Transdisciplinary Approach Using Ensemble Learning Model to Predict and Abort Cyber-Abuse Against Women

Vanathi Selvaraj[1*], Suriyan Karuppiah[2]

[1]Research Scholar, Center for Study of Social Exclusion and Inclusive Policy, Bharathidasan University, Tiruchirappalli, India.

[2]Professor & Director, Center for Study of Social Exclusion and Inclusive Policy, Bharathidasan University, Tiruchirappalli, India. Email: dr.suriyan@rediffmail.com

Correspondence: vanathiselvaraj01@gmail.com

Received 21 June, 2023; Revised 6 August, 2023; Accepted 26 August, 2023
Available online: 26 August, 2023 at www.atlas-tjes.org, doi: 10.22545/2023/00233

Abstract: *Cyber-abuse against women has become a pervasive issue in today's digital age, posing serious threats to their safety and well-being. To combat this problem, proposed an ensemble learning model approach in machine learning technology that aims to detect and abort instances of cyber-abuse targeting women on online platforms. The proposed model combines the machine learning techniques of both Support Vector Machine (SVM) and Random Forest (RF) algorithms to enhance performance and generalization. An ensemble model is constructed by combining the predictions of individual base models using aggregation techniques such as LightGBM (Light Gradient Machines). The trained ensemble learning model is integrated into a real-time monitoring system that continuously analyses social media content. This system identifies and flags potentially abusive or harassing content directed towards women. Combining technological advancements, human expertise, and community engagement, our ensemble learning model approach offers a comprehensive solution to predict and prevent cyber abuse against women, fostering safer and more inclusive online spaces. Additionally, our course emphasizes the importance of jurisdictional considerations and punitive measures implemented in different jurisdictions.*

Keywords: Transdisciplinary, Ensemble Learning, Machine Learning, Cyber-Abuse, Women, Prediction, Prevention, social media, Online Safety.

1 Introduction

Cyberbullying and harassment are serious problems that can have a devastating impact on victims, especially women. Additionally, there seems to be a rising acknowledgement of the necessity to solve this problem in recent years, and several laws and regulations have been enacted to protect victims. However, there are still challenges in detecting and preventing cyberbullying and harassment and ensuring that victims receive justice. This is where ensemble learning models can play a role. Ensemble learning models are a sort of ML method that may be utilised to increase accuracy by combining the findings of many models. This makes them well-suited for tasks such as cyberbullying and harassment detection, where there is a vast amount of data and various types that can be used to identify abusive behaviour. The model used in this paper has made a transdisciplinary approach such as mathematical mode, algorithmic approach and human behaviour in finding the behaviour of the social media users and predicting and aborting the cyber abuse against women.

Cyberbullying and harassment targeting women on social media platforms have emerged as pressing social issues in recent years. The detrimental impact of such online abuse calls for effective measures to combat and prevent it. With the widespread use of social media platforms, women are often subjected to various forms of cyberbullying and harassment, including derogatory comments, threats, doxing, and revenge porn. These acts cause psychological distress and infringe upon their right to participate freely and safely in online spaces. Hence, there is a vital need to develop sophisticated methodologies to identify and mitigate such harmful behaviours.

According to the National Crime Bureau (NCRB) submitted reports in 2020, the number of women experiencing Cyberbullying has increased by 84 per cent from 28,346 incidents in 2018 to 51,026 incidents in 2020. These incidents include extortion, slander, morphing, fabricated identities, and transferring or posting sexually explicit content, among other things.

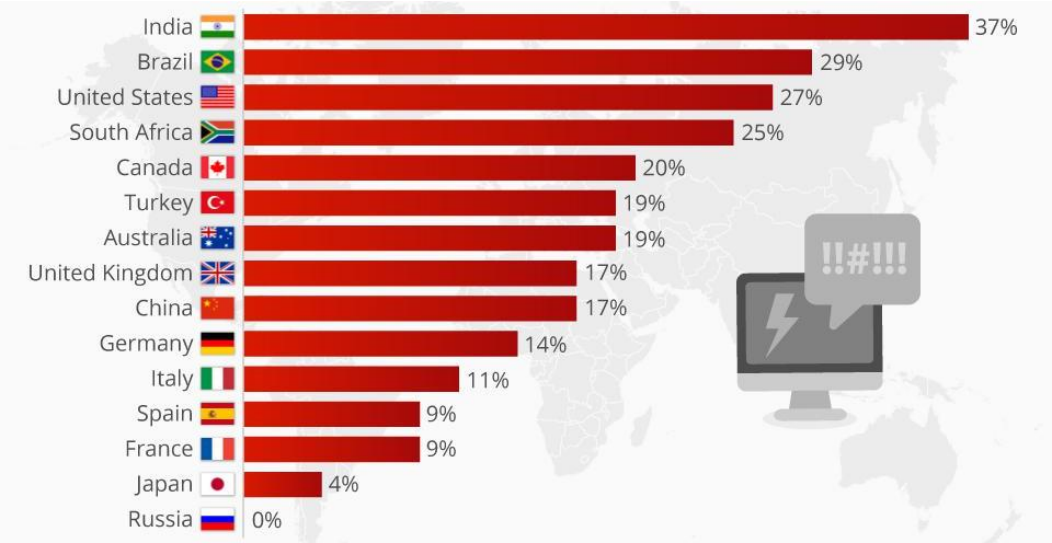

Figure 1. Cyberbullying against women in all over the world [Source: https://ceoworld.biz]

Ensemble learning, which combines the strengths of multiple machine learning models, has proven effective in various domains. By harnessing the collective intelligence of different algorithms,

ensemble models have the prospective to increase accuracy, enhance simplification, and tackle the inherent complexities of cyberbullying and harassment detection. Integrating ensemble learning with the analysis of legal frameworks and punitive measures ensures a holistic approach towards addressing these issues.

In parallel, it is crucial to understand the legal landscape surrounding cyberbullying and harassment. Laws and regulations differ across jurisdictions, and punishments for offenders may vary. By exploring the legal dimensions, can illuminate the alignment between the ensemble model's predictions and the legal consequences imposed on perpetrators. This exploration helps comprehend the effectiveness of existing laws, identify gaps, and support legal reforms in this area.

As per the National Crime Records Bureau reports, cyber-abuse against women on social media has literally increased when compared to previous years. Overall incidents of cyber-abuse against women increased by 16.2% in 2022, according to the most recent data from the National Crime Records Bureau (NCRB), India.(NCRB, 2018)

1.1 Challenges

Developing an ensemble learning model to predict and prevent cyber-abuse against women comes with its fair share of challenges. Some of the key challenges in this endeavor include:

Limited and biased data: Obtaining a comprehensive and diverse dataset of cyber-abuse incidents involving women can be challenging. Data availability and biases in the collected data may affect the generalizability and effectiveness of the ensemble model.

Complex and evolving nature of cyber-abuse: Cyber-abuse against women encompasses a wide range of behaviors, making it challenging to define and identify consistent patterns. Moreover, the landscape of cyber-abuse is constantly evolving, with new techniques and platforms emerging, requiring the model to adapt and remain up to date.

Balancing accuracy and false positives: Striking a balance between accurately predicting cyber-abuse incidents and minimizing false positives is crucial. The model should be capable of accurately identifying abusive cases without flagging non-abusive content, to avoid unnecessary alarm or burden on resources.

