Review on Detection of Cyberbullying in Social Media

Pushplata Pradhan[1] Pranali Sarode[2] Neha Mandlik[3] Komal Ghodeswar[4] Krupali Ambade[5]
[1,2,3,4,5]UG Student
[1,2,3,4,5]Department of Information Technology
[1,2,3,4,5]KDKCE, Nagpur, India

Abstract— While social media offer great communication opportunities, they also increase the vulnerability of young people to threatening situations online. Recent studies report that cyberbullying constitutes a growing problem among youngsters. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. The focus of this paper is on automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders of online bullying. We have presented comprehensive reviews on Natural Language (NLP) based approach to enhance the sentiment classification by adding semantics in feature vectors and thereby using ensemble methods for classification. Adding semantically similar words and context-sense identities to the feature vectors will increase the accuracy of prediction. In addition, SHA-1 authentication mechanism will use to strengthen the authentication.

Key words: Social Network Analysis, NLP, Sentiment Analysis, SHA-1, Text Mining

I. INTRODUCTION

Social Media is a collection of Internet based utilities that build on the ideological and technological foundations of Web 2.0 which allow the users to create and exchanges of user-generated content. People can access the information and can enjoy convenient communication experience over social media and so on. However, social media may have some side effects such as cyberbullying, which may have negative impacts on the life of people, especially children and teenagers [1].

Cyberbullying can be defined as destructive or can be considered as intentional actions performed by an individual or a group of people on social media by sending messages and posting of comments against a victim[2].In traditional bullying, communication occurs face-to-face like in schooling, on other hand cyberbullying on social media can take place anywhere at any time. Anybody can express their feelings to someone without any face to face interaction. For victims, they are easily exposed to harassment since all of us, especially youth, are constantly connected to Internet or social media [3]. The same as traditional bullying, cyberbullying has negative, insidious and sweeping impacts on children. The outcomes for victims under cyberbullying may even be tragic such as the occurrence of self-injurious behaviour or suicides [4][5].

Mining opinions and analyzing sentiments from social network data help in various fields such as event prediction, analyzing overall mood of public on a particular social issue and so on. This helps in exploring potential customers, business demands, and predicting events. To understand the view or opinion implied in the text, the study of sentiment analysis is needed. Text sentiment analysis involves the application of computational powers in understanding sentiment implied in the text. Natural Language Processing involves the task of finding the semantic meaning of the text content and thereby analyzing the text information [6].

II. LITERATURE SURVEY

Rui Zhao and Kezhi Mao [1] presented Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) technique developed via semantic extension of the popular deep learning model stacked denoising autoencoder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. Authors proposed method is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text.

Dinakar and Birago Jones et al. [7] proposed the study of various aspects for social network and solutions to overcome the problem of cyberbullying by giving suggestions on the theoretical basis.

Maral Dadvar et al.[8] incorporated the users characteristics information and post-harassing behavior. Cross-system analyses of the users' behavior- monitoring their reactions in different online environments can lead to more accurate detection of cyberbullying. This suggested technique for detection of cyberbullying performed by text analysis using few words occurrence in the comments or post in social network. Comment made by male or female is also one of the analysis factor included in this research. This research on detection of bullying content using text analysis with suggested techniques has been experimented with only few words.

Yin et al., was the sole submission in the misbehavior detection task of CAW 2.0 [9]. Using three from the five datasets which were provided by the organizers of the content analysis workshop, they proposed a supervised learning approach for detecting harassment with a focus on detecting intentional annoyance. By employing a SVM classifier with the linear kernel and combining tf-idf measure as local features, sentiment features, and contextual features of documents proved that identification of online harassment provide significantly improved performance when tf-idf is supplemented with sentiment and contextual feature attributes.

Dadvar et al., [10] proposed cyberbullying detection based on gender information. Using a supervised learning approach (SVM) in order to detect cyberbullying, they proved that taking gender-specific language features into account and categorizing into male and female groups improves the discrimination capacity of a classifier to detect cyberbullying.

In a recent study on cyberbullying detection, Kontostathis et al., [11] taking a collection of posts from the website Formspring.me, which allows users to post questions

All rights reserved by www.ijsrd.com

anonymously (a question- answer website where users openly invite others to ask and answer questions) proposed a "bag-ofwords" language model, which based on the text in online posts, in order to detect instances of cyberbullying. Moreover, they exploited a supervised machine learning called Essential Dimensions of LSI (EDLSI) approach in order to identify additional terms of cyberbullying in Formspring.me data.

III. PROPOSED METHODOLOGY

The proposed system gathers data from the twitter like social networking site and processes data using NLP techniques to get feature vectors. Then ensemble methods are applied on the training data to form the training model followed by classification of the data. Authentication mechanism is provided with SHA-1 approach to strengthen the authentication process.

