2018 International Conference on Advances in Computing and Communication Engineering (ICACCE-2018)
Paris, France 22-23 June 2018

NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms

Hitesh Kumar Sharma
School of Computer Science
University of Petroleum & Energy Studies
Dehradun, India
hksharma@ddn.upes.ac.in

K Kshitiz
School of Computer Science
University of Petroleum & Energy Studies
Dehradun, India
shailendra15@stu.upes.ac.in

Shailendra
School of Computer Science
University of Petroleum & Energy Studies
Dehradun, India
shailendra15@stu.upes.ac.in

Abstract: **In the era of social media and networking, the usage of bad words and aggressive words has been increased significantly. The young population is playing a major role in it. Cyberbullying affects more than half of the young population using social media. Insults in social media websites create negative interactions within the network. These remarks build up a culture of disrespect in cyberspace. Algorithms and tools used to understand and mitigate it are mostly inactive. Also, current implementations on insult detection using machine learning and natural language processing have very low recall rates. In short, the paper involves determining ways to identify bullying in text by analyzing and experimenting with different methods to find the feasible way of classifying such comments. We proposed a efficient algorithm to identify the bullying test and aggressive comments and analyses these comments to check the validity. NLP and Machine learning is used for analyzing the social comment and identified the aggressive effect of an individual or a group. An effective classifier acts as the core component in a final prototype system that can detect cyberbullying on social media.**

Keywords: *Cyberbullying, Natural Language Processing, Machine Learning.*

I. INTRRODUCTION

Usage of unwanted words or cyberbullying has been increased on cyber platforms by the young community. It is increasing day by day due to actively participation on social media . Users are increasing in an exponential manner and their time devotion is also increasing in the same manner. They are free to say anything on these platform without any rules and regulation. It affects the accessibility of other people for these cyber or social platform. Analyzing these comments which contain cyberbullying words may help to limiting these content on these platform so that a healthy discussion can take place for which these community platform is developed. Availability of a large number of public forum over internet has changed the way of giving own opinion about a subject. It made us easy to put our comments and thoughts about public or individual in a large group of people in a very fast manner. But at the same pace the insulting comments and thoughts are also spread over the public forum which is a matter of consideration. In earlier year cyberbullying was not taken seriously and it was ignored. The reason was low participation of users and it was suggest to screen off or disconnect if you get some kind of bullying comments. But now the scenario is totally changed. In 2017, Half population on social platform is using these comment and these can not be ignored. Major Social platform like, Twitter and Facebook also facing this major problem. Many legal cases has been filed due to these abused word used against and individual and a community.

This research work is done to explore some other methods of developing a prototype that can automatically identify cyberbullying on social media platforms. This work extends current research on cyberbullying and online harassment detection.

II. PROBLEM STATEMENT

With the proliferation of the Internet, cyber security is becoming an important concern. As we know web platforms provides easy, anytime, interactive and

978-1-5386-4485-0/18/$31.00 ©2018 IEEE

anywhere access to the online social network platforms, it also provides an attractive venue for cybercrimes as one of the example is cyberbullying and online harassment.

Cyberbullying is a term used for the content or images placed on online platforms which is not socially accepted. Abusing words, Aggressive words used by a group and individual. Twitter, YouTube, Instagram and many more are such kind of platforms where people used some unsocial words and give their opinion in very rude or aggressive manner. This is also a kind of online harassment. In USA this is considered as online threat.

The major problems in fighting cyberbullying include: finding these kind of words and sentences when it occurred on online platforms; put forward to these cases in Law Agencies and finding the responsible persons. No present online community or social media websites (for example, Facebook and Twitter; where cyberbullying is most common), incorporates a system to automatically and intelligently identify aggression and instances of online harassment on its platform. Due to non-seriousness of this major issue earlier it is not considered the issue of research but now it is in dnagerious phase. No-one can ignore this effect on cyber platform. It require a serious attention by researchers and cyber crime agencies to control this activity.

