UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Real-Time Voice-Based Cyberbullying Detection Using AI in Social Media and Gaming

Acha Roopa[1], N. Navaneetha[2], Dr. G. Vishnu Murthy[3]

[1]M. Tech Student, Department of Computer Science and Engineering, School of Engineering, Anurag University, India
[2]Assistant Professor, Department of Computer Science and Engineering, School of Engineering, Anurag University, India
[3]Professor and Dean, Department of Computer Science and Engineering, School of Engineering, Anurag University, India

Email: *acharoopa01@gmail.com*[1], navaneetha.cse@anurag.edu.in[2], deancse@anurag.edu.in[3]

***Abstract*— The research proposes developing a real-time AI system that detects cyberbullying when it happens over voice chats in online games and on various social media networks. The proposed method here uses the Speech Recognition which removes the noise and disturbance in the audio to make speech conversion into text form more accurate. Then TF-IDF system is used to detect major terms and give extra attention to the most important information from the text. The model was examined to make sure it runs smoothly and fast when it is actually being used in real time. This approach is more accurate, faster and reliable in a different factor. After performing many tests, the LinearSVC has exhibited high detection accuracy, F1 scores, and less false positives. The main advantage of LinearSVC is that it can handle both small and large learning datasets effectively. As a result, the algorithm is more accurate, faster and performs stably than others, so it is helpful for real applications because it provides strong, fast and scalable safety.**

***Keywords*—**Cyberbullying Detection, TF-IDF, LinearSVC, Speech-to-Text, Text Classification, Real-Time Prediction, Social Media Safety, Online Gaming Abuse.

I. Introduction

Detection techniques which are used for the detection in text-based bullying cannot be used to detect audio based because of differences in their approach such as variations in the text and speech patterns, background noise, and accents. These issues can be handled by using the Automatic speech recognition systems like whisper ai, which have models to convert the audio to text format with high accuracy [4][5]. However, the conversion of the audio to text requires more accuracy for the text extraction from the audio which needs the effective feature extraction methods. This shows the use of Term Frequency–Inverse Document Frequency (TF-IDF) to differentiate various terms found in speech models noisy data [22][6]. Since the data may be big and complex, the SVC, in particular the LinearSVC, is the main choice because this allows us to classify data quickly and accurately and suitable for real time deployment [7][8]. To ensure that the model detects voice abuses properly and within the specified delay, tuning and evaluation should be based on precision, recall and F1 score [9]. Current AI is not effective in places like social media and games since it is mostly designed to handle responses later.

If implemented, this way of tackling cyberbullying uses ASR, the TF-IDF method and LinearSVC to monitor recorded voice communications and quickly alert the user of any cyberbullying incidents [10]. Experiments carried out on datasets made available to the public reveal that my classifier performs better and significantly reduces both false positives and the amount of work it requires to run [11].

Still, managing several languages, uncommon slang phrases, texts in more than one language and many accents makes it harder for algorithms to function properly [12]. Later on, I will add techniques that are designed to work well with what each user needs and then integrate the needs of new users. Analyzing a wide range of languages and understanding users' behaviour could improve our ability to detect cyberbullying in modern voice communication [13].

To confirm the quality of the model, the algorithms used in the proposed system are refined using every available setting and default hyperparameter from the Support Vector Classifier. Its objective is to achieve accurate forecasting and remain practical by giving the best results with few resources. We determine how much time and energy will be used by the system while handling the types of operations seen on social media and gaming websites [14].

It is measured against prominent, off-the-shelf machine learning models on voice-to-text cyberbullying detection and this leads to better precision, recall and F1-score measurements [15]. In addition, state-of-the-art methods are used to minimize unwanted features and find the key features linked to abusive language in the spoken speech files [15]. As a result, there is a boost in the model' Comparable performance in different noise situations while also helping distinguish pure classes from those contaminated [16].

Validating a speech recognizer means testing it in real-life environments, with speakers speaking differently, with extra noise around and as slang expressions change [14][16]. For the model to protect against false positives and maintain accuracy in situations involving people talking in different ways, it needs

to be flexible in its algorithms.