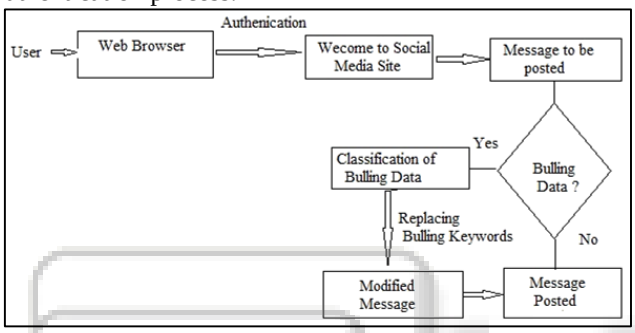

Fig. 1: Flow diagram of Proposed Methodology

The system consists of the following modules

1) Authentication module
2) Data Gathering module
3) Data Processing module
4) Training and Classification module
5) Classification output

Authentication module will created using SHA-1 authentication algorithm. Data Gathering Module will created using sophisticated programming interface with twitter like format. This format helps in extracting specific tweet attributes say user name, geographic location, re-tweet count etc. Once the data is fetched pre-processing of the gathered data is done to extract features.

Data Processing module pre-processes the tweets extracted in the previous step. The data undergo the many processes like Stop Word removal, repeated letters removal, Noise data removal etc.

Stop words are those words which normally do not carry any probable useful information but are added to make grammatically correct sentence like prepositions like on, in, to, above etc., these words do not add any information to the content. But they are always found in abundance in a sentence. So, these words are to be removed from a sentence before framing the feature vector.

Repeated letters removal process removes the extra character available in words as people emphasize their emotional state by repeating the letters of words in the tweets say ''helloooo''. In English any word contains letters repeated twice to the maximum. So, the tweets are carefully examined to look for repeated letters in words. If a letter is repeated more than twice consecutively, the number of its occurrence is reduced to two. Thus ''helloooo'' becomes "hello".

Once the key terms, Synsets, Word Senses are identified the next step is to form the feature vector. The feature vector consists of key terms of the tweets along with the synset words. The identified sense values of the words are also attached to the words. Hence for each key term in the feature vector we have two components one is the term itself and the other is the sense value of it.

Classification Ensemble Methods aims to combine the predictions of several base estimators which are traditional machine learning algorithms in order to build a model that minimizes the biasness of a single machine learning algorithm by introducing generalizability and robustness.

IV. CONCLUSION

The proposed system will collects data from Twitter like social network site and does NLP techniques to extract features out from the tweets. Natural Language (NLP) based approach will enhance the sentiment classification by adding semantics in feature vectors and thereby using ensemble methods for classification. Adding semantically similar words and context-sense identities to the feature vectors will increase the accuracy of prediction. In addition, SHA-1 authentication mechanism will strengthen the authentication.

V. REFERENCES

[1] Rui Zhao and Kezhi Mao, "Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder", IEEE Transactions on Affective Computing, 2016.
[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and metaanalysis of cyberbullying research among youth." 2014.
[3] M. Ybarra, "Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression," National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda, 2010.
[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, "Peer relations in the anxiety–depression link: Test of a mediation model," Anxiety, Stress, & Coping, vol. 23, no. 4, pp. 431–447, 2010.
[5] S. R. Jimerson, S. M. Swearer, and D. L. Espelage, Handbook of bullying in schools: An international perspective. Routledge/Taylor & Francis Group, 2010.
[6] G. Gini and T. Pozzoli, "Association between bullying and psychosomatic problems: A meta-analysis," Pediatrics, vol. 123, no. 3, pp. 1059–1065, 2009.
[7] Dinakar K, Reichart R, Lieberman H,"Modeling the detection of textual Cyberbullying", International Conference on Weblog and Social Media - Social Mobile WebWorkshop, Barcelona, 2011.
[8] Maral Dadvar, F De Jong," Cyberbullying Detection – A step towards a safer internet Yard", Proceedings of the 21st international conference, dl.acm.org, 2012.
[9] Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0

All rights reserved by *www.ijsrd.com*

in CAW 2.0 '09:", Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain,2009.

[10] Dadvar, M., F. de Jong, Ordelman, R. and Trieschnigg, D. 2012. Improved Cyberbullying Detection Using Gender Information, In Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop (DIR2012) (Ghent, Belgium 2012).

[11] Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L.,"Detecting cyberbullying: query terms and techniques", In Proceedings of the 5th annual acm web science conference (pp. 195–204), 2013.

All rights reserved by *www.ijsrd.com*