ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse139942020.pdf
https://doi.org/10.30534/ijatcse/2020/139942020

Detection of Cyber Bullying in Social Media Engineering

M. Maheswari[1], M. Selvi[2], T. Judgi[3], G. Kalaiarasi[4], R. Yogitha[5]
[1,2,3,4,5]Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India

ABSTRACT

In the digital world the individuals in various age bunches have an enthusiasm on correspondence with social media. This increasing trend of users in the social media created a chance for cyber bullying which means harmful words communicate via electronic. From the previous research analysis one of the challenging tasks is to find out the combination of the harmful words in the dataset and it takes a long time process. We propose a technique for detection of cyber bullying is Natural Language processing along with regression technique, finding the accuracy of the statistical measures on Twitter dataset. The results show that accuracy of F1 measures produce better performance on any combination of Bully words.

Key words : Cyber Bullying, Logistic Regression, Natural Language Processing, Semantic Analysis and Twitter.

1. INTRODUCTION

Two coin sides of social media is sharing the knowledge and harmful words .In this paper we are focusing the harmful words. The detection of harmful words by the Natural Language Processing (NLP) and the combination of different classification methods. Logistic regression is the one of the classification method which classifies the sentences based on the bully words. Bully words which are found out, helps to reduce the psycho metric problems. In the internet era, teen aged guys between 11-15 are affected by cyber bullying among 42 countries [1]. In view of these they private three sorts of social use like extraordinary, tricky, and conversing with outsiders on the web [1]. The impact of cyber bullying makes the teen agers and youngsters motivate them to suicide. To protect the victims sentiment analysis make the investigation and similarities between adolescents who are all using internet in school and online. Also, various basic presumptions in regards to on the web or digital harassing were tried. [2].Numerous psychosomatic and psychosocial medical issues follow a scene of harassing exploitation. These discoveries stress the significance for specialists and wellbeing experts to build up in the case of tormenting assumes a contributing job in the etiology of such indications. Besides, our outcomes demonstrate that kids with burdensome manifestations and tension are at expanded danger of being defrauded. Since exploitation could adversy affect youngsters; endeavors to adapt to discouragement or nervousness, it is imperative to consider showing these kids abilities that could make them less powerless against harassing conduct [3]. The rest of this paper discussed the related work in Sec.2. The proposed System discussed about detection of bullying in Sec 3. In Sec. 4 shows the experimental results of the different measures. Finally Sec. 5 discusses the conclusion of the proposed work.

2. RELATED WORK

In the internet era, social media abusing is growing with the teen aged people and also middle aged people. Finding the Frequency of bully words is the problem of the sentimental analysis. S. K. Bharti et al. [4] proposed a Hadoop based system that catches constant tweets and procedure it with a lot of calculations which distinguishes wry notion effectively. It is been seen that the slip by an ideal opportunity for dissecting and preparing under Hadoop based system altogether beats the customary strategies and is more appropriate for constant spilling tweets. Shahid Shayaa1 et al. [5] presented a complete methodical writing survey, intends to examine both specialized part of OMSA (strategies and types) and non-specialized angle as application zones are talked about . Aloufi and Saddik et al. [6] built an assumption classifier which is equipped for perceiving assessments communicated in football discussion. They broadly analyzed on our dataset to look at the presentation of various learning calculations in recognizing the feeling communicated in football related tweets. Their outcomes shown that the methodology is powerful in perceiving the fans' supposition during football occasions. K. Dinakar et al. [7], tried different things with 4500 YouTube remarks about corps and applying two classifiers, for example, paired and multiclass classifiers. The outcome shows that discovery of printed digital tormenting can be handled by building singular theme delicate classifiers.

G. Gini et al. [8] proposed youngsters every now and again

associated with tormenting, especially casualties and menace casualties, experience the ill effects of psychosomatic issues. F. Godin et al. [9] proposed Named Entity Recognition (NER)for apply to micro posts and ordering the named elements to maintain a strategic distance from irregularities. R. M. Kowalski et al. [10] conceived an arrangement, including the necessity for understanding the consistent impact of cyber bullying (great past traditional bothering) on key direct and mental outcomes. Akshikumar et al. [11] proposed an investigation of delicate figuring strategies for assessment examination on Twitter. T. Mikolov et al. [12] considered the nature of vector portrayals of words inferred by different models on an assortment of syntactic and semantic language errands.

