labeled datasets are not sufficient.

The process of cyberbullying detection improves with the various features used to train the system. It is difficult to only claim the nature of the post based on text posted. It is very difficult to differentiate hate speech from profanity [10]. Other features like semantic features, syntactic features, social features, user features etc. also plays an important role and improves the result [26] [9]. Authors Talpur et al., have proved that; use of semantic features improves the detection accuracy [29]. The experimental analysis was performed to check the impact of user personality on cyberbullying detection and the results showed improvement in accuracy when personalities and sentiments were used [3] [33]. Cyberbullying is a repetitive act and its presence is not only identified from the textual comment but also is witnessed from social network features [16] [22]. In the study done by Rosa H. et al, it is mentioned that very few researchers are using social and user features [24]. To include all these features for training the model, the dataset needs to be rich in terms of parameters considered during scraping of data from social networks. It is observed that, most of the datasets are of small size and also considered only 2 columns as text and label [31]. Scraping data from social media platforms is a comparatively simple task, the

challenging part is labeling the collected corpus appropriately. For annotation of dataset till now only 2 approaches have been used: Manual labeling and filtration method [35]. Most of the researchers have used manual labeling done by students or by crowdsourcing e.g. from Amazon Mechanical Turk [36] [24]. Manual labeling is a very hectic, time consuming and costly process and has no scalability. If the dataset size is increased, again the same amount of time needs to be spent on labeling. Manual labeling also depends on human brain interpretation so it may result in ambiguity. In the case of filtration methods very few have contributed and they have focused on identifying the specific bullying words identification only.

A new approach for labeling dataset in an automated way is used, by considering the features which play a vital role in detection of cyberbullying. To improve the accuracy and precision of the detection and prediction, this newly prepared corpus is created by extracting Twitter data. The corpus annotation is done using the novel algorithm which is specifically designed by considering all features extracted by PCA, which participate in identifying the severity of bullying. The newly designed rule based algorithm is capable of creating a labeled dataset. The final dataset consists of 17 lakh entries. Samples of this dataset are also verified from language experts and experts in the field. XGBoost, Random Forest and Decision Tree algorithms are then used for verifying the results generated by the learning models trained with the new dataset. These results are in line with the literature and show improvement in accuracy compared to existing dataset. The dataset is annotated using 5 different labels in terms of identifying the severity: No Bullying, Nasty, Light Bullying, Moderate Bullying and Severe Bullying.

3. Research Methodology

This section describes in detail about the methodology used for the proposed system. Figure 2 explains the flow of the process where data is collected from 2 social media platforms: Twitter and YouTube. For data collection specific fields are selected to have a balanced dataset including bullying and non-bullying posts. After extraction of data, labeling is done using the proposed rule based algorithm and later the machine learning is used for testing the proposed framework.

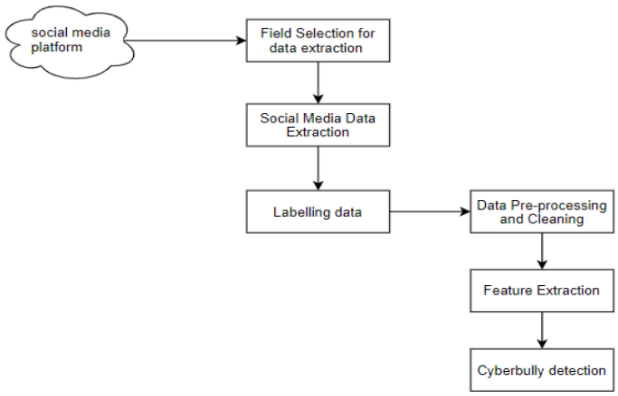


Fig 2. Research Methodology

This section includes: a) The process of collecting, cleaning and preprocessing corpus from two platforms: Twitter and YouTube, b) Data Annotation terminology and c) Features Extraction and Reduction with reasoning.

3.1. Corpus Collection

Out of all social media platforms Facebook, Twitter, YouTube, Ask.fm and Instagram have been listed as top 5 networks with the highest percentage of users reporting experience of cyberbullying. Twitter and YouTube are selected for the work of data extraction and corpus creation. Extracting all the past communications resulted in a very unbalanced data, as most of them were general posts and very few instances were found related to bullying interaction. Filtering bullying posts for having a balanced dataset was becoming difficult. Hence to collect maximum bullying posts it was decided to scrape posts on particular trending topics. For extracting posts from Twitter as well as YouTube; few keywords were finalized: Asian hate, Feminism and Politics and commonly used bullying words like: fatso, ugly, moron, slut. Tweets were extracted along with some social features like:

- Username
- User ID
- Is the account verified
- Followers count
- Media Count
- Reply count
- Retweet count
- like counts

20 lakh tweets were extracted and after preprocessing about 12 Lakh tweets were selected for creating the dataset.

YouTube is another popular online social media platform for sharing videos. Almost 500 hours of videos are being watched every hour on YouTube and the amount of trolling going on is very large on this posted content. Fields which are extracted are Youtuber's Username, video name, Comments, Replies on comments, Likes on replies, time of Comments and reply counts. Total 5 lakh posts are extracted from various YouTube videos.

