International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.111 ISSN: 2581-4621

Detection of Cyberbullying on Social Media Using Machine learning

Shaik Mohammed Akbar Basha[1], K.Kumara Swamy[2]
[1]*MCA Student, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*
[2]*Assistant Professor, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

Abstract
This paper presents an innovative approach for Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has lead to mishappenings like suicide and depression. Regulation of content on Social media platorms has become a growing need. The following study uses data from two different forms of cyberbullying, hate speech tweets from Twittter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning.
Three methods for Feature extraction and four classifiers are studied to outline the best approach. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%. Cyberbullying detection is solved in this project as a binary classification problem where we are detecting two majors form of Cyberbullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyberbullying or not.
Keywords: Cyberbullying, Twitter, Social Media, NLP

Introduction

Modern young people (“digital natives”) have grown in an era dominated by new technologies where communications are pushed to quite a real-time level, and pose no limits in establishing relationships with other people or communities. The fast growing use of social networking sites among the teens have made them vulnerable to get exposed to bullying. Comments containing abusive words effect psychology of teens and demoralizes them. In this work we have devised methods to detect cyberbullying using supervised learning techniques. Cyberbullying is the use of technology as a medium to bully someone. Although it has been an issue for many years, the recognition of its impact on young people has recently increased. Through machine learning, we can detect language patterns used by bullies and their victims, and develop rules to automatically detect cyberbullying content. Moreover, they are a place where people engage in social interaction, offering the existing friendships. On the negative side however, social establish new relationships and maintain existing friendships. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming or sexually transgressive behaviour, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard. The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. It should, however, be noted that cyberbullying detection is intrinsically more difficult than just detecting abusive content. Additional context may be required to prove that an individual abusive message is part of a sequence of online harassment directed at a user for such a message to be labelled as cyberbullying. The growth of cyberbullying activities is increasing as equally as the growth of social networks. Cyberbullying activities poses a significant threat to mental and physical health of the victims. Project about detection of bullying is present but implementation for monitoring social network to detect cyberbullying activities is less. Hence, the proposed system focuses on detecting the presence of cyberbullying activity in social networks using natural processing language.

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.111 ISSN: 2581-4621

Literature Survey

1. M. Di Capua, et al. [1] raises the unsupervised how to develop an online bullying model based on a combination of features, based on traditional textual elements and other "social features". Features were divided into 4 categories as Syntactic features, Semantic features, Sentiment features, and Community features. The author has used the Growing Hierarchical Self Map editing network (GHSOM), with 50 x grid 50 neurons and 20 elements as the insertion layer. M. Di Capua, and others used an integration algorithm k-means to separate the input database and GHSOM in the Formspring database. The effects of this hybrid the unsupervised way surpasses the previous one results. The author then checked the youtube database at 3 p.m. Different Machine Learning Models: Naive Bayes Classifier, Decision Tree Classifier (C4.5), and Support Vector Machine (SVM) with Linear Kernel. It was saw that the combined effects of hate speech turned around so that we have lower accuracy in the youtube database compared to FormSpring tests, as in the text analysis and syntactical features work differently in on both sides. When this hybrid method is used in Twitter Database, led to weak memory and F1 Score. The model proposed by the authors can also be improved used in building constructive mitigation applications cyberbullying problems.

2. J. Yadav, et al. [2] suggests a new approach to this the discovery of internet cyberbullying on social media in using a BERT model with a single line neural and was tested on the Formspring forum and Wikipedia Database. The proposed model provided performance 98% accuracy of spring Form databases and 96% accuracy in a relatively comprehensive Wikipedia database previously used models. The proposed model provided better Wikipedia database results due to its size g without the need for excessive sampling while I The spring data form requires multiple samples.

3. R. R. Dalvi, et al. [3] suggests a way to do this detect and prevent online exploitation on Twitter using Classified supervised machine learning algorithms. In this study, the live Twitter API is used for compilation tweets and data sets. The proposed model tests both Support Vector Machine and Naive Bayes on the data sets are collected. To remove a feature, use it TFIDF vectorizer. The results show that it is accurate of an online bullying model based on Vector Support The machine is about 71.25% better than Naive Bayes was almost 52.75%.