III. LITERATURE REVIEW

Since the research field of online harassment and cyberbullying is still emerging, there is only a limited amount of work available. Over the past few years, several techniques have been proposed to measure and detect offensive or abusive content/behavior on platform like Instagram [1], YouTube [2], 4Chan [3], Yahoo Finance [4], and Yahoo Answers [5]. Chen et al. [11] use both textual and structural features (for example, ratio of imperative sentences, adjective and adverbs as offensive words) to predict a user's aptitude in producing offensive content in YouTube comments, while Djuric et al. [4] rely on word embeddings to distinguish abusive comments on Yahoo Finance. Nobara et al. [6] perform hate speech detection on Yahoo Finance and news data, using supervised learning classification. Kayes et al. [5] find that users tend to flag abusive content posted on Yahoo Answers in overwhelmingly correct way.

Dinakar et al. [7] identified the aggressive and cyberbullying words from YouTube videos and decompose them for better classification. They collect these words mainly from controversial videos. Hee et al. [8] did their study on ask.fm web application. They found cyberbullying from audio content and do its study for aggressiveness. Hosseinmardi et al. [1] did their research work on images. They did it on Instagram and find the abusing content in Images and aggressiveness in representation in visual content.

The current technologies like part of speech, URLs BoW (Bag of Words), lexical features are useful for our study on this context.

Sentiment analysis plays an important role to find the abusement and aggressiveness in list of comments. It help to categories the comments in good or bad categories. In this study we made two main categories bullies and non-bullies and the use of probabilistic sentiment analysis approach is used for filtering in these two categories.

Sentiment Analysis, RNN and other same kind of research ideas helped to some extent to solve such kind of issues. Wang et al. [11] used LSTMs to predict the polarity of tweets and performed comparably to the state-of-the-art algorithms of the time. Huang et al. [12] found that hierarchical LSTMs allow rich context modelling, which enabled them to do much better at sentiment classification. Specifically, they chose to use LSTMs because it solves the vanishing gradient problem. Other researches have used Convolutional Neural Networks in sentiment analysis.

IV. OBJECTIVES

The objective of the research work is to combat online harassment and aggression by developing a prototype that can automatically detect cyberbullying and abusive behavior on social media and online communities by:

1. Extracting, collecting, and labelling the data set.
2. Preprocessing, cleaning, and experiment with various features to improve accuracy.
3. Classification of text, comment, or posts into one of the many classes.
4. Evaluation and analysis of best model.

The motivation for the work is to learn the application and implementation of Natural Language Processing

and Machine Learning in a real-world problem, i.e., cyberbullying and online harassment.

V. METHODOLOGY

We propose a novel methodology employing Natural Language Processing and Machine Learning to analyze texts and predicting abusive behavior. The pipeline involves extraction of a suitable data set from various online sources, preprocessing, ground truth building, feature engineering and selection, classification. Being a supervised learning problem, the goal is to classify a text from an online user which could be in the form of comments, and status/post updates into two categories – "Bully" & "Non-Bully".

The main starting point is to collect relevant data from various online platform. This type of data mainly consist user comment, posts, images, videos and audios.