3.2. Data Cleaning and Preprocessing

Data cleaning and preprocessing plays an important role in machine learning models. Without the cleaning process, the dataset is often a bunch of words that the ML model can't learn from. Especially when it comes to natural language processing, this step is unavoidable and to be done very carefully. Social network data is noisy, thus preprocessing has been applied to improve the quality of the research data and subsequent analytical steps. Following things are covered:

- Stop words removal
- Removal of Punctuations (except full stop, comma and exclamation marks)
- Converting numbers into text
- Converting slang words into respective English words
- Removal of non-alphanumeric symbols
- Removal of URLs
- Part of Speech tagging
- Converting emoticons into text

Emoticons are widely used by the young generation and these graphical representations of emotions typically convey the exact sentiment of the person [14]. Converting emoticons into text helps to count the contribution of emoticons. This conversion of emoticons to text is also included in pre-processing. A dictionary

of emoticons and their meaning in English is prepared and used for the purpose. Dictionary consists of 108 emoticons with meaning in English, Table 1 shows samples of the dictionary.

Table 1. Sample Emoticons with Meaning in English

Emoticon	Meaning in English
;-)	winking happy smiley
:-*	kiss
;^)	smirking smiley
:-)8<	big girl smiley
>:->	devilish smiley
X-(	you are brain dead
:,(	crying smiley
:----}	you lie like Pinocchio
;-)	winking happy smiley

Part of speech tagging is also performed to identify the relation between the noun or pronoun and the profane word in the sentence. Figure 3 and 4 shows the sample or this correlation identified.

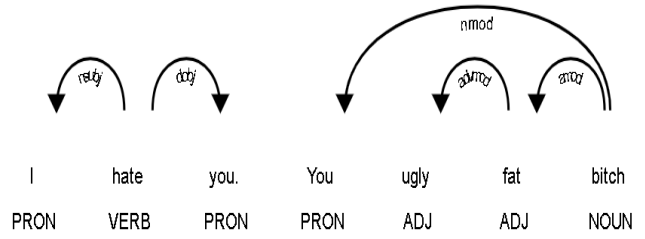


Fig 3. POS Tagging With Parser Sample 1

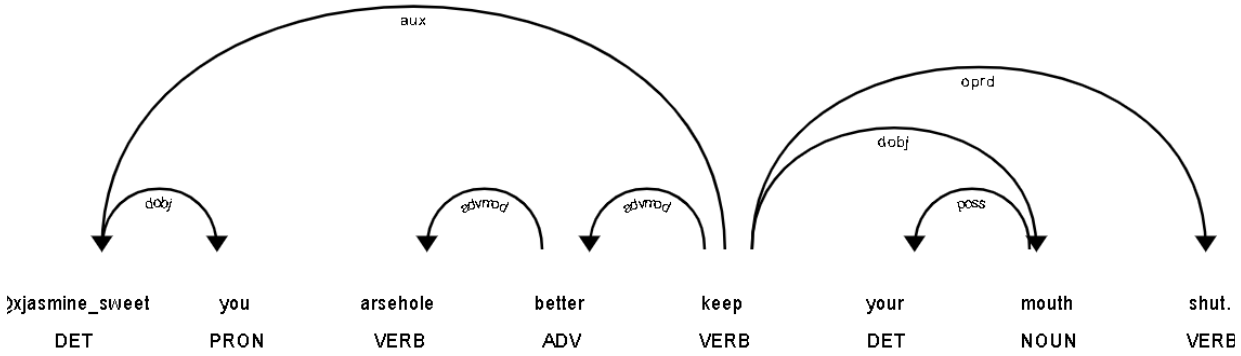


Fig 4. POS Tagging With Parser Sample 2

While removing punctuation a special care was taken for not removing special characters like exclamation marks as it helps in understanding sentiment. As capitalization is one of the indications of aggression, capital letters were not converted to lowercase. Extra-long words are not removed which also are an indication of severity of emotion e.g. pleaseeeeee, soooooo etc.

In natural language processing, to understand the context of the word before categorizing it in any class is very popular and important. Part-of-speech tagging is very popular for identifying words and its corresponding part of speech. In the proposed work identifying the dependency between the noun or pronoun and the profane word is crucial to claim if the sentence is a bullying sentence or not. Only the presence of profane words and pronouns or username does not mean the profanity is intended for a particular person. This can be said only after finding the relation between these entities. The Parser is used for identifying the relation between pronouns and profane words. Identification of this correlation helped improve the annotation accuracy.

3.3. Features Analysis

Selecting appropriate features is very important for training any machine learning model and hence many researchers are focusing on this. Authors Mohammed Ali Al-Garadi et al. have also focused more on data collection and feature engineering [22]. Textual features are mostly considered by the researchers for detecting cyberbullying. In textual features, most of the work is only focusing on profane words and not semantic or syntactic features. Authors in [9] have used Semantic and syntactic features along with textual features and proved it contributes positively in cyberbullying detection. Researchers have proved point wise semantic orientation of each word and phrase also improves the detection of cyberbullying. Authors Q. Huang et al. have claimed that social network features also significantly contribute in detection of cyberbullying [16]. According to the recent research by authors Bayzick et al, presence of bad words along with use of a lot of capital letters, and if all this is addressed to a second person pronoun then it is treated as bullying [12]. From the detailed literature survey it is identified the positively contributing features are profanity, capitalization, and semantic features. Along with this one more impactful feature is considered: sentence length. PCA is used for feature reduction.

A. Profanity

The first and most important parameter is presence of profane or foul or swear or insult or judgmental words. When a sentence does not contain such a word then it's impossible for a machine to identify it as bullying. Profanity factor is a highly rated factor in our algorithm. Many research studied have considered stupid, ugly and idiot as cyberbullying words but these days extreme words are used by netizens [40]. In today's day, youth have a large vocabulary to use for humiliating someone and considering this scenario a dictionary of 2500 profane words is used to match and identify bullying or profane words in the post.

