IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

Cyber Bullying Detection In Hinglish Language On Social Media

Anurag Upadhyay[1*], Manish Maheshwari[2]

[1,2] Department of Computer Science And Application, MCU, Bhopal, India

[*]*Corresponding Author: anurag.upadhyaya@gmail.com, Tel.: +91-98277-22389*

DOI: https://doi.org/10.26438/ijcse/v7si10.8286 | Available online at: www.ijcseonline.org

***Abstract*—** Now a day, most of people are using twitter, face book and micro blogging sites. They share their opinion, feeling for particular topic through comment, review. The volume of data generated daily is very large. So, it is important to analyse the data for gaining information from that. Sentimental analysis is used for mining various types of data for opinion through text analytics. It can be negative, positive or impartial. Twitter became one of the largest platform for people to show their opinion, share their thoughts and consistently updated about any organization, events etc. So, data collected is huge somewhat called big data. To process such a big data we need framework that manages this entire thing. In this paper, we attempt to perform cyber bullying detection in a supervised way by proposing a learning framework. More specifically, we first investigate whether sentiment information is correlated with cyber bullying behavior.

***Keywords*—**Cyber bullying, social media, attacks, security, cyber crime

I. INTRODUCTION

Cyber bullying is an increasingly important and serious social problem, which can negatively affect individuals. It is defined as the phenomena of using the internet, cell phones and other electronic devices to willfully hurt or harass others. Due to the recent popularity and growth of social media platforms such as Facebook and Twitter, cyber bullying is becoming more and more prevalent. It has been identified as a serious national health concern by the American Psychological Association and the White House. In addition to that, according to the recent report by National Crime Prevention Council, more than 40% of the teens in the US have been bullied on various social media platforms Dinakar et al. (2012). The victims of the cyber bullying often suffer from depression, loneliness, anxiety, and low self-esteem Xu et al. (2012).

In more tragic scenarios, the victims might attempt suicide or suffer from interpersonal problems. Since cyber bullying is not restricted by time and place, it has more insidious effects than traditional forms of bullying Squicciarini et al. (2015). Traditional mechanisms to combat cyber bullying behaviours include the development of standards and guidelines that all users must adhere to, employment of human editors to manually check for bullying behaviour, the use of profane word lists, and the use of regular expressions. However, these mechanisms fall short in social media. As a result, the maintenance of these mechanisms is time and labor consuming. Also, they cannot scale well. Therefore, it necessitates the use of a learning framework to accurately detect new cyber bullying instances automatically.

The phenomenon of cyber bullying, referred to as "willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices" [1] has drastically increased in recent years, especially in youth population, mainly due to advances in computerized technology, which can cause tremendous social and financial losses to click-and-mortar organizations in recent years. For instance, Hinduja and Patchin [2] reported that 10-40% of surveyed youth population admitted to have dealt with it either as a victim or as a by-stander where adolescents use technology to harass, threaten, humiliate, or otherwise hassle their peers. Teens have also created web pages, videos, and profiles on social media platforms making fun of others, using the distinguished abilities of camera-enabled devices, which violates universal privacy standards. ScanSafe's monthly "Global Threat Report" found that up to 80% of blogs contained offensive contents and 74% included porn in the format of image, video, or offensive languages. Besides, the open online chat systems and forums have significantly increased the spread of cyber bullying cases. This has negatively impacted organization and damaged economy as a whole. It also put extra pressure on security officers. The latter face increasing challenges for various reasons. First, cyber bullying can happen 24 hours a day, 7 days a week, and reach a kid even when he or she is alone. It can happen any time of the day or night. Second, cyber bullying messages and images can be posted anonymously and distributed quickly to a very wide audience. It can be

difficult and sometimes impossible to trace the source. Third, deleting inappropriate or harassing messages, texts, and pictures is extremely difficult after they have been posted or sent.

