Prevention of Cyber Bullying Using Machine Learning Approach

[1]**Mr. Shivaprasad More**, *Assistant Professor, Sanjay Ghodawat University, Kolhapur, Maharashtra, India.*
[2]**Deveshkumar Mishra**, *Student, Sanjay Ghodawat Group of Institution, Kolhapur, Maharashtra, India.*
[3]**Satyajeet Koule**, *Student, Sanjay Ghodawat Group of Institution, Kolhapur, Maharashtra, India.*
[4]**Mukul Hupare**, *Student, Sanjay Ghodawat Group of Institution, Kolhapur, Maharashtra, India.*
[5]**Deepak Khamkar**, *Student, Sanjay Ghodawat Group of Institution, Kolhapur, Maharashtra, India.*
[6]**Mahesh Khamkar**, *Student, Sanjay Ghodawat Group of Institution, Kolhapur, Maharashtra, India.*

***

Abstract - *The online interaction among people happens mostly using social media. There are recent developments in social media. It has challenges to the research community. The challenge is to analyze the online interactions among people. There are several social networking sites where people can share their views on a particular topic. The recent research reveals that on average 20 to 40 % of all teenagers have been victimized because of online social networking sites.*

In this paper, we mainly focus on the particular form of cyberbullying. This form is nothing but a form of cyber victimization. This can be prevented by adequate detection of such harmful messages. As there is a massive load on the web information, there should be an intelligent system to detect such cyberbullying. The system should identify the potential risk automatically. In this paper, we represent the construction and annotation of a corpus. The fine-grained annotations in cyberbullying such as text categories, insults, abusive words involve cyberbullying. The dataset has the construction of curse words and abusing words. The identification and intimation are done. We present the proof of concept experiments on automatic identification. The fine-grained annotations are used for the identification of categories of cyberbullying.

Key Words*: *Cyberbullying prevention; Text classification, Dataset construction

1. INTRODUCTION

Web 2.0 has risen and has substantially affected the relationship and communication in today's society. The different forums or blogs and social networking sites such as Twitter, WhatsApp are important means of communication. These are especially attracted towards the children or teenagers. As children are using the web, the internet is perfectly safe and enjoyable but there are some risks involved in it. Social media communities can be harmful to teenagers. The youngers can be confronted with threatening situations. The threatening situation can be cyberbullying.

This paper focuses on cyberbullying. This is one of the problems that emerged with the growing popularity of social media. The adoption of these social media sectors in our daily lives. Social media possesses several features. These features make a convenient way for cyberbullies to target their victims. Traditional bullying was limited to schoolyards and youth movements. Cyberbullying can continue at home. Cyberbullies can reach their victims using technology. Technological devices like mobile phones or laptops at any place and at any time of the day. The online content is exposed to a large audience and it's very difficult to remove it. If any message is reposted or liked or shared then it increases the impact of an offensive or hurtful message though it is posted only once. So for many years, cyberbullying becomes a very important problem. This problem needs to handle as it's a challenge to stop cyberbullying.

A recent study among the 2000 Flemish secondary school students shows that almost 11% students of them had been bullied at least once in the six months. The online survey reports named The large-scale EU Kids Online report revealed that 17 % of students.

2. LITERATURE SURVEY

Cyberbullying is the most widely covered research topic. In the past few years, especially in the social sciences, the students have been focused on the conceptualization of cyberbullying and the occurrence of the phenomenon.

There are different types of cyberbullying that have been identified and the consequences of cyberbullying have been investigated.