1.2 Motivation

The motivation behind developing an ensemble learning model to predict and prevent cyber-abuse against women stems from the urgent need to address the growing problem of online harassment and abuse targeting women.

- By developing an effective prediction and prevention model, we aim to safeguard women from online harm and provide them with a safer digital environment.
- Equipping individuals, organizations, and law enforcement agencies with a powerful tool to identify and combat cyber-abuse enables them to take proactive measures, intervene promptly, and potentially prevent abusive incidents from escalating.
- By actively addressing and mitigating online abuse against women, we strive to create a digital space that is inclusive, respectful, and safe for all individuals.

1.3 Objective

The objective of this research is to develop an ensemble learning model that effectively predicts and prevents cyber-abuse against women. Creating an ensemble learning model that combines the strengths of SVM and RF algorithms to improve prediction accuracy and robustness. Conducting rigorous training and validation of the ensemble model with LightGBM using appropriate techniques such as cross-validation and hyperparameter tuning to optimize its performance.

1.4 Our Contribution

This research aims to significantly contribute to combating cyber-abuse against women by leveraging the power of ensemble learning models. Accurately predicting and aborting cyber abuse instances can create safer online spaces for women, protect their rights, and foster an inclusive digital environment that promotes equality and respect. LightGBM algorithm will contribute its unique strengths in capturing different aspects of cyber abuse, leading to a more comprehensive and accurate prediction system.

2 Related Works

The conversion of the students' classrooms to an online environment resulted in a growth in the amount of time spent on social media by the students (Sultan et al., 2023). A large number of youngsters have become victims of cyber harassment, which contains making frightening remarks about young pupils, engaging in sexually torturous behaviour via a digital platform, individuals disparaging each other, and the use of phoney identities to harass others (Abarna et al., 2022). Compared to monolingual situations, incidents involving code-switched users are more difficult to identify as instances of online bullying (Paul et al., 2023). A new code-switched data set by collecting tweets and annotating them with binary labelling as the initial step to allowing the development of methods for detecting cyberbullying. This dataset was acquired from tweets on Twitter.

The application of social media as a weapon to provoke communal violence, promote erroneous propaganda, undermine social cohesion, and denigrate the identity of people or a group in public areas has occurred on several occasions (Sharif & Hoque, 2022). Determine, on an automated basis, the many participant roles involved in textual traces of cyberbullying, including those of bullies, victims, and onlookers (Bai & Malempati, 2023). It describes the building of two cyberbullying corpora, both of which had been manually interpreted with bullying categories and applicant roles, and it goes on to say that run a series of multiclass classification tests to establish whether or not it is possible to identify participating roles in text-based cyberbullying (Jacobs et al., 2022).

The harmful influence on human life is caused by aggressive social media remarks. These kinds of objectionable materials are to blame for depression and other behaviours associated with suicide ideation (Akter et al., 2023). The volume of information promoting hatred is growing in tandem with the expansion of online social networking platforms (Herry & Mulvey, 2022). The most well-known approach to the construction of the cyberstalking recognition model is known as machine learning. Researchers have proposed several recognition processes that use machine learning to manage and combat cyberstalking in web-based media (Gautam & Bansal, 2022).

In this study (Elmezain et al., 2022), the authors present a hybrid model for classifying their own data set photos built on transformer models combined with SVM. The ability to automatically

identify dangerous information that may be found online is a crucial problem for social media sites, governments, and society (Sharma et al., 2022). Another thing that has been seen is that memes may spread all over the world by being repackaged in a variety of languages, and they can also be multilingual and mix together diverse cultures (Hollis, 2021).

It has been brought to light that cyberbullying may be broken down into two distinct categories: the first involves offensive activities that are carried out virtually and are intended directly at a victim, while the second involves actions that are more likely to be carried out indirectly and include bystanders (Guidi et al., 2022). The difficulty of identifying hazardous behaviour in online conversations is the primary subject of this paper (Machová et al., 2022). When individuals are heavily affected by the viewpoints expressed by others on social networks, there is a significant risk of toxicity (Azumah et al., 2023). Enhancement of public health and safety policies via a concentration on expanding one's understanding of cybercrime, women's victimisation, and the pattern of time spent on the internet, as well as sexual harassment and cyberbullying, and the influence of sociodemographic characteristics on cybercrime (Anjum, 2020).

The combination of extensive use of social media with the ability to remain anonymous makes the process of social learning of cyberbullying in social media easier, making cyberbullying more likely (Lowry et al., 2016). This article (Pittaro, 2007) provides a look into the deviant behaviours and strategies that are linked with cyber stalking crimes, as well as legal intervention measures and preventive activities that have been expressly established to cut down on this rising worldwide crime.

This article (Rajbhandari & Rana, 2023) investigated individual coping techniques such as sharing, ignoring, and improving one's self-efficacy to manage technology in a strong and confident manner. It concluded with the implications of collaborative coordination. This article (Shweta Sankhwar and Arvind Chaturvedi, 2018) sheds light on cybercrime and the legal intervention measures that are being taken to combat it. In summary, A paradigm is presented, and within it, specific preventive actions are detailed, especially at reducing instances of cybercrime committed against women and children.

3 Cyber Bullying and Harassment for Women in Social Media

Cyberbullying and harassment targeting women in social media have become pervasive issues that demand urgent attention. The rise of social media platforms has provided a powerful means for communication and interaction, but it has also opened the door to harmful behaviours that disproportionately affect women. In this discussion, the concerning trends of cyberbullying and harassment faced by women in social media and the implications it has on their well-being, safety, and participation in the digital world will explore.

The advent of social media has created virtual spaces where individuals can connect, communicate thoughts and interact with individuals. It has, nevertheless, given birth to negative behaviors that manifest as cyberbullying and harassment. Women are particularly vulnerable to these acts, facing various forms of abuse such as online threats, hate speech, objectification, slut-shaming, and body shaming. The pervasive nature of social media allows these acts to quickly spread, leading to significant emotional distress, mental health issues, and even physical harm for the targeted women.

The impact of cyberbullying and harassment on women extends beyond their individual experiences. It affects their ability to express themselves freely, participate in public discourse, and feel safe within online communities. These behaviors not only perpetuate gender inequality but also create a hostile environment that hinders the progress towards a more inclusive and equal society.

Efforts to combat cyberbullying and harassment in social media require a multi-faceted approach. Technological interventions, such as content moderation algorithms and reporting mechanisms, play a crucial role in mitigating harmful content. However, the complex and ever-evolving nature of these behaviors calls for continuous research and innovation in developing effective solutions.