Twitter provides free access to 1% of its tweets to open source developers through its Streaming API which could be used in case of absence of a relevant data set. After data collection and extraction, the second step involves preprocessing or cleaning of the data set – noise reduction, lowercasing, tokenization, stemming, lemmatization, stop words removal, discarding URLs and punctuations, normalization, removal of spam content and handling missing values. If the gathered data set is not labelled/classified, each sample will be labelled or categorized into "Bully" & "Non-Bully". "Bully" represents an aggressive/harassing text/comment/post update with possible signs of cyber bullying and "Non-Bully" otherwise. Depending upon the data set, there could be multiple labels, for example, "Bully", "Aggressor", "Spammer", and "None", making it a multi-classification problem. A learning algorithm learns better when a large amount of data is provided to it and it learn even better when additional information about the data is fed. In order to increase the accuracy of model, the next step would be feature engineering – extracting user, textual, and network features. Some possible features could be – Lexical Syntactic Features, TF-IDF (Term Frequency – Inverse Document Frequency), count of offensive words in a sentence, count of positive words in a sentence, count of second person pronoun in a sentence, Character 4-gram and 5-gram count, Word/Document Vectors. The final step is to perform classification using the (extracted) features and the ground truth. Naturally, different machine learning techniques can be used for this task, including probabilistic classifiers (for example, Naïve Bayes), Decision Trees, Ensembles (for example, Random Forest) or Artificial Neural Network. For evaluation and analysis of the result, various accuracy metrics – model accuracy score, test accuracy score, cross-validation score, precision, recall, sensitivity, specificity, AUC score will be noted. Based on the evaluation, a prototype of automatic detection and flagging of comments on a dummy social community would be created to demonstrate the final data product.

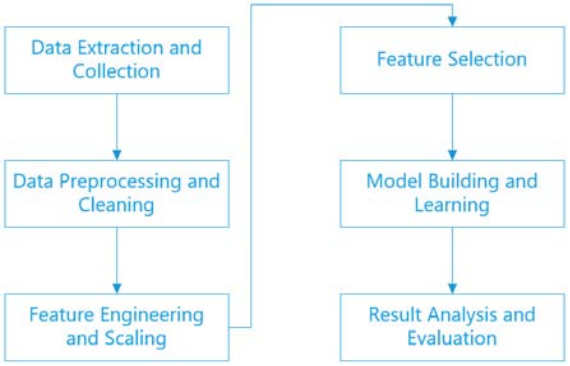

Figure 1. Data Product Pipeline

Data Extraction and Collection:

As discussed earlier, the first step towards detection of cyber bullying in the form of hate speech and insults is to get raw data sets. Data sets for cyber bullying usually consists of user comments, posts, images, and videos on social networking and social media. There are multiple sources to obtain vast amount of data sets – UCI Machine Learning Repository which houses thousands of open source data sets for data analysis purpose, Kaggle wherein individuals and businesses contributes data for research and competition, etc. Of the multiple data sets available, MySpace group data crawled in 2011, Formspring me data crawled in 2010 and those contributed by Impermium on Kaggle was collected among many other. Twitter constitutes a large ocean of data in the form of tweets and user information available for public using Twitter Streaming API and Twitter Rest API. API stands for Application Programming Interface. It is a tool that makes the interaction with computer programs and

web service easy. Many web service provides APIs to developers to interact with their services and to access data in programmatic way. The Streaming APIs give access to (usually a sample of) all tweets as they publish on Twitter. On average, about 6,000 tweets per second are posted on Twitter and developers get a small proportion (less than 1%) of it. Rest APIs are more suitable for singular searchers, such as searching historic tweets, reading user profile information, or posting tweets. The Streaming API only sends out real-time tweets, while the Search API (one of the popular REST APIs) give historical tweets to up to about a week with a max of a couple of hundreds. Streaming API is used for the purpose of the work. API Key, API secret, Access Token and Access Token secret is required to connect to the Streaming API which is obtained by creating a developer account using personal Twitter account.

