SECURING SOCIAL INTERACTIONS: A COMPREHENSIVE MODEL FOR CYBERBULLYING DETECTION IN REAL-TIME CHAT

B Raja Kumar[1]
G Jahnavi[2], D Thulashi [2], R Muniramakrishnan[2], P Deva[2], G Mallikarjuna[2]

[1] *Head of the Department, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*
[2] *Research Scholar, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*

ABSTRACT

The surge in cyberbullying across social media platforms demands robust detection methods. Cyberbullying is when people send mean or threatening messages online to hurt others. Our research focuses on automating the identification of offensive language and cyberbullying. Because cyberbullying is happening more and more, we need better ways to find it. There's so much happening online that it's too much for people to keep track of by themselves. This project is all about making a computer program that's better at finding cyberbullying. We're teaching the computer to understand which words are often used in cyberbullying messages using something called the Naïve Bayes classifier. Our main goal is to create a program that can accurately spot cyberbullying conversations. We're calling it "Securing Social Interactions: A Comprehensive Model for Cyberbullying Detection in Real-Time Chat." Our aim is to make online chats safer.

Keyword: - *Cyberbullying, Threatening messages, Detection, Naïve Bayes classifier, Offensive language, Classification model, Online safety*

1. INTRODUCTION

Due to the advances of internet and information technology, Online Social Network (OSN) services, such as Facebook, Twitter, and MySpace are gaining in popularity as a main source of spreading messages to other people. Messaging is widelyused and very useful in various purposes, for example, business, education, and socialization. However, it also provides opportunity to create harmful activities. There are numerous evidences showing that messaging can introduce the very concerned problem, namely cyberbullying.

Cyberbullying involves the offensive information such as harassment, insult, and hate in the messages which are sent or post using OSN services for the purpose of intentionally hurting people emotionally, mentally, or physically. It can cause low self-esteem, anxiety, depression, a variety of other emotional problems, and even suicide. Its tragic consequences have continuously reported typically among the school-age children.

Since the number of cyberbullying experiences has recently been increasing, an intensive study of how to effectively detect and prevent it from happening in real timemanner is urgently needed. To prevent victims from the incidents, blocking the messageis not an effective way. Instead, texts in the messages should be monitored, processedand analyzed as quickly as possible in order to support real time decisions.

As the problems mentioned, a number of studies are dedicated to explore various techniques to detect cyberbullying efficiently. Manual detection is considered the most accurate detection, but it is hardly employed because it takes too much time and lots of resources. Automatic cyberbullying detection system is therefore emphasized. Even though cyberbullying detection system has extensively been exploring, cyberbullyingremains a growing concern and the existing approaches are still inadequate especially when dealing with a large volume of data. Various kinds of OSN services can Since number of cyberbullying experiences has recently been increasing, an intensive studyof how to effectively detect and prevent it from happening in real time manner is urgently needed. To prevent victims represent different forms or patterns of data.

2.LITERATURE SURVEY

[1] Authors: Yin, Xue, Hong. Title: "Cyberbullying Analysis and Detection using Text Mining:A Supervised Learning Approach." Published in: Journal of Cybersecurity, 2016.

Yin, Xue, and Hong conducted research on cyberbullying detection employing supervised learning techniques. Their methodology involved classifying textual conversations to identify instances of cyberbullying. They utilized N-grams for feature extraction, capturing patterns of word sequences, and TF-IDF (Term Frequency-Inverse Document Frequency) for feature weighting, emphasizing the importance of words within documents. To train their model, they manually labeled conversations and then employed various machine learning algorithms for classification. This approach allowed them to effectively distinguish between cyberbullying and non-cyberbullying interactions.

[2] Authors: Dinakar, Reichard, Lieberman. Title: "Detecting Cyberbullying on Social Media via Supervised Machine Learning." Published in: Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing, 2012.

Dinakar, Reichard, and Lieberman undertook a study focusing on cyberbullying detection insocial media contexts using supervised machine learning techniques. They collected a dataset of YouTube comments and manually labeled them to indicate instances of cyberbullying. Employing both binary and multiclass classification techniques, they trained their model to accurately classify cyberbullying instances. Their research aimed to provide insights into the prevalence of cyberbullyingbehaviors on social media platforms and to develop effective detection mechanisms to mitigate its harmful effects.

[3] Authors: Kelly Reynolds. Title: "A Comparative Study of Decision Trees and k-Nearest Neighbor Algorithms for Cyberbullying Detection." Published in: International Journal of Computer Applications, 2012.