For matching the word with a dictionary word fuzzy string matching using Levenshtein distance is used. The number of profane words (P) present in the sentence are calculated to finally calculate the profanity ratio.

$$Rp_i = \frac{\sum Profane\ Words\ in\ i}{\sum\ Words\ in\ i} \quad (1)$$

Where,
Rp = Profanity Ratio
i = i^{th} entry from Dataset

B. Semantic Correlation

Semantic features are considered by many researchers in their work for confirmation of bullying. Authors have detected 'personal' and 'targeting to specific person' comments by finding presence of 1st and 2nd person pronouns in the comment [8][9][38]. Existence of profanity without the presence of any 2nd or 3rd person pronoun indicates only nastiness and not bullying. If the bad word is intended for someone then there is a chance that the sentence is a bullying sentence. Many researchers have used this concept of pronouns for justifying bullying in their work. Authors Maral Dadvar et al. have used profanity windows of different sizes to confirm bullying [7].

Only the presence of pronouns is not sufficient. Present pronouns appearing near profanity also do not justify the relation between profane word and pronoun. If the correlation between profane word and pronoun is proved then only it can be said the bully has used these words intentionally to insult or harass the person referred to in the post. The parser is used to find this relationship between profane words and pronouns or named entity or username.

Many times while posting comments on social media, people don't use any pronoun but the sentence is still intended to the person, on whose wall the comment is posted. It is observed that such sentences are smaller in length and just have bad words in them.

For example: What a suckass, BITCH WHORE, DUMBASS WHORE, BITCH etc.

It is observed that these sentences are severe bullying sentences. Sentences are smaller in length, containing a maximum of 5 to 6 words and have high profanity. If the identification of bullying is kept dependent on only the presence of pronoun or noun then these sentences were missed. So such sentences are taken into consideration.

Method checks the presence of profane words in the sentence. Then, identify the pronoun or user name or named entities (N). Then it is identified if N has any dependency with the profane word. If there is dependency then the semantic correlation is assigned with value 1. Else the length of sentence is checked. If the sentence length is very small the semantic correlation is assigned with value 1. If profanity is present, N is false but the sentence is very small then still the severity of the sentence is calculated. For small sentences approximately 30 letters in the sentence are considered.

for, N = pronoun / username / named entity
Sl = sentence length ratio
if (N has dependency with profane word)
    return 1
else
    if $(1 - Sl) \geq 98$ && $isPronoun \neq True$
        return 0
    else:
        return 1

C. Capitalization

Use of capital letters while posting any message represents the aggressiveness of the user sentiment. According to the authors, capital letters mean intense feelings which can be positive or negative [7] [11]. If this intense feeling is accompanied by profane words then it is one indication of bullying. Authors Samghabadi and Niloofar S. have also claimed from their survey that users with bullying intentions sometimes try to be more robust by using capital letters in their comments [27]. Authors Kouadri et al. in their research have stated if all characters of a polar word are in upper case, the polarity intensity increases [17]. As per the survey presented, Capitalization is treated as intensifiers by many researchers [12]. Bayzick et al. argued that excessive use of capital letters was an indication of hostile communication [5]. There are many instances which prove how capitalization increases the intensity of the sentence. While calculating severity 2nd priority is given to capitalization. The capital word ratio is calculated using the following formula.

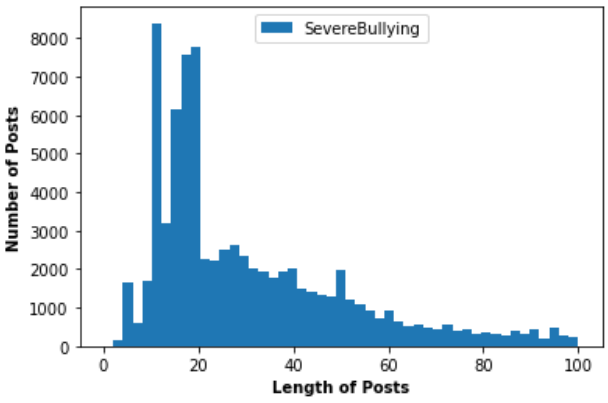
$$Cw_i = \frac{\sum Capital\ letters\ in\ i}{\sum Letters\ in\ i} \quad (2)$$

Where,
Cw = Capital word ration
i = i^{th} entry from Dataset

D. Sentence Length

Another factor which contributes in confirmation and severity calculation of bullying is length of sentence. It is observed that when people are talking about general things or giving their opinion they usually write longer sentences. When they are bullying someone or commenting on someone the length of the sentence is comparatively shorter or moderate. Researchers H. Herodotou et al. [13] from their survey and experimentation have identified length of sentence is smaller in abusive posts.

After analyzing standard labeled dataset of twitter it is confirmed that smaller sentences have more chances of being abusive. Figure 5 represents the graph plotting length of posts vs. number of posts for severe and light bullying sentences.



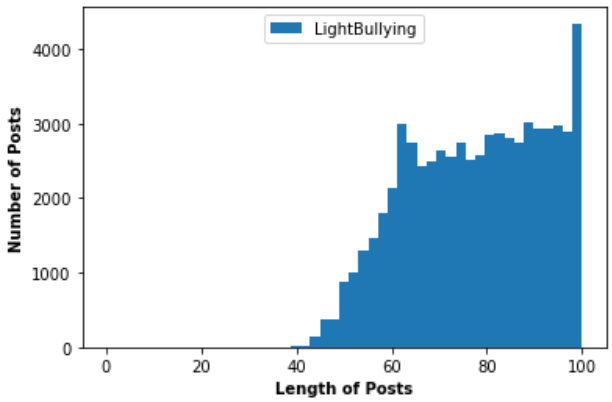


Fig 5. Sentence Length vs. Bullying Severity

Following is the observations from standard dataset analysis:

- Light bullying comments on an average are having larger length whereas the ill-intentional i.e. severe bullying comments are smaller in length.

