IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 7, July 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

Detection of Cyberbullying on Social Media Using Machine Learning

Anusha S Suregaonkar[1], Chethan GM[2], Neha RP[3],Anusha MN[4], Dr.Poornima B[5]

U.G. Student, Department of Information Science And Engineering, BIET Engineering College, Shamanur, Davangere, Karnataka, India[1,2,3,4]

Professor& Head of Department, Department of Information science and Engineering, BIET Engineering College, Shamanur, Davangere, Karnataka, India[5]

ABSTRACT: Harassment by cyberbullies is a significant phenomenon on the social media. Existing works for cyberbullying detection have at least one of the following three bottlenecks. First, they target only one particular social media platform (SMP). Second, they address just one topic of cyberbullying. Third, they rely on carefully handcrafted features of the data. We show that Machine learning-based models can overcome all three bottlenecks. Knowledge learned by these models on one dataset can be transferred to other datasets. We performed extensive experiments using three real-world datasets: Formspring (~12k posts), Twitter (~16k posts), and Wikipedia(~100k posts). Our experiments provide several useful insights about cyberbullying detection. To the best of our knowledge, this is the first work that systematically analyzes cyberbullying detection on various topics across multiple SMPs using Machine learning- based models and transfer learning.

I. INTRODUCTION

1.1 DESCRIPTION With the advent of internet and technology, social media has emerged as a major part of our life. It helps us keep in touch with one another with the use of different applications with just a few taps and/or swipes. It is a constant source of entertainment. People have started feeling more sociable despite their current situation, even if they are at home or at work. With our smartphones and tablets the social media platforms are easily accessible, there has been a rise in the number of users over the past few years. The global digital report created by Dave Chaffey in 2018 [1] indicates the following statistics related to internet user – there are around 4.021 billion Internet users, 3.196 billion social media users and 5.135 billion mobile phone users. However, social media has its own difficulties and challenges. For example, social media may contain a lot of antisocial behavior, including cyberbullying, cyber stalking, and cyber harassment. These behaviors have now become part our lives and are not only bounded to juveniles, but any person can be a victim of it. immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuffs they do is bullying each other over the internet. Often this internet fight results into real life threats for some individual. Some people have turned to suicide. It is necessary to stop such activities at the beginning.

II. RELATED WORK

This chapter discusses the problem of cyberbullying and presents an overview of previous work done by others in this domain to detect bullying.

2.1 Research Related to Cyberbullying Detection Many researchers have studied and analyzed effects of cyberbullying on society. “The emotional and psychological consequences of cyberbullying have been broadly studied by Hinduja et al. in” [16]. Many surveys have been carried out each year to study the nature of cyberbullying and generate guidelines that can benefit victims to accord with the problem [6][7]. In some instances, the expert guidance is to keep a watch on children’s social media activity [17]. Although these approaches seem useful, they are lacking to address the overall problem of automatic cyberbullying detection, as children report minimal number of incidents to their parents or social media directly [18]. Moreover, regardless of these procedures, cyberbullying behaviour is growing in today’sworld[19].

2.2 Content Filtering Software There are many content filtering software systems available in the market including BullyBlocker [20], SafeChat [21], Facebook WatchDog [22], and Rethink [23]. These software systems can be used by parents to monitor the social media activity of children and thus can help detect and prevent cyberbullying. A few instances of such software systems are listed below: Bully Blocker: They created a computational model to detect and

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

compute the intensity of cyberbullying in social media platforms. This app detects cyberbullying on Facebook and notifies a parent/ teacher when cyberbullying occurs. SafeChat: This is a software tool for Facebook, which protects children's communication from explicit messages. They have developed a model which look for specific words in communication and drops such messages. Facebook WatchDog: This software detects threats and cyberbullying on social media. They detect threats using social media analytics, image analysis, and text mining techniques. Rethink: This app can be installed on any device. As soon as a user types some word, this app runs in thebackground and looks for abusive word. It then shows a warning message, if that word belongs to bullying category. Although these software can provide a method for parents to keep an eye on children's activity, but it often fails to fully exploit the potential of such software as it can be easily bypassed by witty children [24]. Further, there is no collaboration between the different software of similar type. In addition, parent tends to fail to utilize the software, so such instances go unpunished and without reported to trusted authority.