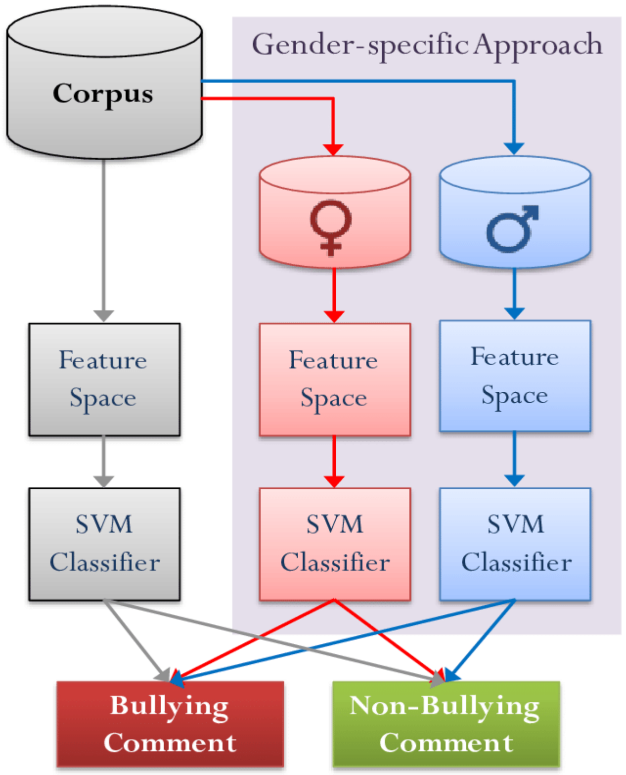
Recent studies have been focusing on the use of NLP techniques. The NLP is nothing but its Natural Language processing. The NLP focused on the use of NLP techniques for the detection and prevention of cyberbullying. The involvement of the supervised machine learning approach can be studied in the prevention of cyberbullying. Yin et al. [2] applied a supervised machine learning approach for the automatic detection of cyber harassment.

They combined local to-IDF features with sentiment features and features capturing the similarity between several posts and obtained an F-score of 0.44. Dadvar [15] applied a hybrid approach combining supervised machine learning models with an expert system that incorporates knowledge from a sociological and psychological point of view to recognize cyberbullying. They showed that combining user information and expert views with lexical features, yields fairly good results (F = 0.64). Reynolds et al. [17] applied rule-based learning to develop a model for detecting cyberbullying based on textual features.

3. LIMITATIONS OF THE EXISTING SYSTEM.

The existing system is not providing alerts of cyberbullying. The current system has limitations of the dataset. The wide variety of dataset, fine-grained annotation type of dataset is not used. Also, the system is not going to update the dataset. The old data has the limitation of updating it.

4. PROPOSED SYSTEM MODEL/ ARCHITECTURE



In the paper, we have tried to explore the feasibility of automatic classification of cyberbullying events. The binary classifier was developed for the differentiation of cyberbullying from non-cyber bullying posts. The more fine-grained text categories are related to cyberbullying. The binary classifiers are built for each of these categories. The support vector machine (SVM) as used for high-skew text classification tasks

5. METHODOLOGY

5.1 Data Collection:

We have collected data from different social media sources such as Twitter. The data may have annotations or fine-grained annotations. We have collected datasets or tweets of the Twitter website which is a social media source. We have collected tweets that have bullying words.

5.2 Data Cleaning:

Once the data is collected, data is preprocessed before using it. The empty rows and unnecessary data are removed from the dataset. There is some library in python like pandas or numpy which are used for data cleaning. Once the data is cleaned, we need to preprocess the data. There are some processing libraries in python like Scikit-learn.

5.3 Training model:

The data is split into training and testing. Generally, we keep a 70: 30 ratio or 80:20 ratio of training and testing respectively. The data is generally divided using the Scikit-learn library. Once the data is trained using the SVM algorithm, it is tested on the remaining data.

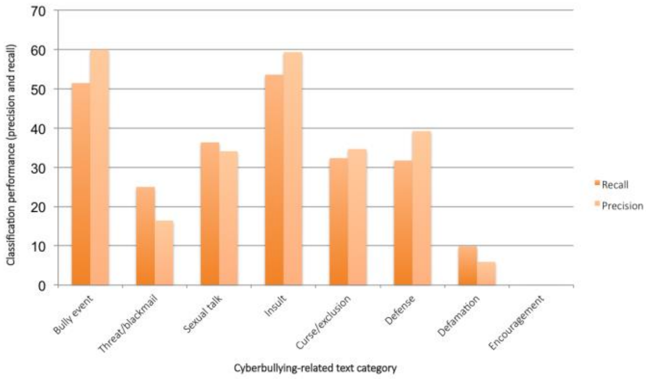
5.4 Test Model:

The model is tested for accuracy with the real-time dataset. Once the dataset is trained and tested successfully, the model is created and is used in the application. We have designed an application on the web and it's deployed. The model is tested on the new data and the accuracy is checked.