It is concluded from the study that; if the sentence length is larger, then it has a greater chance that it is a discussion on some topic or personal harsh opinion and not harassment of a person. Hence another factor for confirmation of bullying is considered as length of sentence.
The length of the sentence with respect to the total length allowed is calculated. Smaller the length indicated more possibility of bullying.

3.4. Data Annotation

To label the extracted data, instead of depending on manual labeling, a new rule based algorithm is proposed and implemented. This algorithm can annotate enormous amounts of data as per the need of research. Manual labeling is time consuming and also the labeling depends upon the human brain interpretation hence it may result in ambiguity. Automated systems are also not always completely reliable and need verification of the results. 3 levels of verification and validation of the final annotated dataset have been done.

3.4.1. Annotation Categories

The work is focused on detection and prediction of cyberbullying traces present on the social media platform. Prediction of future occurrences of cyberbullying is possible by early detection of these traces. This can be done only if the process of cyberbullying is initiated. Early stage of cyberbullying is called cyber aggression [27]. Hosseinmardi H. et al., have defined and differentiated between cyberbullying and cyber aggression [15]. The difference between cyberbullying and Cyber aggression is also explained in detail by authors [26]. Author Christopher P. Barlett has shown with results that early cyberbullying behavior was a strong predictor of later cyberbullying behavior [4]. Bullying detection or prediction can't be so clear as black and white; that is, the sentence can't always be either bullying or not bullying. Many times it is just towards bullying, i.e. aggressive but is not actually bullying and sometimes it is intensely bullying. Hence only classifying the posts as bullying or not bullying as done by most of the researchers is not sufficient for our work. It was required to identify the level of severity of bullying for the appropriate prediction. The proposed algorithm is labeling data using 5 different labels based on severity of text: Non Bullying, Nasty, Light Bullying, Medium Bullying, and Severe Bullying.

Non Bullying: When the post does not contain any profane word it is treated as a non-bullying post. For identifying the profane word, a predefined list of profane words is used. This list consists of a total 1437 stemmed words declared as profane/ bad words.

Nasty: Youth these days use a lot of slang words in their casual communication. Presence of such slang language or bad words is not always bullying. Many people use such profanity while expressing their opinion about social issues. Even these types of posts are not intended for bullying anyone. Identifying such posts and separating them from bullying posts is very important. Existence of profanity without the presence of any named entity or pronoun indicates nastiness and not bullying. Authors Maral Dadvar et al, have used profanity windows of different sizes to confirm bullying [7]. Here the category nasty indicates profanity is present but there is no named entity or pronoun directly correlated to it in the sentence.

Light Bullying: When bullying is just initiated and not very harsh in that case the post falls in this category. After calculating severity the threshold is decided to label sentence as light bullying as: $5 < \text{severity} < 33.5$

Medium Bullying: Moderate bullying is where the sentence is having more profanity than light bullying. Confirmation of bullying cannot be only dependent on profane words. Other parameters also represent the severity of bullying. Parameters like sentence length, capitalization are the parameters which have an impact on the severity as proved in the literature [27] [13]. By considering all the relevant parameters the severity is calculated. After experimentation, analysis and verification the threshold for deciding medium severity is finalized between 33.5 and 42.
Severe Bullying: Severe bullying is where the harassment is of extreme level and very harsh. As an example in many posts the bully uses extremely abusive and dirty words for victims. These types of words create extremely negative impressions on the heart and mind of the victim. These types of posts are considered severe bullying. To label the post as severe bullying the threshold selected is: $\text{severity} > 42$.

4. Proposed Algorithm for Dataset Annotation

The proposed algorithm uses a rule-based method for annotation. Following is the algorithm and figure 7 represents the flowchart for the proposed novel algorithm for dataset annotation. This algorithm labels data with 5 categories: Non Bullying, Nasty, Lightly bullied, moderately bullied, and highly bullied.

Step 1. Start
Step 2. Read comment
Step 3: Find Profane / swear / insult / judgmental words
Step 4: Calculate number of profane words P and profanity ratio (Rp)
Calculate semantic correlation (SCo)
Calculate capital words ratio (Cw)
Calculate length of sentence w.r.t. total length allowed (Sl)
Using severity detection formula calculate severity (S)
Step 6: if (P = 0)

```
Label = Non Bullied
else
        if ( SCo = 0)
                Label = Nasty
        else
                if (5 < severity < 33.5 )
                        Label = Light Bullied
                else if (33.5 <= severity < 42)
                        Label = Moderate Bullied
                else
                        Label = Highly Bullied
```

Step 7: Stop

Severity Detection:

$$S = (\alpha \times Rp) + (\beta \times Cw) + \big(\gamma \times (1 - Sl)\big) + (\delta \times SCo) \quad (3)$$

Where,

α = 0.5, β =0.2, γ =0.1, δ =0.2

Rp = (Number of profane words in Comment) / (Total number of words in comment)

Cw = (Number of Capital words in Comment) / (Total number of words in comment)

SLen = (Total number of characters in comment) / 280

SCo = if presence of 2nd or 3rd person pronoun SCo = 1

if no 2nd 3rd person pronoun but (1-Sl) >=0.98 then SCo = 1

else SCo = 0

Values for α, β, γ, δ are decided on an experimental basis. 5 different cases are considered by varying values of all coefficients. With every case it is manually identified if the algorithm is able to label the posts properly or not. Based on the experimental study changes are made in the threshold and the coefficients. Figure 7 gives the sample of the labeled dataset.