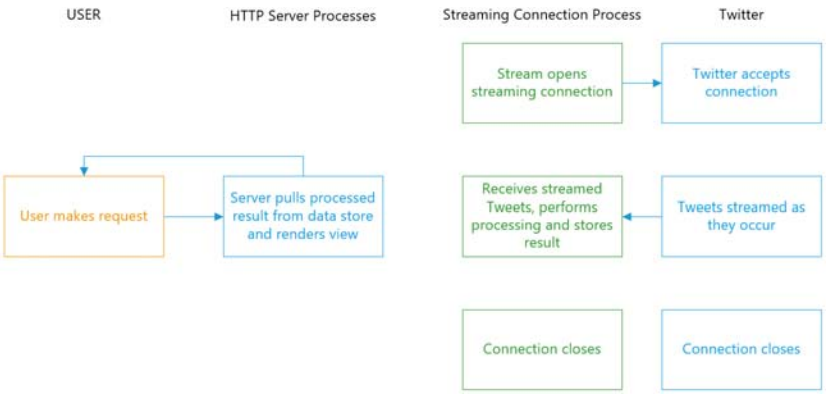

Figure 2. Architecture of Twitter Streaming API

Instances of cyberbullying and hate speech are ever increasing on the popular video streaming website YouTube.com. The research work considers extracting comment threads from popular YouTube videos that are suspected to potentially ignite hate speech. The method uses a non-API approach due to recent instability of YouTube Data API v3. YouTube comments are parsed and extracted in delimited JSON format.

Data set Description and Labelling:

Among all the collected data sets, data set on cyberbullying detection contributed by Imperium on Kaggle is selected to build the ground truth. The data consists of two attribute fields and an identifier. First Attribute shows the timestamp of the comment posted. . There are multiple null instances which means the real and accurate timestamp is not possible . It is in the form "YYMMDDHHMMSS" followed by a Z character. The time is recorded in 24 hrs format as per the local time of the location of comment posted. The next attribute is the actual content in double quotes which is shown in Unicode text. Posted content is mostly posted in English language with little bit formatting. Total number of samples in the data set is 2235. There is a small amount of noise (less than 1%) in the data set as it is not meticulously cleaned.

The gathered data set was manually labelled because it was not readily labelled. Data set labelling was the most time consuming and labor intensive. It For the purpose of detecting cyberbullying instances through hate speech and insults, each and every unit of textual data is carefully read, understood and classified. List of the possible classes – "Bully" and "Non-bully" which constitutes a binary classification problem; "Bully", "Aggressor", "Spammer" and "None" which constitutes a multi-class classification problem; bully and non-bully comments are classified as "1" and "0" respectively.

In Simple terms the data taken for machine learning is labeled into two levels "1" for meaning insulting

comment and "0" means neutral comment. The final result should lie in the range of [0,1]. If it is "1" then the comment is 100% an insulting comment. If it is 0 then comment is 100% neutral comment. Labelling the data set into two classes makes it a binary classification problem.

The data set is labelled on the basis of following guidelines:

1. Those comments which are giving an insulting impression on many blogs/forums are labelled as insulting and "1" value is given to those.
2. Non-Participants like Celebrities, public-figures, etc. related comments are labelled as neutral comments.
3. Comments which contain profanity, racial slurs, or other offensive language are labelled as insulting comments.
4. The comments related to racial slurs and not related to individual person are omitted and considered as neutral comment.
5. The insulting nature of the comment are obvious, and not subtle.

	id	Insult	Date	Comment
0	1	0	20120603163526Z	"like this if you are a tribe fan"
1	2	1	20120531215447Z	"you're idiot....."
2	3	1	20120823164228Z	"I am a woman Babs, and the only "war on women...
3	4	1	20120826010752Z	"WOW & YOU BENEFITTED SO MANY WINS THIS YEAR F...
4	5	1	20120602223825Z	"haha green me red you now loser whos winning ...
5	6	0	20120603202442Z	"\nMe and God both hate-faggots.\n\nWhat's the...
6	7	1	20120603163604Z	"Oh go the of a goat....and you DUMMY...
7	8	0	20120602223902Z	"Not a chance Kid, you're wrong."
8	9	0	20120528064125Z	"On Some real Shit LIVE JASMIN!!!"
9	10	1	20120603071243Z	"ok but where the hell was it released?you all...

