IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Automatic Method To Prevent Andclassify Cyberbullying Incidents Using Machine Learning Approach

[1]Mrs. A. Aruna, [2]P.Bhargavi, [3]G.Guru Rasagna, [4]G.Charvika, [5]A.Pavan Durga Prasad, [6]V.Srimannarayana,
[1]Professor, [2,3,4,5,6]Students, Department of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Vijayawada, India

Abstract: This project establishes an analytical system to decode the complex nature of cyberbullying and link it to relevant legal offenses. By amalgamating attributes like platform, content, relationship, and intensity, detailed event profiles are constructed. A comprehensive list of cyberbullying-related crimes, from harassment to extortion, is developed, and a deep neural network, trained on labeled data, classifies these offenses by intent, harm, and legal codes. The system identifies patterns of repeated crime, demonstrating the alignment of cyberbullying with legal violations, and examines the offline consequences of online conduct. This work aims to support interventions in law enforcement, education, and social work by providing informed, targeted responses. By tracing crimes back to specific cyberbullying events, the system offers deeper insight into cybercrime, ultimately enhancing our knowledge and providing a foundation for improved prevention and reduction strategies.

***Index Terms* -** Cyberbullying,SocialMedia,DeepLearning,Detetction System,Classification System

I. Introduction

Social media provides live global communication and information sharing and has served as a useful channel for research and government information. Social media also made bad behaviors like scamming, mis-information, and cyberbullying possible. Cyberbullying as one of the harassing acts conducted online comprises offensive remarks posted on the internet and online messages sent to embarrass and discredit their victims. Cyberbullying comprises behaviors such as exclusion, stalking, and impersonation and targets victims with the intention of causing harm and rendering them defenseless and vulnerable.

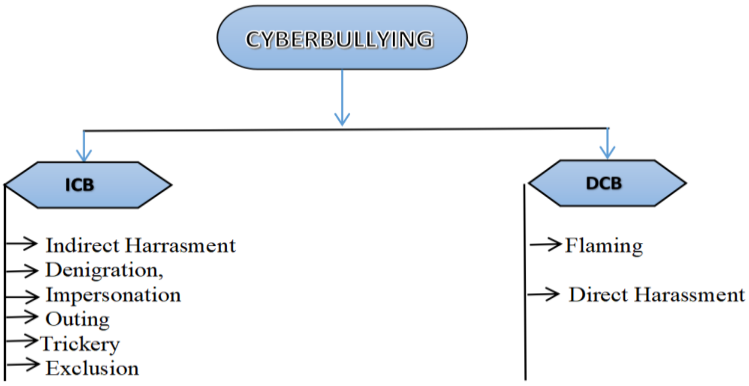


Fig-1 Types of Cyberbullying

Types of cyberbullying	Details
Flaming	Imagine a mean fight in real life, but compounded by the anonymity of the web. People say things they never would in real life in ordinary situations.
Direct Harassment	"That's absolutely cyberbullying. It's when a person targets another person specifically with negative or threatening messages, such as in texts or emails."
Indirect Harassment	That's cyberbullying, but they're doing it behind their back. Instead of standing up to them, they're doing it in a way that's meant to hurt them using other people.
Denigration	"It's like an online character assassination. They're trying to assassinate the way people think about someone by making them look bad with fake facts."
Impersonation	"That's when you act like you online, like make a fake profile, and then act bad to sound terrible."
Outing	"That's putting someone's personal info out there without them, just to embarrass or hurt them."
Trickery	"That's when you make someone trust you by acting like you're friends with them, and then use the trust to get their personal info and betray them."
Exclusion	"That is literally excluding people from online groups or communities and cu tting them out."

Table-1 Details of Types of Cyberbullying

Anonymity provided by Web 2.0 tools renders cyberbullies hard to track, and the victims may feel that they are being invaded inescapably. Timely intervention requires effective detection, but human moderation tends to be inconsistent, biased, and closed, intruding possibly upon freedom of speech. Identifying the bully and the victim can be problematic, particularly in communities that live as close-knit groups. Cyberbullying continues on social media despite increased awareness owing to sheer content volumes generated by users.

The real-time identification of cyberbullying is also made more complex by the multiform nature of online postings, such as memes, videos, and picture-based posts. Anonymity and context-less utterances are sources of ambiguity. Cultural differences, linguistic subtleties, and the incorporation of non-standard typographical devices also present linguistic challenges.

Researchers are in the process of creating techniques to detect and counter cyberbullying. Complex analytical and computational frameworks are needed to analyze and process negative material in different media. Automated detection has made considerable progress, helping eliminate toxicity in the online space. Text-based cyberbullying studies have advanced, yet real-time detection demands complex semantic analysis. Conventional techniques tend to be based on manual feature extraction or lexicon-based approaches, which are non-efficient and narrow in scope.

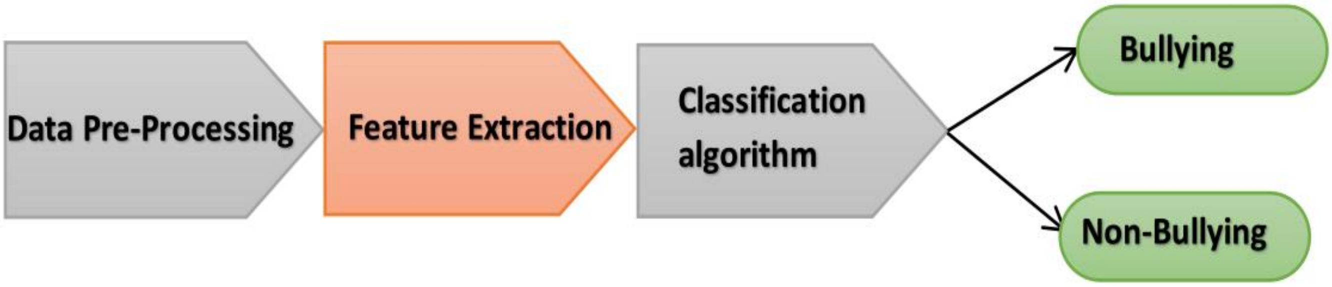


Fig-2 Generic cyberbullying detection process

The pre-processing stage involves cleaning the data obtained by eliminating unwanted URL's or strings etc., dealing with missing values, correcting words etc. and then converting it into a representation thatfeature-extractable. Then features like keywords representing bad/nasty/rude/abusive/hateful/attacking words, N-grams, pronouns, skip-grams are extracted. The following phase employs supervised learning algorithms to classify the messages as either with bullying or without bullying content.

1.1DEEP LEARNING

The number and diversity of user-created content on intricate social media platforms have increased the challenges to identify cyberbullying in real time. The flood of content poses the difficulty to timely control online expression. Furthermore, the anonymity and context-independence of statements in online postings may be misleading or vague. More recently, as memes, internet videos and other image-based, inter-textual material have become the norm in social streams; typo-graphic and info-graphic visual material has also become a significant part of user-generated information.

