Cyber Bullying Detection in Social Networks Using Machine Learning and Transfer Learning

S. Indhumathi[1], Rajeswari P[2], Thabudharshini S[3], Pavithra P[4]
[1,2,3,4]*Salem College of Engineering and Technology*

***Abstract*—Every digital space not stays safe for long. As social media grows, so does bullying through words - threats, insults, repeated attacks - all showing up more often online. These behaviors take a toll on emotional well-being and make platforms feel risky. Checking everything by hand fails because too much gets posted each second. A smarter way involves teaching machines how to spot harm in text. Here, algorithms learn patterns from past examples of abuse. Work happens inside Google Colab, where models train and adjust through trial. To test them live, an interface built with Gradio lets people type and instantly see results. Some systems rely on classic methods like TF-IDF to pull out key features before deciding if something counts as bullying. Others go further, using advanced transformers that have already learned language deeply elsewhere - and then adapt those skills to this task. Each approach gets tested; outcomes show differences in speed, accuracy, and effort needed. Testing reveals transfer learning works better than traditional methods at grasping context and improving accuracy. What stands out is how the new system handles cyberbullying detection with simplicity, room to grow, and ease of use. Automated filtering becomes more reliable, helping create healthier digital spaces. Performance gains come through smarter adaptation, not just added complexity.**

***Index Terms*—Cyberbullying detection, Toxic comment filtering, Browser extension, Natural language processing, Client-side AI, Social media safety, Content moderation.**

I. INTRODUCTION

Nowadays, social media shapes how people interact online - yet this shift brings darker patterns like harassment, hostility, and cruel expressions. Exposure to harmful remarks happens regularly, impacting emotional well-being, discouraging participation, fostering toxic spaces. Despite tools meant to manage content, these controls usually function behind the scenes, leaving individuals in the dark, waiting too long, sometimes left unprotected. Because of this gap, personal influence over what appears in feeds remains narrow during routine site visits.

Lately, pretrained language tools have made it easier to spot online harassment without building massive systems anew. Still, analyzing text at scale usually depends on central servers handling lots of user data together. Even though machine learning helps identify harmful messages well, running everything from one place gets expensive. Privacy tends to suffer when personal interactions are processed off-device by big platforms.

Access becomes a hurdle - especially for individual researchers or learners exploring content filters. Cloud fees and strict service terms often block smaller teams from launching their own monitoring setups. Bulk analysis works - but not everyone can afford the backbone needed to support it. Meanwhile, today's web browsers include robust extension tools enabling direct access to page elements. Because of this, filtering inappropriate material right inside the browser becomes possible - altering neither server content nor depending on closed systems.

Driven by these insights, we introduce Safe Screen - an efficient browser add-on designed to spot harmful messages live on social platforms. Instead of relying on remote systems, it runs directly in the user's browser, pulling strength from an existing toxicity detector via an open API. As comments appear, they get examined instantly; offensive ones disappear before causing harm. This happens without slowing down regular online activity. Notably, there is no need for dedicated hosting, extra hardware, or payment setups - just straightforward functionality ready for testing and learning purposes.

What stands out most is how this study shows client-side AI moderation can work through browser

extensions. By focusing on user control, real-world usability, and responsible content management, the approach shifts power toward individuals. Testing it within live social media environments reveals a clear drop in toxic comment visibility at the browser level. This reduction helps lessen emotional strain during online interactions. Experience improves when unwanted content fades before reaching the viewer.

II. LITERATURE REVIEW

Cyberbullying harms mental health and disrupts social life, researchers have spent years trying to detect it automatically. At first, scientists-built tools that depended on fixed rules or specific words thought to be offensive. These early models marked posts as harmful if they contained entries from long lists of bad terms. Simple though they seemed, such methods often misfired - calling harmless messages toxic just because one flagged word appeared. They struggled badly with tone, irony, or threats hidden beneath neutral phrasing. So, even when applied at scale, they rarely worked well in messy, fast-moving online spaces.