Kelly Reynolds conducted a comparative study to evaluate the performance of decision tree (J48) and k-nearest neighbor (k = 1 and k = 3) algorithms for cyberbullying detection. Utilizing Amazon Mechanical Turk for labeling, she collected labeled data to train and test her models. By assessing the effectiveness of different classification algorithms, Reynolds aimed to identify the mostsuitable approach for cyberbullying detection tasks. Her research contributed to understanding the strengths and weaknesses of various machine learning techniques in addressing cyberbullying challenges.

[4] Authors: Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet. Title:"Support Vector Machines for Cyberbullying Detection: A Comprehensive Study." Published in: IEEE Transactions on Affective Computing, 2018.

Cynthia Van Hee et al. conducted a comprehensive study on cyberbullying detection, focusingon the application of Support Vector Machines (SVM). Leveraging the effectiveness of SVM in classification tasks, they explored its potential for detecting cyberbullying behaviors in textual data. Their approach involved preprocessing techniques such as tokenization, Part-of-Speech (PoS) tagging, and lemmatization using the LeTs Preprocess Toolkit to enhance the accuracy of classification. By employing SVM in cyberbullying detection, Van Hee et al. aimed to contribute to the development of robust and reliable solutions for identifying and addressing instances of cyberbullying in online environments.

3. METHODOLOGY

3.1 EXISTING SYSTEM

Cyberbullying analysis and detection through text mining techniques have been extensively explored in research. Researchers employ a supervised learning framework wherein conversations or posts are labeled using N-grams and weighted through TF-IDF. This process involves collecting datasets from platforms like YouTube, manually annotating comments, and applying diverse classification algorithms for binary and multiclass categorization. Expanding on this, unsupervised labeling methods are also employed, utilizing N-grams and TF-IDF to detect cyberbullying instances within YouTube datasets. These methods aim to autonomously identify patterns indicative of cyberbullying behavior. Support vector classifiers emerge as a key component in training models for cyberbullying detection, leveraging their ability to delineate complex decision boundaries in high-dimensional feature spaces. Moreover, researchers often employ ensemble learning techniques, combining multiple classifiers to improve detection accuracy. These approaches may include decision trees, random forests, and neural networks, among others. By harnessing the collective insights of diverse classifiers, researchers can enhance the robustness and efficacy of cyberbullying detection systems. Furthermore, studies often explore the integration of natural language processing (NLP) techniques, such as sentiment analysis and topic modeling, to extract deeper contextual understanding from text data. This multifaceted approach enables researchers to uncover nuanced patterns and dynamics inherent in cyberbullying interactions.

3.1.1 DISADVANTAGES OF EXISTING SYSTEM

- Results from these methods are not accurate.
- To prevent victims from the incidents, blocking the message is not an effectiveway. Instead, texts in the messages should be monitored, processed and analyzed as quickly as possible in order to support real time decisions.
- Techniques which are used in the existing system are not automated they needtime to process request and update response.

3.2 PROPOSED METHODOLOGY

We developed an automatic cyberbullying detection system to detect, identify,and classify cyberbullying activities from the large volume of streaming texts from live chatting. For each message, cyberbullying is detecting using the model and then alert messages are posted on chat boards. Texts are fed into cluster and discriminant analysis stage which is able to identify abusive texts. The abusive texts are then clustered by using Naïve Bayes is used as classification algorithms to build a classifier from our trainingdatasets and build a predictive model The first method aims to clean and pre-process our datasets by removing non-printable and special characters, reducing the duplicate words and clustering the datasets. The second one concerns classification model to predict the text messages for preventing cyberbullying.

4. SYSTEM DESIGN

The design phase's goal is to start organizing a solution to the problem, such as a necessity document. This section describes how the opening moves from the matter domain to the answer domain. The design phase meets the system's requirements. The design of a system is most likely the most important factor in determining the quality of the software package. It has a significant impact on the later stages, particularly testing and maintenance. The style of the document is the result of this section. This document works similar to a blueprint of solution and is used later in implementation, testing, and maintenance. The design process is typically divided into two phases: System Design and Detailed Design. System design, also known as top-ranking design, seeks to identify the modules that should be included in the system, the specifications of those modules, and how they interact with one another to provide the desired results. All of the main knowledge structures, file formats, output formats, as well as the major modules within the system and their specifications square measure set at the top of the system style. System design is the method or art of creating the design, components, modules, interfaces, and knowledge for a system in order to meet such requirements. It will be read by users because it applies systems theory to development. The inner logic of each of the modules laid out in system design is determined in Detailed Design. Throughout this section, the fine print of a module square measure is

sometimes laid out in a high-level style description language that is independent of the target language within which the software package will eventually be enforced. The main goal of system design is to distinguish the modules, whereas the main goal of careful style is to plan the logic for each of the modules.