Cyber bullies can have a mask by being anonymous on the chat rooms. Many forums and chat rooms don't require a real name to be registered as a user. This makes cyber bullies even more violent and brave. Anonymity and the lack of meaningful supervision in the electronic medium are the two factors that have aggravated this social menace. Besides, different from physical bullying, cyber bullying is "behind the scenes" as the messages, if posted on public forum, can stand for ages, creating a continuous frustration and harm to the victim, and potentially to many other users. Negative consequences of cyber bullying victim are devastating. It can have a huge affect on the growing up process of a child. The child will lose confidence, feel depressed, become anti-social and many more negative consequences that will harm a child mentally and these often affect the victims until adulthood. Some serious cases might lead to a child committing suicide but more than often resulting in tragic outcomes [3]. Boyd [4] identified four aspects of the Web that can significantly magnify the impact and damage of bullying: persistence, searchability, replicability and invisible audiences. Several attempts to deal with cyber bullying and offensive content have been reported, including many commercial products. For instance, few social networks have an "Online Safety Page" that leads to resources such as the anti-bullying sites of the government or other organizations, where the bullying issue is handled primary as a response to explicit complaints. However, the method soon becomes obsolete as the rate of daily received complaints overwhelms the ability of small groups of complaint handlers to deal with them. Other commercial solutions imitate and accommodate the spam-detection filter technology. In this respect, Appen and Internet Security Suites [5] have been endowed with moderate ability to detect and filter out online offensive contents by simply blocking web-pages and paragraphs that contained dirty words. Nevertheless, such word-based approaches fail to identify subtle offensive messages and affect web-site readability. For instance, the sentence "you are such a half-intelligent person" will not be identified as offensive content, because none of its words is included in general offensive lexicons. Besides, the approach often yields high false positive rate due to the word ambiguity problem where the word can have multiple meanings. Moreover existing methods treat each message as an independent instance without tracing the source of offensive contents.

Mishna et al. [6], among other social science researchers, explored the short and long term consequences of cyber bullying on school education, parenting and social workers. Dinakar et al. [7] in their survey paper noticed that the current detection efforts of cyber-bullying problems are largely absent or extremely naïve, while intervention efforts were largely offline and fail to provide specific actionable assessment and advice. Therefore, it becomes crucial to seek for advanced automatic cyber bullying systems. Nevertheless the potential effectiveness of such approach is still to be validated. Especially, will advanced linguistic analysis improve the accuracy and reduce false positives in detecting message-level offensiveness? Is the conceptual textual analysis efficient or secondary / third party information will be required to achieve sufficient classification result? This motivates the work highlighted in this paper where a new prototype for automatic cyber bullying detection using a combination of natural language processing technique and ad-hoc based approach. The feasibility and performance of the proposal have been assessed using some manually labelled corpus. Especially, WordNet lexical database is employed in order to identify semantically related words and evaluate the similarity with selected cyber bullying terms. On the other hand, a classification based approach is put forward in order to identify genuine cyber bullying cases.

In this paper, we attempt to perform cyber bullying detection in a supervised way by proposing a learning framework. More specifically, we first investigate whether sentiment information is correlated with cyber bullying behaviour. Then, we discuss how to deal with short, noisy, unstructured content and how to properly leverage sentiment information for cyber bullying detection. Afterwards, we present a novel optimization framework for cyber bullying detection in Hinglish language called (BDHL). To the best of our knowledge, this is the first attempt to leverage sentiment information to detect cyber bullying behaviours in hinglish like language with a learning model. To summarize, we make the following contributions:
Formally define the problem of sentiment informed cyber bullying detection in social media;

- Verify that there exists a difference of sentiment between normal posts and bullying posts by comparing the sentiment score distribution;
- Present a novel framework that leverages sentiment information of the post to detect cyber bullying in social media; and
- Perform extensive experimental studies on many real-world, publicly available social media datasets.

