ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 6, July 2017

A Framework for Automatic Detection and Prevention of Cyberbullying in Social Media

Shruthi G[1], Prof. Mangala C N[2]

M.Tech[SCS,2nd Year], Department of Computer Science & Engineering, EWIT, Bengaluru, India[1]

Associate Professor, Department of Computer Science & Engineering, EWIT, Bengaluru, India[2]

ABSTRACT: Making use of technology as a medium to bully someone is called cyberbullying. Social networking sites provide a fertile medium for bullies, and teens and young adults who use these sites are vulnerable to attacks. Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content. The method named semantic enhanced Marginalized Stacked Denoising Autoencoder (smSDA) exploits the hidden feature structure of the bullying information.

KEYWORDS*:* Cyberbullying, Sentiment analysis, Polarity, SVM, TFIDF, LSF, EBoW.

I. INTRODUCTION

In the current era, with the popularity of Web 2.0. people are highly depended on social networking sites. With the recent development of social media, people have adopted new ways of spreading hate speech through sites such as Twitter, Facebook, Myspace which finally lead to cyber crime. Therefore, any form of bullying through the internet can be detected by extracting from microblogs, social media sites performing sentiment analysis on them. Sentiment analysis is generally denoted as techniques used to determine the pre disposition of text, usually expressed in free text form. Subjective information in source materials is recognized and extracted by the means of natural language processing, text analysis, and computational linguistics. It is used to determine an author's attitude, with respect to a particular topic or the overall contextual polarity in the text. Our work covers all features from mining texts from social media, applying sentiment analysis based on opinion of the people that is expressed on social media to finally assigning polarity to them as positive, negative or neutral. We are working on Twitter as our social network. Twitter is one of the most popular online social networks to date, where users post their opinions in short text called “tweets”. Twitter also provides the feature of Retweet (RT), which allow user to share content posted by another user. This aspect will help us know the degree spread of a hate message.

Paper is organized as follows. Section II describes the Related Works done on cyberbullying detection. Proposed Method is described in Section III. Section IV presents System Architecture. Section V presents related algorithm for detecting cyberbullying. Section VI presents the Results and discussion. Finally, Section VII presents conclusion and future work.

II. RELATED WORK

Yin et al. 2009 showed that a supervised learning approach could be used to classify harassment. They used test data from three sources: Kongregate, Slashdot and Myspace. A Support Vector Machine (SVM), a supervised learning method learning from examples to classify new data into one of two categories, trained on a Term Frequency - Inverse Document Frequency (TFIDF) model coupled with contextual and sentiment features achieved up to 40 percent precision, 60 percent recall and 45 percent F-measure. TFIDF is a way of calculating how important every word is in a document.

Dinakar et al. 2011 determined that it was possible to get better results by first labeling bullying into categories and then using a binary classifier for every category. They used sexuality, race/culture and intelligence as categories. A decision tree using JRip (an implementation of the propositional rule learning algorithm RIPPER) achieved the best accuracy while an SVM using the Sequential Minimal Optimization (SMO) algorithm for training was the most reliable method. The test data consisted of comments on Youtube videos.

A different approach was suggested in 2012 by Chen et al. They proposed a Lexical Syntactic Feature (LSF) which determines the offensiveness of a sentence based on the offensiveness of the words and the context. The offensiveness of words were measured from two lexicons. To get the context they parsed the sentence grammatically to identify dependencies between words. When a bad word could be grammatically related to a username or another bad word the offensiveness of the sentence was adjusted. The LSF method achieved 98.24 percent precision and 94.34 percent recall in detecting offensive sentences in a data set of Youtube comments. Offensive sentences were defined as sentences containing vulgar, pornographic or hateful language.

Dadvar et al. produced another paper in 2013 where they used content-based, cyberbullying- specific and user-based features to improve classification results. Using activity history of users, they trained an SVM to classify bullying in youtube comments. The results were 77 percent precision, 55 percent recall and 64 percent F-measure.

Huang et al. determined in 2014 that considering social relationships between users could improve results for classification. The experiment used regular textual analysis combined with social network features to classify bullying in a dataset from Twitter.

In 2016 Zhao et al.used a set of features they eventually named EBoW, consisting of a bag of words model combined with Latent Semantic Analysis and word embeddings by calculating word vectors. They then trained an SVM using these features to classify a biased data set from Twitter. The dataset only consisted of tweets containing one of the keywords ”bully”, ”bullying” or ”bullied”. We can see that there has been significant progress over these few years. Two different methods appear to have achieved the best results when looking at previous research. The LSF method and the weighted TFIDF features fed to an SVM both achieved remarkable levels of precision and recall.