Researchers across the globe have been attempting to create novel methods for the detection of cyber bullying, handling it and lowering its occurrence on social media. Sophisticated analytical techniques and computational models for effective processing, analysis and modeling for the detection of such nasty, taunting, abusive or negative content on images, memes or text messages are the need of the hour.

Deep learning (DL) is viewed as being a subset of the larger class of ML as data representation learning,distinguished from task-specific algorithms and where learning can be supervised, semi-supervised or unsupervised. DL encompasses techniques such as deep neural network (DNN), recurrent NN, CNN, deep-belief networks etc., whereas NN is being one of the sub-categories of SC techniques which includes feed forward; MLP ; deep NN (DNN); radial-basis etc.

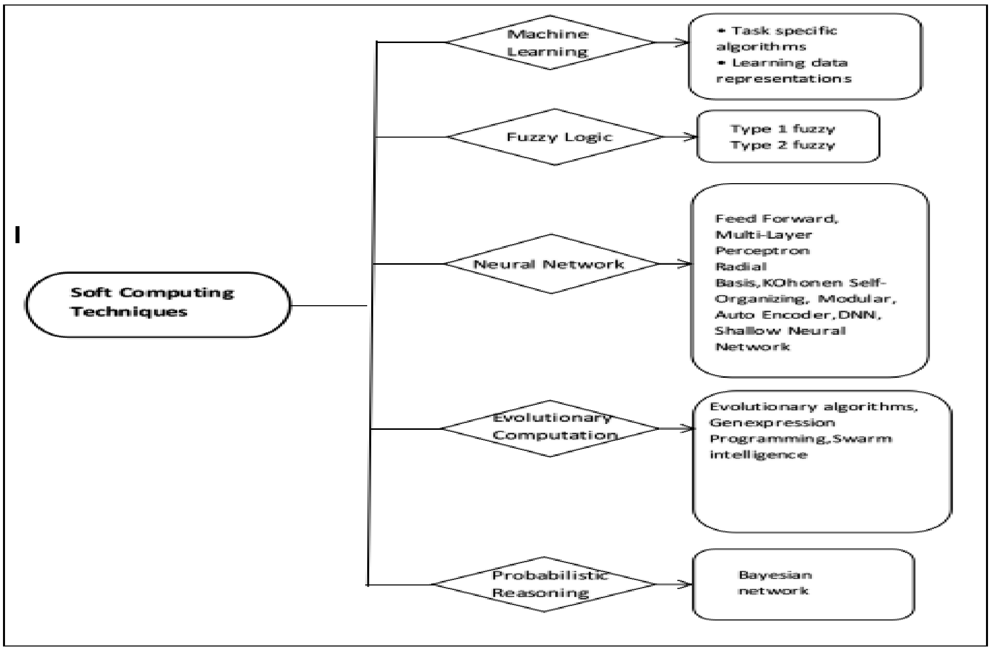

Soft Computing Techniques
Machine Learning
• Task specific algorithms
• Learning data representations
Fuzzy Logic
Type 1 fuzzy
Type 2 fuzzy
Neural Network
Feed Forward, Multi-Layer Perceptron Radial Basis,KOhonen Self-Organizing, Modular, Auto Encoder,DNN, Shallow Neural Network
Evolutionary Computation
Evolutionary algorithms, Genexpression Programming,Swarm intelligence
Probabilistic Reasoning
Bayesian network

Fig-3 Flow chart of soft computing Techniques

These deep models possess better representation learning capacity. They automatically extract features for the wanted outcomes and are also effective. The measurement of social media user-generated content can be useful for automatic cyberbullying detection using the deep neural architectures. This research demonstrates the practicability, scope & applicability in implementingDL models for identifying CB in social media websites. Application of deep learning algorithms to detect cyberbullying on social media is an area of future research scope for identification, investigation and analysis of extentability of human-centric expressions.

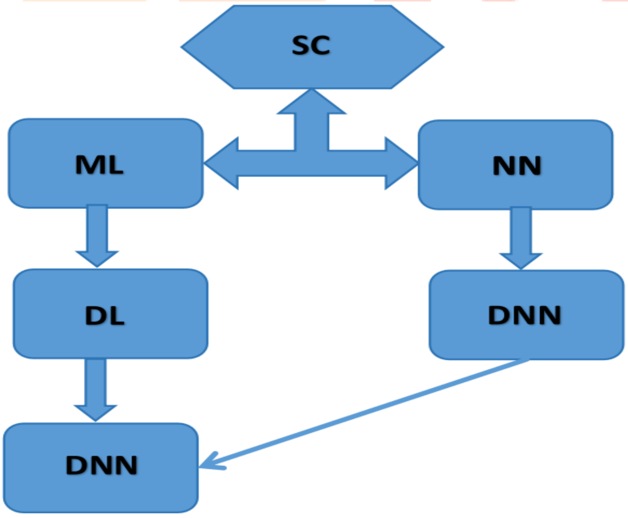


Fig-4 Relation between SC, ML and DL

1.2OBJECTIVE

•To successfully come up with a system to enable social media platforms to detect and identify abusive comments.

•In order to utilize Deep Neural Network, a regression machine learning algorithm for effective detection of abusive posts.

1.3ARCHITECTURE

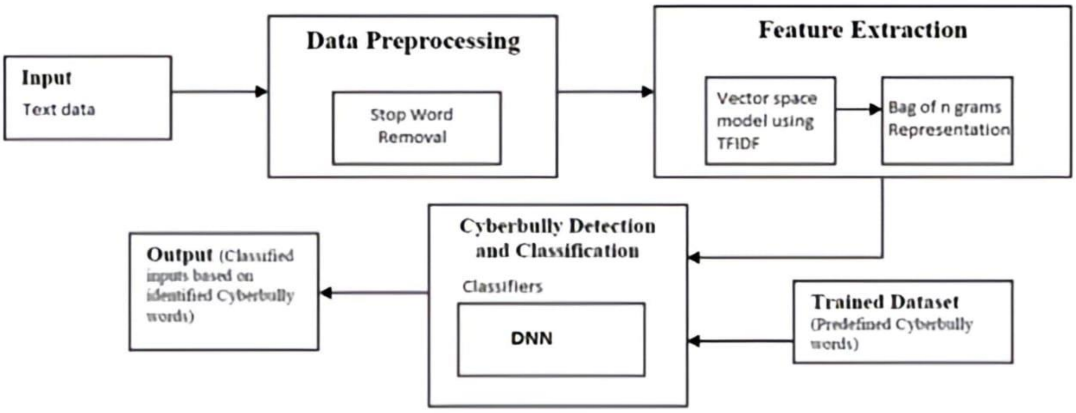

Input
Text data
Data Preprocessing
Stop Word Removal
Feature Extraction
Vector space model using TFIDF
Bag of n grams Representation
Cyberbully Detection and Classification
Classifiers
DNN
Output (Classified inputs based on identified Cyberbully words)
Trained Dataset (Predefined Cyberbully words)

Fig-5 Architecture of Cyberbullying Detection

WORKING:

This project seeks to identify abusive social media posts through data mining and machine learning.
Data Collection: Collect varied datasets of abusive posts from the web.
Data Preparation: Prepare and divide datasets into training and test sets.
Preprocessing: Classify and preprocess the datasets.
Model Training: Use a Deep Neural Network (DNN) regression algorithm to train an abusive comment detection model.
Evaluation: Use the test dataset to measure the performance of the model.
Application: Run the model to identify and mark abusive comments and enhance the social media experience.