2.3 Cyberbullying Detection Techniques Cyberbullying detection approaches mainly fall under two categories: machine learning techniques and lexicon-based techniques.

2.3.1 Machine Learning Techniques In this approach, different supervised learning and unsupervised learning algorithms are used to detect cyberbullying. Common steps followed in this approach are: gathering a dataset and tagging it, preprocessing on the dataset, training the machine learning model and testing it on some portion of the dataset. Here we describe prominent approaches that mainly deal with multiple languages. In [25], Haider et al. conducted a survey on multilingual cyberbullying detection. In the survey, they found that most of the work in this domain is focused on English texts. They attempted cyberbullying detection in the Arabic language in [26]. In their work, they used the ML learning approach to detect cyberbullying. Their dataset contains 32K tweets out of which 1800 tweets were bullying ones. They used Support Vector Machine (SVM) and Naïve Bayes (NB) algorithms to detect cyberbullying and got 92% and 90% F1 scores respectively. Ting et al., in [27], gathered a dataset from 4 popular social sites in Taiwan. They used Social Network Mining (SNM) technique to detect cyberbullying. In this, they identify three features from the data: Keywords, Social Network Analysis, and Sentiment. They have identified that sentiment is the most important feature to detect cyberbullying, as it helpsto understand the intent of a user when he posts messages on social media. They used precision and recall as performance metrics and their results show the precision and accuracy are around 0.79 and the recall is 0.71Nurrahmi et al., in [30], proposed a model, that not only detects cyberbullying but also keeps track of bullying between users, which can help to determine the credibility of the user. Their aim is to detect cyberbullying for the Indonesian Language. For this, they gathered around 700 tweets out of which 300 were bullying ones. They developed a machine-learning model using SVM and got an F-1 score of 67%. To determine a user's credibility, they keep track of number of bullying tweets and non-bullying tweets sent by that user. Based on this data, they calculated the user's behavior and then classified it as a bullying actor if the abnormal behavior is >60%. All the above-mentioned related work using machine-learning techniques have these limitations: i) Most of the work is done in only the English language, and ii) Machine- learning approaches do not take language inputs such as grammar and negation handling into the consideration.

2.3.1 Lexicon based techniques These methods are based on the simple Bag-of-Words (BoW) technique. In this approach, a corpus of delicate, abusive, and unpleasant words is created. At that point, algorithms use this corpus to check the occurrences of these words in messages to detectbullying. In [33], Chen et al. presented a method called Lexical Syntactic Feature (LSF) for detecting cyberbullying. LSF highly relies on BoW, for message-level abuse recognition. They achieved a precision of 98.24% and recall of 94.34% in sentence level abuse detection. In [34], Kontostathis et al. have analyzed a certain set of words that are used by cyberbullies and their context. These words are then used to form a query which can analyze cyberbullying. All these lexicon-based techniques have the following limitations: i) they depend on the dictionary of words and weights associated with it; and ii) people use slang language/misspell the word while texting and those words might not appear in the dictionary and hence, chances are high that score of the entire text message change to exact opposite

III. METHODOLOGY

INTRODUCTION

System design thought as the application of theory of the systems for the development of the project. System design defines the architecture, data flow, use case, class, sequence and activity diagrams of the project development.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

