Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, No. 2, May : 2023

CYBER-SAFE COMMENTING

Mrs. Sureka. K, Assistant Professor, Dept. Of Computer Science and Engineering, SSM Institute of Engineering and Technology. sindureka@gmail.com
Akash. V. S, **Jebarson. S**, **Aravindhan. G**, Student, Dept. Of Computer Science and Engineering, SSM Institute of Engineering and Technology

Abstract

Cyberbullying has become an increasing concern on social media platforms, where individuals can easily hide behind anonymity and use digital tools to harass, intimidate, or humiliate others. To address this issue, researchers have developed automated methods for detecting cyberbullying on social networks. This paper proposes a machine learning approach that uses random forest classifier and gradient boosting algorithm to detect cyberbullying words present in comments. The model is trained and tested to choose the one with higher accuracy, which is then integrated into the comment section of the website. Upon posting, the model automatically scans comments for cyberbullying words. If detected, a warning message prompts the user to edit or remove the offending language. If the user tries to repost without removing cyberbullying words, the comment is automatically deleted. However, comments without cyberbullying words are posted normally. Real-time detection and prevention of cyberbullying promotes a safer online environment and respectful interactions.

Keywords: Cyberbullying, Machine learning, Comment

I. Introduction

Cyberbullying has emerging as a developing difficulty in recent years, particularly on social media platforms in which individuals can effortlessly disguise in the back of anonymity and use virtual gear to annoy, intimidate, or humiliate others. The outcomes of cyberbullying can be devastating, leading to despair, anxiety, and even suicide. As a result, there was an increasing want to broaden effective strategies for detecting and stopping cyberbullying. In reaction to this want, researchers were working to expand automatic methods for detecting cyberbullying on social networks. One promising technique is to use gadget studying strategies to analyse the language utilized in messages and discover capabilities that distinguish among cyberbullying and non-cyberbullying content material. via education device mastering models on massive datasets of cyberbullying and non-cyberbullying content material, these models can learn how to apprehend patterns in language that are associated with cyberbullying behaviour. The proposed system studying approach has the ability to significantly enhance the potential of social media structures to hit upon and save you cyberbullying. by using automating the detection process, social media structures can fast perceive and take away cyberbullying content before it causes damage. moreover, by decreasing the prevalence of cyberbullying on social networks, the technique can assist promote a safer and healthier online surroundings. ordinary, using gadget gaining knowledge of for detecting cyberbullying is an important step closer to growing a extra advantageous and supportive on-line network. As era keeps to enhance, it is probable that we can see even more innovative approaches to combatting cyberbullying and promoting wholesome online interactions.

II. LITERATURE SURVEY

1. Cyberbullying Detection on Social Networks using LSTM Model

Authors: ArshaDass M A, Deepa K Daniel

This research paper proposed an efficient algorithm to identify the bullying test and aggressive comments and analyses these comments to check the validity. NLP and Machine learning is used for analyzing the social comment and identified the aggressive effect of an individual or a group. An effective classifier acts as the core component in a final prototype system that can detect cyberbullying on social media.

2. Detecting A Twitter Cyberbullying Using Machine Learning
Authors: Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe

A machine learning model is proposed to detect and prevent bullying on Twitter. Two classifiers i.e., SVM and Naïve Bayes are used for training and testing the social media bullying content. Both Naive Bayes and SVM (Support Vector Machine) were able to detect the true positives with 71.25% and 52.70% accuracy respectively. But SVM outperforms Naive Bayes of similar work on the same dataset. Also, Twitter API is used to fetch tweets are passed to the model to detect whether the tweets are bullying or not.

3. Cyberbullying Detection on Social Networks using LSTM Model
Authors: ArshaDass M A, Deepa K Daniel

In the era of social media and networking, the usage of bad words and aggressive words has been increased significantly. Cyberbullying affects more than half of the young population using social media. Insults in social media websites create negative interactions within the network. These remarks build up a culture of disrespect in cyberspace. Algorithms and tools used to understand and mitigate it are mostly inactive. Also, current implementations on insult detection using machine learning and natural language processing have very low recall rates. In short, the paper involves determining ways to identify bullying in text by analyzing and experimenting with different methods to find the feasible way of classifying such comments. We proposed a efficient algorithm to identify the bullying test and aggressive comments and analyses these comments to check the validity. NLP and Machine learning is used for analyzing the social comment and identified the aggressive effect of an individual or a group. An effective classifier acts as the core component in a final prototype system that can detect cyberbullying on social media.