2.LITERATURE SURVEY: -

1) Abdhullah-Al-Mamun, Shahin Akhter[2019]

"This work aims to detect social media cyberbullying in Bangla text through the use of machine learning. Towards this end, a specially tailored dataset was prepared by employing a Java program that was used to scrape and harvest Bangla text conversations from leading online platforms, including Facebook and Twitter. The work utilizes a Convolutional Neural Network (CNN), a sophisticated algorithm used extensively in text-based analysis for its effectiveness, to detect and classify instances of cyberbullying in the data harvested. This work is an attempt to meet the emerging need for detection of online harassment in the Bangla language using advanced computational techniques."

2) Md Faisal Ahmed, Zalish Mahmud, Zarin Tasnim Biash[2018]

"The study, titled 'Bangla Text Dataset and Exploratory Analysis for Online Harassment Detection,' aimed to build a resource for identifying online harassment in Bangla text. To enhance the dataset's depth, the researchers incorporated gender classification, labeling the authors and targets of comments. This addition allowed for a more nuanced analysis of harassment patterns. However, it was noted that the distribution and quantity of information within each classification presented a limitation, resulting in a potentially less robust learning source. This suggests that while the dataset's inclusion of gender is valuable, further expansion and balancing of the data would be beneficial for improved model training and accuracy

3) Bandeh Ali Talpur [2020]

"This research provides a machine learning approach for determining cyberbullying severity. A model was built that used a feature-based approach, consisting of a set of features based on tweets. These features are network features, patterns of user behavior, user profile information, and the text within the tweets themselves. However, the model also experienced data imbalanced problems, specifically those of over-sampling and under-sampling. This data imbalance can lead to discrepancies in the model's predictions, and therefore can potentially affect the model's capacity for distinguishing the severity of cyberbullying incidents."

4) Christian Reuter[2018]

"This study performed a fifteen-year retrospective analysis of social media use during crises, seeking to offer directions for future research in crisis informatics.[1]To study the enormous body of social media data, researchers used the Stanford Network Analysis Platform (SNAP). It should be noted that the Snap library itself is open source, and this offers a rich collection of functions.".[2]Hence enabling the researcher to be capable of anaylzing network data. This research explored how social media has become an important tool in emergency response, and how its use can be enhanced.

5) Stefan Stieglitz[2018]

"This study examined the sense-making process on social media during times of extreme events. The study tried to accomplish this through several Natural Language Processing (NLP) methods, with the aim of extracting useful information and trends from the usually disorganized and overwhelming data generated when crises erupt. But the researchers noted that the tools and platforms used were not as easy to use, and that there is a need for more accessible and more intuitive interfaces to allow for efficient analysis and sense-making in emergency environments."

3. METHODOLOGIES:-

3.1. Data Acquisition and Preparation:

- **Data Collection:** Data is collected through prompt-based scenarios, social media (Twitter, Reddit), online forums, questionnaires, and interviews. The collection is varied in an attempt to cover a vast majority of cyberbullying incidents.
- **Dataset Description:** The dataset is labeled text data (tweets, comments, posts) which specifies the prevalence and type of cyberbullying. Publicly available datasets such as Twitter, Reddit, and Formspring.me cyberbullying datasets are used.
  - **Text Preprocessing:** Raw text goes through numerous preprocessing operations:
  - **Tokenization:** Breaking text down into words.
  - **Lowercasing:** Lowercase the text to maintain consistency.
  - **Stop Word Removal:** Removing popular, non-descriptive words.
  - **Stemming/Lemmatization:** Reducing words to their base word.
  - **Feature Extraction:** Preprocessed text is converted into machine learning numerical data:
  - **TF-IDF (Term Frequency-Inverse Document Frequency):** Assigns word relevance in documents.
  - **Word Embeddings:** Positions word meanings in compact vectors preserving semantic relationships.
  - **Text Normalization and Cleaning:** Unnecessary characters, URLs, and HTML tags are stripped off. Text formatting is normalized.
  - **Feature Engineering:** New features are created from existing ones to enhance model performance.
  - **Encoding Categorical Responses:** Categorical responses (e.g., type of cyberbullying) are encoded using one-hot or label encoding.

3.2. Analysis and Modeling:

- **Sentiment Analysis:** Identifies the emotional sentiment of the text to determine negativity or aggression.
- **Classification:** Categorizes cyberbullying cases by severity and type.

Machine Learning Models

- **Naive Bayes:** Probabilistic classifier.
- **Support Vector Machines (SVM):** Classification and regression algorithm.
- **Logistic Regression:** Binary classification model.
- **RandomForest:** Ensemble-based decision tree learning algorithm.

Deep Learning Models:

- **Convolutional Neural Networks (CNNs):** For text and image data.
- **Recurrent Neural Networks (RNNs):** For sequential data such as text.
- **Long Short-Term Memory (LSTM) Networks**: For learning long-term text dependencies.

Text Representation:

- **TF-IDF:** Weights words based on frequency and inverse document frequency.
- **Bag of Words (BoW):** Text is represented as word count vectors.
- **Stress Level Classification:** The cases of cyberbullying are categorized into mild, moderate, or severe stress classes depending on their effect.

3.3. Model Evaluation:

Evaluation Metrics:

- **Accuracy:** number of correctly classified cases / total number of instances.
- **Precision:** true positives / predicted positives.
- **Recall:** true positives / actual positives.
- **F1-Score:** harmonic mean of precision and recall.
- **Precision Formula:** An example of Python code based on sklearn.metrics.precision_score is given.
- **Model Training and Prediction:** Labeled datasets are trained in machine learning models to learn patterns of cyberbullying and make predictions about cyberbullying in unseen data.
- **Accuracy Calculation:** Accuracy is calculated per-class label or is averaged, based on research requirements.