MODEL DESCRIPTION

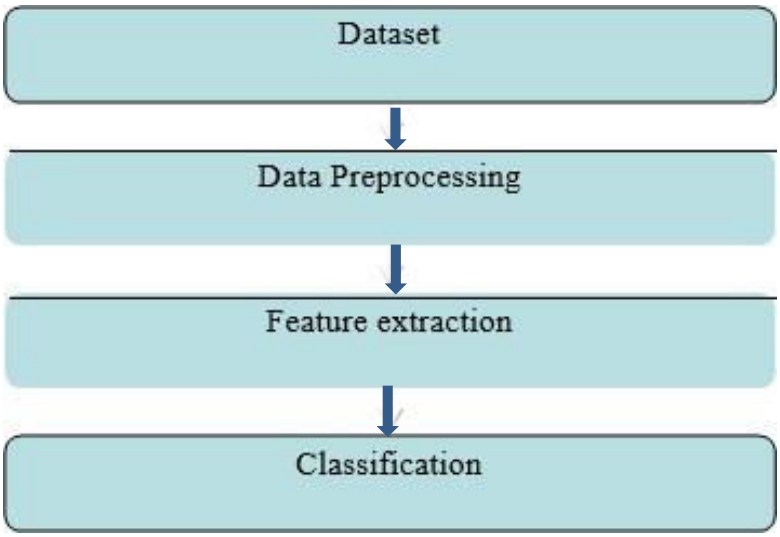

Fig. 3.2.1. Methodology diagram

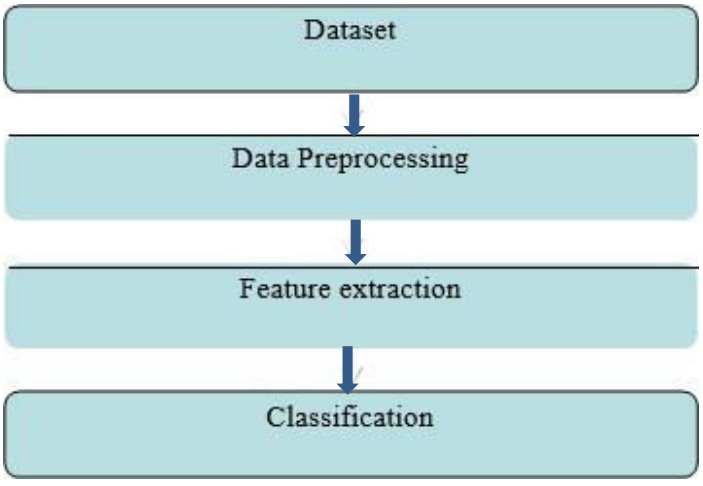

Fig. 3.2.2. Data Pre-processing

3.2.1Detailed description of the methodology

Data Extraction

Extraction of dataset from Kaggle data source (Twitter dataset). The Twitter Dataset is combined from two datasets containing hate speech:
Hate Speech Twitter Dataset by Waseem, Zeerak and Hovy, Dirk [11] which contains 17000 tweets labelled for sexism or racism. The tweets are mined using the annotations .5900 tweets are lost due to accounts being deactivated or tweet deleted.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

Hate Speech Language Dataset by Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar [12]. It contained 25000 tweets obtained by crowdsourcing.

Data Pre-processing

- **Tokenization:** In tokenization we split raw text into meaningful words or tokens. For example, the text "we will do it" after tokenization becomes, "we", "will", "do". "it".
- **Stemming:** Stemming is the process of converting a word into a root word or stem. Eg: for three words 'eating', 'eats', 'eaten', the stem is 'eat'. Since all three branch words of root 'eat' represent the same thing it should be recognized as similar. NLTK offers 4 types of stemmers: Porter Stemmer, Lancaster Stemmer, Snowball Stemmer and Regex p Stemmer. The following project uses Porter Stemmer.
- **Stop word removal:** Stop words are words that do not add any meaning to a sentence Eg: some stop words for English language are: what, is, at, a etc. These words are irrelevant and can be removed.
  NLTK contains a list of English stop words which can be used to filter out all the tweets.

Feature Extraction

Bag of Words (BoW): The BoW that is bag of words model is a simple method of extracting features from documents that uses occurrence of words within a document. Bag of Words model has two important parts:

- A vocabulary of words(tokens) derived from all documents.
- A way of measuring all these words as features in each document.