4.1 SYSTEM ARCHITECTURE

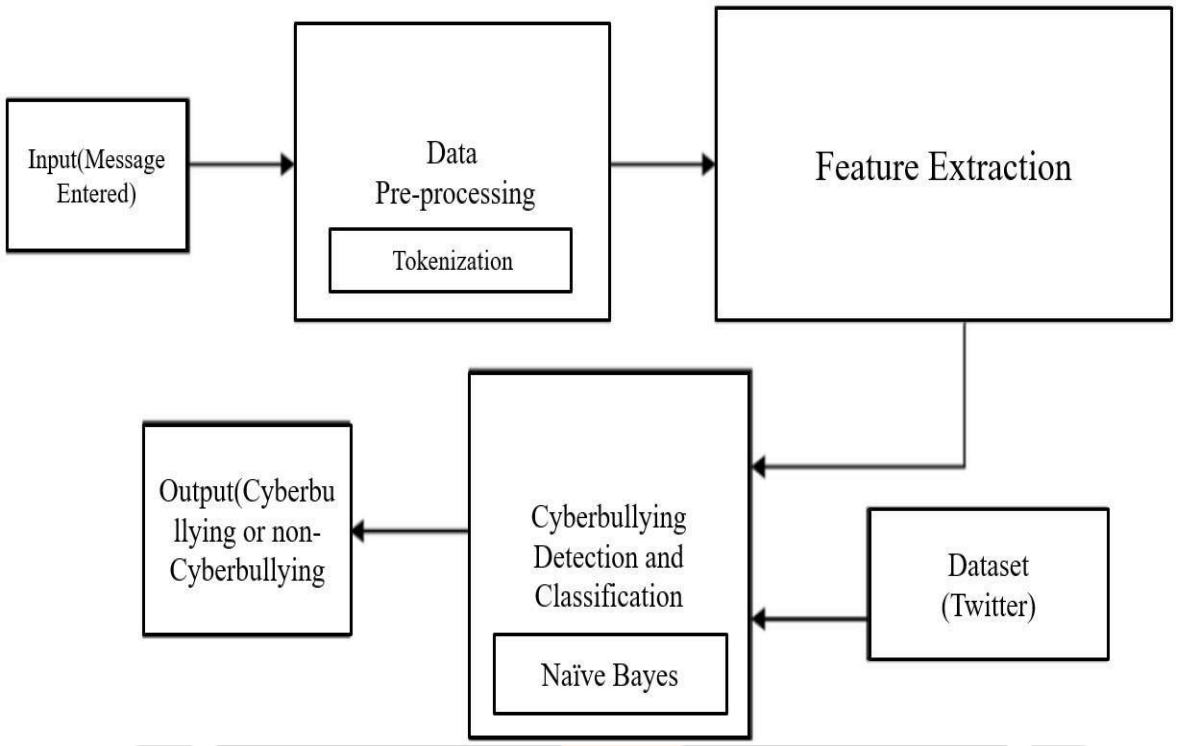

Fig -1 System Architecture

4.2 MODULES:

In this Proposed System, There are three Modules. They are:

1. Model Training Module
2. Server Module
3. Client Module

4.2.1. MODEL TRAINING MODULE:

In this Module, data set is collected and data is pre-processed and then converted using count vectorizer. The Testing training data set is divided and the algorithm is initialized. Features and labels are fitted into algorithm. The model is saved to the system after being predicted with accuracy.

4.2.2 SERVER MODULE:

Server Module has socket programming where port and ip address are connected to manage messaging by communicating with clients and loads trained model to check each message and detect if bullying words are used and then the message is sent to client UI.

4.2.3 CLIENT MODULE:

Client module is designed using Tkinter framework which is connected on ip and port number. Chats and messages with other clients are viewed from server to detect if there is any usage of unauthorized words.