II. RELATED WORK

In the same spirit as natural language processing challenges tasks, e.g., misbehaviour detection task of CAW 2.0, the cyber bullying detection task is primarily focused on the content of the conversations (of the text written by the participants, both the victim and the bully), regardless the known features and characteristics of those involved. Building on some social science and psychiatry studies

(see,e.g., Mishnaa et al. [8], Hinduja and Patchin [9]), one hypothesizes that any cyber bullying case involves both Insult/Swear wording and Second person or Person name. We hypothesize when the association Insult/Swear wording and Person Name / Second person is validated then, the occurrence of cyber bullying case is enabled. Nevertheless such reasoning is not straight from natural language processing as it can see from the examples below.

"You are an idiot" is a typical example of cyber bullying as it contains both Insult/Swear word "Idiot" and Second person "You" as well as a clear association between the word and Second person. "This computer is stupid" contains only an Insult/Swear and naturally it does not promote the sentence to a cyber bullying case. "This computer is stupid despite you are hard-working person" contains both Insult/Swear word and Second person but it is not a cyber bullying case as the association between the two is not established. "I know you are not stupid" contains both Second person, Insult/Swear word and there is an established connection between the two, but it is not a cyber bullying case.

In other words, the presence of the aforementioned conditions for cyber bullying case is only a necessary condition but it does not systematically entail cyber-bullying because of the variety of natural language modifiers to express negation and opposition. The paragraph "I found you nice today. Idiot" is a cyber-bullying case despite the second sentence "Idiot" contains only an Insult/Swear wording and no Second person or Person entity, but since it refers to previous sentence, the link can easily established from an operator perspective. The above few examples demonstrate the complexity of the task of identification of cyber-bullying case using standard natural language processing tools, which requires investigating all the textual information of the phrase.

There are different features present which is used to detect sarcasm efficiently. Bharti et al[10] proposed different types of feature available to detect sarcasm easily. *First,* Lexical feature is used to detect sarcasm in only text data in which uni-grams, bi-grams and n-grams parameters used to detect sarcasm. Bi-grams and n-grams have more impact on sentimental analysis. *Second,* Hyperbole feature is used to emphasize meaning of text. In that, Interjection words have more tendencies to become sarcastic. So interjection words play important role to detect sarcasm. Another features under hyperbole are punctuation mark, quotes, intensifier is used to improve performance of system. For example, "excellent marks" has high impact rather than "good marks". So, intensifier makes task easy to detect sarcasm. *Third,* pragmatic feature is used to express emotions more accurately using smiles, emoticons, replies. So, we need to identify which type of feature is used so that accordingly algorithm is applied. In our research, we are hybrid two features that are lexical and hyperbole to improve accuracy of system.

Negation words have impact on sentimental analysis. We have to consider it to detect sarcasm because it reverses the polarity of sentence. Here, we are considered two feature lexical, hyperbole and hybrid them to improve the accuracy of system. We are considered negation feature to improve precision of sarcasm detection. Map-reduce is used to reduce execution time. It is parallel computing platform to build reliable, cost-effective, flexible application.

There are different types of sarcasm are present:
(1) Confliction between negative situation and positive sentiment. For example, *"I feel great being ignored".*
(2) Confliction between positive situation and negative sentiment. For example, *"I hate south Africa team because it always wins".*
(3) Tweet starts with an interjection word. For example, *"Wow, there is huge amount of discount but I don't buy anything!!".*
(4) Likes and Dislikes contradiction.
(5) Tweet conflicting ubiquitous facts.
(6) Tweet contains positive sentiment with antonym pair.
(7) Tweet conflicting a fact that is time sensitive.