III. PROPOSED SYSTEM

Training and testing random forest classifier and gradient boosting algorithm to detect cyberbullying words present in comments. Choose the model with higher accuracy to integrate it into the comment section of the website. Model automatically scans comments for cyberbullying words upon posting attempt. If detected, warning message prompts user to edit or remove offending language. If user tries to repost without removing cyberbullying words, comment is automatically deleted. Comments without cyberbullying words are posted normally. Real-time detection and prevention of cyberbullying promotes safer online environment and respectful interactions.

IV. METHODOLOGY

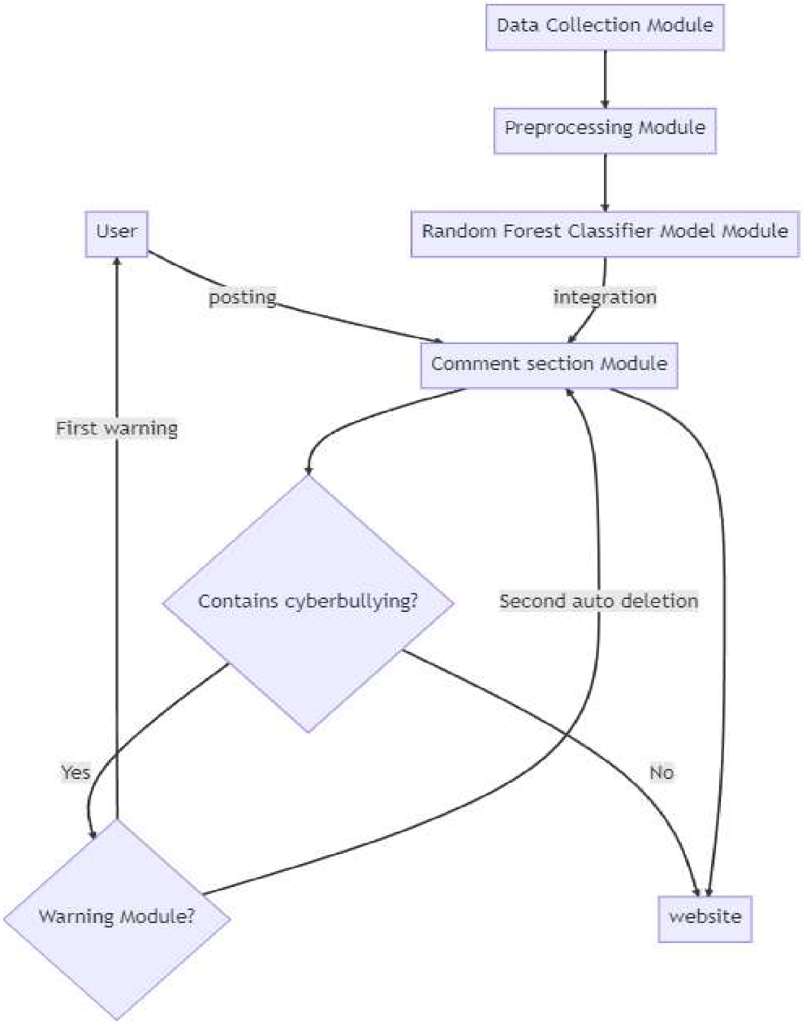
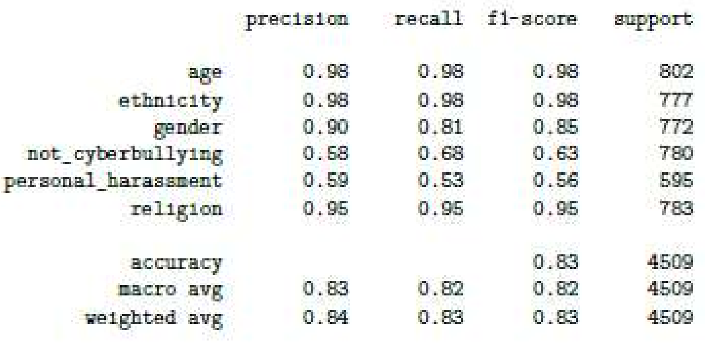


Figure 1: Flow diagram

Data Collection – The Kaggle dataset with the name "cyberbullying_tweets" was used.

Model Creation – A Random Forest Classifier model and a Gradient Boosting Model were developed.

Model evaluation – The accuracy and performance of the two models are compared, and the model that performs better is chosen to be integrated into the website's comment section.