4. IMPLEMENTATION: -

4.1First Approach

Preparation of the Information (The Foundation)

Obtaining the Comments:

We needed a lot social media comments. We were pretty much collecting datasets from everywhere we could find. Our aim was to have all types of comments: nice ones, horrible ones, and those that fall in between. We also wanted to know how people communicate in different regions, so instead of consolidating comments from a single region, we extracted information globally. We also had to identify all the comments that were genuinely misconducts. That required sifting through the data and tagging which was… well, let's just say a lot of valuable work went into it.

Dataset Collection
Datas Collected from Kaggle website
These datas are used for Training Using Machine Learning Algorithm

Fig-6 Dataset Collection

4.2Second Approach

Comment Clustering and Abusive Comment Detection Methodology

The project aims at detecting abusive comments using a structured approach of data preparation, model training, and testing.

1. Data Preparation and Clustering:

Data Collection: A large dataset of social media comments was collected.

Clustering: The gathered comments were divided into two groups, each with fifteen folders.

Training Set: Utilized to train the machine learning model to identify abusive language.

Testing Set: For the purpose of model performance evaluation.

Data Splitting: The typical 80/20 ratio was used, dividing 80% of data into the training set and 20% into the test set. The scikit-learn library's function for data splitting was used.

Random Seed: An explicit random seed was applied in order to guarantee reproducibility by making consistent data splits at different runs, ensuring identical initial conditions every time.

Data Cleaning: Raw comment data were rigorously cleaned in order to rectify inconsistencies:

Elimination of unnecessary characters.

Correction of typos.

Conversion of everything to lowercase.

Text Vectorization: In order to allow machine learning algorithms to analyze text data, the comments were converted into numerical data representations via the TF-IDF method. This method gives different words weights according to their significance in the dataset.

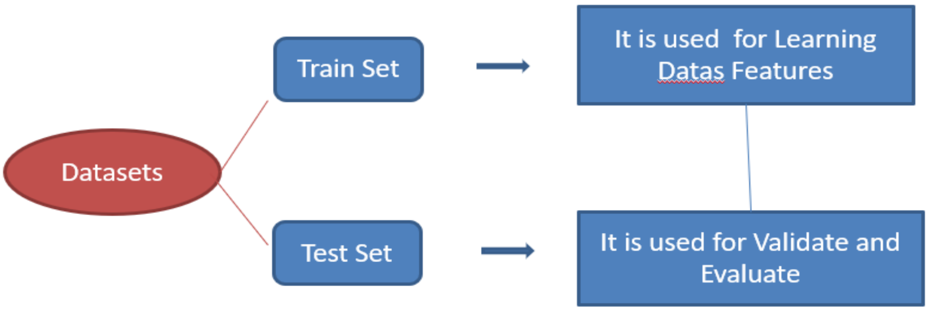

Datasets
Train Set
Test Set
It is used for Learning Datas Features
It is used for Validate and Evaluate

Fig-7 Types of Datasets

2. Model Training and Evaluation:

Model Selection: Random Forest algorithm was utilized because of its efficiency in classification problems. This algorithm works by averaging the predictions of ensemble decision trees.

Model Training: scikit-learn library was utilized to create the Random Forest model. The model was trained with the set training dataset.

Model Evaluation: The performance of the trained model was tested using the test dataset. Primary metrics were utilized:

Accuracy: Percentage of comments that were classified correctly.

Precision: Ratio of accurately identified abusive comments to those that were predicted as abusive.

Recall: Ratio of accurately identified abusive comments to all true abusive comments.

ROC Curves (Receiver Operating Characteristic): To visually evaluate the model's capacity for discriminating abusive from non-abusive comments.

Model Optimization: If the performance of the model was not good enough, iterative tuning was conducted:

Parameter tuning of the Random Forest algorithm.

Investigation of other machine learning algorithms.

Adding to the training set with more data.

This trial-and-error process enabled the tuning of the model to get the best results.

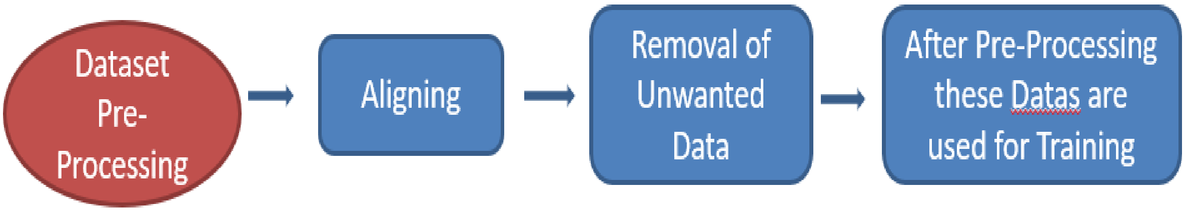


Fig-8 Dataset Pre-Processing

4.3Website & UI Integration

Deployment

Creating an Interface:An API was created so that other applications could send comments to the model and receive a reply. The API was built using a framework like Flask or FastAPI. Every effort was made to ensure security and speed.

Looking Forward:

In the future, we hope to see these features implemented in social media sites and apps. Consider the implications if platforms were able to automatically erase hateful comments using this technology. That would be remarkable. We also believe that creating browser extensions that preemptively warn users before they make abusive comments would be a fantastic feature.