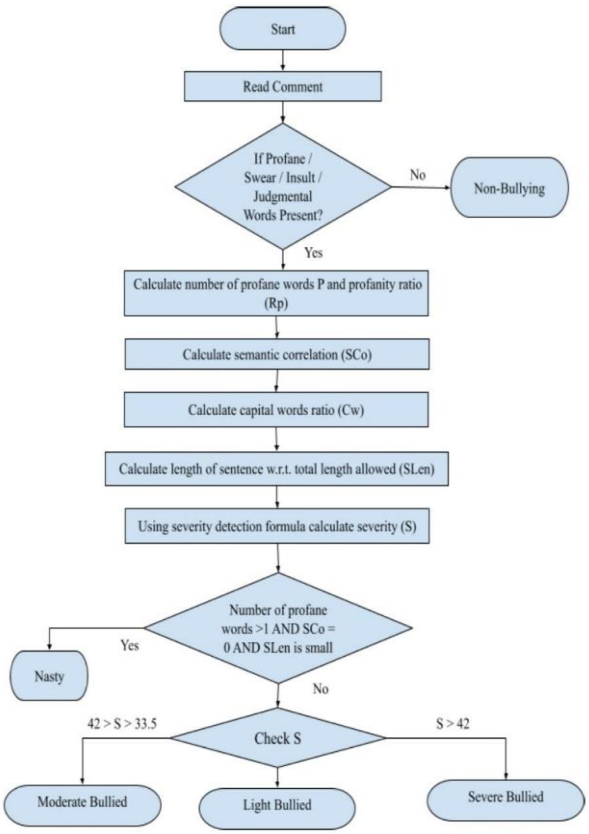


Fig 6. Proposed Dataset Annotation Methodology

The algorithm is capable of annotating large size corpus. The time taken for annotation is extremely less compared to the manual labeling. The large size dataset once labeled properly can be used for training machine learning or deep learning models.

	Text	Followers Count	Friends Count	Media Count	Reply Count	Retweet count	Like count	Quote Count	Gender	algo_label	Severity value	Profanity Ratio	Sentence length Ratio	Semantic correlation	Capital word ratio
1	@DracoYTOffical1 IDK MAN I WAS WATCHING RIO AND THE FCKIN OKD UGLY BIRD HELD A WHUTE CUTH UP TO A MANS MOUTH AND HE PASSED OUT	219	379	30	0	0	0	0	0	Severe Bullying	43.48367	7.692308	54.64286	100	70.86614
2	i was sayin this you dont deserve to be a parent if you gone act ugly	1199	1138	146	0	0	0	0	0	Light Bullying	30.66071	6.25	75.35714	100	0
3	Good night to only ugly people	34391	897	2225	14	0	49	0	0	Nasty	17.87135	16.66667	88.92857	0	3.225806
4	Thanks ugly i hate you more	424	326	715	0	0	1	0	0	Medium Bullying	38.10979	16.66667	90.35714	100	3.703704
5	face reveal please be nice i know im ugly	533	549	925	6	0	7	1	0	Nasty	14.09127	11.11111	85.35714	0	0
6	i have such an ugly laugh	161	168	6	0	1	4	0	0	Nasty	17.40476	16.66667	90.71429	0	0
7	@AbeehaTariqArt @ATrench93 And whats good is that shes a dark skinned black woman who isnt painted as masculineugly	42	243	222	0	0	2	0	0	Nasty	9.714113	5.555556	58.92857	0	5.217391
8	this janitor at my job be following me around like sir and than he ugly fuck ughh	713	625	286	1	0	0	0	0	Light Bullying	32.9895	11.76471	71.07143	100	0
9	@lavendercowboy the kill steal bestie what else in my bio looks ugly Instant Message in crisis	46	667	1656	1	0	1	0	0	Nasty	13.53116	12.5	66.42857	0	3.191489
10	@alexinwonland Good Theyre ugly and suggest the wearer is too lazy to put in shoes	23	81	86	0	0	0	0	0	Light Bullying	30.89257	6.666667	70.71429	100	2.439024

Fig 7. Sample of Labeled Dataset

5. Results and Analysis

Preparation of dataset is when done in an automated way or by using filtration method, according to the authors Fatma E. et al. it is still useful to have a human annotator involved to verify the final labels [35]. Experts also mention that identifying incorrect and correct labeling needs to be done by linguistic experts [10]. Hence the prepared dataset with automated annotation was verified using 3 steps and later tested with machine learning algorithms.

5.1. Verification and Validation of Annotated Dataset

After the finalization of the algorithm, coefficients and threshold

values, the verification process was initialized. To prepare the standard dataset and then make it available for researchers; verification and validation of the labels given by algorithm was the most important phase. The verification is completed using 3 steps: 1. Outsourcing sample dataset labeling, 2. Standard dataset was re-labelled, 3. Verification by language experts.

5.1.1. Outsourcing Sample Dataset Labeling

The final dataset is having 107833 posts extracted from twitter. Ten thousand entries were randomly selected from this dataset and given to the expert for labeling. Later this labeled dataset was relabeled using the algorithm and the results were compared. The matching accuracy is 98.79 percent, which is above the expectations. Figure 8 shows the results and analysis:

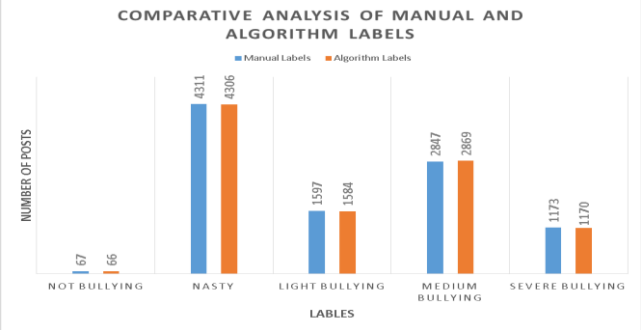


Fig 8. Comparative Analysis of Manual and Algorithm Labels

5.1.2. Standard Dataset was Re-labelled

The standard multiclass dataset is collected from Kaggle. The dataset is a multiclass and multilabel dataset which consists of a total of 159472 posts labeled using 6 different labels: toxic, severe_toxic, obscene, threat, insult and identity_hate. This dataset is relabeled using the proposed algorithm. As all of these labels are not relevant with the proposed work; only relevant labels are considered for the comparative analysis: obscene, insult and identity_hate. Total 143785 entries are matching out of 159472. Which comes to an accuracy of 90.16%.