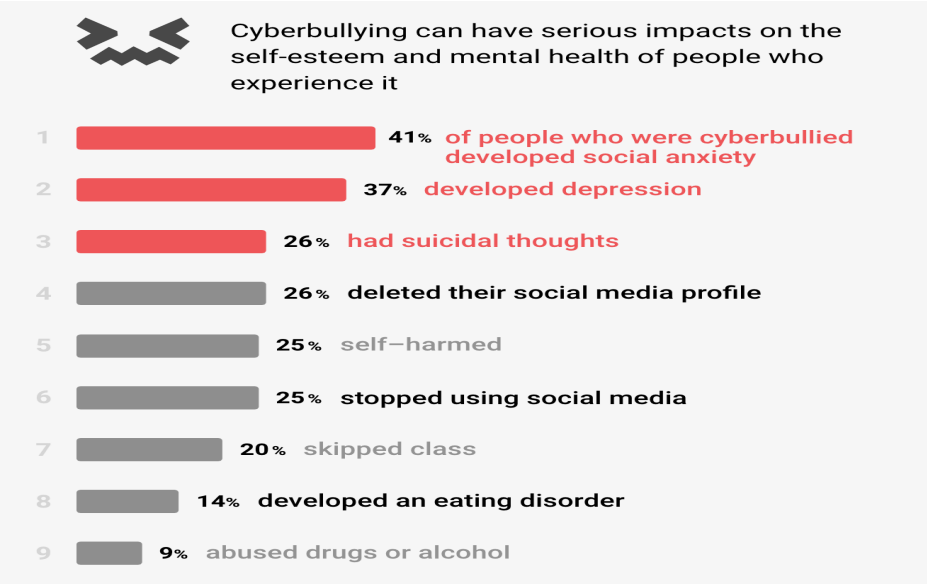

Figure 2. Impact of Cyberbullying in Social Media against Women
[source: https://www.ditchthelabel.org]

Legal frameworks also play a vital role in addressing cyberbullying and harassment. Many countries have enacted laws and regulations to hold perpetrators accountable and provide legal remedies for victims. However, challenges exist in effectively enforcing these laws, given the global and borderless nature of social media platforms.

Educational initiatives that promote digital literacy and raise awareness about the consequences of cyberbullying and harassment are equally important. By equipping individuals with the necessary knowledge and skills to navigate social media responsibly, foster a culture of respect, empathy, and digital citizenship.

4 Ensemble Learning Model for Text-based Classification

The objective of this research is to develop an ensemble learning model specifically tailored for predicting and aborting cyber-abuse against women. The model will be trained on a comprehensive and representative dataset that encompasses various forms of abusive instances and the overall structure of proposed model in shown in fig 3. By incorporating features such as linguistic patterns,

user-related information, network characteristics, and temporal dynamics, the ensemble model will learn to recognize and differentiate between abusive and non-abusive content.

Cyberbullying and harassment targeting women in online spaces have become pervasive and concerning issues. Detecting and combating such harmful behavior is crucial to ensure the safety and well-being of individuals in the digital realm. In this context, ensemble learning models combined with text-based classification techniques offer a powerful approach to identify instances of cyberbullying and harassment targeting women. In this discussion, explore the application of ensemble learning models in text-based classification for addressing cyberbullying and harassment against women.

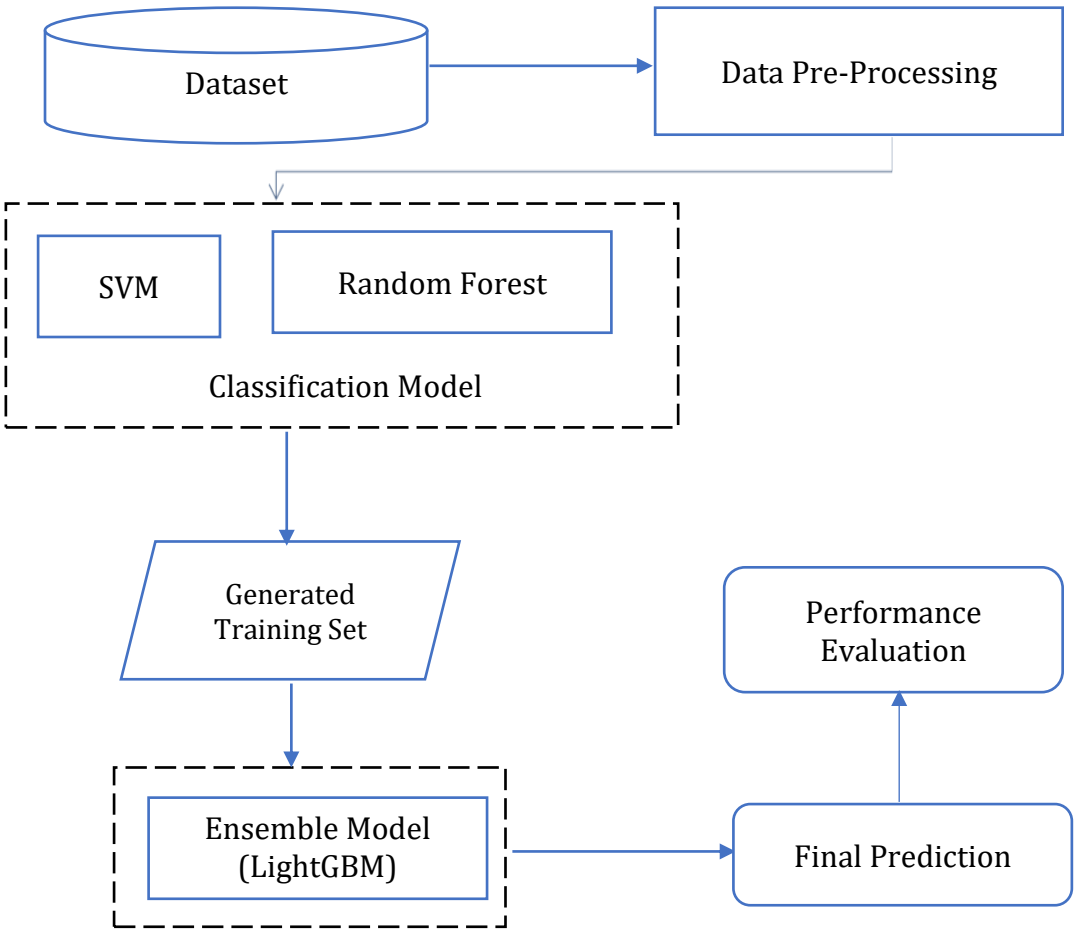

Figure 3. Overall Cyberbullying Prediction Process.

To build an ensemble learning model for cyberbullying and harassment detection, two base models can be employed, such as SVM, RFs. Each base model brings unique capabilities for capturing different linguistic features and contextual information, enhancing the overall performance of the ensemble. Finally, LightGBM Ensemble classifier has combined with our framework that is known for its high speed and low memory usage. It uses a novel tree-growing algorithm and supports parallel learning, making it particularly suitable for large-scale datasets.

4.1 Data Pre-processing

Data pre-processing is an important step in analyzing cyberbullying and women harassment in social media. It involves cleaning and preparing the collected data to ensure accurate and meaningful analysis as shown in fig 4.

Text Cleaning:

- Remove any unnecessary characters, symbols, or special characters that they were do not help to the assessment.
- To guarantee uniformity, convert everything to lowercase.
- Remove URLs, mentions, or any other user-specific information that may not be relevant to the analysis.
- Handle abbreviations, acronyms, or slang words by expanding or normalizing them for better understanding.

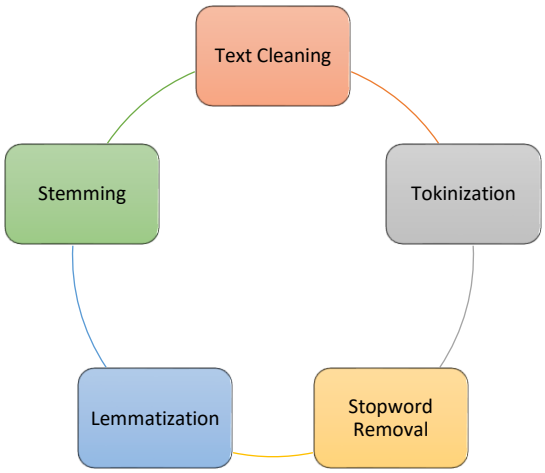

Figure 4. Data Pre-processing Steps

Tokenization:

- Separate the written content into token or sentences. This can be done using techniques like whitespace tokenization or more advanced methods like word tokenization using libraries like NLTK or spaCy.