T. Mikolov et al.[13] presented a few augmentations that improved both the nature of the vectors and the preparation speed. By sub inspecting of the incessant words it has been acquired huge speedup and furthermore learn more standard word portrayals. They portrayed a basic option in contrast to the various leveled delicate max called negative testing. Divya shree et al.[14] presented positioning calculation to get to most elevated visited interface and furthermore give age confirmation before get to the specific web-based social networking. The analyses showed viability of the methodology used. Veeramallu Naga Srinivas et al.[15] proposed Semantic-redesigned Marginalized Stacked Denoising Auto encoder can take in amazing features from BoW depiction in a gainful and convincing way . H. C. Wu et al.[16] introduced epic probabilistic recovery model. It framed a premise to decipher the TF-IDF term loads as settling on pertinence choices. It reproduced the neighborhood importance dynamic for each area of a report, and joins these "nearby" significance choices as the "record wide" pertinence choice for the archive.

J.- M. Xu et al.[17] presented benchmark results on these assignments utilizing off-the-rack NLP arrangements M. R. Zhao et al. [18] proposed a Semi-Random Projection (SRP) system, which took the value of arbitrary element inspecting of RP, however utilizes learning instrument in the assurance of the change lattice. Jiameng Lia et al. [19] decided the predominance and hazard elements of customary tormenting and digital harassing in Chinese center younger students, and investigated the relationship among harassing and psychosomatic side effects [20,21].

Our Proposed system can analyze all this information and Parallely model the method based on F1 measures to detect the bullies in social media dataset.

3. PROPOSED SYSTEM

Number The architecture diagram described below shows the detailed view of working process and data flow.

The datasets are taken from different sources then preprocessed next analyzed by processing techniques like NLP(Natural processing language) and classified by using different regression techniques like logistic regression. Finally, the features are generated and build the regression model.

3.1 Preprocessing

.All the special characters like (@#$%^&*), extra spaces, numbers, links, punctuations, stopwords, capitalization, names, are removed from the dataset. At the point when we investigate explicit kinds of loathe, some can be much scarcer, for example, 'bigotry' and as referenced previously, the outrageous instance of 'both'. This has two ramifications. Initial, an assessment measure, for example, the miniaturized scale F1 that takes a gander at a framework's presentation on the whole dataset paying little mind to class contrast can be one-sided to the framework's capacity of distinguishing 'non-loathe'. As it were, a speculative framework that accomplishes practically immaculate F1 in recognizing 'prejudice' Tweets can in any case be eclipsed by its poor F1 in distinguishing 'non-loathe', and the other way around. Second, contrasted with non-loathe, the preparation information for despise Tweets are rare. This may not be an issue that is definitely not hard to address as it shows up, since the datasets are accumulated from Twitter and mirror the veritable thought of data ponderousness at the present time. Therefore to comment on all the more preparing information for derisive substance we will more likely than not need to burn through fundamentally more energy explaining non-despise. Figure 1 shows the data flow diagram.

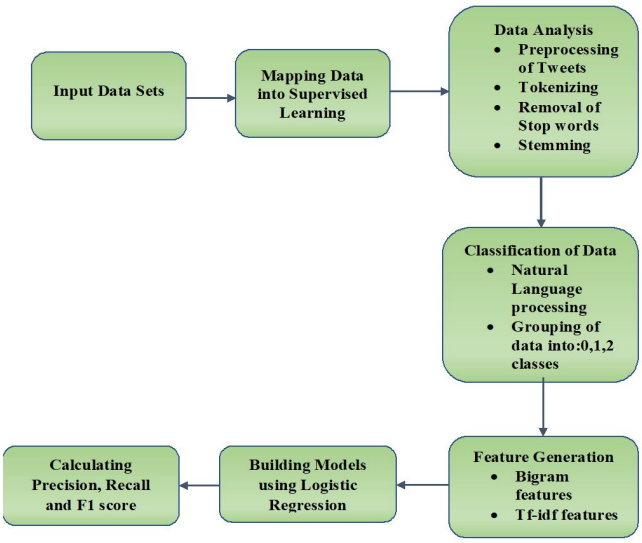

Figure 1: Data Flow Diagram

3.2 Natural Language Processing

Since we need to classify the data into abuse and non-abusive words we use NLP(Natural Language Processing).Machines cannot understand the humans language so the developers developed NLP.NLP algorithm can do many things like summarization of block texts, automatically generated keywords tags, identify the type of

entity, sentiment analysis, reduce words to the roots. The twitter dataset is full of tweets and some invalid values which the NLP doesn't understand so the data is preprocessed.

