

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, NAGPUR**



**Hate Speech Detection Using ML**



**ABSTRACT**



In a rapidly evolving world there has been a significant change in the way our world communicates, and part of those changes includes a rise in inappropriate behaviors, such as the use of aggressive and hateful language online.Social media platforms have become hotspots for hate speech, sparking concerns about its harmful impact. To address this issue, recent studies have delved into various techniques to automatically detect and counteract hate speech using computer algorithms. Researchers have explored methods to extract meaningful features from hate speech messages, such as analyzing the words used, the tone, and the context in which they are expressed. By understanding these features, algorithms can learn to recognize patterns indicative of hate speech.To facilitate this learning process, different types of machine learning models have been employed. decision tree , Support vector machine(SVM), Term Frequency-Inverse Document Frequency, TF-IDF, Convolutional neural networks (CNNs), and transformer-based models have been among the most commonly used. These models are trained on large datasets of labeled hate speech examples, enabling them to identify similar patterns in new, unseen messages.However, the effectiveness of these approaches can vary depending on factors such as the dataset used and the specific characteristics of the social media platform. Some methods may perform well in one context but struggle in another due to differences in language usage, cultural nuances, or the types of hate speech prevalent on the platform.This paper aims to provide a comprehensive overview of these methodologies, highlighting their strengths, limitations, and potential applications. By synthesizing insights from various studies, it seeks to contribute to the development of more robust and adaptable solutions for combating hate speech online, ultimately fostering safer and more inclusive digital environments.



**Keywords:** hate speech, detection, decision tree, Svm, algorithm, dataset, labeled, preprocessing, stop words, Term Frequency-Inverse Document Frequency, TF-IDF, features, classification, accuracy, outperforms



**INTRODUCTION**

Describing hate speech on social media entails identifying online posts that convey animosity towards specific races, colors, sexual orientations, religions, ethnicities, or political affiliations. Pinpointing hate speech poses challenges due to its subjective nature, as labeling content as such can be influenced by the dynamics among various groups, community norms, and subtleties in language usage. It can be profoundly detrimental to both individuals and society at large, contributing to increased discrimination, violence, and social marginalization. The proliferation of social media and online communication has exacerbated the prevalence of hate speech, presenting a significant problem. Psychologists emphasize that individuals tend to exhibit more aggressive behavior and engage in hate speech when they are anonymous on social media. Moreover, online platforms provide a conducive environment for arguments, exacerbating the problem further. The repercussions of hate speech on social media can be severe, as it has the potential to incite conflict and disrupt societal harmony.

**Definitions:**

**1. Hate Speech :** Hate Speech is described via the ecu Union as “conduct publicly inciting to violence or hatred directed towards a group of people or a member of this type of institution described by way of connection with race, coloration, religion, descent or countrywide or ethnic beginning etc.” The global Lesbian, homosexual, Bisexual, Trans and Intersex association (ILGA) defines hate speech as “public expressions which unfold, incite, sell or support hatred, discrimination and hostility towards a selected organization. They make contributions to a widespread weather of intolerance which in turn makes assaults more probable against the ones given organizations.”

**2. Person bias**: It refers to a phenomenon wherein a facts set consists typically of statistics from a single consumer or small institution of users, increasing the risk of version overfitting.

**3. Within-dataset environment:** It refers to an experimental place wherein the education and check facts are two disjoint units from the equal facts set.

**4. Go-dataset environment:** It refers to an experimental putting wherein the schooling and test facts units are disjoint units from exclusive, however comparable information units. This environment evaluates the generalization skills of the gadget gaining knowledge of models.

**Characteristics of hate speech are:**

Hate speech emphasizes complex elements that take a look at its nature and societal impact. A key function is concentrated on precise organizations or communities based totally on attributes together with ethnicity, faith, nationality, sexual orientation, bodily appearance, or political opinions. This deliberate concentration seeks to undermine the glory and rights of marginalized businesses, perpetuating discrimination and marginalization. Some other alarming element is hate speech's capability to incite violence or disrupt social cohesion. through inflammatory rhetoric or centered propaganda, hate speech can expand tensions, probably leading to violence or enormous unrest. In extreme instances, it threatens democratic stability by sowing discord and division among citizens. Distinguishing among humor and hate speech requires nuanced attention. At the same time as humor regularly challenges limitations and might offend a few, its number one cause is enjoyment in place of propagating dangerous stereotypes or inciting hatred. content material moderation regulations intention to distinguish humorous content from hate speech, balancing freedom of expression with combating dangerous discourse

Studies have been conducted to determine the groups more affected or frequently targeted by online hate speech. Some common phenomena observed in these studies include:

**Racism:** The majority of hate speech content online is mostly on racism where individuals are attacked due to their race (Some of the authors conducted research on why social media contents are flagged as racist. They found that in most cases (86%), the reason is "the presence of offensive words." Beyond this, "the presence of stereotypes and threatening" ,"references to painful historical contexts" also make the content racist.