Stopword Removal:

- Remove common stopwords (e.g., "the," "and," "is") that do not have a substantial bearing on what they signify in the analysis. These can be obtained from libraries like NLTK or spaCy, or you can create a custom list based on your specific context.

Lemmatization or Stemming:

- Consolidate variants of the same term by reducing them to their foundation or root form. This can be achieved through lemmatization or stemming techniques.

- Lemmatization maps words to their dictionary form, while stemming applies a set of rules to remove prefixes or suffixes.

Pseudocode for Data pre-processing

```
Input: Raw data collected from social media
def preprocess_data(data, stopwords, additional_stopwords):
    preprocessed_data = []
        for text in data:
        cleaned_text = clean_text(text) # Clean text
         tokens = tokenize_text(cleaned_text)       # Tokenization
        tokens = remove_stopwords(tokens, stopwords) # Stopword removal
        tokens = remove_stopwords(tokens, additional_stopwords)
tokens = apply_lemmatization(tokens) # Lemmatization or stemming
         tokens = apply_stemming(tokens)
        preprocessed_text = ' '.join(tokens)
        preprocessed_data.append(preprocessed_text)
    return preprocessed_data
preprocessed_data = preprocess_data(data, stopwords, additional_stopwords)
```

In this pseudocode, the preprocess_data() function takes the raw data collected from social media (data), a list of stopwords (stopwords), and additional stopwords specific to the analysis (additional_stopwords) as input. It iterates over each text in the data, applies various pre-processing steps (vacuuming, tokenization, Stopword removal, lemmatization or stemming, handling negations or emphasis, handling emoji and emoticons), and returns a list of preprocessed texts.

4.2 Classification Models

4.2.1 SVM Model

SVM classification aims to identify a hyperplane that divides the data values into two groups. The hyperplane is chosen so that the distance between it and the data elements immediately next to it on each side is minimized to the greatest extent feasible. This part of the page is called the margin. The data points that are located in the closest proximity to the hyperplane are known as support vectors.

The mathematical model for SVM entails constructing an optimization problem to identify the best hyperplane that distinguishes data points from distinct classes by the greatest margin. Here's a summary of the key equations and steps involved:

Given a labeled training dataset is given in eq 1 : $\{ (x_1, y_1), (x_2, y_2), \ldots, (xn, yn) \}$ (1)

where x_i represents a feature vector and y_i represents the corresponding class label (-1 or +1).

Objective function: SVM aims to find a hyperplane with weights w and bias b It maximizes the difference from the two groups while minimizing the classification error. This can be formulated as an optimization problem is given in eq 2 and 3:

$$minimize: ½ * ||w||^2 + C * \Sigma\xi_i \quad (2)$$

$$subject\ to: y_i * (w \cdot x_i + b) \geq 1 - \xi_i\ for\ all\ i,$$

$$and\ \xi_i \geq 0\ for\ all\ i. \quad (3)$$

The first term in the objective function represents the margin regularization, and the second term represents the classification error regularization. C is the regularization parameter that determines how much of an emphasis is placed on either margin maximization and classification error.

Dual problem: The optimization problem is often converted into its dual form, which simplifies the solution and allows for the use of the kernel trick to handle nonlinearly separable data. The dual problem involves maximizing the following equation 4 and 5:

$$maximize: \Sigma\alpha_i - ½ * \Sigma\Sigma\ \alpha_i * \alpha n * y_i * yn * (x_i \cdot xn) \quad (4)$$

$$subject\ to: \Sigma\ \alpha_i * y_i = 0, and\ 0 \leq \alpha_i \leq C\ for\ all\ i. \quad (5)$$

The α_i's are the Lagrange multipliers associated with the training samples, and they represent the weights assigned to the support vectors.

Decision function: Once the optimal Lagrange multipliers (α_i) are obtained, the classification decision value for a new data point x may be computed as follows in eq 6:

$$f(x) = sign(\Sigma\ \alpha_i * y_i * K(x_i, x) + b) \quad (6)$$

where $K(x_i, x)$ is the function of kernel that estimates the differences and similarities between x_i and x. Kernel functions that are often used includes linear, polynomial, RBF, and sigmoid.

Support vectors: The support vectors are the training samples that lie on or within the margin boundaries. They are the data points that have non-zero Lagrange multipliers ($\alpha_i > 0$).

These equations summarize the key mathematical components of SVM. Solving the optimization problem involves techniques such as quadratic programming or convex optimization algorithms. The high-quality of the kernel function and tuning of the regularize parameter (C) significantly impact the performance of SVM.

4.2.2 Random Forest Model

A random forest model is a sort of ensemble learning model that makes predictions using decision trees. It is a method of learning under supervision, and it may be utilized in classification in addition to regression analysis. To create a forecast, the random forest model averages the outputs of each decision tree in the forest. The class with the greatest average prediction is the class predicted by the model. While doing regression tasks, the mean or averaged prediction of each of the different trees is presented.

Each Random Forest decision tree is created individually on a random subset of the training data and characteristics. Here's a summary of the mathematical model for Random Forest:

Given a labeled training dataset is given in eq 7:

$$\{ (x_1, y_1), (x_2, y_2), \ldots, (xn, yn) \} \quad (7)$$

where x_i represents a feature vector and y_i represents the corresponding class label or target value.

where xn represents a feature vector and yn represents the corresponding class label or target value.

Pseudocode for random forest algorithm

```
def train_random_forest(X_train, y_train, num_trees, max_depth):
    forest = []
   for _ in range(num_trees):
        tree = build_decision_tree(X_train, y_train, max_depth)
        forest.append(tree)
        return forest
def build_decision_tree(X, y, max_depth):
    tree = DecisionTree()
    tree.build(X, y, max_depth)
        return tree
def predict_random_forest(forest, X_test):
    predictions = []
   for tree in forest:
   tree_predictions = tree.predict(X_test)
predictions.append(tree_predictions)
    final_predictions = majority_voting(predictions)  # Or averaging(predictions)
return final_predictions
random_forest = train_random_forest(X_train, y_train, num_trees, max_depth)
random_forest_predictions = predict_random_forest(random_forest, X_test)
```

In this simplified pseudocode, the train_random_forest() function takes the training features (X_train), training labels (y_train), the amount of trees present in the random forest (num_trees), and the greatest possible depth of every decision tree (max_depth) as input. It creates a random forest (forest) and trains each decision tree by calling the build_decision_tree() function.

The build_decision_tree() function creates a decision tree object (tree) and recursively builds the tree by calling its build() method with the training data and maximum depth.

The predict_random_forest() function takes the trained random forest (forest) and test features (X_test) as input. It makes predictions using each decision tree in the forest by calling each tree's predict() method. The individual predictions are then combined using majority polling or averaging to attain the final calculations.

4.2.3 LightGBM Model

LightGBM is based on the gradient boosting algorithm, which combines multiple learners to create a strong predictive model. The mathematical equations for LightGBM involve the loss function, gradients, and the update rule for the tree construction.