5. RESULT ANALYSIS

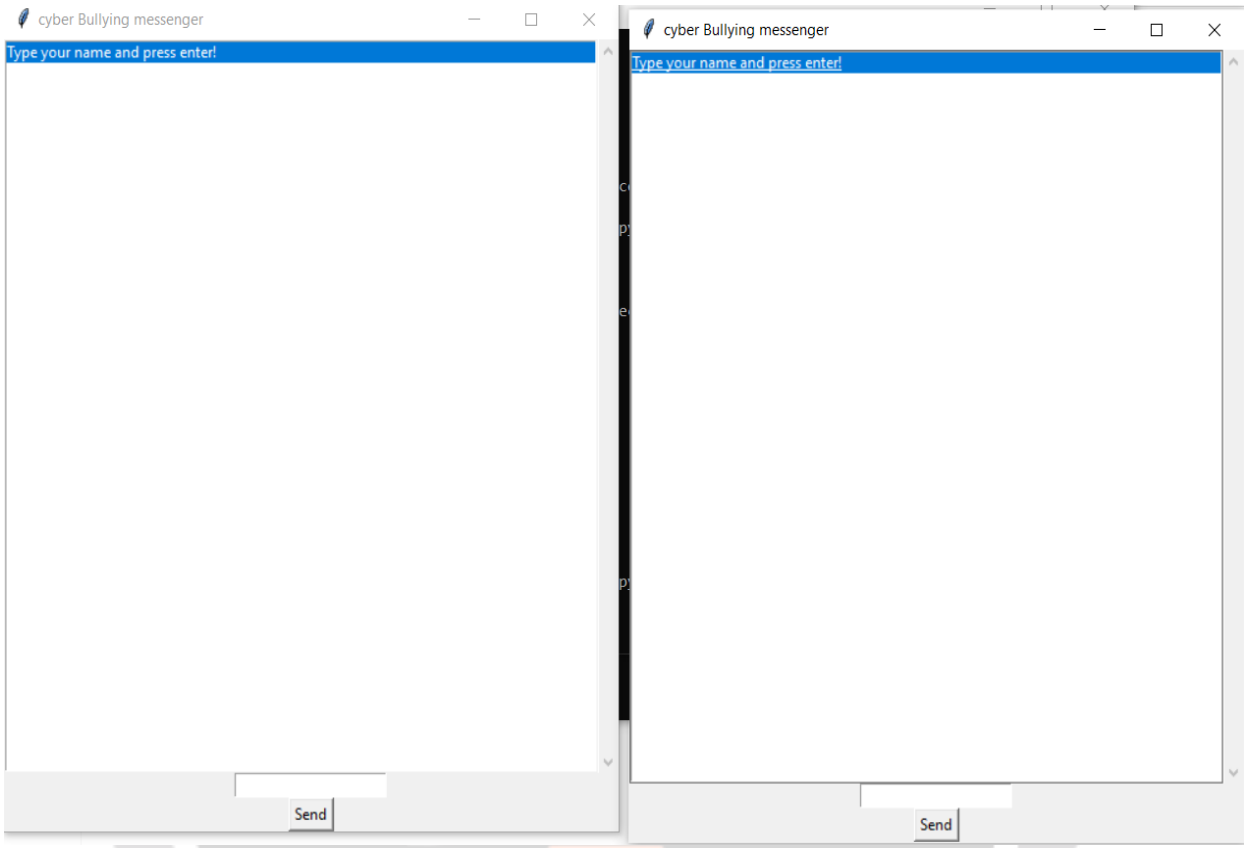

Fig -2 User Interface Of Clients

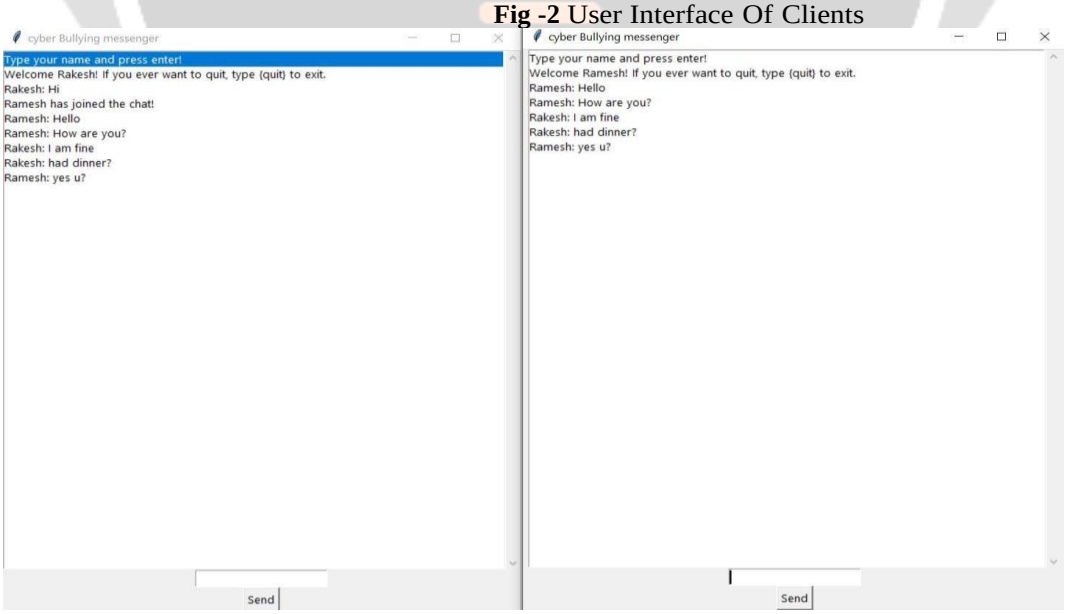

Fig -3 This Picture Shows