5.1.3. Verification by Language Experts

After labeling the dataset using our algorithm it was decided to get it verified from linguistic experts. Two experts were ready to help in this work. Total 1000 entries were selected for each expert. These 1000 entries were selected in such a way to have a balanced dataset. Approximately 200 posts from each category were selected and given to the language experts. After receiving verified dataset comparison was done with algorithm annotations and corrections suggested. Following are the results:

Results:

The analysis of the dataset verified by both the experts is done by comparing the labels given by algorithm and verified labels given by experts.

Expert 1: Figure 9 shows the accuracy of the labels given by algorithm and by expert 1.

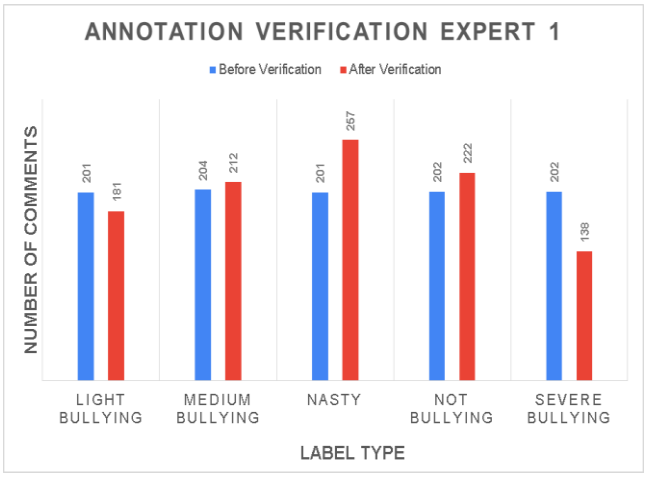


Fig 9. Annotation Verification Comparative Analysis (Expert 1)

Expert 2: Total 902 entries are matching out of 1010 entries. Accuracy Score = 89.30%. Figure 10 shows the comparative analysis of labels given by algorithm and by the expert 2.

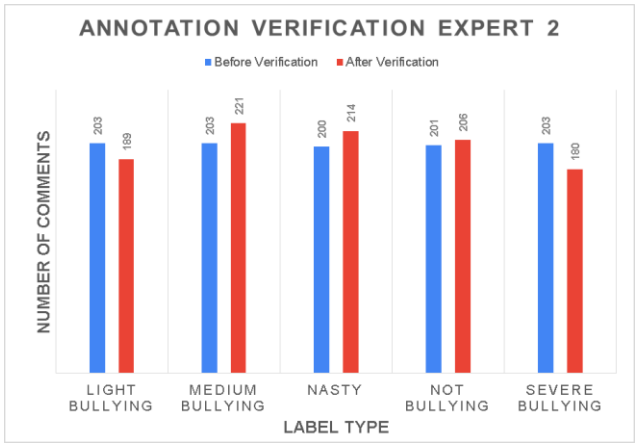


Fig 10. Annotation Verification Comparative Analysis (Expert 2)

5.2. Performance of ML Models with Designed Dataset

The newly formed dataset was then tested with machine learning models. The models which according to the literature are giving good results for cyberbullying detection are selected for the purpose [36] [28] [3] [38]. These models are tested with existing multiclass dataset and then with the newly formed dataset. Testing was also done by considering only comments as an input as well as other parameters which the new dataset is considering. A visible improvement is observed in the performance of all algorithms after considering different parameters. From the figure 11 it can be observed XGBoost is giving best results with 95.36% with the newly formed multiclass datasets. Random forest stood second with accuracy 94.1% and Decision tree with accuracy 92.57.

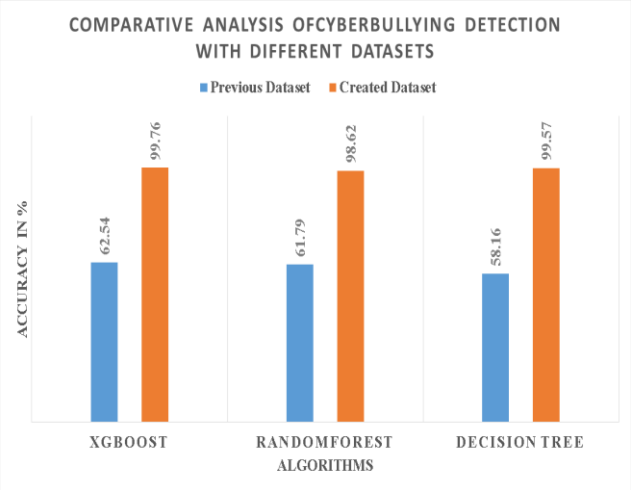


Fig 11. Accuracy Comparison of ML Models with Existing and Newly Created Datasets

The accuracy achieved through machine learning algorithms is also better than the accuracy with state of the art deep learning algorithms.