III. PROPOSED METHOD

Our proposed Semantic-enhanced Marginalized Stacked Denoising Auto encoder is able to learn robust features from BoW representation in an efficient and effective way. These robust features are learned by reconstructing original input from corrupted (i.e., missing) ones. The new feature space can improve the performance of cyberbullying detection even with a small labeled training corpus. Semantic information is incorporated into the reconstruction process via the designing of semantic dropout noises and imposing sparsity constraints on mapping matrix. In our framework, high-quality semantic information, i.e., bullying words, can be extracted automatically through word embeddings. Finally, these specialized modifications make the new feature space more discriminative and this in turn facilitates bullying detection. Comprehensive experiments on real-data sets have verified the performance of our proposed model.

IV. SYSTEM ARCHITECTURE

Users first enters a text messages, and these text messages are stored in database. From the database, all the messages are extracted and the classification of the messages is done, that is messages are classified into bullying words and normal words. Once classification of the messages is done then the sentiment prediction is done using NLP, that is the attitude of the person who is sending the message is determined by correlating between the other words in the message. For the input messages mining rules are applied to determine the frequency of the occurrence of the words in the messages. Neurons are trained using the training dataset. Finally result of the messages is predicted that is whether the entered messages are positive, negative or neutral.

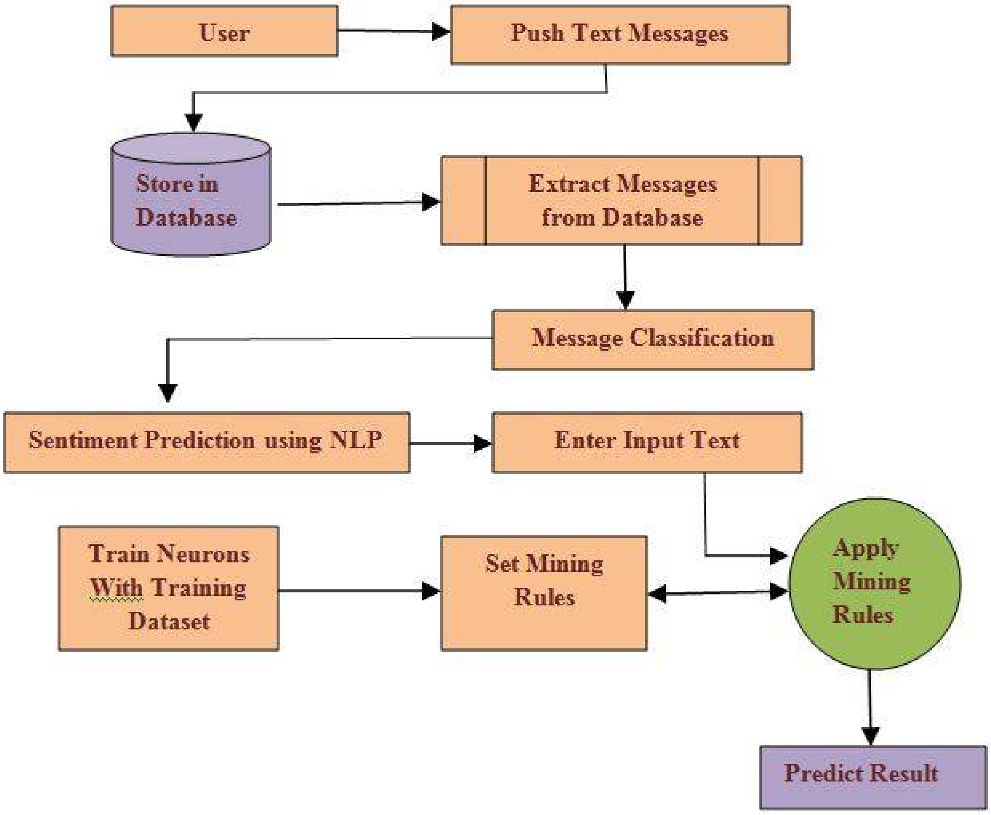


Fig 1: An Overview Architecture For Cyberbullying Detection.

V. ALGORITHM

When a person enters a comment or post in social media it goes into text pre-processing for analysing sentiment in later step. Using sentiment analysis the polarity percentage of the text can be determined. The polarity from sentiment analysis is distinguished in five ways as Very-Positive (VP), Positive (P), Neutral (NU), Negative (N) and Very Negative (VN). This algorithm helps in differentiating the text from the normal and bullying text, which in turn in protecting an individual from cyberbullying. As said, this algorithm plays a major role in categorizing the text. The algorithm with input and output is mentioned below as stepwise .