Figure 3. Manually Labelled Data set

Data Preprocessing and Cleaning:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

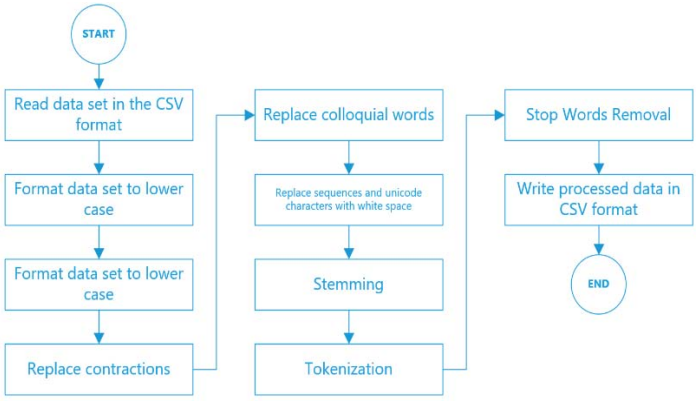

Figure 4. Data Cleaning Process

Building Machine Learning Model

All the extracted feature vector sets are stacked together into a single feature set consisting of count vectors and TF – IDF vectors of both words and characters as tokens with an n-gram sequencing of up to 5 level. The shape of the final feature set for training dataset is (6594, 4600) where 6594 is the number of samples and 4600 is the number of features or predictors. Similarly, test dataset consists of (2235, 4600) features.

For the purpose of building machine learning model, four kinds of classifiers are used – the most basic Logistic Regression, Support Vector Machine, and the two most popular ensemble methods, Random Forest Classifier and Gradient Boosting Machine. Logistic regression and support vector machine requires an input of sparse matrix of feature set whereas random forest and gradient boosting machine requires dense matrix. Thus, the sparse matrix of feature vectors is appropriately converted to dense matrix for training purpose. To improve the efficiency of the learning, various hyper parameters are studied and tuned. For instance, "C", a parameter of logistic regression and support vector machine, is the inverse of regularization strength. Smaller values in SVM specify stronger regularization. Other parameters include "number of trees in the forest" in random forest, "learning rate", "number of subsamples" in gradient boosting machine, etc.

Scikit-learn's implementation of the above classifiers in Python is used to build the models ad classify the examples. The Support Vector Machine was trained with a linear kernel on the training data. Since the training dataset is not balanced, it contains mainly negative examples, the cost-factor, a factor representing how much the cost of an error on a positive example should outweigh an error on a negative example. Tuning the parameters was done by hit and trial method where "C" assumed the values [0.002, 0.02, 0.003, 0.03, 300] ad "J", the cost factor, assumed the values [10, 30, 100]. Training time varied greatly for all the four models. Logistic regression took a total training time of 0.060 seconds whereas SVM took 6.668 seconds. Both the ensemble methods took a much larger time, 196.236 seconds and 324.447 seconds for random forest and gradient boosting machine respectively.

After training the models, they were applied to the test dataset provided on Kaggle. Lastly, the predictions generated for test dataset is written down on a file. All the values of prediction lie

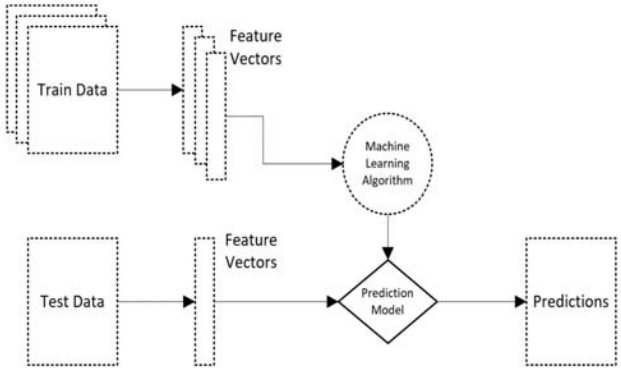

Figure 5 Building machine learning model

between 0 and 1, where a score ranging from 0 to 0.5 denotes a "non-insulting" comment whereas a score between 0.5 to 1 denotes "insulting" comment.

I. RESULTS & CONCLUSION

The results of classifying the gathered training dataset and test dataset provided by Impermium on Kaggle are displayed in Table 4. The table shows various metrics of evaluation of the performance used after training the dataset and validating it with a test dataset. Training accuracy varied from 75% - 90% for all the four classifiers while the test accuracy lies between 50% - 55%.