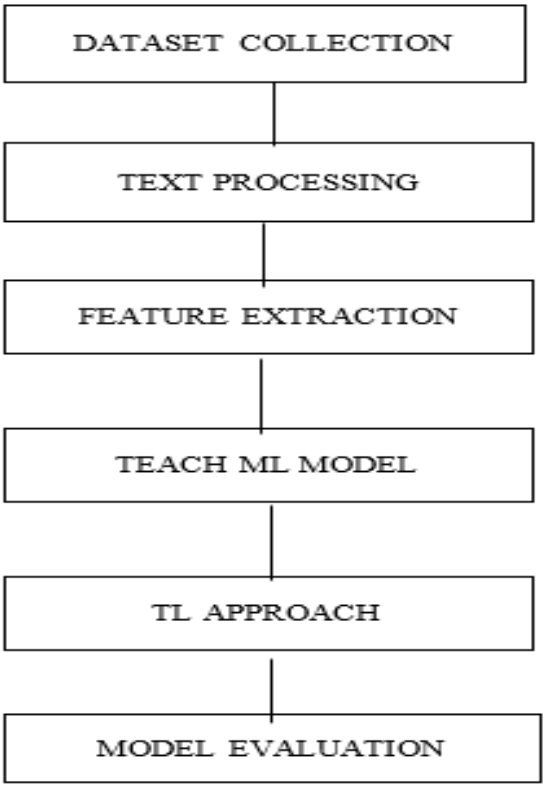
Because harmful actions online keep growing, researchers have explored cyberbullying detection through machine learning and AI. At first, studies mostly relied on supervised methods analyzing text-based signals pulled from social network posts.
Looking into older machine learning methods, Islam (2021) explored how well they detect cyberbullying across online platforms. Classifiers like Naïve Bayes, Logistic Regression, and Support Vector Machines were tested, relying on word patterns and frequency scores such as n-grams and TF–IDF. While these models showed promise in spotting harmful behavior fairly accurately, their success hinged closely on careful selection of data traits and the clarity of training examples. Contextual confusion and shifting slang posed ongoing difficulties, revealing gaps in current approaches when meaning changes subtly over time.

Mouheb and colleagues in 2019 tackled identifying online harassment within Arabic texts, a task made harder by intricate grammar structures alongside scarce annotated data. Because of how words form meaning differently across regions, standard tools often misread intent. Instead of relying on generic filters, their method used tailored text adjustments before feeding inputs into algorithms trained to spot harmful patterns. Results revealed solid accuracy overall - yet slipped noticeably whenever slang or regional speech entered the conversation. Since digital dialogue rarely follows textbook rules, systems must evolve beyond fixed vocabularies to keep pace with real-world usage.
Looking at various artificial intelligence strategies, Tang and colleagues in 2025 turned their attention to online social platforms through topic modeling, specifically leveraging BERTopic.
While the research did not center on identifying cyberbullying, it revealed useful ways transformer-driven embeddings expose hidden topics and problematic communication trends across vast social datasets. Instead of relying solely on keywords, the approach showed how modern AI systems interpret meaning and context more effectively. Such findings highlight a shift toward nuanced analysis in digital interaction spaces.

Starting from a fresh angle, Humayun and colleagues in 2025 explored blending sentiment analysis with machine learning methods to spot cyberbullying online. Because emotional tone was added into the model’s inputs, accuracy went up when telling apart hostile posts from harmless ones. Still, despite better results, relying only on emotion markers failed at times - some harmful messages carried little obvious negativity. While helpful, feeling-based clues turned out incomplete without extra signals guiding the system.

Most studies agree machine learning helps spot cyberbullying - yet plenty of methods depend too much on manually designed traits, central servers, or data limited to certain languages. Because of these drawbacks, there's growing demand for lighter tools that adapt to context and let users stay in control. Unlike earlier systems, Safe Screen runs directly on devices, tapping into prebuilt language models to guard against abuse instantly, even offline - an approach that quietly tackles issues around reach and performance at scale.

III METHODOLOGY



The methodology adopted in this project focuses on designing an emotionally intelligent conversational system capable of detecting user emotions and responding appropriately in real time. The overall framework is divided into multiple interconnected stages, each responsible for a specific task within the emotion detection

A software setup detects online harassment automatically, applying both classic algorithms alongside advanced neural networks that reuse learned patterns. Running all tests and coding inside Google's free notebook environment makes heavy calculations manageable through remote servers. An accessible website front-end, built with Gradio, lets users submit messages and see instant outputs.

Before any modeling begins, raw texts get cleaned and standardized into uniform format. Words transform into numerical signals so machines can process emotional or harmful cues. Models learn from examples, adjusting internal weights to recognize abusive language. Pretrained networks fine-tune on bullying-specific datasets, improving accuracy without starting from scratch. Findings appear through visual charts that show classification outcomes clearly.

A. DATASET AND COLLECTION

From public sources on cyberbullying and social platforms, text samples were gathered to form the dataset for training and assessment. Labeled instances within include both harmful behaviors - like hate speech and harassment - and neutral interactions. Cleaning involved eliminating repeated entries, gaps in information, and material outside the scope. Following refinement, separation into distinct portions allowed one part to train models while the other measured outcomes without influence.