4. Trana R.E., et al. [4] The goal was to design a machine learning model to minimize special events including text extracted from image memes. Author include a site that contains approximately 19,000 text views have been published on YouTube. This study discusses the operation of three learning machines equipment, Uninformed Bayes, Support Vector The machine, as well as the convolutional neural network used in on the YouTube website, and compare the results with existing details of the Form. They do not write continuously investigate cyber bullying algorithms sub-sections within the YouTube website. Naive Bayes beat SVM and CNN in the next four categories: race, nationality, politics, and general. SVM passed well with the inexperienced Naïve Bayes again CNN is in the same gender group, with all three algorithms show equal performance with the middle body group accuracy. The results of this study provided inaccurate data used to distinguish between incidents of abuse and non-violence. Future work can focus on the construction of a two-part separation system used to test text taken from photos to see that the YouTube website provides a better context for aggressionrelated collections. 5. N. Tsapatsoulis, et al. [5] detailed review of introduced cyberbullying on Twitter. The importance of identifying the various abusers on Twitter is provided. Kwe paper, various practical steps required the development of an effective and efficient application for Internet traffic detection is well defined. I styles involved in data classification and recording platforms, machine learning models and feature types, and model studies using such tools explained. This paper will serve as the first step in the process project on acquisition of cyberbullying technology using the machine reading.

5. G. A. León-Paredes et al. [6] they described I the development of an online bullying detection model is used Native Language Processing (NLP) and Mechanics Reading (ML). Spanish Cyberbullying Prevention The system (SPC) was developed by installing the machine learning strategies Naïve Bayes, Support Vector Machine, and Logistic Regression. Database used this study was posted on Twitter. Plurality 93% accuracy was achieved with the help of third parties

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.111 ISSN: 2581-4621

techniques used. Charges of online exploitation were found with the help of this system presented the accuracy of 80% to 91% on average. Stemming and lemmatization techniques in NLP can be used continuously increasing system accuracy. Such a model can and used for adoption in English and local languages if possible.

6. P. K. Roy, et al. [7] details about creating a request for hate speech on Twitter via the help of a deep neural convolutional network. Machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN) have it used to identify tweets related to hate speech in Twitter and features have been removed using tf-idf process. The leading ML model was SVM but it managed predict 53% hate speech tweets on a 3: 1 database to be tested train. The reason behind the low forecast scale was unequal data. The model is based on the prediction of hate speech tweets. Advanced learning based methods on the Convolutional Neural Network (CNN), Long-Term Memory (LSTM), and its Content LSTM combinations (CLSTM) have the same effects as separate distributed database. 10 times cross confirmation was used in conjunction with the proposed DCNN model once you got a very good rate of memory. It was 0.88 hate speech and 0.99 non-hate speech. Test results confirmed that the k-fold opposite verification process is a better resolution with unequal data. In the future, the the current database can be expanded for better performance accuracy.