Loss Function (L): The loss function quantifies the variation among the expected values and the actual labels. Frequent loss functions in categorization included binary cross-entropy and multi-class cross-entropy. The function of loss is typically defined as in eq 8:

$$L = \Sigma(y_i, p_i) + \Omega(f) \quad (8)$$

where y_i represents the true label, p_i represents the predicted probability or score, and Ω(f) denotes the regularization term associated with the ensemble of trees.

Gradients (G) and Hessians (H): The gradients and Hessians are the first and second products of the loss function w.r.t. the predicted values. For binary classification, the gradients (G) and Hessians (H) can be calculated as in eq 9 and 10:

$$G\_i = \partial L / \partial p\_i = p\_i - y\_i \quad (9)$$

$$H\_i = \partial^2 L / \partial p\_i^2 = p\_i * (1 - p\_i) \quad (10)$$

where p_i is the predicted probability for the i-th sample, and y_i is the true label.

Tree Construction: LightGBM constructs decision trees in a leaf-wise manner, selecting the split points that maximize the reduction in the loss function. To find the optimal split, the following quantities are calculated as in eq 11:

$$Gain = 0.5 * (G\_L^2 / (H\_L + \lambda) + G\_R^2 / (H\_R + \lambda) - (G\_L + G\_R)^2 / (H\_L + H\_R + \lambda)) - \gamma \quad (11)$$

where G_L and G_R represent the gradients of the left and right child nodes, H_L and H_R represent the Hessians of the left and right child nodes, λ is the regularization term, and γ is the parameter controlling the minimum gain for a split.

Update Rule: The update rule is used to adjust the predicted values based on the gradients and the learning rate (η). For each tree, the predicted value (p_i) is updated as in eq 12:

$$p\_i \leftarrow p\_i - \eta * \sum(G\_j) / (\sum(H\_j) + \lambda) \quad (12)$$

where G_j and H_j are the gradients and Hessians for the samples in the leaf node j.

These equations summarize the key mathematical components of LightGBM. However, it's important to note that LightGBM employs additional optimizations and techniques, such as feature bundling and categorical feature handling, to improve efficiency and accuracy. The exact

implementation and equations may vary depending on the specific version and enhancements of LightGBM.

Pseudocode for LightGBM

```
Input: X_train: Training features; y_train: Training labels; params: Parameters for LightGBM model
def train_lightgbm(X_train, y_train, params):
    lgb_train = lgb.Dataset(X_train, label=y_train)    # Create LightGBM Dataset
    model = lgb.train(params, lgb_train)    # Train the LightGBM model
 return model
model = train_lightgbm(X_train, y_train, params)
```

In this pseudocode, the lgb.Dataset() function is used to create a LightGBM dataset from the training features (X_train) and labels (y_train). The dataset is then used to train the LightGBM model using the lgb.train() function, which takes the parameters (params) and the dataset as input. The trained model is returned as the output.

5. Laws and Punishments

The laws and punishments for cyber bullying and women harassment in social media in India are as follows:

Section 509 of the Indian Penal Code (IPC): This section addresses the act of offending a woman's humility. In addition to possible monetary penalties, the perpetrator of this offence might face incarceration for a period of time ranging from one to three years.

Section 67 of the Information Technology Act, 2000: This section addresses the penalties for posting or sending obscene information through electronic means. This offence carries a potential sentence of a maximum of five years in jail in addition to a monetary punishment.

Section 66A of the Information Technology Act, 2000: This section addresses the penalties for delivering offensive remarks through digital interaction. The offender of this offence might face incarceration for a period of time ranging from one to three years.

The punishment for cyber bullying and women harassment in social media can be severe. The accused can be imprisoned for a term of up to three years and also fined. In addition, the accused can also be held liable for damages to the victim.

In India, there are a number of laws that can be used to punish cyber bullying and women harassment in social media. These laws include:

The Information Technology Act, 2000: This act affords for punishment for cyber-crimes, including cyber bullying and harassment. The punishment for cyber bullying and harassment under the IT Act ranges from Penalty ranges from imprisonment for up to 3 years to a penalty of equal to Rs. 2 lakhs.

The Indian Penal Code, 1860: This act also provides for punishment for a number of offences that can be committed through social media, Defamation, harassment, and obscenity are only a few examples. Under the IPC, the penalty for these crimes may vary from imprisonment for up to 2 years to a penalty of up to Rs. 20,000.

The Protection of Children from Sexual Offences (POCSO) Act, 2012: This statute imposes severe penalties on individuals who commit child sexual neglect or abuse, particularly internet child sexual abuse. According the POCSO Act, the penalty for internet child sexual abuse can range from imprisonment for up to 7 years to a penalty of up to Rs. one lakh.

In addition to these laws, there are also a number of self-regulatory guidelines that have been issued by social media platforms, such as Facebook and Twitter. These guidelines set out the standards of behavior that are expected from users of these platforms, and they also provide for a number of mechanisms for reporting and dealing with incidents of cyber bullying and harassment.

5.1. Preventing Cyber-abuse Against Women

Preventing cyber-abuse against women requires a comprehensive approach that involves various stakeholders, including individuals, communities, online platforms, and policymakers. Here are some strategies and measures that can help in preventing cyber-abuse against women:

Education and awareness: Promote education and awareness about cyber-abuse, its impact on women, and responsible online behavior. Conduct workshops, training sessions, and awareness campaigns to equip women with knowledge and skills to protect themselves and recognize signs of cyber-abuse.

Online safety guidelines: Develop and disseminate clear online safety guidelines that specifically address cyber-abuse against women. These guidelines should include information on privacy settings, secure password practices, recognizing and reporting abusive content, and steps to take if targeted by cyber-abuse.

Stronger legislation: Advocate for stronger legal frameworks and policies that explicitly address cyber-abuse against women. Encourage governments to enact laws and regulations that protect individuals from online harassment and provide legal recourse for victims.

Reporting mechanisms: Encourage social media platforms and online service providers to implement robust reporting mechanisms that allow users to easily report instances of cyber-abuse. Ensure that these platforms have dedicated teams to handle reports promptly and take appropriate action against perpetrators.

User empowerment: Empower women to take control of their online presence by teaching them about privacy settings, blocking and reporting features, and how to maintain a positive digital footprint. Encourage them to report abusive content and support them throughout the process.

Technological solutions: Invest in developing and deploying advanced technologies, such as machine learning algorithms, natural language processing, and image recognition, to automatically detect and filter out abusive content targeting women. Collaborate with technology companies to integrate such solutions into their platforms.

Community support: Foster a supportive and inclusive online community by promoting respect, empathy, and digital citizenship. Encourage bystanders to intervene when they witness cyber-abuse and discourage the sharing or spreading of abusive content.

Engaging men and boys: Involve men and boys in efforts to prevent cyber-abuse against women. Promote gender equality and challenge harmful stereotypes that contribute to online harassment. Encourage them to be allies and stand against cyber-abuse.

Collaboration and partnerships: Foster collaboration among various stakeholders, including government agencies, NGOs, technology companies, and online communities.

Continuous monitoring and evaluation: Regularly monitor the effectiveness of prevention efforts and adapt strategies as needed. Collect data on the prevalence and nature of cyber-abuse against women to inform policy decisions and interventions.