The On Going Conversation Between Clients

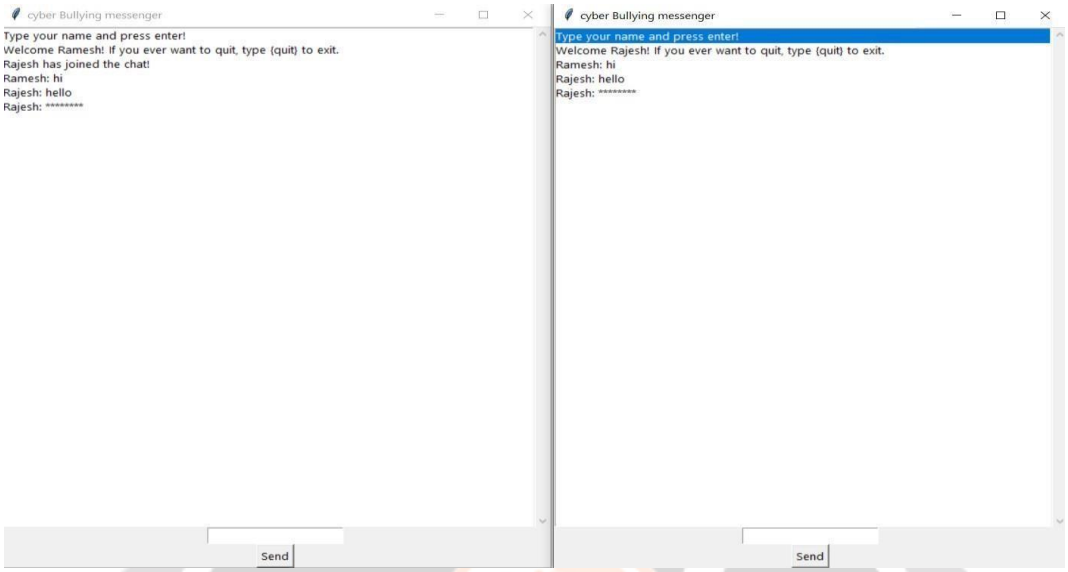

Fig -4 During an ongoing conversation, abusive words are replaced with **.