For classification of data the NLP algorithm uses sentimental analysis. Sentiment analysis is process of identifying and extracting information that underlines the text. After the preprocessing of dataset and tokenize the data text blocks into different sentences and words. Using the parts of speech tagging the tokenized data is tagged. Import the data into the model and train the model after tagging the first tweets, the model will start making its own predictions which class the data belongs to Make the corrections if the answer is wrong. Test the model by passing more data improve the accuracy of the model by checking all the false positives and false negatives and re-tag the incorrect ones. Now the model is ready pass the twitter dataset and get the classified data.

3.3 Logistic Regression

Basically Logistic regression is a statistical model for generating a binary dependent variable by using logistic function. Although there are many more complex extensions. In regression analysis, logistic regression measures the parameters of the logistic model in a form of binary regression. In Figure 2 shows the logistic regression measures.

	Positive Prediction	Negative Prediction
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Figure 1: Logistic Regression Measures

Finding the precision using

Precision = TP / (TP + FP)

Finding the recall using

Recall = TP / (TP + FN)

Finding f 1 score is

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

3.3.1 Steps of Regression

1.After the NLP the data consist of n features

2.Apply the precision, recall and f1 measures , features are modeled as 3 classes (Hate, offensive, neither)

3. Using the constructed regression model, calculate the following feature

Bigram Feature

It is a combination of two words alternatively.

TF-IDF without Additional Feature

TF = (Frequency of a term in the document)/(Total number of terms in documents)
Inverse Document Frequency(IDF) = log((total number of documents)/(number of documents with term t))

3.3.2 TF-IDF with Additional Feature

TF=(Frequency of a term in the document)+(Number of occurrences of same word)/ (Total number of terms in documents)

In this regression the classification is based on harmful word and normal word. Prediction is based on three class fields such as Hate, offensive, neither.

Here the classification is based on negative which means the hate words is the majority and the positive is normal tweets.

4. EXPERIMENTAL RESULTS

4.1 Dataset

4.1.1 Bigram Feature

In Figure 3 shows the statistical measures like Precision, Recall, F-Measure find out the probability metrics of occurrences of these three classes using the Bigram features

	precision	recall	f1-score	support
0	0.43	0.41	0.42	164
1	0.97	0.84	0.90	1905
2	0.59	0.97	0.74	410
accuracy			0.83	2479
macro avg	0.66	0.74	0.69	2479
weighted avg	0.87	0.83	0.84	2479

Accuracy Score: 0.8342073416700282

Figure 3: Bigram Features

4.2 TF-IDF without Additional Feature

Figure 4 shows the results of the TF-IDF features based on the three classes

	precision	recall	f1-score	support
0	0.56	0.12	0.19	164
1	0.90	0.97	0.93	1905
2	0.84	0.81	0.83	410
accuracy			0.88	2479
macro avg	0.77	0.63	0.65	2479
weighted avg	0.87	0.88	0.86	2479

Accuracy Score: 0.8846308995562727

Figure 4: TDIDF without Additional Features

4.3 TF-IDF with Additional Feature

Figure 5 shows the result of the TD-IDF with additional features which means the frequency of the repetitive bully words with the combination of the three classes. The result shows TF-IDF with additional features gives the better accuracy compared to the previous bigram and TF-IDF without additional features.

	precision	recall	f1-score	support
0	0.58	0.12	0.19	164
1	0.90	0.97	0.94	1905
2	0.86	0.85	0.85	410
accuracy			0.89	2479
macro avg	0.78	0.64	0.66	2479
weighted avg	0.87	0.89	0.87	2479

Accuracy Score: 0.8918918918918919

Figure 5: TDIDF with Additional Features

Finally Figure 6, the confusion matrix shows the offensive word of tweets occurs high compared with hate and neither.

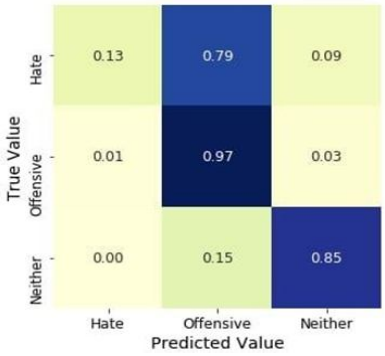

Figure 6: Confusion Matrix

5. CONCLUSION

This paper proposed a system for cyber bullying detection using novel based representation learning method, which is the combination of logistic regression with NLP and the bullying features. Prediction is based on the Bullying features which can be classified as Hate, offensive and neither. For a real world Twitter corpus the efficacy of the proposed system has been proven experimentally.