Remember, preventing cyber-abuse against women is an ongoing effort that requires the collective commitment of individuals, organizations, and society as a whole.

6 Results and Discussions

6.1 Dataset Description

This set of data is a compilation of datasets relating to the automated identification of cyberbullying from various sources. The information comes from many social media sites such as Kaggle, Twitter, Wikipedia Talk pages, and YouTube. The data contains content that has been labelled as bullying or not. The data comprises several forms of cyberbullying such as hate speech, aggressiveness, insults, and toxicity.

6.1.1 Training Model

Consider, Online source tagged over 25,000 Tweets in the dataset. This dataset is compiled and label by Crowd Flower. According to the total number of people who labelled such Twitter posts, each is classified into one of three categories: hate speech, offensive language, or none.

ID	Label	Comment
1	T	@offthis_ way to be a asshole
2	T	They are ugly
3	F	T
4	F	Wooow
5	F	They stink don't even try
6	F	@emilyreeves8 YOU SUCK go step on a lego
7	F	@naildiys ummmm, ok?? i dont know you. no need to be a jerk
8	F	Well.... @emilyreeves8 maybe you should tell people to stop cursing k,K! Thx love u bestieeee
9	F	@naildiys im really confused by me telling people to stop cursing. and i wouldnt be bestiessss with someone who told me to step on a lego, so ya. idk why you find the need to be rude when i was telling someone not to be mean. so ya
10	F	I didn't put that my partner did so booomm and were being scarcastic duh! So boom @emilyreeves8

Figure 5. Comments from twitter

The information from both sources is combined. Here the tweets were divided into two categories: True (T) or False (F) of Hate Speech occurrence as shown in fig 5 and the dataset distribution is given in fig 6.

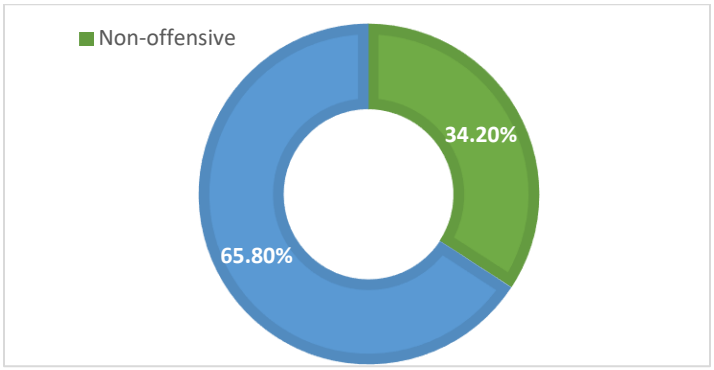

Figure 6. Distribution of Datasets

6.1.2 Testing Model

By using our proposed model, the testing model predicts that the given comment is offensive or non-offensive as shown in fig 7 and 8.

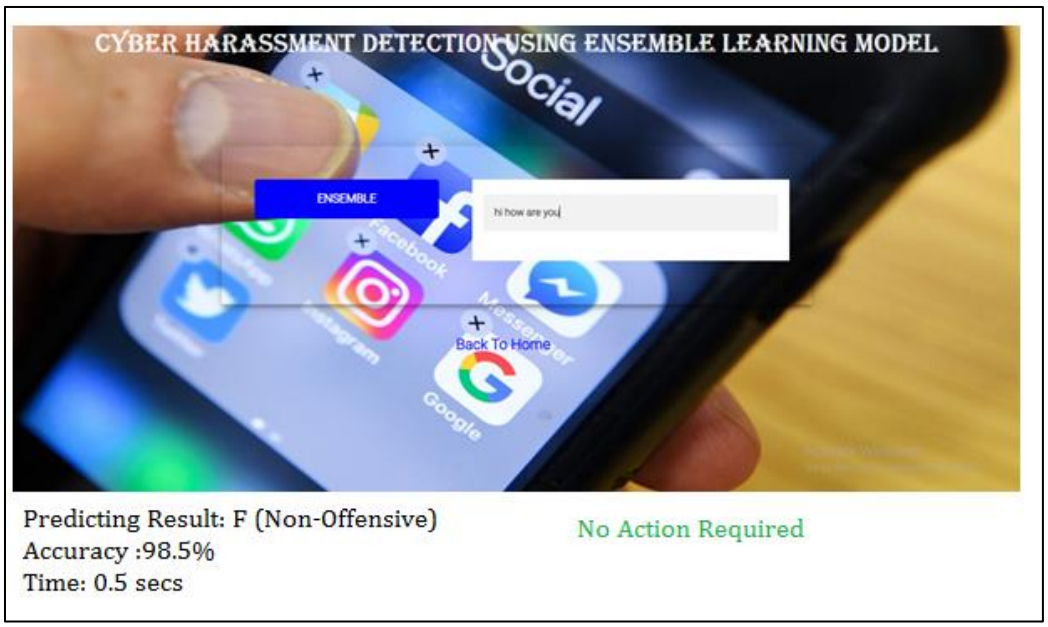

Figure 7. Non-Offensive Message

In the above figure 7 and figure 8, the predicting results have been shown. If the stranger is sending any messages to the women's inbox, immediately the proposed algorithm will read the content and gives a warning about the message sent by the stranger. If that message contains any offensive words, immediately it will alert the user to abort the chat from the stranger. If the message is non-offensive words, then the proposed method will allow the user to continue the chat with the stranger.

Figure 8. Offensive Message

6.2 Classification Metrics

Accuracy: The fraction of accurately anticipated cases in relation to the total number of instances as given in eq 13. Although it may be deceptive, it is a popular statistic for balanced datasets.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (13)$$

Precision: The fraction of real positive assumptions out of all anticipated positives is calculated in eq14. It is useful when the focus is on minimizing false positives.

$$Precision = TP / (TP + FP) \quad (14)$$

Recall (Sensitivity or True Positive Rate): The fraction of genuine positive forecasts among all real positives. It is important when the goal is to minimize false negatives as calculated in eq15.

$$Recall = TP / (TP + FN) \quad (15)$$

F1 Score: Precision and memory are balanced by the harmonic mean. It delivers a balanced metric that incorporates both measurements as given in eq 16.

$$F1\,Score = 2 * (Precision * Recall) / (Precision + Recall) \quad (16)$$

An ensemble algorithm model compares with different algorithms as given in table 1 and graphically represented in fig 8-11. The proposed ensemble algorithm is compared with the other algorithms like Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes, Ensemble (gradient boosting), Ensemble (Xgboosting), Ensemble (LightGBM). The propose system has used the parameters lie Accuracy, precision, recall and F1-Score to obtain the best results.