OUTPUTS FOR TEXT

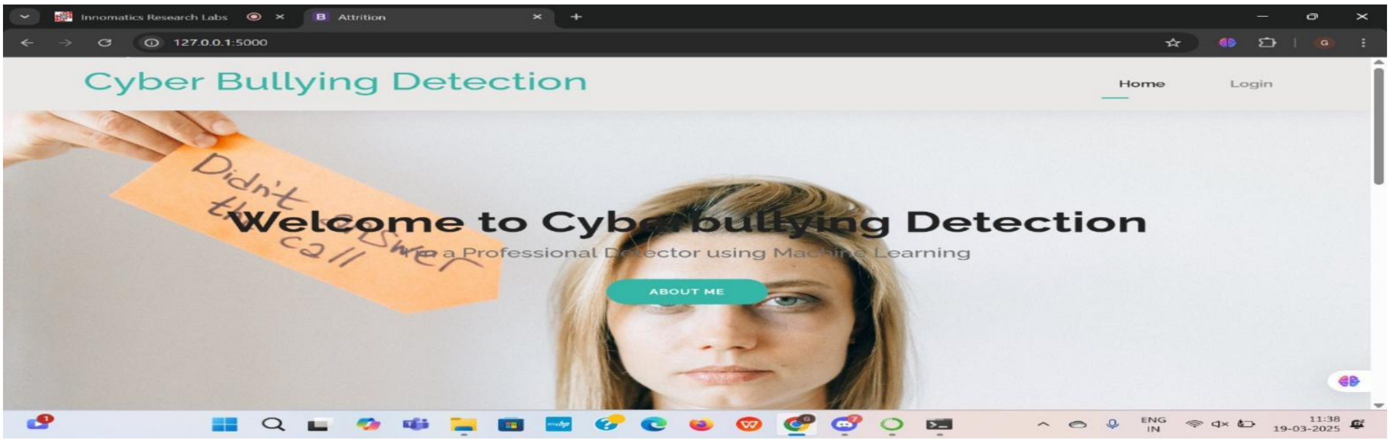
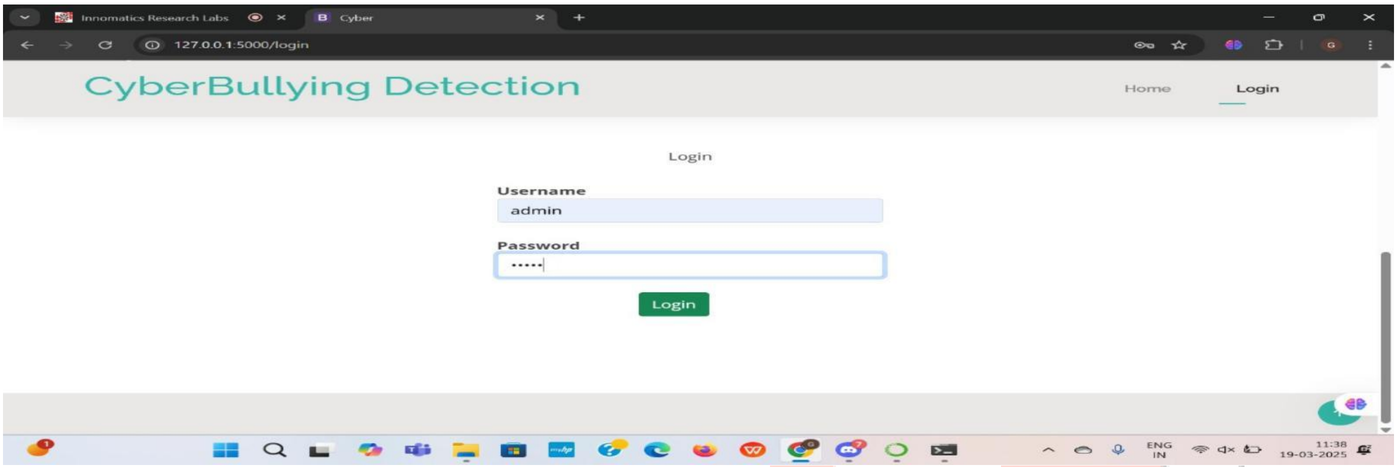