Algorithm : Polarity Count Algorithm

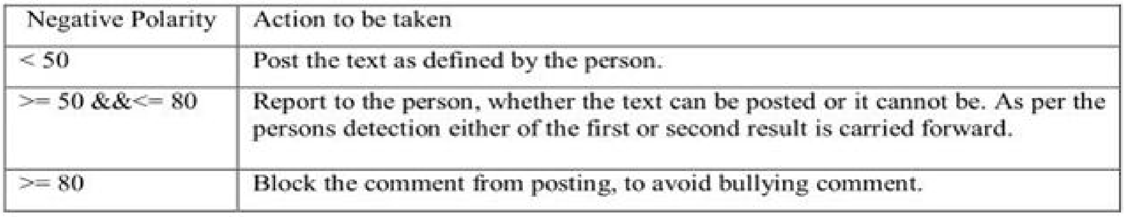
Input: Polarity values from corpus {VP, P, NU, N and VN}
Output: Identifies whether to block or report or to post the text.

```
Step 1: Neg := VN + N;
Step 2: Pos := VP + P;
Step 3: If ( Neg>= 80)
Step 4: Block;
Step 5: Else If ( Neg>= 50)
Step 6: Report;
Step 7: Else
```

Step 8: Post;
Step 9: End;

VI. RESULTS AND DISCUSSION

The polarity value goes into the polarity count algorithm, according to the algorithm the result of the process is defined The result can be either of the three followings:



Negative Polarity	Action to be taken
< 50	Post the text as defined by the person.
>= 50 &&<= 80	Report to the person, whether the text can be posted or it cannot be. As per the persons detection either of the first or second result is carried forward.
>= 80	Block the comment from posting, to avoid bullying comment.

VII. CONCLUSION AND FUTURE WORK

From the proposed idea, it is clearly known 60-70% of text cyberbullying can be avoided from posting and also reporting the person helps us to improve that people does not miss something they want to know. This helps in preventing a person from getting bullied and also protect the internet from cyberbullying crimes. Using sentiment analysis polarity of the text has been defined and also analysing the text from corpus helps in identifying the most used cyberbullying text. In addition, word embeddings have been used to automatically expand and refine bullying word lists that is initialized by domain knowledge. The performance of our approaches has been experimentally verified through two cyberbullying corpora from social Medias: Twitter and MySpace. As a next step we are planning to further improve the robustness of the learned representation by considering word order in messages.

ACKNOWLEDGEMENT

First and foremost, we express our sincere thanks to Dr. Arun Biradar, Head, Department of computer science and engineering for his moral support. We express our sincere gratitude to our principal Dr. K Chennakeshavalu for his constant support and encouragement, we also thank all the faculties of East West Institute of Technology for their co-operation and support.

REFERENCES

[1] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of social media,” Business horizons, vol. 53, no. 1, pp. 59–68, 2010.
[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, “Bullying in the digital age: A critical review and meta-analysis 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] A. Kontostathis, L. Edwards, and A. Leatherman, “Text mining and cybercrime,” Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.
[8] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, “Learning from bullying traces in social media,” in Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, 2012, pp. 656–666.
[9] Q. Huang, V. K. Singh, and P. K. Atrey, “Cyber bullying detection using social and textual analysis,” in Proceedings of the 3rd Inter-National Workshop on Socially-Aware Multimedia. ACM, 2014, pp.3–6.

[10] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, “Detection of harassment on web 2.0,” Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[11] K. Dinakar, R. Reichart, and H. Lieberman, “Modeling the detection of textual cyberbullying.” in The Social Mobile Web, 2011.

[12] V. Nahar, X. Li, and C. Pang, “An effective approach for cyberbullying detection,” Communications in Information Science and Management Engineering, 2012.

[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, “Improved cyberbullying detection using gender information,” in Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop(DIR2012). Ghent, Belgium: ACM, 2012.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, “Improving cyberbullying detection with user context,” in Advances in Information Retrieval. Springer, 2013, pp. 693–696.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

[16] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” Unsupervised and Transfer Learning Challenges in Machine Learning, Volume 7, p. 43, 2012.

[17] M. Chen, Z. Xu, K. Weinberger, and F. Sha, “Marginalized denoising autoencoders for domain adaptation,” arXiv preprint arX-iv:1206.4683, 2012.

[18] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” Discourse processes, vol. 25, no. 2-3, pp. 259–284, 1998.

[19] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” Proceedings of the National academy of Sciences of the United States of America, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.