Results from comparing the current system against other text-based cyberbullying frameworks have found that the proposed system can detect cyberbullying more efficiently [16].

Future efforts will concentrate on giving the system the ability to work with various languages and to support adaptive learning [16]. Moreover, considering how people use the service will help improve detection accuracy, allow for prompt action and help understand the behavior involved in voice-based cyberbullying. The plan is to create a moderate-related system that can be adapted to many services and help increase safety in voice-centric social media and games [14][15][16].

II. Literature Survey

This field involves using artificial intelligence, speech recognition and natural language processing to immediately detect cases of cyberbullying via voice chat. While traditional cyberbullying detection uses texts, now that chatting via voice is getting more popular on platforms, voice-based systems are being adopted [1].

The basics of voice-based detection include the ability to turn speech into text. Lately, OpenAI's Whisper and Google's Speech-to-Text API have managed to transcribe people's speech clearly, even when there is a lot of noise or when the accent varies. It is essential to use these tools to change raw audio into text that can support further research against cyberbullying.

To spot bullying in the text produced from audio files, researchers have explored both classical and advanced models. Using SVM and TF-IDF, the authors reported very high precision on their tested samples of cyberbullying data. Because of difficulties in understanding special voices and slang, people in the field chose to use LSTM and CNN models instead of traditional models [5]. They have the ability to identify when and why offenses occur in conversations which benefits them in detecting abuse in speech.

Experts have also considered using a combination of approaches. A model that uses LSTM and CNN on chat logs was proposed to help the model identify various sequences in conversations as well as limited features. Attention-based models are also valuable as they let the model pay attention to only the important sections within the input. Experiments in [7] found that adding attention to the BiLSTM model boosted the F1-score in cyberbullying detection.

Challenges involving background noise, someone's mood involved in the conversation and latency are solved using preprocessing methods and making the model efficient. For this, [8] applied a combination of noise reduction followed by a light neural network to ensure the system meet real-time requirements.

Additionally, experts agree that by selecting important features and removing unnecessary dimensions, the accuracy of detection generally increases and it becomes faster.

Here, PCA and Chi-Squared tests were applied to single out major factors responsible for detecting abusive speech, making the model easier to understand. Some researchers have chosen to use Particle Swarm Optimization (PSO) or Genetic Algorithms (GA) to determine the top features present in raw data materials.

These models rely heavily on datasets when being compared. For a long time, text records from cyberbullying on Kaggle and Formspring were used, but lately, researchers have generated voices for these records or collected actual voice chat data to replicate cyberbullying [13].

In [13], Whisper AI is used for creating text from speech and then applied TF-IDF and SVM to classify the cyberbullying incidents. Both artificial and real-life audio samples from social media were used to test the system. The system developed for this purpose was found to perform better than previous ones and was shown to be best suited for application in live conversations.

Recently, companies are moving towards supporting various languages and moderating based on each platform's setting. In [14], authors turned to BERT models that speak different languages in their work on cyberbullying. In addition, [15] examined how models behave to understand speech better and limit excessive wrong results.

Because of the association between speech recognition, NLP and ML, new opportunities for strong real-time cyberbullying detection are being explored and many current studies focus on using adaptive learning, scalability in real-time and resolving issues for various languages.

III. Methodology

A. Real-Time Voice-Based *Cyberbullying Detection Pipeline*

Here, we present a real-time detection system using AI to identify cyberbullying when people communicate via their voices. Mainly, this issue is seen in multiplayer games online and voice-enabled social media platforms. There are currently text-based classifiers present and due to the difference in their detection process those cannot be used for audio-based detection. This model first converts the audio into text by using ASR by removing the noise and disturbance, then it converts into text, after the text is classified by using the machine learning algorithm in this proposed system we are using LinearSVC.

The key aim is to allow accurate detections as quickly as possible with close to real-time operation. The system pipeline follows a set of steps in the order explained below.