Fig-9 Website Home page for text detection



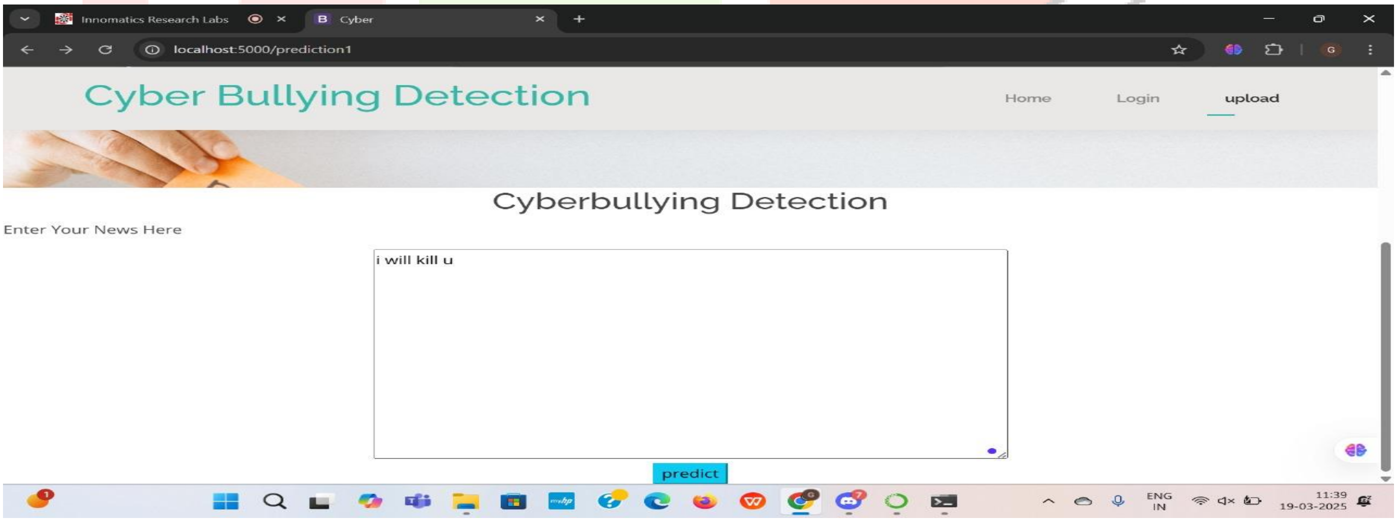
Login

Username
admin

Password
•••••

Login

Fig-10 Text Website loginpage



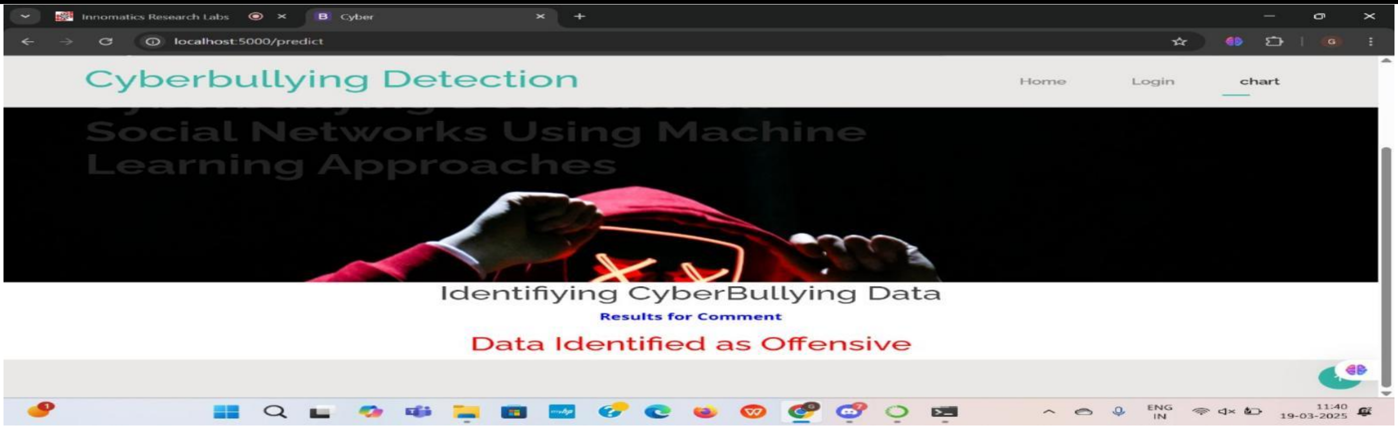
Cyberbullying Detection

Enter Your News Here

i will kill u

predict

Fig-11 Giving Random comment



Social Networks Using Machine Learning Approaches

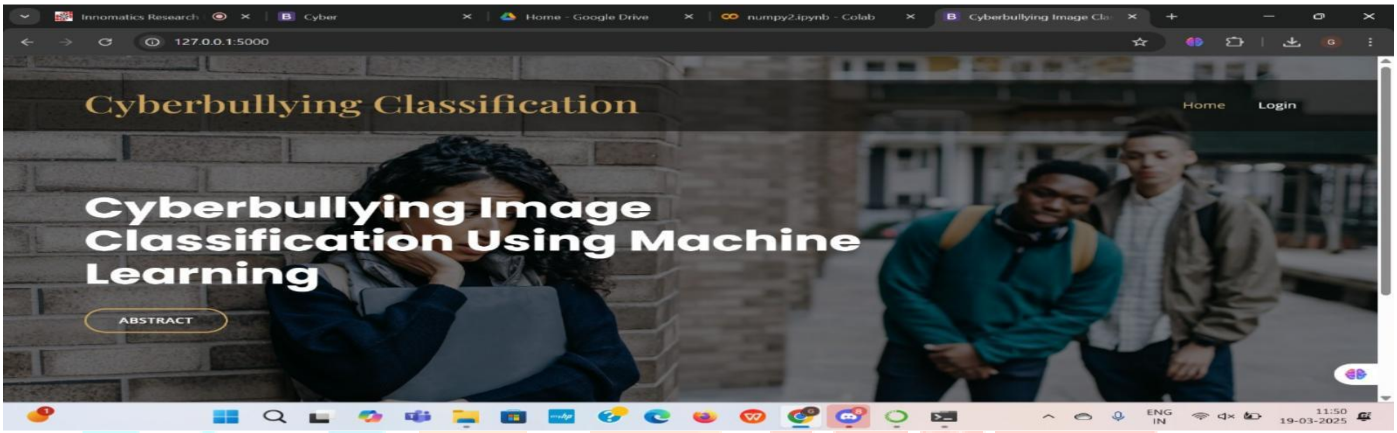
Identifiying CyberBullying Data

Results for Comment

Data Identified as Offensive

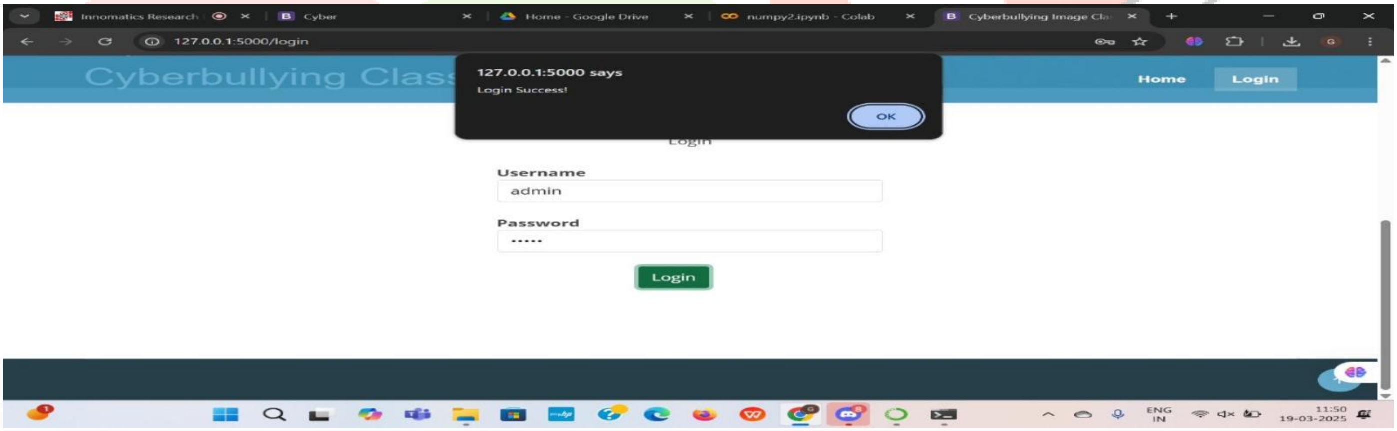
Fig-12Detected text bullying or not

OUTPUTS FOR IMAGE



Cyberbullying Image Classification Using Machine Learning

Fig-11 Image Detection Website Homepage



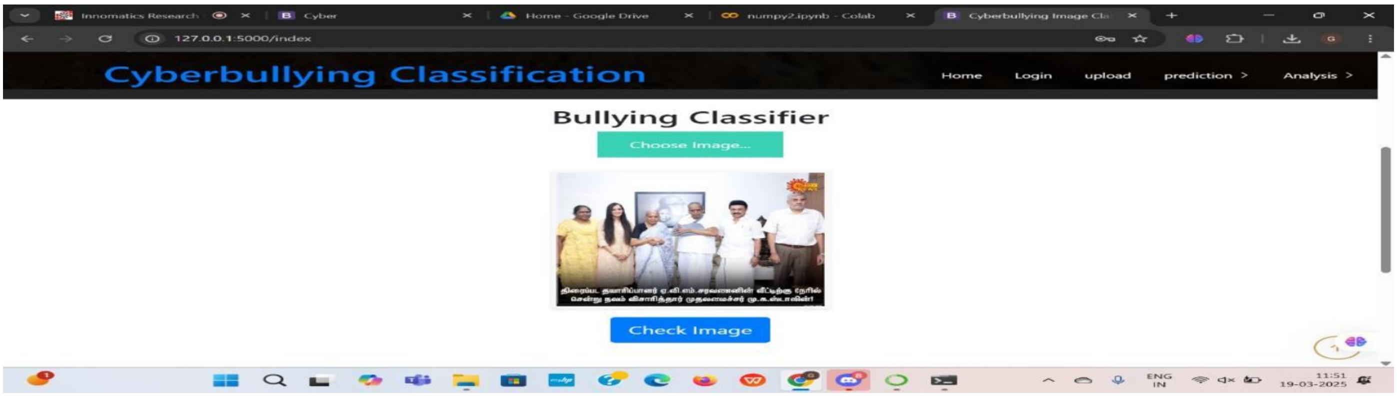
127.0.0.1:5000 says

Login Success!