6. IMPLEMENTATION DETAILS

We explored the feasibility of automatic classification of cyberbullying events and more fine-grained text categories related to cyberbullying. To this end, binary classifiers were built for each of these categories. For our experiments, we used Support Vector Machines (SVM) as the classification algorithm, since they have been proven to work well for high-skew text classification tasks similar to the ones under investigation. We used linear kernels and experimentally determined the optimal cost value c to be 1. All experiments were carried out using Pattern.

7. RESULT ANALYSIS



To verify the results, the classification task is carried out in two ways. One is cyberbullying event detection and the other is the classification of fine-grained classification.

For the result analysis, the evaluation is done using 10-fold cross-validation our classifier is an F-score classifier and it gives a score of 55.39 %. The F-score for the fine-grained classification of cyberbullying varies considerably. The insult classifier yields an F-score of 56% and the classification performance for the categories of encouragement and defamation is significantly lower with an F-score of 0.12 % and 7 % respectively.

8. CONCLUSION

Both positive and negative experiences occur on the web. The children and youngsters are vulnerable groups and those are harmful for communication. In this paper, we have constructed a dataset of social media messages containing cyberbullying. We have proposed and evaluated the methodology for adequate annotation of this data. We have also explored the feasibility of automatic cyberbullying detection. Our initial results show that cyberbullying detection is not a trivial task.

9. FUTURE WORK

In the future, there is scope for focusing more on the final grained categories. As the ultimate goal of automatic cyberbullying detection, the technique is to reduce manual monitoring, so in the future, an alert can be added to the cyberbullying messages. We can also explore the extent author's role information to enhance cyberbullying detection. We can also use syntactic patterns, semantic information to enhance the features of cyberbullying detection techniques.

REFERENCES

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

[2] J. J. Dooley and D. Cross, "Cyberbullying versus face-to-face bullying: A review of the similarities and differences," Journal of Psychology, vol. 217, 2010, pp. 182–188, ISSN: 0044-3409.

[3] K. Van Cleemput, S. Bastiaensens, H. Vandebosch, K. Poels, G. Deboutte, A. DeSmet, and I. De Bourdeaudhuij, "Zes jaar onderzoek naar
cyberpesten in Vlaanderen, Belgi¨e en daarbuiten: een overzicht van de bevindingen. (Six years of research on cyberbullying in Flanders, Belgium and beyond: an overview of the findings).

[4] "EU Kids Online: findings, methods, recommendations."2014,URL:http://eprints.lse.ac.uk/60512/ [accessed: 2015-07-30].

[5] J. Juvonen and E. F. G, "Extending the school grounds?-Bullying experiences in cyberspace," Journal of School Health, vol. 78, 2008, pp. 496–505, ISSN: 1746-1561.

[6] R. S. Tokunaga, "Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization," Computers in Human Behavior, vol. 26, 2010, pp. 277–287, ISSN: 0747-5632.

[7] F. Dehue, C. Bolman, and T. Vollink, "Cyberbullying: Youngster's Experiences and Parental Perception," CyberPsychology, vol. 4, 2006, pp. 148–169.

[8] Q. Li, "New Bottle but Old Wine: A Research of Cyberbullying in Schools," Computers in Human Behavior, vol. 23, 2007, pp. 1777– 1791, ISSN: 0747-5632.

[9] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: its nature and impact in secondary school pupils," Journal of Child Psychology and Psychiatry, vol. 49, 2008, pp. 376–385.

[10] M. Price and J. Dalgleish, "Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People." Youth Studies Australia, vol. 29, 2010, pp. 51–59, ISSN: 1038-2569.