It is referred to as 'bag' because the model only concerns with the word rather than its order of occurrence in the document. The intuition for this method is that similar documents have similar words in them.

IV. EXPERIMENTAL RESULTS

4.1 OUTPUT SNAPSHOTS

Fig. 4.1.1. Display page

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

Fig. 4.1.2. Admin login page

Fig. 4.1.3. Admin page

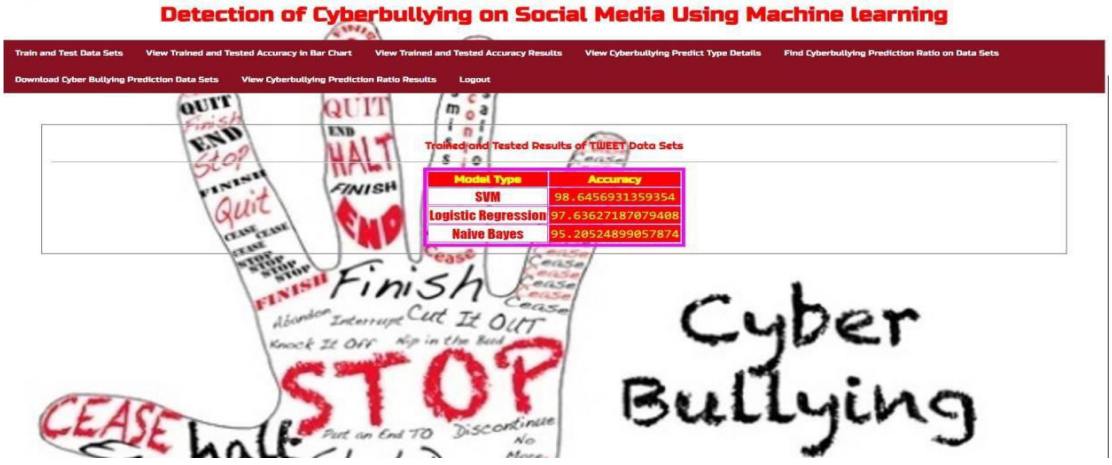

Fig. 4.1.4. Training results

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

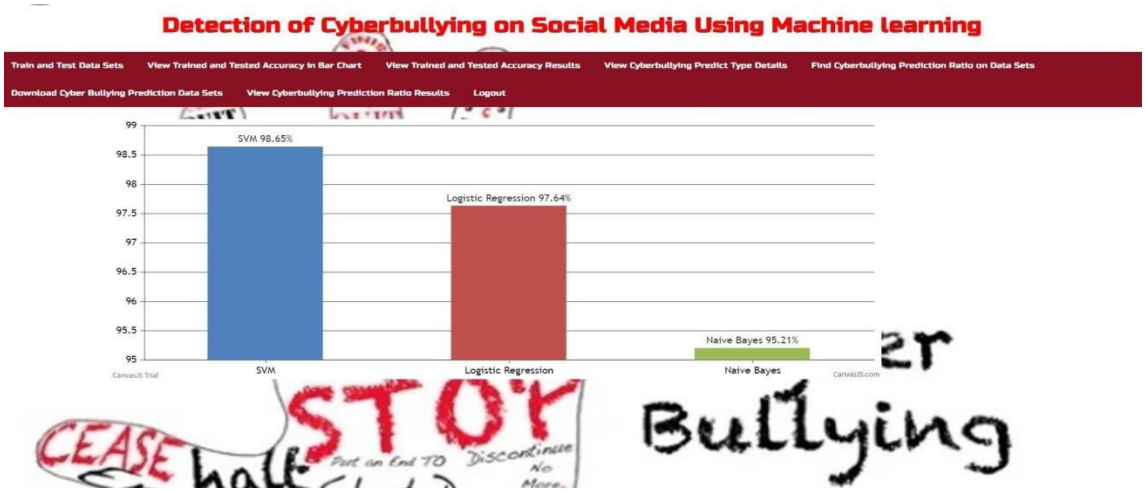

Fig. 4.1.5. Bar graph showing the accuracy of three models

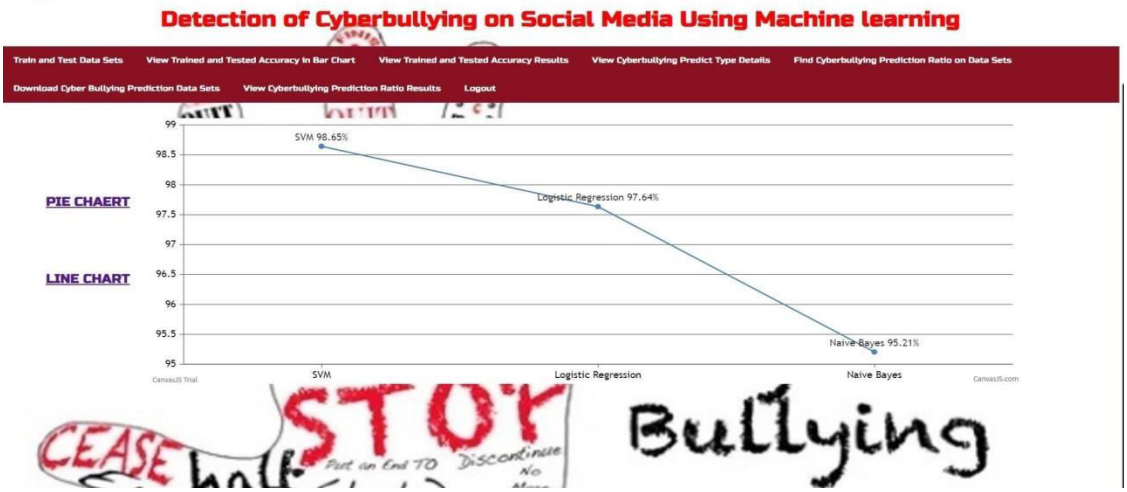

Fig. 4.1.6. Line graph showing accuracy of three mode