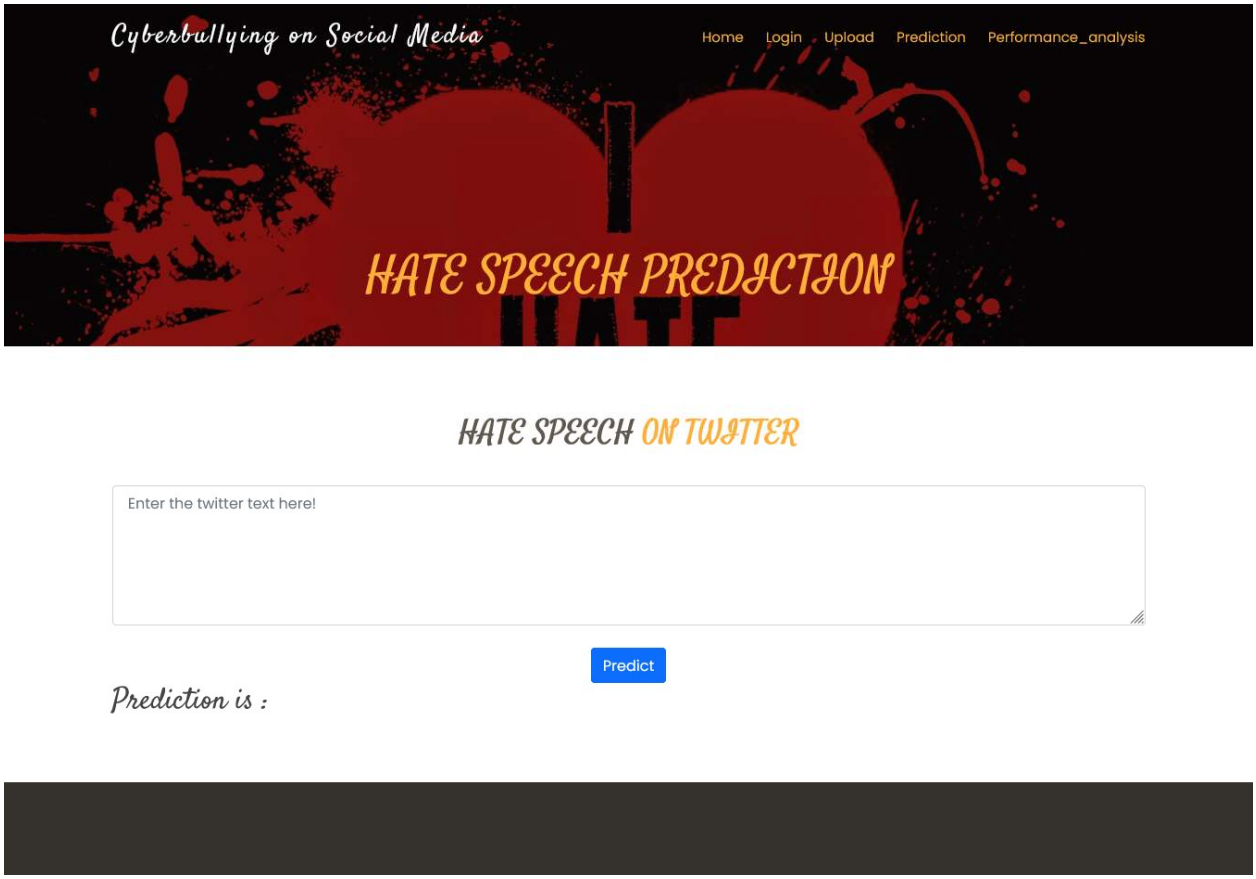
Results

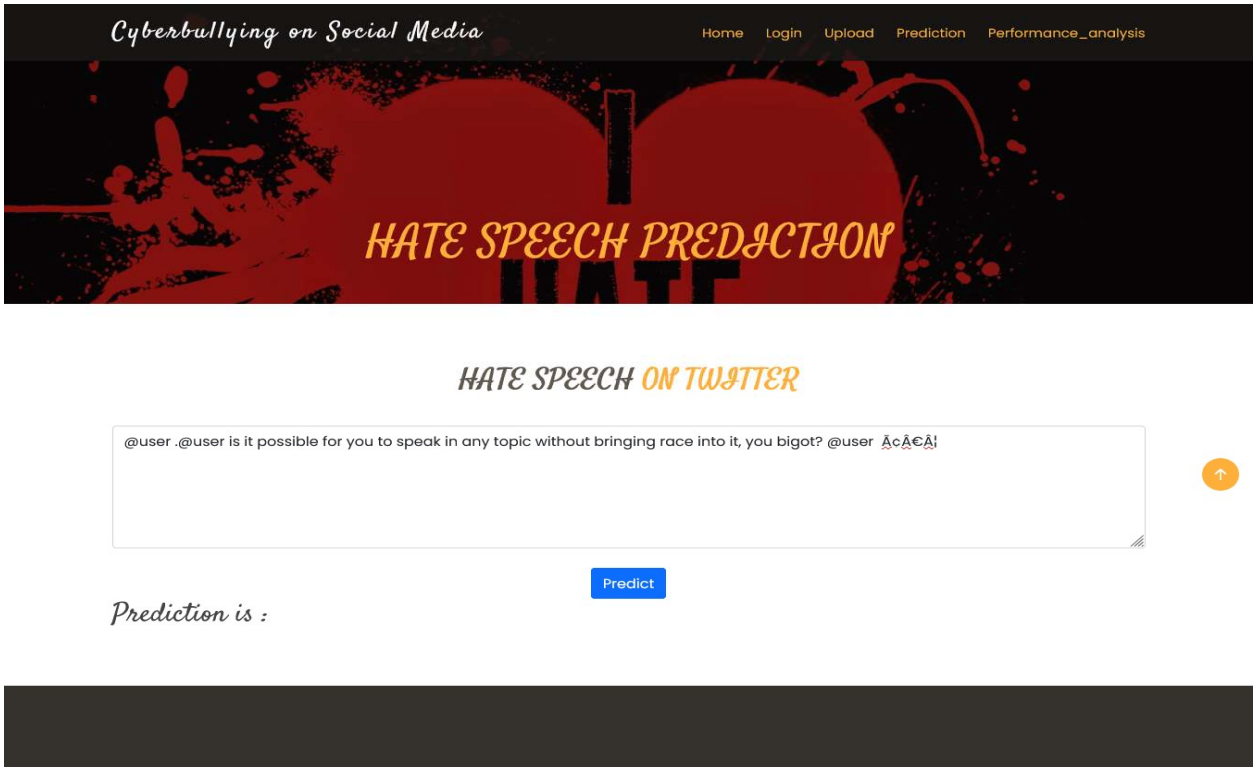
In this screenshots we have shown the detection plane shows an example of bullying content that is detected as bullying in the result," This is a type of bullying content", and if the text is not the content of cyberbullying there will be a output is "This is not a type of bullying content". As the framework is done in the Django show that " NAME ,USERNAME ,PASSWORD " for register option and if the user have the account then they can login with already created account with " LOGIN" option. Cyberbullying research has often focused on detecting cyberbullying 'attacks' and hence overlook other or more implicit forms of cyberbullying and posts written by victims and bystanders. However, these posts could just as well indicate that cyberbullying is going on.The main contribution of this paper is that it presents a system to automatically detect signals of cyberbullying on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders. We evaluated our system on a manually annotated cyberbullying corpus for English and Dutch and hereby demonstrated that our approach can easily be applied to different languages, provided that annotated data for these languages are available.