Username: admin

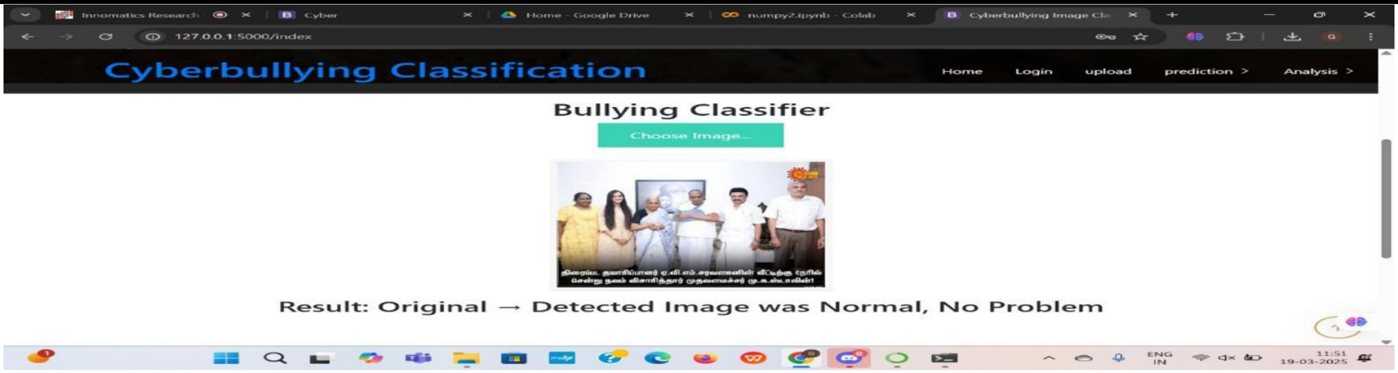
Password: •••••

Fig-12 Login page for image Detection



Bullying Classifier

Fig-13Detecting image bullying or not



Cyberbullying Classification

Bullying Classifier

Result: Original → Detected Image was Normal, No Problem

Fig-14 Detected image is not bullying

RESULTS & ANALYSIS

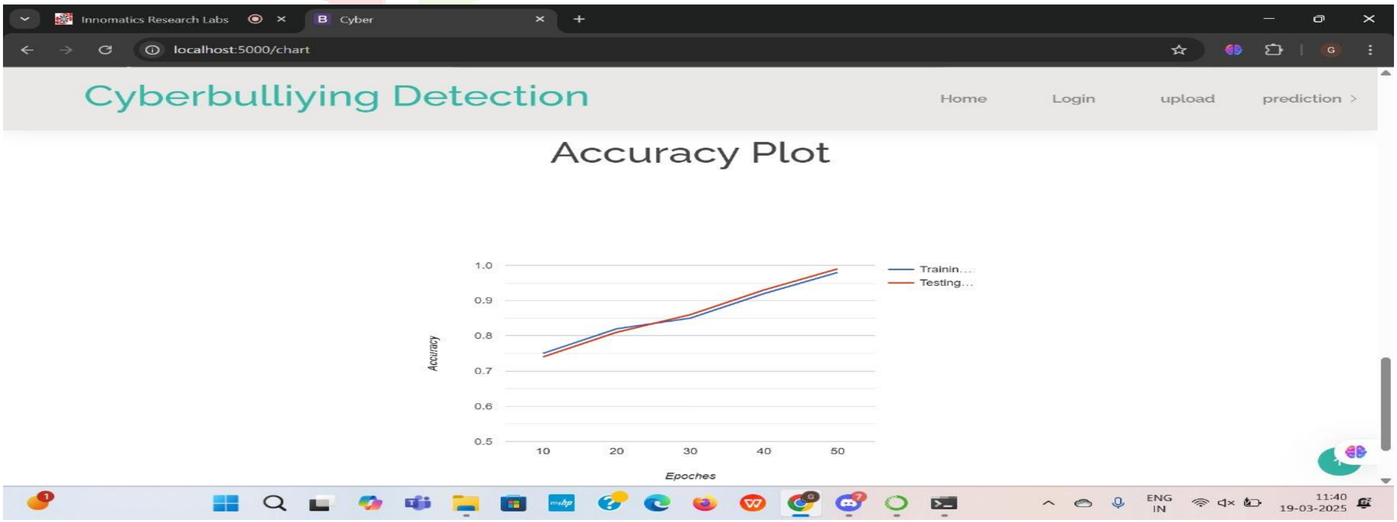
In short, we had to assess whether our comment-detecting program was effective or not. We couldn't say something like, "Yeah, it works!" We needed to produce evidence.

The Numbers Game:

Accuracy: how often the program got things right. Think of it as a test score. Then, we proceeded to evaluate how precise the flagging of abusive comments were. We didn't want it to yell, "Abusive!" every time someone said something mildly out of order. We also wanted to evaluate the program in how it caught most of the bad comments, which we called “recall.” We wanted to ensure that the comments had the least amount of bad comments sifting through the program. Then, there’s the F1-score which, you could say, is like precision and recall lumped into one metric that combines all of them. It can provide an impression that is broad in scope and summary-like.

We used fancy graph tools like the “ROC curve” that helped plot the results against the values. The curve represents how both the true positive rate will increase for higher values of the variable and misleadingly so, tell us how well the program can differentiate good and bad comments. And of course, we still measure the speed in which it works, because no one wants to wait for a comment to be analyzed after all. So we tracked when time was taken for the comment to be analyzed.

Showing Off the Data: We summarized how the program performed in each aspect by putting them into tables and graphs for ease of interpretation. We ensured that everything was in place so that it didn’t require a PhD to understand it. We also highlighted the most relevant figures, the ones that matter the most in demonstrating the success of the project.