There are many approaches available for sarcasm detection. Different authors consider various feature and approaches to improve accuracy of system. There are mainly two approaches available: (1) Machine Learning (2) Rule based method. The machine learning is a method of analysis that forms a model to predict, arrange or classify data through the statistical process. Meanwhile, rule-based approach is a technique which exploits semantic, syntactic and stylistic properties of sentences in any language such as phrase pattern, lexical and structural attributes to analyse the sentiment of a sentence. Bouazizi and Ohtsuki et al. [11] proposed supervised machine learning approach. They focus on importance of proposed set of feature to detect sarcasm and for each feature they identified different set of parameters to train the data set and tested them. Sentiment, punctuation, syntactic, semantic, pattern based feature are considered to train classifier. For classification, Random forest, maximum entropy, SVM, naïve Bayes is used. Rajadesingan et al. [12] aims to raise the difficulty in sarcasm detection on Twitter. They are utilizing behavioral viewpoints of users for expressing mockery. They employ theories from psychological and behavioral studies to construct a behavioral modelling framework for detecting sarcasm. SCUBA (Sarcasm classification using behavioral modelling approach) Model is used. Different forms of sarcasm like Sarcasm as a confliction in sentiments, complexity in expression, representation of emotion, possible function of familiarity, written expression are considered.
Tungthamthiti et al. [13] use concept level knowledge to recognize confliction between situation and sentiment. For

example, *"I like going to do job on holidays"* has positive sentiment *love* but it is actually sarcastic sentence. So, apply concept level knowledge that is holidays have relaxed situation while work has stressful situation so contradiction between them present and it considered as sarcastic. Also, focus on coherency that is correlation among sentences while multiple sentences are present to detect sarcasm. Bharti et al. [10] proposed algorithm for different types of sarcasm and also considered lexical and interjection feature to detect sarcasm. They captured and processed real time tweets using Apache Flume and Hive under the Hadoop framework, proposed a set of algorithms to detect irony in tweets with map reduce function and proposed another set of algorithms to detect mockery in tweets without map-reduce function. Riloff et al. [14] proposed bootstrapping algorithm. This algorithm automatically learns phrases which have positive sentiment and phrases having negative situation. They use tweets containing #sarcasm as a clue for the learning process. After that, Learned lists of situation and sentiment phrases are used to detect sarcasm in new tweets as a confliction of sentiment and situation context. Peter et al. [15] apply string matching algorithm which check the presence of positive sentiment and interjection lexicons. If both are present then tweet is classified as sarcastic. By focusing only on the positive sentiment, which would suggest a negative feeling, those tweets which contained negative sentiment and therefore positive feeling were ignored. Additionally, the use of interjections is not unique to sarcastic texts and many tweets may contain them where an author wishes to enhance the expressed sentiment.

Vijayalaksmi et al. [16] proposed different semi-supervised algorithm like lexical Analysis with N-grams approach, Knowledge extraction, contrast approach, emoticon based approach and hyperbole approach to propose a new rule based Hybrid approach for sarcasm detection.

III. METHODOLOGY

Our work is also closely related to detecting abusive language or harassment on the web.

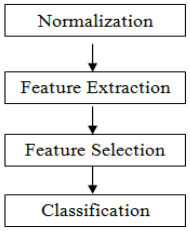

Figure 1: Proposed steps

Normalization
A raw source cannot be used directly as an input. The data from our dataset has to be modified before it can be used for Bully detection. It also helps to reduce unnecessary computation. It includes following:

- First we will remove the unwanted strings like some HTML tags or some English words that were not translated. We also remove words that have single occurrence or come too frequently as they have no effect on insult but they are necessary for grammar.
- The second task will be reducing words to their root. There are many words that have similar meanings but due to grammar or usage are modified. Since this result in unnecessary increase in number of features, we will reduce them to their root. This helps to reduce the number of features and enables the code to produce good results on lesser data.

Feature Extraction
The words have no meaning for a computer and thus have to be translated into a vector form to perform operations on it. We convert the strings into vectors which are then used by Supervised Machine Learning algorithms. Following task will be performed:

Tokenizing
Split the data into tokens. The tokens can be characters, n-grams or words. The code uses words as tokens and builds 2,3,4,5 n-grams for feature vector.

Counting
Enumerate the tokens generated in previous step for each text string. This way a matrix (generally sparse) is created, representing our data (text strings) where the number of occurrence of each token is a feature for that string. The size of matrix is S x F, where S is the size of training data and F is size of the vocabulary.