REFERENCES

1. Wendy Craig , Meyran Boniel-Nissim, Nathan King, Sophie D. Walsh , Maartje Boer , Peter D. Donnelly, M.D. f , Yossi Harel-Fisch, , Marta Malinowska-Cieslik , Margarida Gaspar de Matos, Alina Cosma, Regina Van den Eijnden, Alessio Vieno , Frank J. Elgar , Michal Molcho , Ylva Bjereld, and William Pickett, " Social Media Use and Cyber-Bullying: A Cross-National Analysis of Young People in 42 Countries", Journal of Adolescent Health, 66 (2020) ,S100eS108. https://doi.org/10.1016/j.jadohealth.2020.03.006
2. J. Juvonen and E. F. Gross. Extending the school grounds?ˆaA˘Tbullying experiences in cyberspace. ˘ Journal of School health, 78(9):496–505, 2008.
3. M. Fekkes, F. I. Pijpers, A. M. Fredriks, T. Vogels, and S. P. Verloove-Vanhorick. Do bullied children get ill, or do ill children get bullied? a prospective cohort study on the relationship between bullying and health-related symptoms. Pediatrics, 117(5):1568–1574, 2006.
4. S. K. Bharti, B. Vachha, R. K. Pradhan, K. S. Babu, and S. K. Jena, ‘‘Sarcastic sentiment detection in tweets streamed in real time: A big data approach,’’ Digit. Commun. Netw., vol. 2, no. 3, pp. 108–121, 2016.
5. Shahid Shayaa , Noor Ismawati Jaafar , Shamshul Bahri , Ainin Sulaiman , Phoong Seuk Wai , Yeong Wai Chung , Arsalan Zahid Piprani , And Mohammed Ali Al-Garadi ‘‘Sentiment analysis of big data: Methods, applications, and open challenges,’’ IEEE Access, vol. 6, pp. 37807–37827, 2018.
6. Samah Aloufi And Abdulmotaleb El Saddik , (Fellow, Ieee), ‘‘Sentiment identification in football-specific tweets,’’ IEEE Access, vol. 6, pp. 78609–78621, 2018.
7. K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In The Social Mobile Web, 2011.
8. G. Gini and T. Pozzoli. Association between bullying and psychosomatic problems: A meta-analysis. Pediatrics, 123(3):1059–1065, 2009. https://doi.org/10.1542/peds.2008-1215
9. F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle. Named entity recognition for twitter microposts using distributed word representations. In Proceedings of the Workshop on Noisy User-generated Text, pages 146–153, Beijing, China, July 2015. Association for Computational Linguistics.
10. 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.
11. Akshi kumar,Arunima jaiswal, ‘‘Systematic literature review of sentiment analysis on Twitter using soft computing techniques,’’ Concurrency Comput., Pract. Exper., p. e5107, 2019. doi: 10.1002/cpe.5107.
12. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
13. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
14. Divyashree, Vinutha H, Deepashree N S" An Effective Approach for Cyber bullying Detection and avoidance", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2016
15. Veeramallu Naga Srinivas, Veerendra Bethimeedi," Detection of Text based Cyberbullying using Semantic Enhanced Marginalized Denoising Autoencoder Learning Model ",International Journal of Computer Science and Mobile Computing, Vol.6 Issue.8, August-2017, pg. 89-94
16. H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok. Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3):13, 2008. https://doi.org/10.1145/1361684.1361686
17. 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, pages 656–666. Association for Computational Linguistics, 2012.
18. R. Zhao and K. Mao. Semi-random projection for dimensionality reduction and extreme learning machine in high-dimensional space. Computational Intelligence Magazine, IEEE, 10(3):30–41, 2015.
19. Jiameng Lia, Aissata Mahamadou Sidibea, Xiaoyun Shena, Therese Hesketha, "Incidence, risk factors and psychosomatic symptoms for traditional bullying and

cyberbullying in Chinese adolescents",2019,Children and Youth Services Review,107,104511.

20. Mohammad Rasmi Al-Mousa, "Analyzing Cyber-Attack Intention for Digital Forensics Using Case-Based Reasoning", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.6, November – December 2019, pp. 3243-3248. https://doi.org/10.30534/ijatcse/2019/92862019
21. Reem K. Alqurashi, Mohammed A. AlZain, Ben Soh*, Mehedi Masud, Jehad Al-Amri, "Cyber Attacks and Impacts: A Case Study in Saudi Arabia", International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, No.1, January – February 2020, pp. 217-224. https://doi.org/10.30534/ijatcse/2020/33912020