Table 1: Model Comparisons

Models	Accuracy	Precision	Recall	F1-Score
SVM	94	95	99	87
RF	93	93	99	87
Naïve Bayes	89	89	97	85
Ensemble (gradient boosting)	94.6	95.2	98	89
Ensemble (Xgboosting)	95.4	95.6	99	90
Ensemble (LightGBM)	98.4	98.1	97.8	93

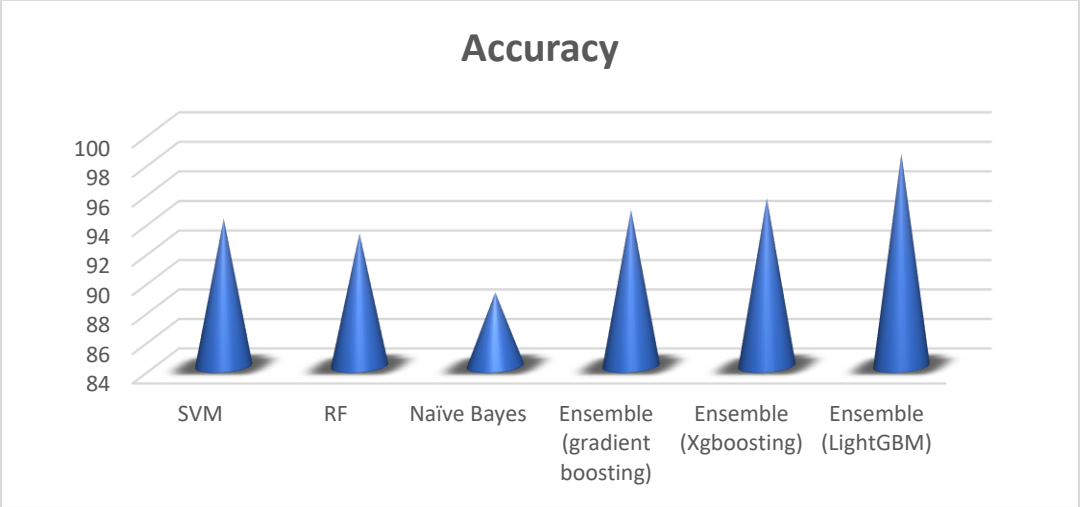

Figure 8. Accuracy Comparison

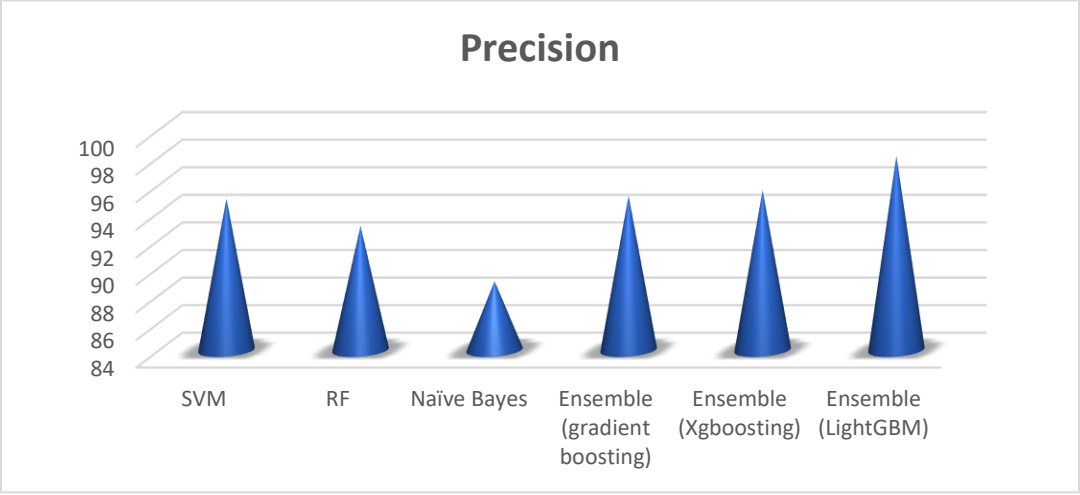

Figure 9. Precision Comparison

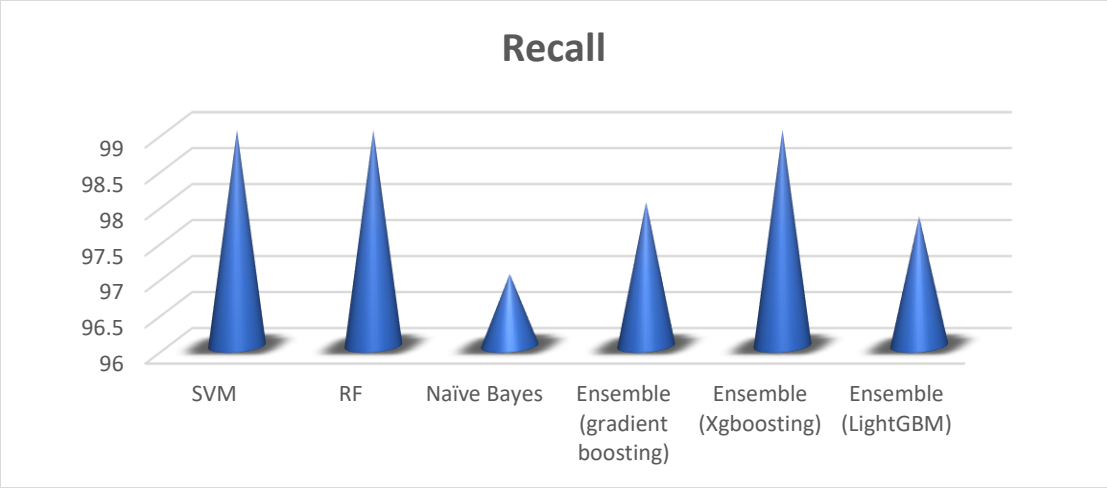

Figure 10. Recall Comparison

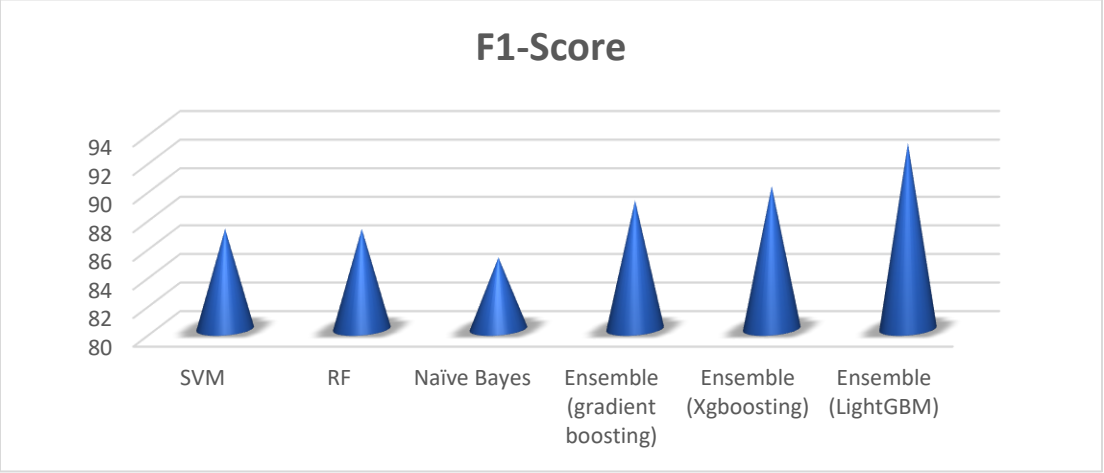

Figure 11. F1-score Comparison

The above results clearly show that our proposed Ensemble learning model performs better than other existing models in cyber-abuse prediction model.