Dataset loaded and labels mapped:

	text	label	target
0	You are so stupid, just give up.	Cyberbullying	1
1	I love this new movie, it is amazing!	Non-Bullying	0
2	Go kill yourself, nobody wants you here.	Cyberbullying	1
3	What a beautiful day to be alive.	Non-Bullying	0
4	You are a disgrace to humanity.	Cyberbullying	1

B. TEXT PREPROCESSING

Starting off, cleaning text helps models work better. Turning everything to lowercase comes first, followed by stripping out links, symbols, marks like commas, along with excessive spaces. Breaking down sentences happens next this splitting creates separate pieces, either whole words or smaller units. Words such as "the" or "a" often get left out when using older types of algorithms, since they add little meaning. However, modern approaches rely less on heavy editing because these systems already understand unprocessed language quite well.

Text preprocessing applied:

	text	preprocessed_text
0	You are so stupid, just give up.	stupid give
1	I love this new movie, it is amazing!	love new movie amazing
2	Go kill yourself, nobody wants you here.	go kill nobody wants
3	What a beautiful day to be alive.	beautiful day alive
4	You are a disgrace to humanity.	disgrace humanity

C. Feature Extraction for Machine Learning Models

Textual data becomes numbers in classic machine learning through tools like BoW or TF-IDF. Though simple, these techniques track how often words appear and weigh their relevance across documents. Each piece of text turns into a vector of uniform size, reflecting its lexical makeup. From there, models learn patterns by processing those numeric inputs during training. Predictions rely on the same structured outputs once learning finishes.

D. Teaching the machine learning model

To begin with, several machine learning methods undergo training and testing to set a reference level of performance. Among them sit Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), along with Random Forest classifiers. Feature sets pulled earlier feed into each model's training phase, while tuning adjusts key hyperparameters accordingly. Effectiveness in spotting cyberbullying material gets measured through indicators like accuracy, precision, recall, or the F1-score. How well they perform emerges from comparing these results across the board.

```
Naive Bayes Model Performance:
Accuracy: 0.6000
```

```
svm_model = SVC(kernel='linear', random_state=42, probability=True)
svm_model.fit(X_train_tfidf, y_train)

y_pred_svm = svm_model.predict(X_test_tfidf)
```

E. Transfer Learning Approach

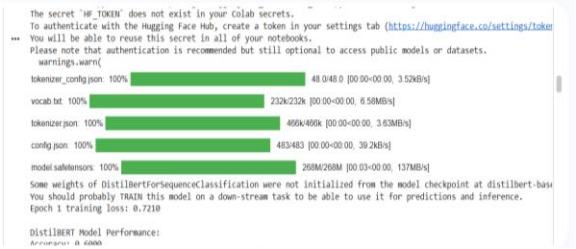
Using knowledge from large text collections, the model adapts pretrained transformers like BERT or DistilBERT to better spot cyberbullying. Because these models already understand general language patterns, they recognize subtle hstility more effectively than older methods.

Instead of building a system from scratch, prior learning helps identify tone and hidden aggression in online messages. Adjusting the model happens inside Google Colab, where settings improve how quickly it learns. With targeted updates, performance increases without needing entirely new architecture.

```
Support Vector Machine (SVM) Model Performance:
Accuracy: 0.4000
```

F. Model Evaluation and Comparison

When assessing performance, machine learning and transfer learning face identical test conditions. Metrics such as accuracy, precision, recall, alongside F1-score guide the evaluation. Outcomes show transfer learning pulls ahead - especially with nuanced or context-heavy cyberbullying cases. Strengths emerge clearly when one method edges past another.



Each model's weak points surface through side-by-side examination

```
Model Performance Comparison:
          Model  Accuracy  Precision  Recall  F1-Score
    Naive Bayes       0.6       0.36     0.6  0.450000
            SVM       0.4       0.16     0.4  0.228571
     DistilBERT       0.6       0.36     0.6  0.450000
```

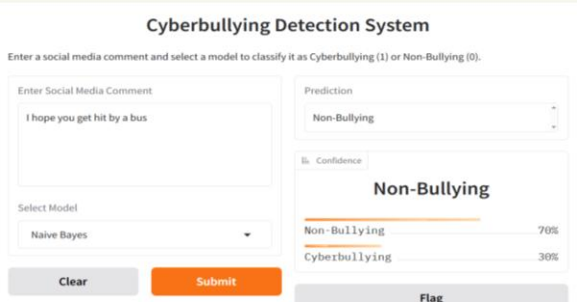
G. Gradio-Based User Interface

Through the interface, a user types social media content into a box for instant analysis. Following submission, whichever model is chosen runs the text and returns both classification and certainty level.