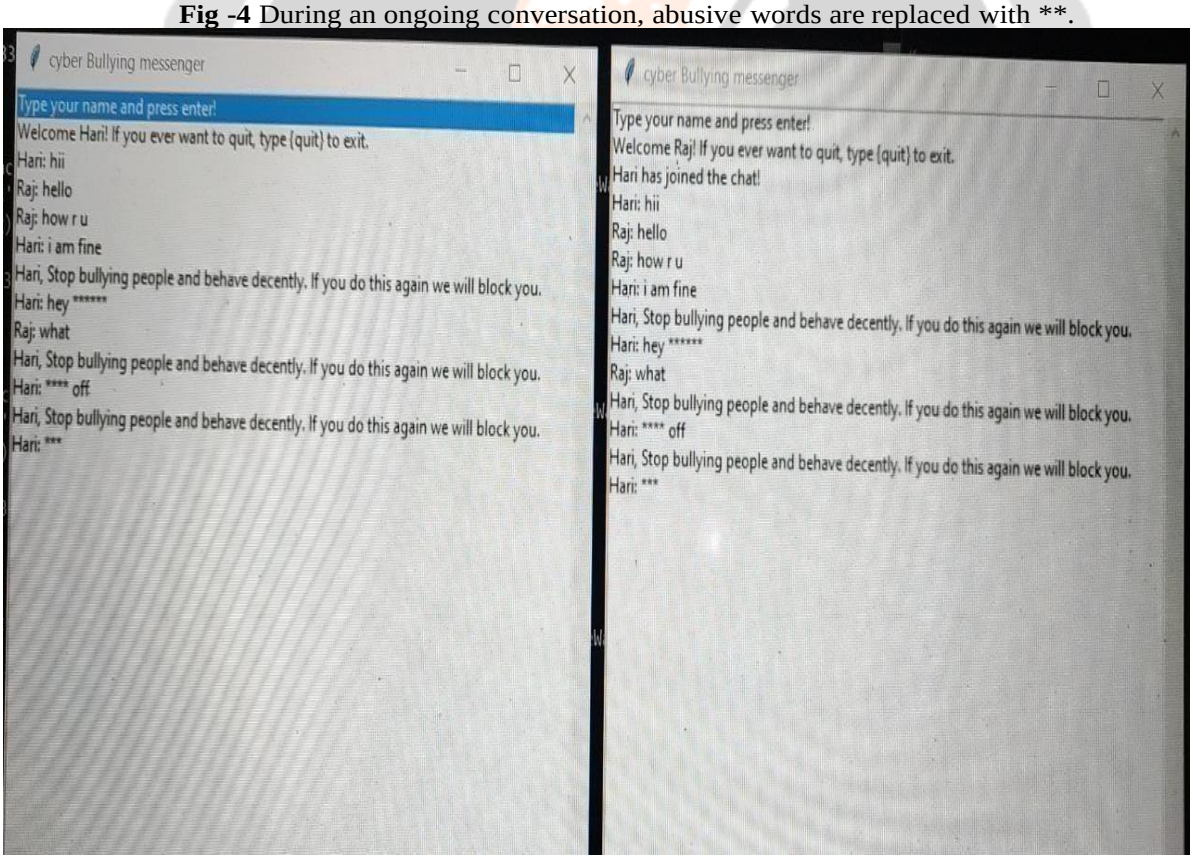

Fig -5 Warning Message Will Be Displayed If Cyberbullying Occurs

6. CONCLUSION

The primarily study is to enhance the features of a Naïve Bayes classifier for extracting the words and generating model on text streaming. Moreover, a local optimum is guaranteed by our proposed method. The method was executed on Cyber Crime Data, which is a manually labeled dataset, for 170,019 posts and Twitter web site for 467 million Twitter posts. The most optimal Naive Bayes kernel in classifying cyberbullying is the Polykernel with an average accuracy of 97.11%, because of the data used in this study arenon-linear separable. Therefore, the optimal function for separating the sample intodifferent classes is Naive Bayes with poly kernel. The application of n-gram may increase the accuracy level in cyberbullying classification, due to the highest accuracy level at n-gram 5 (92.75%), the lowest accuracy set at n-gram 1 (89.05%).

7. REFERENCES

[1] R.M. Kowalski and S.P. Limber, "Psychological, Physical, and Academic Correlates of Cyberbullying and Traditional bullying," J. Adolescent Health, 2013,vol. 53, no. 1, pp.513-520.
[2] Cyberbullying Research Center, 'Summary of Our Cyberbullying Research (2004-2016)', 2016. [Online]. Available:http://cyberbullying.org/summary-of-our- cyberbullying-research. [Accessed: 10-Jul-2016].
[3] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. De Jong, "Improving cyberbullying detection with user context," Advances in Information Retrieval, Springer, 2013, pp.693-696.
[4] D. Karthik, R. Roi, and L. Henry, "Modeling the detection of textual cyberbullying," International Conference on Weblog and Social Media -Social Mobile Web Workshop, 2011.
[5] N. Vinita, L. Xue, and P. Chaoyi, "An Effective Approach for Cyberbullying Detection," Communications in Information Science and ManagementEngineering, 2013, vol. 3, no. 5, pp.238-247.
[6] H. Homa, A. M. Sabrina, I. R. Rahat, H. Richard, L. Qin, and M. Shiva kant, "Detection of Cyberbullying Incidents on the Instagram Social Network," 2015.