7 Conclusions

Our proposed Ensemble learning models can be valuable for predicting cyberbullying and preventing cyber-abuse against women in online platforms. By harnessing the power of ensemble learning, this approach combines multiple algorithms to enhance the accuracy and robustness of detection systems. By leveraging the combining SVM and RF models, ensemble models can often improve performance and provide more robust predictions. Through the collection and labelling of Crowd Flower datasets, the approach captures various forms of cyber-abuse against women, enabling the model to learn and identify nuanced patterns and abusive content. The trained ensemble learning model provide proactive detection, allowing for timely intervention and prevention of cyber-abuse incidents with accuracy of 98.4%. The integration of human moderation and intervention is crucial to review flagged content and make informed decisions, complementing the outputs of the ensemble learning model. Additionally, a strong emphasis on education and awareness programs fosters a culture of responsible online behavior and gender equality,

empowering women to navigate online spaces safely. Collaboration among stakeholders, including individuals, communities, online platforms, and policymakers, is essential for the success of this approach. By advocating for strong legal frameworks, encouraging online platform responsibility, and providing support networks for victims, people can create a safer digital environment for women.

8 Future Scope

In future, expanding the analysis beyond text-based data to include other modalities, such as images, videos, and audio, can provide a more comprehensive understanding of cyber-abuse incidents. Allowing users to define their own thresholds for abusive content or providing mechanisms to report false positives can enhance the model's performance and address individual user needs.

Authors' Contribution: Authors contributed equally to this work.

Funding Statement: There was no external funding received for the conduction of this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

Abarna, S., Sheeba, J. I., Jayasrilakshmi, S., & Devaneyan, S. P. (2022). Identification of cyber harassment and intention of target users on social media platforms. *Engineering Applications of Artificial Intelligence*, *115*, 105283. https://doi.org/10.1016/j.engappai.2022.105283

Akter, M. S., Shahriar, H., Ahmed, N., & Cuzzocrea, A. (2023). *Deep Learning Approach for Classifying the Aggressive Comments on Social Media: Machine Translated Data Vs Real Life Data*. https://doi.org/10.1109/BigData55660.2022.10020249

Anjum, U. (2020). Cyber Crime In Pakistan; Detection And Punishment Mechanism. *Sted journal*, *2*(2). https://doi.org/10.7251/STED0220029A

Azumah, S. W., Elsayed, N., ElSayed, Z., & Ozer, M. (2023). *Cyberbullying in Text Content Detection: An Analytical Review*. https://doi.org/https://doi.org/10.48550/arXiv.2303.10502

Bai, Z. S., & Malempati, S. (2023). An Ensemble Approach for Cyber Bullying: Text Messages and Images. *Revue d'Intelligence Artificielle*, *37*(1), 179–184. https://doi.org/10.18280/ria.370122

Elmezain, M., Malki, A., And, I. G., & Atlam, E.-S. (2022). Hybrid Deep Learning Model–Based Prediction of Images Related to Cyberbullying. *International Journal of Applied Mathematics and Computer Science*, *32*(2), 171–269.

https://doi.org/https://doi.org/10.34768/amcs-2022-0024

Gautam, A. K., & Bansal, A. (2022). A Review on Cyberstalking Detection Using Machine Learning Techniques: Current Trends and Future Direction. *International Journal of Engineering Trends and Technology*, *70*(3), 95–107. https://doi.org/10.14445/22315381/IJETT-V70I3P211

Guidi, S., Palmitesta, P., Bracci, M., Marchigiani, E., Di Pomponio, I., & Parlangeli, O. (2022). How many cyberbullying(s)? A non-unitary perspective for offensive online behaviours. *PLOS ONE*, *17*(7), e0268838. https://doi.org/10.1371/journal.pone.0268838

Herry, E., & Mulvey, K. L. (2022). Gender-based cyberbullying: Understanding expected bystander behavior online. *Journal of Social Issues*. https://doi.org/10.1111/josi.12503

Hollis, L. P. (2021). Human Resource Perspectives on Workplace Bullying in Higher Education. In *High-Tech Harassment: A Chi-Squared Confirmation that Workplace Cyberbullying Disproportionally Affects People of Color and the LGBQ Community in Higher Education*. Routledge. https://doi.org/10.4324/9781003051923

Jacobs, G., Van Hee, C., & Hoste, V. (2022). Automatic classification of participant roles in cyberbullying: Can we detect victims, bullies, and bystanders in social media text? *Natural Language Engineering*, *28*(2), 141–166. https://doi.org/10.1017/S135132492000056X

Lowry, P. B., Zhang, J., Wang, C., & Siponen, M. (2016). Why Do Adults Engage in Cyberbullying on Social Media? An Integration of Online Disinhibition and Deindividuation Effects with the Social Structure and Social Learning Model. *Information Systems Research*, *27*(4), 962–986. https://doi.org/10.1287/isre.2016.0671

Machová, K., Mach, M., & Adamišín, K. (2022). Machine Learning and Lexicon Approach to Texts Processing in the Detection of Degrees of Toxicity in Online Discussions. *Sensors*, *22*(17), 6468. https://doi.org/10.3390/s22176468

NCRB. (2018). National Crime Records Bureau. *Ministry of Home Affairs*. https://ncrb.gov.in/en

Paul, S., Saha, S., & Singh, J. P. (2023). COVID-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. *Multimedia Tools and Applications*, *82*(6), 8773–8789. https://doi.org/10.1007/s11042-021-11601-9

Pittaro, M. L. (2007). Cyber stalking: An Analysis of Online Harassmentand Intimidation. *International Journal of Cyber Criminology*, *1*(2), 180–197. https://doi.org/https://doi.org/10.5281/zenodo.18794

Rajbhandari, J., & Rana, K. (2023). Cyberbullying on Social Media: an Analysis of Teachers' Unheard Voices and Coping Strategies in Nepal. *International Journal of Bullying Prevention*, *5*(2), 95–107. https://doi.org/10.1007/s42380-022-00121-1

Sharif, O., & Hoque, M. M. (2022). Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, *490*, 462–481. https://doi.org/10.1016/j.neucom.2021.12.022

Sharma, S., Alam, F., Akhtar, M. S., Dimitrov, D., Martino, G. D. S., Firooz, H., Halevy, A., Silvestri, F., Nakov, P., & Chakraborty, T. (2022). *Detecting and Understanding Harmful Memes: A Survey*. https://doi.org/arXiv:2205.04274v2

Shweta Sankhwar and Arvind Chaturvedi. (2018). Woman Harassment In Digital Space In India. *International Journal of Pure and Applied Mathematics*, *118*(20), 595–608.

Sultan, D., Omarov, B., Kozhamkulova, Z., Kazbekova, G., Alimzhanova, L., Dautbayeva, A., Zholdassov, Y., & Abdrakhmanov, R. (2023). A Review of Machine Learning Techniques in Cyberbullying Detection. *Computers, Materials & Continua*, *74*(3), 5625–5640. https://doi.org/10.32604/cmc.2023.033682

About the Authors

Vanathi Selvaraj is a Research Scholar from the Centre for Study of Social Exclusion and Inclusive Policy, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India. She has 9 years of experience as an Assistant Professor. Her major area of research includes Physical and Sexual Abuse of Adolescents and Cyber Abuse against Women.

Dr. K. Suriyan is a Professor and Director from the Centre for Study of Social Exclusion and Inclusive Policy, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India. He has 19 years of teaching experience. His major areas of research include Sociology of Marginalized – Social Exclusion, Training and Development (HR), Organizational Behavior and Corporate Social Responsibilities, Social Work and Sociology. He has more than 30 publications in refereed journals. He is the reviewer of many refereed journals and also acted as an advisory member in various conferences.