preview

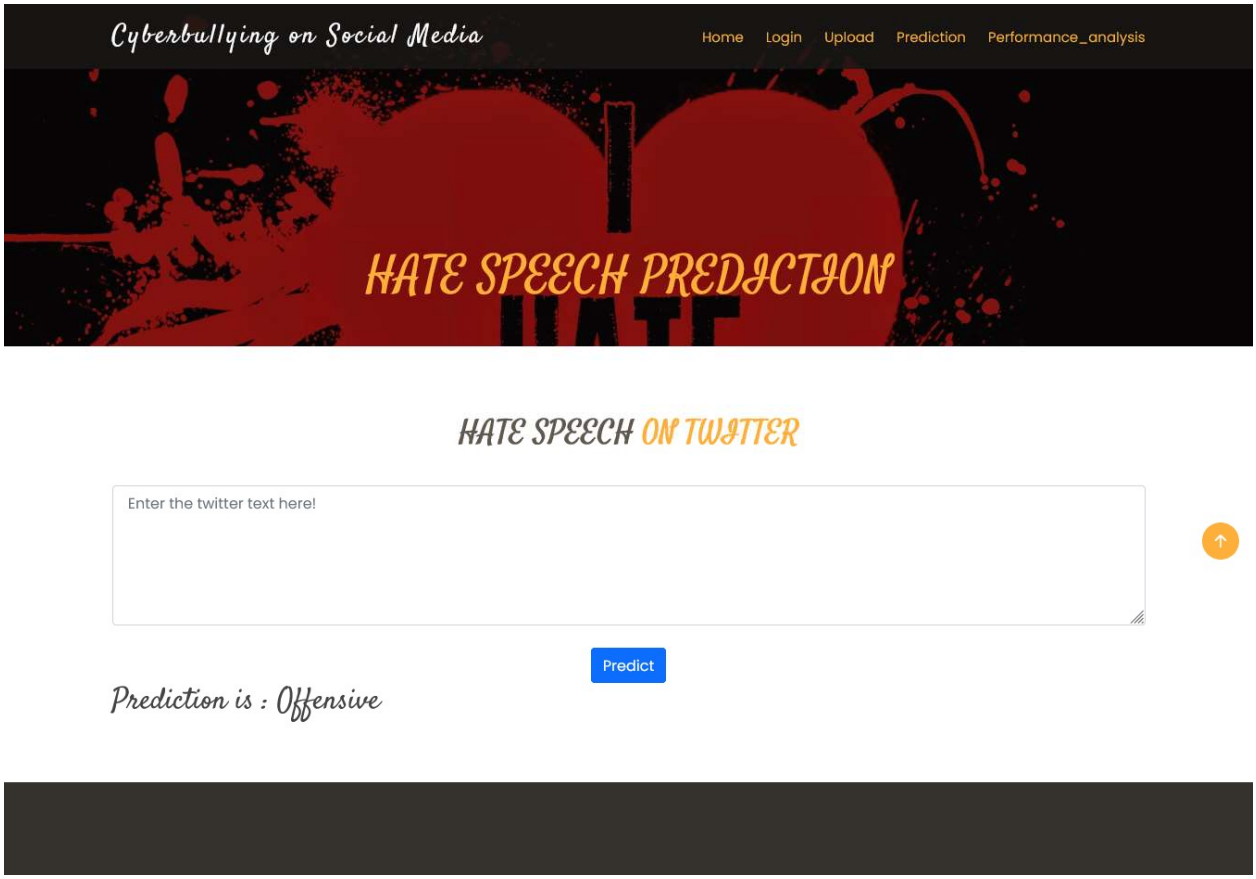
	label	
Id		
1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
3	0	bihday your majesty
4	0	#model i love u take with u all the time in urÃ°ÂÂÂ±!!! Ã°ÂÂÂÃ°ÂÂÂÃ°ÂÂÂÃ°ÂÂÂÃ°ÂÂÂ¦Ã°ÂÂÂ¦Ã°ÂÂÂ¦
5	0	factsguide: society now #motivation

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.111 ISSN: 2581-4621





International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.111 ISSN: 2581-4621



Conclusion

Cyber bullying across internet is dangerous and leads to mishappenings like suicides, depression etc and therefore there is a need to control its spread. Therefore cyber bullying detection is vital on social media platforms. With avaibility of more data and better classified user information for various other forms of cyber attacks Cyberbullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation. We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia. For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 percent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable. Due to this it gives better results with BoW and Tf-Idf models rather than Word2Vec models However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly.Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptrons.

References

[1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, “Towards the detection of cyberbullying based on social network mining techniques,” in Proceedings of 4th International Conference on Behavioral, Economic, and Socio Cultural Computing, BESC 2017, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.

DOI:10.46647/ijetms.2025.v09i02.111 ISSN: 2581-4621

[2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.
[3] A. Mangaonkar, A. Hayrapetian, and R. Raje, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.
[4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.
[5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.
[6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.
[7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.
[8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," arXiv. 2018.
[9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," arXiv. 2018.
[10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.
[11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.
[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.
[13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.
[14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.
[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.