6. Conclusion

The quality and size of the dataset contributes largely to machine learning. A good corpus is the priority requirement for enhancing the process of machine learning and improving the performance of ML models. The work focused on preparing the standard unbiased and balanced dataset with appropriate labels for identifying bullying severity of the social media posts. The framework was designed and implemented for dataset annotation, by considering all the features having an impact on cyberbullying detection. Annotation done by proposed new rule based automated approach that does not contain human mind error and ambiguity. Time required for the huge size data annotation is almost negligible which in terms of manual labeling is very high. The verification process has proved the trustworthiness of the annotation provided by the proposed algorithm. The dataset of size10 lakh was annotated in the process and the future plan is to even increase the size of dataset by scraping more posts from different social media platforms, based on current trending nasty topics and hashtags. The same algorithm can be modified as per the requirement of different research areas and social media platforms for annotation of large size data. Testing after feature reduction of the models have shown improved accuracy of machine learning models over the previous dataset. Increasing the size of the dataset will also help the deep learning models to learn more and predict the outcome more accurately.

Author contributions

Madhura Vyawahare: Conceptualization, Methodology, Design, Software, Field study, Writing-Original draft preparation.
Dr. Sharvari Govilkar: Data validation, Investigation, Visualization and representation validation, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Aggarwal Akshita, Kavita Maurya, and Anshima Chaudhary. "Comparative Study for Predicting the Severity of Cyberbullying Across Multiple Social Media Platforms." In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 871-877. IEEE, 2020

[2] Agrawal Sweta, and Amit Awekar. "Deep learning for detecting cyberbullying across multiple social media platforms." In European conference on information retrieval, pp. 141-153. Springer, Cham, 2018

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

[4] Barlett Christopher P. "Predicting adolescent's cyberbullying behavior: A longitudinal risk analysis." Journal of adolescence 41 (2015): 86-95.

[5] Bayzick J., Kontostathis, A., & Edwards, L. (2018). Detecting the presence of cyberbullying using computer software. (Distinguished Honors), Ursinus College

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

[7] Dadvar Maral, and Kai Eckert. "Cyberbullying detection in social networks using deep learning based models; a reproducibility study." arXiv preprint arXiv:1812.08046 (2018)

[8] Dadvar Maral, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. "Improving cyberbullying detection with user context." In European Conference on Information Retrieval, pp. 693-696. Springer, Berlin, Heidelberg, 2013.

[9] Schmidt, Anna, and Michael Wiegand. "A survey on hate speech detection using natural language processing." In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, April 3, 2017, Valencia, Spain, pp. 1-10. Association for Computational Linguistics, 2019.

[10] Di Capua Michele, Emanuel Di Nardo, and Alfredo Petrosino. "Unsupervised cyber bullying detection in social networks." In 2016 23rd International conference on pattern recognition (ICPR), pp. 432-437. IEEE, 2016.

[11] Malmasi, Shervin, and Marcos Zampieri. "Challenges in discriminating profanity from hate speech." Journal of Experimental & Theoretical Artificial Intelligence 30, no. 2 (2018): 187-202

[12] Foong Yee Jang, and Mourad Oussalah. "Cyberbullying system detection and analysis." In 2017 European Intelligence and Security Informatics Conference (EISIC), pp. 40-46. IEEE, 2017.

[13] Fortunatus Meisy. "Classifying cyber aggression in social media posts." PhD diss., Lincoln University, 2019.

[14] Herodotou Herodotos, Despoina Chatzakou, and Nicolas Kourtellis. "A Streaming Machine Learning Framework for Online Aggression Detection on Twitter." In 2020 IEEE International Conference on Big Data (Big Data), pp. 5056-5067. IEEE, 2020.

[15] Hogenboom, Alexander, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska De Jong, and Uzay Kaymak. "Exploiting emoticons in polarity classification of text." Journal of Web Engineering (2015): 022-040.

[16] Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. "Examining the consumption of radical content on YouTube." Proceedings of the National Academy of Sciences 118, no. 32 (2021): e210196711

[17] 8.Huang Qianjia, Vivek Kumar Singh, and Pradeep Kumar Atrey. "Cyber bullying detection using social and textual analysis." In Proceedings of the 3rd International Workshop on Socially-aware Multimedia, pp. 3-6. 2014.

[18] Kumari, S. S. ., and K. S. . Rani. “Big Data Classification of Ultrasound Doppler Scan Images Using a Decision Tree Classifier Based on Maximally Stable Region Feature Points”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 76-87, doi:10.17762/ijritcc.v10i8.5679.

[19] Kouadri Wissam Mammar, Mourad Ouziri, Salima Benbernou, Karima Echihabi, Themis Palpanas, and Iheb Ben Amor. "Quality of sentiment analysis tools: The reasons of inconsistency." Proceedings of the VLDB Endowment 14, no. 4 (2020): 668-681.

[20] Whittaker, Elizabeth, and Robin M. Kowalski. "Cyberbullying via social media." Journal of school violence 14, no. 1 (2015): 11-29.

[21] Kowalski, Robin M., Susan P. Limber, and Annie McCord. "A developmental approach to cyberbullying: Prevalence and protective factors." Aggression and Violent Behavior 45 (2019): 20-32.

[22] Sudhakar, C. V., & Reddy, G. U. . (2022). Land use Land cover change Assessment at Cement Industrial area using Landsat data-hybrid classification in part of YSR Kadapa District, Andhra Pradesh, India. International Journal of Intelligent Systems and Applications in Engineering, 10(1), 75–86. https://doi.org/10.18201/ijisae.2022.270

[23] Mahlangu, Thabo, Chunling Tu, and Pius Owolawi. "A review of automated detection methods for cyberbullying." In 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC), pp. 1-5. IEEE, 2018.