Skip-Grams
We can have long distance related features in the input data. So in addition to n-grams we use skip grams and thus increase the size of our feature matrix. This feature has a parameter which determines the number of words to skip between two words.

Negation Feature
We also added a feature that considers the words with negative implications which inverts the meaning of a string. We then give extra weight to such sentences. This has significant improvement in results.

Pragmatic Features
A pragmatic feature is used to express emotions more accurately using smiles, emoticons, replies. So, we need to identify which type of feature is used so that accordingly algorithm is applied.

Feature Selection

Since the number of features generated is high, it will be inefficient to compute directly on all of them. They might not be as important in deciding if a string is an insult or not. We will use feature selection.

Classification

This is the final step. We now apply machine learning algorithms to learn a classifier. We are using SVM and Logistic Regression to train our classifier and then combine the results of both algorithms to obtain a final classifier. This classifier is then used to classify whether a given string is an insult or not.

IV. CONCLUSION

The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

REFERENCES

[1] Ditch the Label. The annual cyberbullying survey. http://ditchthelabel.org/downloads/cyberbullying2013.pdf. [Online; accessed January-2016].

[2] S. Hinduja and J. W. Patchin, "Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying," Youth Violence And Juvenile Justice, vol. 4, 2006, pp. 148–169.

[3] H. Cowie. Cyberbullying and its impact on young people's emotional health and well-being. The Psychiatrist, 37(5):167-170, 2013.

[4] D. Boyd, *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. MacArthur Foundation Series on Digital Learning, Youth, Identity, and Digital Media, David Buckingham, Ed., MIT Press, Cambridge, MA, 2007

[5] Cybersafetysolutions.com. What can i do if i am cyber bullied. http://www.cybersafetysolutions.com.au-/factwhat- to-do-if-i-am-bullied.shtml. [Online; accessed January-2016]

[6] F. Mishra, M. Saini, and S. Solomon, Ongoing and online: Children and youth's perceptions of cyber bullying. *Children Youth Services Rev. 31*, 12, 1222–1228, 2009

[7] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In The Social Mobile Web, 2011.

[8] F. Mishna, M. Saini, and S. Solomon, Ongoing and online: Children and youth's perceptions of cyber bullying. Children Youth Services Rev. 31, 12, 1222–1228, 2009.

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

[10] Bharti, S. K., Babu, K. S., & Jena, S. K, "Parsing-based sarcasm sentiment recognition in twitter data," 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining(ASONAM), Paris, 2015, pp. 1373-1380.

[11] M. Bouazizi and T. Otsuki Ohtsuki, "A Pattern-Based Approach for Sarcasm Detection on Twitter," in *IEEE Access*, vol. 4, pp. 5477- 5488, 2016.

[12] Rajadesingan, A., Zafarani, R., & Liu, H. (2015). "Sarcasm detection on twitter: A behavioral modelling approach." 2015 WSDM - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 97-106.

[13] Tungthamthiti, P., Shirai, K., & Mohd, M. (2014). "Recognition of sarcasm in tweets based on concept level sentiment analysis and supervised learning approaches." 28th Pacific Asia Conference on Language, Information and Computation, PACLIC 2014, pp. 404-413.

[14] Riloff, Ellen & Qadir, A & Surve, P & De Silva, L & Gilbert, N & Huang, R. "Sarcasm as contrast between a positive sentiment and negative situation." Proceedings of EMNLP 2013, pp. 704-714.

[15] Clews P. & Kuzma J.(2017). "Rudimentary Lexicon Based Method for Sarcasm Detection." International Journal of Academic Research and Reflection, 5(4), 24-33.

[16] N.Vijayalaksmi, Dr. A.Senthilrajan. "A hybrid approach for Sarcasm Detection of Social Media Data." International Journal of 0Scientific and Research Publications (IJSRP), Volume 7, Issue 5, May 2017.