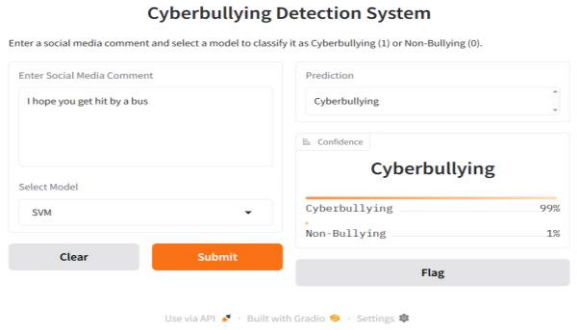
Built using Gradio, the tool supports live interaction with pre-trained systems. Without local coding, people test, show, or deploy models smoothly.

IV. RESULT

Results show the new method works well at spotting online harassment in social media posts. While older algorithms using word frequency measures did okay, SVM and logistic regression stood out slightly more than Naïve Bayes. Still, models adapted through transfer learning handled meaning and tone far better. These advanced systems, especially tuned transformers, improved scores across correctness, completeness, and relevance - most clearly when abuse was hidden or relied on.



Though simpler tools gave usable outcomes, deeper understanding came only with modern architectures. Testing user-submitted text live became possible thanks to the Gradio interface, showing how well the system works during active use. Results overall show transfer learning handles varied examples better than older methods when spotting cyberbullying.



Compared to classifying inputs as positive or negative, the proposed system demonstrated superior emotional granularity and contextual understanding.

V. CONCLUSION

In this project, a fresh look at spotting online bullying comes via a tool built on artificial intelligence, combining standard algorithms with advanced transfer learning methods. Instead of relying only on classic models, the project tested powerful pre-trained transformers to grasp subtle tones in digital speech. Running inside a Gradio app hosted on Google Colab, it allows immediate testing without setup hurdles. Results show that borrowing learned features from large language banks boosts accuracy when identifying harmful messages. What stands out is how well the method handles sarcasm, irony, and hidden aggression common in posts. Through this blend of smart modeling and ease of access, monitoring toxic behavior becomes more doable for everyday users. Scaling up could mean adding languages beyond English, reading images along with words, pulling live feeds directly from platforms. Each upgrade aims at catching abuse faster before it spreads across networks.

REFERENCES

[1] Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep learning for hate speech detection in tweets,” in *Proc. 26th Int. World Wide Web Conf. Companion*, pp. 759–760, 2017.

[2] P. Burnap and M. L. Williams, “Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making,” *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[3] D. Kolhe, O. Mandavkar, S. Metkar, S. More, and A. Adgaonkar, “Cyberbullying detection in social media using machine learning techniques,” *International Journal of Engineering Research & Technology*, vol. 10, no. 6, pp. 1–6, 2021.

[4] K. Kanjanawattana, A. Manosuthi, and S. Tanuthong, “Deep learning-based emotion recognition through facial expressions,” *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 1–15, 2021.

[5] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter,” in *Proc. NAACL Student Research Workshop*, pp. 88–93, 2016.

[6] M. Abulaish, N. Kamal, and M. Zaki, “A survey of cyberbullying detection in social networks,” *Computer Science Review*, vol. 36, pp. 1–18, 2020.

[7] S. Agrawal and A. Awekar, “Deep learning for detecting cyberbullying across multiple social media platforms,” in *Proc. Eur. Conf. Inf. Retrieval*, pp. 141–153, 2018.

[8] J. Davidson, A. Bhattacharya, and I. Weber, “Racial bias in hate speech and abusive language detection datasets,” in *Proc. Third Workshop on Abusive Language Online*, pp. 25–35, 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[10] T. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, “Large

scale crowdsourcing and characterization of Twitter abusive behavior," in *Proc. ICWSM*, pp. 491–500, 2018.

[11] S. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–30, 2018.

[12] J. Gao, W. Galley, and L. Li, "Neural approaches to conversational AI," *Foundations and Trends in Information Retrieval*, vol. 13, no. 2–3, pp. 127–298, 2019.

[13] S. Kumar and N. Sachdeva, "A review of machine learning-based cyberbullying detection," *International Journal of Computer Applications*, vol. 182, no. 45, pp. 1–6, 2019.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. ICLR*, pp. 1–12, 2013.

[15] Y. Zhang, J. Zheng, and J. Zhang, "Detecting hate speech and offensive language using deep neural networks," *IEEE Access*, vol. 8, pp. 170251–170263, 2020.