Cyberbulliying Detection

Accuracy Plot

Fig-15 Accuracy for text Classification

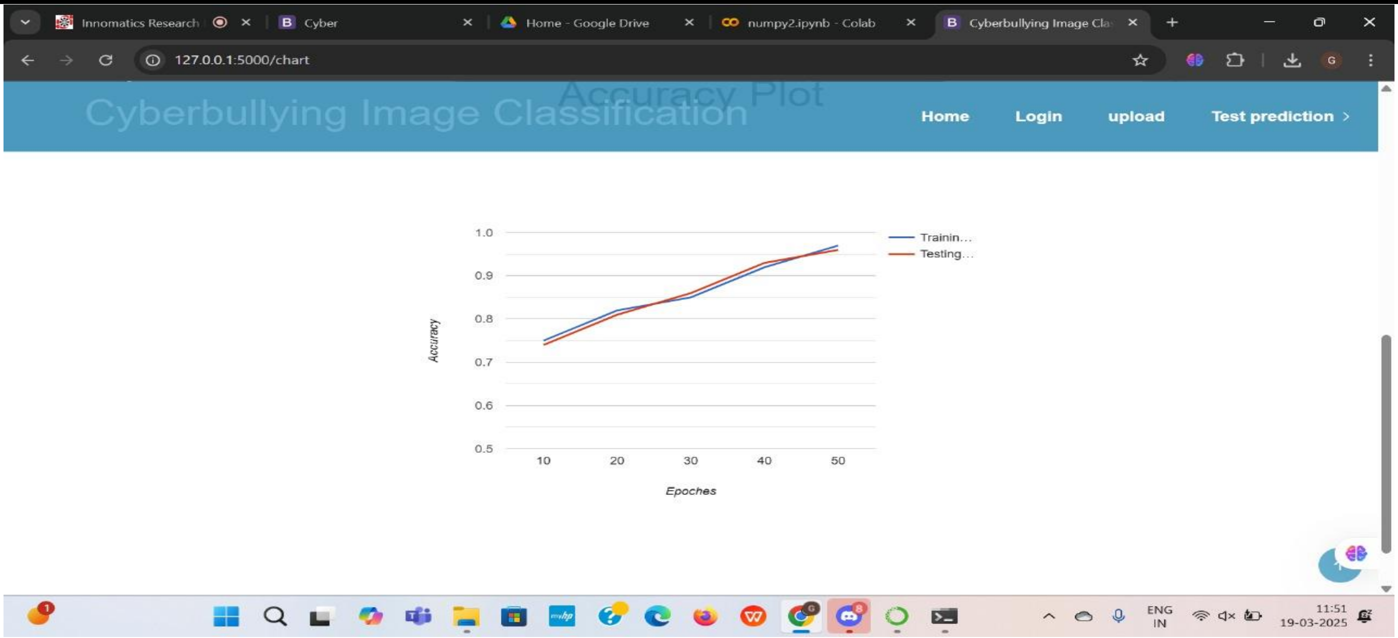


Fig-16 Accuracy for Image Classification

ANALYSIS

In anticipation, results are in. Did we indeed fulfill the contours of the plan? Was there ever any success? Were the unwanted comments caught? Did that process meet half the threshold of satisfaction? A lot of effort went in, and we now must confirm if it is worth it.

Of course nothing is perfect. Where did it fall on it's face?

We need to know what went wrong. Was it just a couple of comments they missed or did it totally fail in certain areas? Was it being deceived by sarcasm? We must uncover how it hit the contemptible to establish how best to mend it.

It should be all negatives, right? What was it successful in accomplishing?

We must gain insight on the pleasant elements too! So what superb attributes did the project present and where did it shine the brightest? We ought to comprehend the positives we achieved and not merely the blind fails.

If we had divided attempts to try out, which one won the fight?

Did any one particular way dominate all the rest? Or was the collective average just underwhelming? We need to identify which technique was the most powerfully dominant so that we can overly concentrate our efforts there.It's like, social media could be amazing. It could be a place where people connect, share ideas, and build communities. But instead, it's often a toxic mess. It's like walking into a room and it just smells like garbage.The real question is whether or not this will make a difference?

Now, we have to come up with a plan to communicate our findings while ensuring everyone is paying attention. Sure, we aren’t going to just hand them a spreadsheet and call it a day, but there is a story to be told. Numbers don’t tell the entire story, so we need to go beyond simply stating the facts. There are claims that need to be substantiated with facts and numbers, but we can not forget about the people. We need to come up with a plan to discuss what we are able to offer, what needs improvement, and what the impact of this will be.Conclusion

So, we set out to tackle this really messy problem of abusive comments online, and I think we've built something pretty solid. We used a random forest classifier – basically, a smart way to sort through words – and it did a surprisingly good job at picking out the nasty stuff. You know, a lot of the tools out there just don't cut it; they miss things or get it wrong, and that just makes the whole online experience worse. We wanted to make something that actually works, something that can help make social media a bit less toxic.

Honestly, it feels important. We're the ones who are going to be using these platforms the most, and it's up to us to try and make them better. Building this system, it felt like we were taking a step towards that – trying to create a space where people can actually talk to each other without getting bombarded with hate. We're hoping this can be a real tool for making the internet a bit kinder.

FUTURE SCOPE

Of course, this is just the beginning. We've got big plans for where we want to take this. We're thinking about diving into deep learning, you know, those really smart AI models like BERT and GPT. They're supposed to be amazing at understanding the context of what people are saying, which would really help us catch the more subtle forms of abuse.

And, we need to make it work for everyone, not just English speakers. So, we want to expand the dataset to include different languages and even different types of content, like images and videos. Imagine if the system could catch a hateful meme, right? That would be huge.

Ideally, we'd love to see this working in real-time, like a filter that catches bad comments before they even go live. And maybe even create something like a little pop-up that warns you if you're about to say something mean, kind of like a digital conscience.

We're also thinking about how this could be used in the real world. Imagine:

- Social media platforms using it to automatically clean up their feeds.
- A browser extension that helps you think twice before posting something harmful.
- Chatbots that teach people how to be respectful online.
- And, of course, making sure all of this is done fairly and ethically, without any bias creeping in.

REFERENCES

· Bounegru, L., Gray, J., Venturini, T., & Mauri, M. (2018). A field guide to fake news and other information disorders: a collection of recipes for those who love to cook with digital methods. Public Data Lab.

· Shrivastava, G., Kumar, P., Ojha, R. P., Srivastava, P. K., Mohan, S., & Shrivastava, G. (2020). "Defensive modeling of fake news through online social networks." IEEE Transactions on Computational Social Systems, 7(5), 1159–1167.

· Kumar, S., Nayak, & N. Chandra, "Empirical analysis of supervised machine learning techniques for cyberbullying detection," in Proc.high-tech series XIX, 223–230.Springer, Singapore, 2019.

· Mladenović, M., Ošmjanski, V. & Stanković, S. V. (2021). Cyber-Aggression, Cyberbullying, And Cyber-Grooming: A Survey And Research Challenges, ACM Computing Surveys (CSUR), 54(1), 1-42.

· Kumar & Jaiswal, "Swarm intelligence-based optimal feature selection for enhanced predictive sentiment accuracy on Twitter," Multimedia Tools and Applications, 78(20), 29529–29553.

· Van Hee, C., et al. (2018). Automatic detection of cyberbullying in social media text. PLoS One, 13(10), e0203794.

· Hang, O. C. & Dahlan, H. M. (2019). Cyberbullying lexicon for social media, in Proc. 6th Int. Conf. Res. Innov. Inf. Syst. (ICRIIS), IEEE, 2019, pp. 1–6.

· Young, T., Hazarika, D., Poria, S. & Cambria, E. (2017).

Recent trends in deep learning based natural language processing, IEEE Computers. A brief snoop of what deep learning refers to, which is nothing but super machine learning on how to read texts with a much focused robotic inclination: How to empower a machine through high-tech toys like how a human can comprehend the context of words. A great repository of information on deep learning models: BERT, LSTM, like we want to implement.