[24] Garg, D. K. . (2022). Understanding the Purpose of Object Detection, Models to Detect Objects, Application Use and Benefits. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 01–04. https://doi.org/10.17762/ijfrcsce.v8i2.2066

[25] Al-Khater, Wadha Abdullah, Somaya Al-Maadeed, Abdulghani Ali Ahmed, Ali Safaa Sadiq, and Muhammad Khurram Khan. "Comprehensive review of cybercrime detection techniques." IEEE Access 8 (2020): 137293-137311.

[26] Al-Garadi, Mohammed Ali, Mohammad Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak, and Abdullah Gani. "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges." IEEE Access 7 (2019): 70701-70718.

[27] Terizi, Chrysoula, Despoina Chatzakou, Evaggelia Pitoura, Panayiotis Tsaparas, and Nicolas Kourtellis. "Modeling aggression propagation on social media." Online Social Networks and Media 24 (2021): 100137.

[28] Rosa, Hugo, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. "Automatic cyberbullying detection: A systematic review." Computers in Human Behavior 93 (2019): 333-345.

[29] Lee, Yeungjeom, Michelle N. Harris, and Jihoon Kim. "Gender Differences in Cyberbullying Victimization From a Developmental Perspective: An Examination of Risk and Protective Factors." Crime & Delinquency (2022): 00111287221081025.

[30] Mladenović, Miljana, Vera Ošmjanski, and Staša Vujičić Stanković. "Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges." ACM Computing Surveys (CSUR) 54, no. 1 (2021): 1-42.

[31] Samghabadi Niloofar Safi. "Automatic Detection of Nastiness and Early Signs of Cyberbullying Incidents on Social Media." PhD diss., University of Houston, 2020.

[32] Sugandhi, Rekha, Anurag Pande, Abhishek Agrawal, and Husen Bhagat. "Automatic monitoring and prevention of cyberbullying." International Journal of Computer Applications 8 (2016): 17-19.

[33] Talpur Bandeh Ali, and Declan O’Sullivan. "Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter." In Informatics, vol. 7, no. 4, p. 52. Multidisciplinary Digital Publishing Institute, 2020.

[34] Van Bruwaene, David, Qianjia Huang, and Diana Inkpen. "A multi-platform dataset for detecting cyberbullying in social media." Language Resources and Evaluation 54, no. 4 (2020): 851-874.

[35] Vyawahare Madhura, and Madhumita Chatterjee, ‘‘Taxonomy of Cyberbullying Detection and Prediction Techniques in Online Social Networks.” In Data Communication and Networks, pp. 21-37. Springer, Singapore, 2020.

[36] Wiguna, Tjhin, R. Irawati Ismail, Rini Sekartini, Noorhana Setyawati Winarsih Rahardjo, Fransiska Kaligis, Albert Limawan Prabowo, and Rananda Hendarmo. "The gender discrepancy in high-risk behaviour outcomes in adolescents who have experienced cyberbullying in Indonesia." Asian journal of psychiatry 37 (2018): 130-135.

[37] N. A. Libre. (2021). A Discussion Platform for Enhancing Students Interaction in the Online Education. Journal of Online Engineering Education, 12(2), 07–12. Retrieved from http://onlineengineeringeducation.com/index.php/joee/article/view/49

[38] Choi, Yoon-Jin, Byeong-Jin Jeon, and Hee-Woong Kim. "Identification of key cyberbullies: A text mining and social network analysis approach." Telematics and Informatics 56 (2021): 101504.

[39] Murshed, Belal Abdullah Hezam, Jemal Abawajy, Suresha Mallappa, Mufeed Ahmed Naji Saif, and Hasib Daowd Esmail Al-Ariki. "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform." IEEE Access 10 (2022): 25857-25871.

[40] Elsafoury, Fatma, Stamos Katsigiannis, Zeeshan Pervez, and Naeem Ramzan. "When the timeline meets the pipeline: A survey on automated cyberbullying detection." IEEE Access 9 (2021): 103541-103563.

[41] Z. Zhang, D. Robinson, and J. Tepper, ‘‘Detecting hate speech on twitter using a convolution-GRU based deep neural network,’’ in The Semantic Web, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham, Switzerland: Springer, 2018, pp. 745–760.

[42] Maity, Krishanu, and Sriparna Saha. "BERT-Capsule Model for Cyberbullying Detection in Code-Mixed Indian Languages." In International Conference on Applications of Natural Language to Information Systems, pp. 147-155. Springer, Cham, 2021.

[43] Al-Garadi, Mohammed Ali, Kasturi Dewi Varathan, and Sri Devi Ravana. "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network." Computers in Human Behavior 63 (2016): 433-443

[44] Yuvaraj, N., K. Srihari, Gaurav Dhiman, K. Somasundaram, Ashutosh Sharma, S. M. G. S. M. A. Rajeskannan, Mukesh Soni, Gurjot Singh Gaba, Mohammed A. AlZain, and Mehedi Masud. "Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking." Mathematical Problems in Engineering 2021

[45] Perera, Andrea, and Pumudu Fernando. "Accurate cyberbullying detection and prevention on social media." Procedia Computer Science 181 (2021): 605-611

[46] Sultan, Daniyar, Shynar Mussiraliyeva, Aigerim Toktarova, Marat Nurtas, Zhalgasbek Iztayev, Lyazzat Zhaidakbaeva, Lazzat Shaimerdenova, Oxana Akhmetova, and Batyrkhan Omarov. "Cyberbullying and Hate Speech Detection on Kazakh-Language Social Networks." In 2021 7th IEEE Intl Conference on Big Data Security on Cloud, IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), pp. 197-201. IEEE, 2021.