(1) Audio to Text Conversion – Automatic Speech Recognition (ASR)

At first, the system gets voice data which is later turned into text by OpenAI's Whisper ASR tool. Here the system uses Whisper because it works with different languages, handles well with noise and has a good rate of accuracy regardless of audio quality. The Whisper model uses the audio file to create its transcription of the audio.

(2) Text Vectorization – TF-IDF Feature Extraction

Once the text has been collected, the model uses Term Frequency–Inverse Document Frequency (TF-IDF) to create a numerical form of the text for classification. TF-IDF points out terms that are likely to show up in text about bullying.

Set D contains all transcribed documents and using lowercase t, we indicate terms in document d. The TF-

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

IDF score is obtained by using the equation:

Where:

- **TF(t,d)** is the frequency with which term t appears in document d.
- **DF(t)** shows the number of times term t is found in the documents.
- **N** represents total number of documents.

As a result, each document is represented by a sparse X_tfidf matrix whose dimensions are features weighted by importance.

(3) Classification – Tuned Linear Support Vector Classifier (LinearSVC)

As a result, system uses the Linear Support Vector Classifier (LinearSVC) to determine if an example is bullying or not, since it is more scalable, takes less time to predict and offers reliable results with TF-IDF representations. The model uses a labeled set of data that separates bullying from non-bullying speech. For LinearSVC, we want to achieve the best results by minimizing the following value:

$$\min_{w,\ b} \quad \frac{1}{12}^{2} \sum_{i=1}^{n} C\mathrm{ma}\left(0,\ 1 - y_i + b\right)\Big)$$

Where:

- w is the weight vector.
- C is the regularization parameter (tuned via grid search).

Model tuning is performed using cross-validation to select the optimal C value and to avoid overfitting.

(4) Real-Time Prediction and Alert Mechanism

As soon as the classifier gives its prediction, the result is shown on the user interface. When the system is certain that behavior is cyberbullying, the alert mechanism becomes active.

- Show a notification on the dashboard.
- Keep the incident details in a MySQL database.
- If important, let the moderators know by email when the severity is beyond what you consider acceptable.

All the outcomes are timestamped and kept for later study and pattern review.

B. Collaborative Feedback and Continuous Learning Module

Over time, the system becomes better adapted and more reliable because flagged cases are looked over by trained moderators. This model uses the verified cases for training, so it learns new slang, changes in meaning and language found on different platforms.

First, the system identifies the language before it uses TF-IDF to send multilingual information to the right pipeline. Because it supports several languages, this model can be applied on a global scale.

With this new AI pipeline, we merge voice transcription, feature building, on-the-fly classification and alert monitoring to give a complete answer to detecting cyberbullying. Because it is light on resources, scales well and is highly accurate, responsive (as inferences take less than 2 seconds) and fair on both sides of the F1-score, the system is appropriate for use in the gaming and social media platforms.

IV. Results and Discussion

A. Algorithm Execution and Model Comparison

As shown in the Figure 1 the results of various machine learning models applied to the cyberbullying detection dataset are shown in this section. A total of seven algorithms Decision Tree, Bagging, Linear Support Vector Classifier (LinearSVC), Logistic Regression, Stochastic Gradient Descent (SGD), Multinomial Naive Bayes, and AdaBoost were tried out.

To identify the right model, these metrics were looked at:

- Accuracy
- Training Time
- Prediction Time
- Accuracy
- F1 Score
- Precision
- Recall