Table 4. Accuracy score of train and test dataset.

Model	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0 900	0 537	0 577	0 659
SVM	0 766	0 523	0 578	0 620
Random Forest	0 905	0 545	0 579	0 653
Gradient Boost	0 774	0 532	0 537	0 647

Real-time extraction of tweets using Twitter Streaming API. The tweets are filtered on the basis of keywords, in this case “Modi”, “NTPC”, and “Yogi”. The extracted tweets are formatted into a JSON file.

Although the stream also give access to a number of meta-data, a query to extract only the name of username and the tweet will exclude every other data.

```
Forgoten_Indian b"RT @pbhushan1: Gujarat ranks 25/29 in infant mortality &amp;amp; underweight children  with 39% malnourished. Thi
s is Modi's fabled Gujara "
NikhilGoyal46 b'  \nhttps://t.co/fRJnMxTm89'
rodadams46 b'RIDICULOUS #MAGA #Modi #auspol @younglabourUK https://t.co/CYhx54BeO3'
aganpat b"RT @SurajPrSingh: PM @narendramodi got Sardara Singh in a semi hug and Sardara has his hand on Modi's shoulder..@Ra_T
HORe\n\nThis is th "
YogaSharer b'RT @MuscIeFitness: RT if yoga helps your mind, body, and soul! Check out the Free eBook below! \n#Namaste #yogi #y
ogalife https://t.co/NUS6P'
allahabadwali b'RT @allanmlobo: @BJP4India @narendramodi Modi ji personally checking if they have 1000 &amp;amp;500 nots Our Indian
 Swamjiee are filthy rich look a'
reinergdrs b'@etzelgdrall direi di s, mi hanno soprannominato in tutti i modi, ma mai cos'
ranaalikash b'At a rally in Valsad district, Rahul Gandhi said PM Modi had failed to generate jobs for the people\n\nNana Pondh
a,  https://t.co/PHIuCIqhq9'
OfficeOfAS_ b'   !!! https://t.co/Zan8GmfSvu'
SultanK38669495 b'RT @HindiNews18:           \nhttps://t.co/8rJReeCnl0'
ParthkaSarthi b'RT @SushantBSinha:      8   550 , 23  .. 1400      https://t.co/T8l9NYxtFl'
rameshagrawal95 b'RT @narendramodi177: Gujarat Is Delighted To Warmly Welcome PM Modi ji, The Pride of Gujarat. #PMAtAkshardham
 https://t.co/oaLCVykihp'
Alikhansdaring b'#Happiest5WordSentence modi and company are prisoned'
siyawardas b'RT @TrueIndology: @thewire_in @rkarnad Notice the dirty tactics of leftists at work \n\nYogi Adityanath is respons
ible for oxygen thief '
allahabadwali b'RT @Saurabh83111542: @BJP4India @narendramodi Har har modi ji'
WithCongKerala b'RT @sidmtweets: If there was an election in UP today,anyone has any doubt that PM would have visited NTPC vict
ims rather than a temple prgr'
RajaShu83401000 b'@BJP4India @narendramodi @nsitharaman @PiyushGoyal V nyc modi g'
```

Figure 6. Extracting real-time tweets using Twitter Streaming API.

YouTube comments are extracted without using a API key through parsing the HTML and CSS data of a queried YouTube video using its video ID. All the comments including replies in the thread are extracted and saved in a text file in delimited JSON format.

Supervised Machine Learning is used in cyberbullying involves training a machine learning model. This work is mostly focuses on feature and pattern engineering, i.e., finding features that can separate bullying comments from non-bullying comments. It is difficult to find good features. The features used in one platform may be different in another platform. YouTube and twitter use different feature in their own platforms.