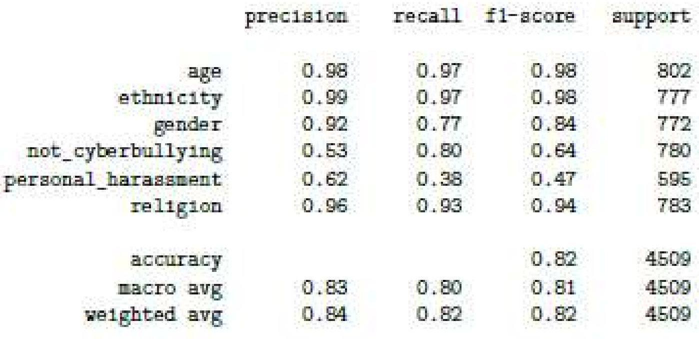


```
                     precision    recall  f1-score   support

                age       0.98      0.98      0.98       802
          ethnicity       0.98      0.98      0.98       777
             gender       0.90      0.81      0.85       772
  not_cyberbullying       0.58      0.68      0.63       780
personal_harassment       0.59      0.53      0.56       595
           religion       0.95      0.95      0.95       783

           accuracy                           0.83      4509
          macro avg       0.83      0.82      0.82      4509
       weighted avg       0.84      0.83      0.83      4509
```

Figure 2: Random Forest Classifier



```
                     precision    recall  f1-score   support

                age       0.98      0.97      0.98       802
          ethnicity       0.99      0.97      0.98       777
             gender       0.92      0.77      0.84       772
  not_cyberbullying       0.53      0.80      0.64       780
personal_harassment       0.62      0.38      0.47       595
           religion       0.96      0.93      0.94       783

           accuracy                           0.82      4509
          macro avg       0.83      0.80      0.81      4509
       weighted avg       0.84      0.82      0.82      4509
```

Figure 3: Gradient Boosting

Model Integration – The Random Forest Classifier is successfully integrated into a website based on the test results.
Warning and Deletion – Then, the method of alerting the user when they post an offensive remark for the first time and automatically deleting offensive comments for the second time is coded in the website's backend.

V. CONCLUSION

The proposed solution using an RFC model to detect and prevent cyberbullying in online interactions is a promising approach to creating a safer online environment. By automatically scanning comments for cyberbullying words, we can detect and prevent instances of cyberbullying in real-time. This not only promotes respectful and positive interactions but also encourages individuals to think carefully about the impact of their words and to promote a culture of mutual respect and kindness.

References

[1] Agarwal A, Xie B, Vovsha I, Rambow O and Passonneau R (2011). “Sentiment analysis of Twitter data”. In Proceedings of the Workshop on Languages in Social Media LSM ’11.

[2] Chen. Y, Zhou Y, Zhu S, and Xu H (2012). “Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”. In PASSAT and SocialCom.

[3] Hosseinmardi H, Mattson S. A, Rafiq R. I, Han R, Lv Q and S. Mishra. (2015) “Analyzing Labelled Cyberbullying Incidents on the Instagram Social Network”. In SocInfo.

[4] Hine G. E, Onaolapo J, De Cristofaro E, Kourtellis N, Leontiadis I, Samaras R, Stringhini G, and Blackburn J (2017). “A Measurement Study of 4Chan’s Politically Incorrect Forum and its effort on the web”. In ICWSM.

[5] John Hani Mounir, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, Ammar Mohammed (2019.), “Social Media Cyberbullying Detection using Machine Learning”, (IJACSA) International Journal of Advanced Computer Science and Applications Vol. 10, pages 703-707.

[6] Kelly Reynolds, April Kontostathis, Lynne Edwards (2011), "Using Machine Learning to Detect Cyberbullying”, 10t h Int ernational Conference on Machine Learning and Applications volume 2, pages 241–244. IEEE.

[7] Patchin J, Hinduja, S (2006). “Bullies move beyond the schoolyard; a preliminary look at cyberbullying.”. Youth violence and juvenile justice. 4:2. 148-169.

[8] Saif H, Fernandez M, He Y, and Alani H (2014), “On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter,” Proc. 9th Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, pp.80-817.

[9] Wang, Xin (2015). “Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory”. In ACL.

[10] Yin. D, Davison B. D, Xue Z, Hong L, Kontostathis A, and Edwards L. (2009), “Detection of Harassment on Web 2.0”. In Proceedings of CAW2.0 Workshop.