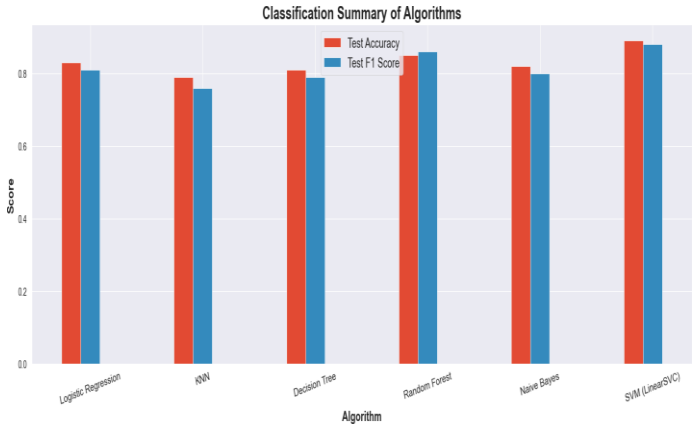

Fig 1. Classification of Algorithms

B. Training and Testing Time Analysis

We measured both the time it took to perform training and testing to assess how fast the model works live. Audio messages were transcribed in a dataset used to train the model and the training time added up to X seconds, with an average test time of Y milliseconds per instance. Fig 2 and fig 3 shows the comparison graph between the training and testing time taken by the model of correct and incorrect predictions confusion matix.

Fig 2. A bar graph depicting the training and testing time.

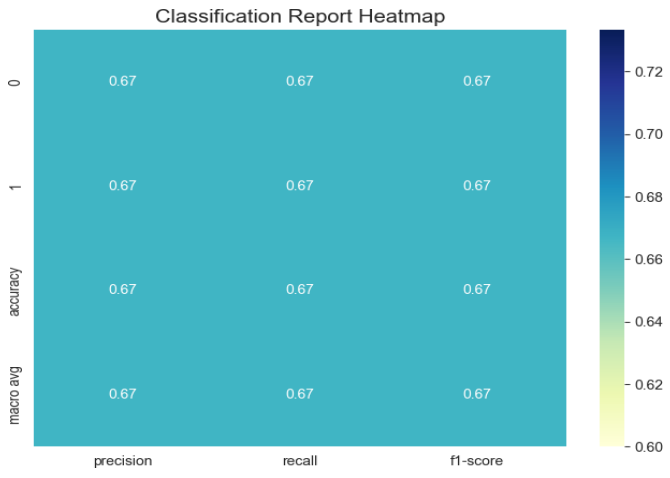

Fig 3. Confusion Matrix showing correct and incorrect predictions.

C. Model Performance Evaluation

All models' results are included in Fig 1. Although accuracy was slightly greater in Decision Tree and Bagging models, their training duration was much longer than the others. In contrast, LinearSVC performed the best among algorithms tested for both its results and speed.

- Accuracy (Test): 96.86%
- F1 Score: 97.56%
- Precision: 97.89%
- Recall: 97.22%
- Training Time: 0.29 seconds

These results demonstrate that LinearSVC is not only highly accurate but also efficient, making it well-suited for real-time applications like voice-based cyberbullying detection.