It is noticed that the model gave good results training dataset, producing a score between 77% - 90% but it is fail to generalize the test dataset. This is a case of over-fitting or large amount of variance in which the model tries its best to fit the training dataset but cannot classify the test dataset correctly. There are a number of possible reasons for such behavior:

Size of the dataset. Any machine learning algorithm performs well on a dataset containing a huge number of samples. Whereas the training dataset used for training the algorithm contains a very limited number of samples.

Differences in the dataset due to mixing social comments from two different social media platforms.

The dataset requires further data cleaning and preprocessing. Upon looking at the normalized dataset after data preprocessing step, it is found that although the preprocessing did a good job in normalizing the dataset, a lot of samples still remain inconsistent. A large number of abusive words and insults is missed out from the vocabulary because of its vastness of usage in many different forms. Apart from the abusive words, there were Unicode characters that remained in the preprocessed data. All these factors contributed greatly towards the poor performance of the model.

There were comments present in foreign languages like French and Spanish. The model only learned to classify English comments.

Requirement of different kinds of features, for example, Latent Dirichlet allocation (LDA), Latent Semantic Analysis (LSA), Predictive Word Embeddings like Word2Vec features and Doc2Vec features, etc.

The conclusion of the experiments and overall project work is that out of the method that were have evaluated, Logistic Regression and Random Forest Classifier trained on the feature stack performed better than Support Vector Machine and Gradient Boosting Machine in this particular case.

REFERENCES

[1] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing Labelled Cyberbullying Incidents on the Instagram Social Network. In *In SocInfo*, 2015

[2] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In PASSAT and SocialCom, 2012.

[3] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. A Measurement Study of 4Chan's Politically Incorrect Forum and its effort on the web. In *ICWSM*, 2017.

[4] N. Djuric, J. Zhou, R. Morris, M. Grbvoic, V. Radosavljevic, and N. Bhamidipati. Hate Speech Detection with Comment Embeddings. In *WWW*, 2015.

[5] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The Social World of Content Abuser in Community Question Answering. In *WWW*, 2015.

[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive Language Detection in Online User Content. In *WWW*, 2016.

[7] K. Dinakar, R. Reichart, and H. Liebrman. Modelling the Detection of Textual Cyberbullying. *The Social Mobile Web*, 11, 2011.

[8] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic Detection and Prevention of Cyberbullying. In *Human and Social Analytics*, 2015.

[9] V. Nahar, S. Unakard, X. Li, and C. Pang. Sentiment Analysis for Effective Detection of Cyberbullying. In *APWeb, 2012.*

[10] J. M. Xu, X. Zhu, and A. Bellmore. Fast Learning for Sentiment Analysis on Bullying. In *WISDOM,* 2012.

[11] Wang, Xin. Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. *ACL*, 2015.

[12] Huang, Minlie, Y. Cao, and C. Dong. Modelling Rich Contexts for Sentiment Classification with LSTM. *arXiv preprint arXiv: 1605.01478*, 2016.

[13] D. Santos, C. Nogueira, and M. Gatti. Deep Convolutional Neural Network for Sentiment Analysis of Short Texts. *COLING*, 2014.

[14] Divyashree, H. Vinutha, N. S. Deepashree. An Effective Approach for Cyberbullying Detection and Avoidance. *International Journal of Innovative Research in Computer and Communication Engineering*, 2016.

[15] L. Engman. Automatic Detection of Cyberbullying on Social Media. *UMEA UNIVERSITY*, 2016.

[16] R. Sugandhi, A. Pande, A. Agrawal, H. Bhagat. Automatic Monitoring and Prevention of Cyberbullying. In *International Journal of Computer Applications*, 2016.

[17] T. Chu, K. Jue. Comment Abuse Classification with Deep Learning. *Stanford University*, 2017.

[18] P. Ravi. Detecting Insults in Social Commentary. *University of Illinois at Urbana Champaign*, 2016.

[19] K. Heh. Detection of Insults in Social Commentary. *Stanford University,* 2013.

[20] R. K. Amplayo, J. Occidental. Multi-level Classifier for the Detection of Insults in Social Media. In *ResearchGate*, 2015.