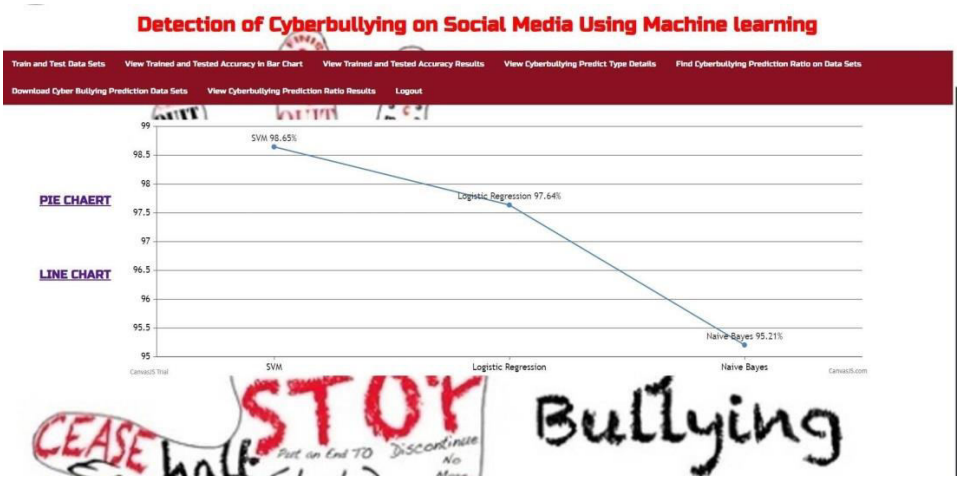

Fig. 6.1.6. Line graph showing accuracy of three models

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

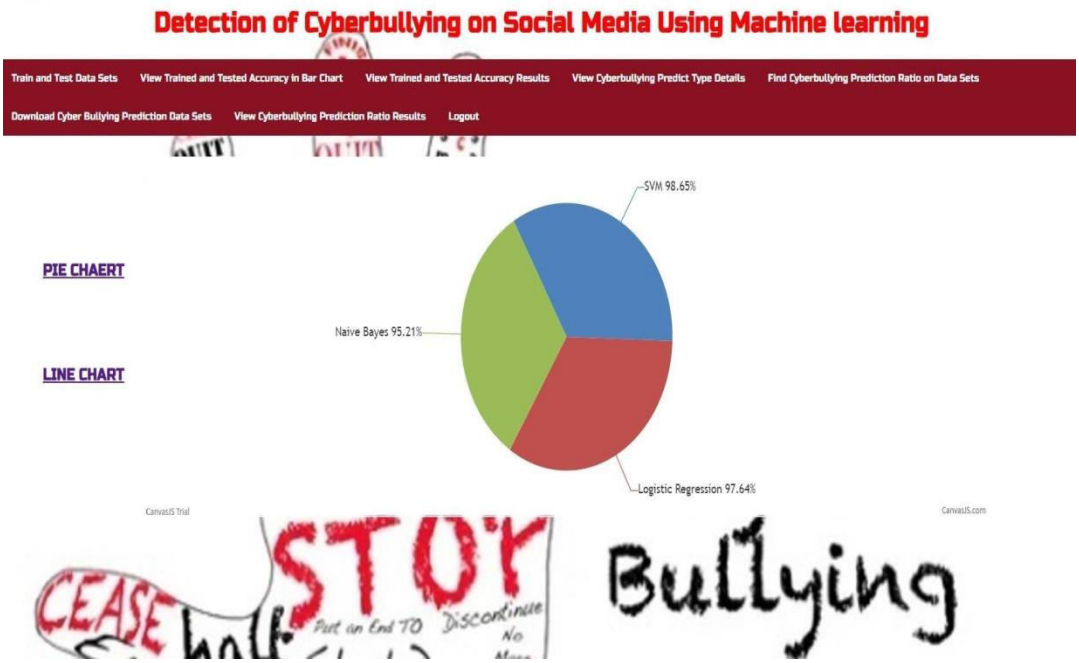

Fig. 6.1.7. Pie chart showing the accuracy of the three models

Tweet Message	Cyber Bullying Prediction Type
studiolife aislife requires passion dedication willpower to find newmaterials	Non Offensive or Non Cyberbullying
studiolife aislife requires passion dedication willpower to find newmaterials	Non Offensive or Non Cyberbullying
studiolife aislife requires passion dedication willpower to find newmaterials	Non Offensive or Non Cyberbullying
hey guys tommorow is the last day of my exams i m so happy yay	Non Offensive or Non Cyberbullying
thought factory bbc neutrality on right wing fascism politics media blm brexit trump leadership gt 3	Offensive or Cyberbullying
chick gets fucked hottest naked lady	Offensive or Cyberbullying
chick gets fucked hottest naked lady	Offensive or Cyberbullying
finally at peace sad news so many lives lost hatred no answer haiku haikuchallenge micropoetry poetry finally	Non Offensive or Non Cyberbullying
everybody hates the white crayon	Offensive or Cyberbullying
emma stone on hollywood they ve given my jokes away to male co stars via	Offensive or Cyberbullying
some people are just too committed to their own disfunction truth tired	Non Offensive or Non