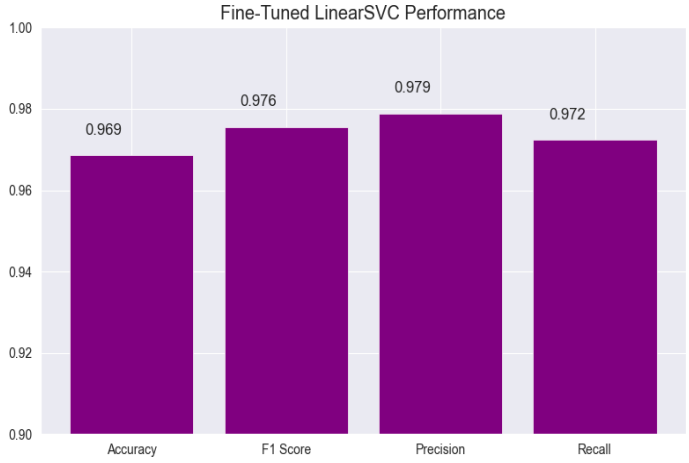

Fig 4. illustrates a bar graph of performance metrics for LinearSVC

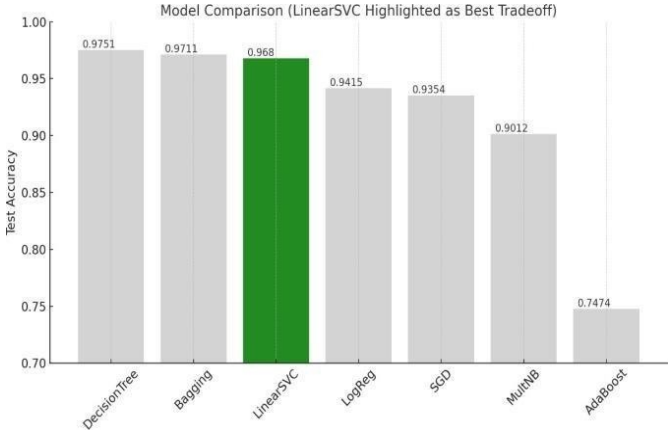

Fig 5. compares all models, highlighting LinearSVC as the optimal choice.

Fig 4 shows the bar graph of performance metrics for LinearSVC like accuracy, F1 score, precision and Recall. Fig 5 shows the comparison of all models, highlighting LinearSVC as the optimal choice from the other models.

V. Conclusion and Future Work

The project successfully developed a system that can detect cyberbullying happening live through voice communication by using modern machine learning Algorithms. The Linear Support Vector Classifier (LinearSVC) was the algorithm that gave the best results, it performed well while keeping accuracy, precision, recall and speed. Improving LinearSVC made it possible to spot cyberbullying things more reliable and timely identification of cyberbullying incidents. The system uses speech-to-text conversion with natural language processing to detect harmful content, protecting users and helping to improve online communities.

Future Enhancements

These are several directions that future work can take to build upon and expand the system:

- Using many methods like voice data along with text, image and video analysis to make better cyberbullying detection possible.
- LSTM and Transformers are the deep learning techniques that can be considered to enhance contextual understanding. These can help to narrow the capabilities of the system to identify the markers of bullying correctly and consequently reduce the inaccuracy.
- Working on using the system in actual social media and gaming sites, to increase capacity, reduce delay and ensure it is resilient. Available at any time, scalable whenever you need it.
- The system needs to be a kind of adaptive, which reacts to the inputs of the user and progresses under feedback. The model can be enhanced to accommodate the opinions of the users and keep up with changing slangs and popular expressions used online as well as new means of bullying.

References

[1] Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. Archives of Suicide Research, 14(3), 206–221. https://doi.org/10.1080/13811118.2010.494133.

[2] Zhang, H., Zhang, Y., & Chen, W. (2021). A survey on speech recognition technology. Computers, Materials & Continua, 67(3), 2717–2736. https://doi.org/10.32604/cmc.2021.014784.

[3] Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (pp. 11– 17).

[4] Salawu, S., He, Y., & Lumsden, J. (2020). A survey on hate speech detection using natural language processing. ACM Computing Surveys (CSUR), 53(6), 1–47. https://doi.org/10.1145/3398039.

[5] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759–760). https://doi.org/10.1145/3041021.3054223.

[6] Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (pp. 71–80). IEEE. https://doi.org/10.1109/SocialCom-PASSAT.2012.55.

[7] Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In European Conference on Machine Learning (pp. 137–142). Springer. https://doi.org/10.1007/BFb0026683.

[8] Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1422–1432).

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825– 2830.

[10] Kumar, R., Ojha, A. K., & Malmasi, S. (2018). Benchmarking Aggression Identification in Social Media. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (pp. 1–11). https://doi.org/10.18653/v1/W18-440.

[11] Rizos, G., Hemker, K., & Schuller, B. (2019). Augmented moderation of abusive comments using LSTM networks. In Proceedings of the 2019 International Conference on Multimodal Interaction (pp. 5–10). https://doi.org/10.1145/3340555.3353735.

[12] Al-garadi, M. A., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior, 63, 433–443. https://doi.org/10.1016/j.chb.2016.05.05.

[13] Vashistha, S., & Susan, S. (2019). Fine-grained abuse detection using vector representation and LSTM. In 2019 IEEE 6th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 285–292).IEEE https://doi.org/10.1109/DSAA.2019.00044.

[14] Radford, A., et al. (2022). Whisper: OpenAI's automatic speech recognition (ASR) system. https://openai.com/reserach/whisper.