Fig. 6.1.8. Test dataset showing prediction results

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

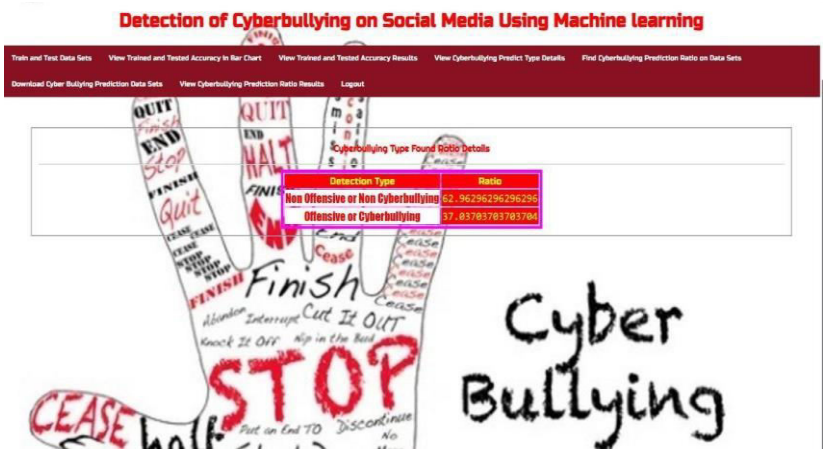

Fig. 6.1.9. Test data prediction ratio

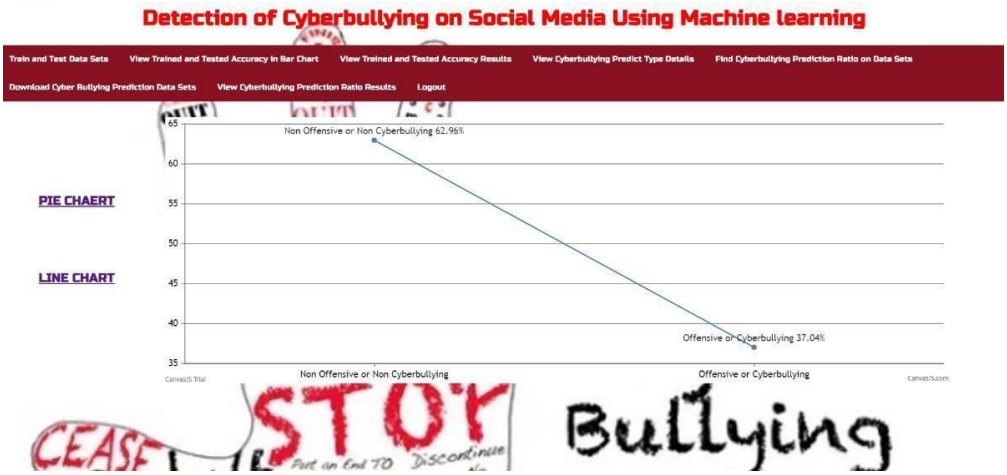

Fig. 6.1.10. line graph of the test data prediction

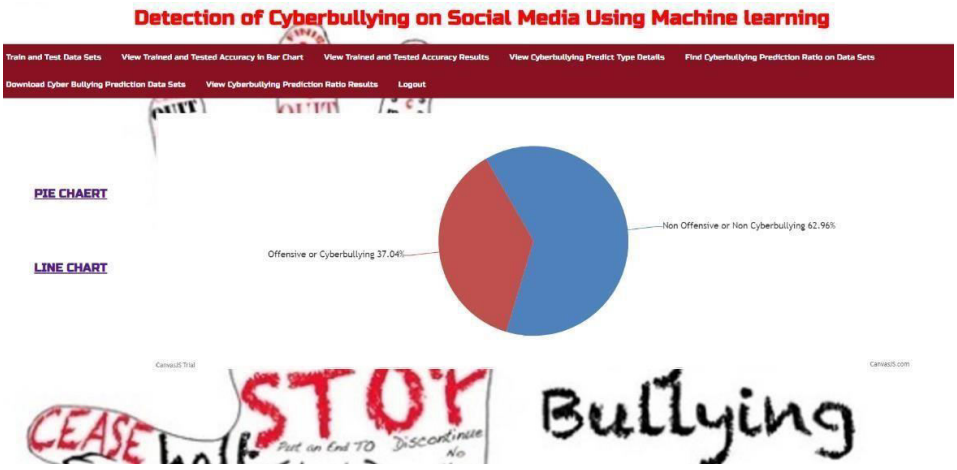

Fig. 6.1.11. pie chart showing test data prediction percentage

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

1) Fig. 6.1.12. User's prediction page

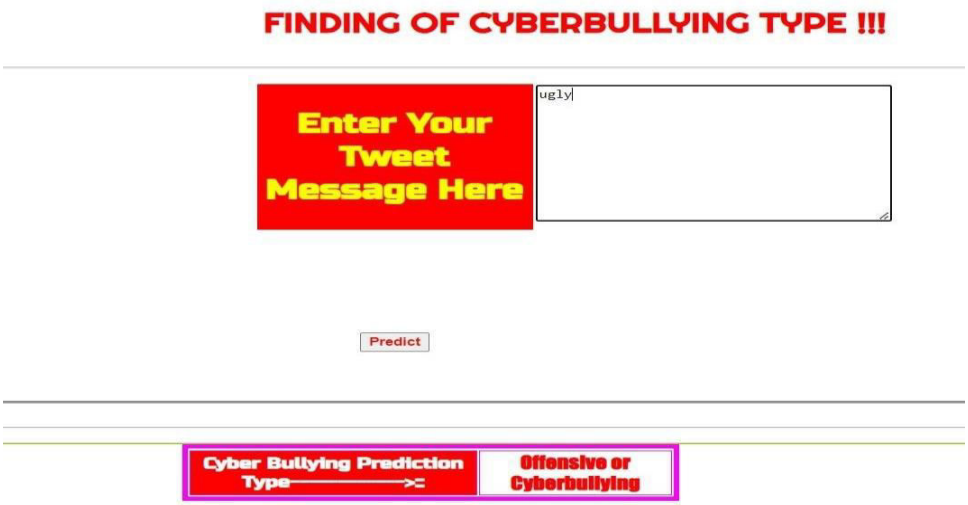

Fig. 6.1.13. Prediction result from user's end

V. CONCLUSION

In particular, cyberbullying has become more common and has begun to raise significant social issues with the rising prevalence of social media sites and increased social media use by teenagers. There needs to design automatic cyberbullying detection method to avoid bad consequences of cyber harassment. Considering the significance of cyberbullying detection, in this work, we studied the automated identification of posts on social media related to cyberbullying. In future we are planning to design a framework for automatic detection and classification of cyberbullying from texts using deep learning algorithms.

REFERENCES

[1] N. Selwyn, "Social media in higher education," *The Europa world of learning*, vol. 1, no. 3, pp. 1– 10, 2012.

[2] H. Karjaluoto, P. Ulkuniemi, H. Kein¨anen, and O. Kuivalainen, "Antecedents of social media b2b use in industrial marketing context: customers' view," *Journal of Business & Industrial Marketing*, 2015.

[3] W. Akram and R. Kumar, "A study on positive and negative effects of social media on society," *International Journal of Computer Sciences and Engineering*, vol. 5, no. 10, pp. 351– 354, 2017.

[4] D. Tapscott *et al.*, *The digital economy*. McGraw-Hill Education, 2015.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107127 |

[5] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. Desmet, and I. De Bourdeaudhuij, "Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully," *Computers in Human Behavior*, vol. 31, pp. 259–271, 2014.

[6] D. L. Hoff and S. N. Mitchell, "Cyberbullying: Causes, effects, and remedies," *Journal of Educational Administration*, 2009.

[7] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of suicide research*, vol. 14, no. 3, pp. 206–221, 2010.

[8] 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.K. Dinakar, R. Reichart, and H.

[9] Lieberman, "Modeling the detection of textual cyberbullying," in *In Proceedings of the Social Mobile Web*. Citeseer, 2011.

[10] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine learning and applications and workshops*, vol. 2. IEEE, 2011, pp. 241–244.

[11] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," *Computers & Security*, vol. 90, p. 101710, 2020.

[12] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European Conference on Information Retrieval*. Springer, 2018, pp. 141–153.

[13] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 759–760.M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection,"
International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, 2018.

[14] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[15] T. A. Buan and R. Ramachandra, "Automated cyberbullying detection in social media using an svm activated stacked convolution lstm network," in *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, 2020, pp. 170–174.

[16] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 479–486.

[17] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[18] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2090–2099.

[19] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2018, pp. 1–7.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details