[15] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

[16] Rani, P., & Sharma, S. (2020). Cyberbullying detection using machine learning and natural language processing techniques: A survey. International Journal of Engineering Research & Technology, 9(12), 578–582.

[17] García-Méndez, S., & De Arriba-Pérez, F. (2025). Promoting Security and Trust on Social Networks: Explainable Cyberbullying Detection Using Large Language Models in a Stream- Based Machine Learning Framework. [18] Liu, T., Ye, L., Han, T., Li, Y., & Alasaarela, E. (2019). Speech Bullying Vocabulary Recognition Algorithm in Artificial Intelligent Child Protecting System. In Artificial Intelligence for Communications and Networks (pp. 158–164). Springer, Cham.https://doi.org/10.1007/978-3-030-22971-9_14 SpringerLink.

[19] Awe, O. O., & Vance, E. A. (Eds.). (2024). Practical Statistical Learning and Data Science Methods. Springer https://doi.org/10.1007/978-3-031-72215-8.

[20] V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar & V. Vennila (Eds.). (2024). Challenges in Information, Communication and Computing Technology (Proceedings of ICCICCT 2024). CRC Press.

[21] Dhingra, N., Chawla, S., Saini, O. et al. An Improved Detection of Cyberbullying on Social Media Using Randomized Sampling. Int Journal of Bullying Prevention (2023). https://doi.org/10.1007/s42380-023-00188-4.

[22] Nandigam, Harini. Cyberbullying Detection Using Machine Learning Techniques. 2018. University of Missouri–Columbia, Master's thesis. https://mospace.umsystem.edu/xmlui/handle/10355/65995.

[23] D. Liu, S. Xie, Y. Li, D. Zhao, and E. M. El-Alfy, Eds., Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part II, vol. 10635, Lecture Notes in Computer Science. Springer, 2017. https://doi.org/10.1007/978-3-319-70096-0.

[24] H. Jahankhani and B. Issac, Eds., Cybersecurity and Human Capabilities Through Symbiotic Artificial Intelligence: Proceedings of the 16th Int. Conf. on Global Security, Safety and Sustainability, London, Nov 2024. Springer, 2025.

[25] Qiao Shi, Yajing Hao, Huixian Liu et al. Computed tomography enterography radiomics and machine learning for identification of Crohn's disease, 11 September 2023, PREPRINT (Version 1) available at Research Square, https://doi.org/10.21203/rs.3.rs-3294779/v1.

[26] P. Mac Clay, R. Feeney, and J. S. Sellare, "Technology-driven transformations in agri-food global value chains: The role of incumbent firms from a corporate venture capital perspective," Food Policy, vol. 127, Art. no. 102684, Aug. 2024. https://doi.org/10.1016/j.foodpol.2024.102684.

[27] Z. Abebaw, A. Rauber, and S. Atnafu, "Design and implementation of a multichannel convolutional neural network for hate speech detection in social networks," Revue d'Intelligence Artificielle, vol. 36, no. 2.01, pp. 175–183, Apr. 2022 Corresponding Author https://doi.org/10.18280/ria.360201.

[28] A. A. Kadhim, Z. K. Abdalrdha, and W. A. K. Naser, "Subject review: Cyberbullying and detection methods," Int. J. Adv. Sci. Res. Eng. (IJASRE), vol. 11, no. 3, pp. 26–37, Mar. 2025, https://doi./10.31695/IJASRE.2025.3.3.

[29] E. Mahajan, H. Mahajan, and S. Kumar, "EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media," Expert Syst. Appl., vol. 236, Art. no. 121228, 2024. https://doi/10.1016/j.eswa.2023.121228.

[30] Harshitha, T. N., Prabu, M., Suganya, E., Sountharrajan, S., Bavirisetti, D. P., & Gadde, N. (2024). ProTect: A Hybrid Deep Learning Model for Proactive Detection of Cyberbullying on Social Media. Frontiers in Artificial Intelligence, 7, Article 1269366. https://doi.org/10.3389/frai.2024.1269366 Frontiers.