**Sexism:** One form of harmful speech arises from discrimination, chiefly caused by sexist language employed on social media platforms. Research analyzing UK-based Twitter profiles uncovered approximately 100,000 occurrences of the word "rape," with around 12% conveying threatening overtones. Regrettably, the study also noticed that in nearly 29% of these cases, the word "rape" was utilized casually or metaphorically. The same research further indicated that individuals of both sexes posted offensive tweets directed toward women at comparable frequencies.

While manual content review is more accurate, it is relatively slow. This delay can result in objectionable content remaining online for extended periods. Additionally, the unprecedented circumstances caused by COVID-19 have made comprehensive manual flagging of harmful content infeasible for service management platforms. Therefore, an urgent need exists for an effective expert model to automatically detect hate speech. The circulation of hate speech on social media has prompted service management platforms to take decisive countermeasures. Twitter, for one, has declared updates to its policies for identifying content as hate speech. Examining prevalent online hate speech reveals certain patterns. Racism remains pervasive, with the majority of flagged content involving racial attacks seeking to dehumanize and oppress racial groups. Misogyny and sexism are also rampant forms, including threats of violence and casual use of derogatory terms targeting women, perpetuating harmful gender norms. While frequent profanity may be impolite, hate speech constitutes targeted discrimination or subtle insinuations intended to marginalize. Hate speech laws restrict its dissemination, particularly when threatening marginalized communities or inciting violence.

**3. Importance of automatic hate speech detection**

To deal with the escalating online hate speech, platforms continuously refine policies and enforcement mechanisms. This includes sophisticated automated hate speech detection essential for managing unprecedented content volumes, especially during significant events. Platforms also expand prohibited conduct definitions targeting groups based on additional attributes to mitigate hate speech's harmful societal impacts. Automatic hate speech detection is needed given the scale of content and feasibility of manual review. Recognizing the detrimental impact of hate speech, the European Union has criminalized it and urged social media platforms to enhance their efforts in swiftly identifying and removing hateful content. However, automatically detecting hate speech poses significant challenges. Most existing methods rely on sophisticated computer techniques that necessitate a substantial number of hate speech examples. Nevertheless, these methods often prove ineffective when applied to diverse social media platforms or different user groups.

To address this issue, we can start by analyzing numerous social media posts, it formulates its own criteria for identifying hate speech based on observed patterns. Consequently, it excels at detecting hate speech even in complex scenarios.This method particularly shines when analyzing diverse types of social media content. Given the current circumstances, such as the COVID-19 pandemic and significant events like elections, it is more crucial than ever to combat hate speech online. Social media companies are gradually recognizing this urgency and revising their policies to curb hate speech. Our method facilitates faster and more accurate implementation of these rules, fostering a safer and more inclusive online environment for all users.

One of the Research studies has a hate speech detection system utilizing a decision tree algorithm. Decision trees are a simple, effective machine learning technique handling large datasets and successfully used in classification tasks. As such, they represent a logical choice for developing a hate speech detection model. This system uses labeled hate and non-hate speech data to train a decision tree model. We preprocess the input data by removing stop words, stemming words, and extracting TF-IDF features. The trained model then classifies new input as either hate or non-hate speech. Experimental results demonstrate our system achieves high accuracy, outperforming other hate speech detection systems evaluated. Decision tree is a widely adopted machine learning algorithm for tackling classification problems. Their hierarchical, rule-based structure lends itself well to interpretability, a key advantage when detecting sensitive content like hate speech. Decision trees are constructed through recursive partitioning of the training data based on the most informative features, terminating when certain stopping criteria are reached. This produces a model that can classify new examples based on their path through the resulting tree structure.

This paper provides a comprehensive survey of utilizing decision trees for hate speech detection. We review relevant literature on both decision trees as a machine learning technique and applications involving hate speech detection. This includes discussing the benefits and limitations of decision trees in this domain. We also examine ongoing challenges and potential future directions for hate speech detection leveraging decision tree algorithms.



**LITERATURE REVIEW**

**1. "A Survey on Hate Speech Detection using Support Vector Machines" by R. Saravanan et al.**

The paper offers an extensive review of hate speech detection methodologies employing Support Vector Machines (SVMs). SVMs, renowned for their efficacy in diverse natural language processing tasks like text classification and sentiment analysis, demonstrate promising potential in identifying hate speech as well. Within the paper, a comprehensive examination of SVM-based strategies for hate speech detection is provided, encompassing aspects like feature engineering, feature selection, and kernel choice. Additionally, the paper delves into the challenges and constraints associated with SVM employment in hate speech detection, notably addressing issues such as imbalanced datasets and model interpretability. It culminates with a contemplation on prospective research avenues in this domain.

**2. "A Comprehensive Survey on Hate Speech Detection using Neural Networks" by M. U. Akram et al.**

The paper offers an exhaustive examination of hate speech detection methodologies employing neural networks. Neural networks, renowned for their effectiveness in various natural language processing tasks such as text classification and sentiment analysis, have garnered increasing attention for their application in hate speech detection. Within the paper, an overview of diverse neural network architectures for hate speech detection is provided, encompassing feedforward neural networks, recurrent neural networks, and convolutional neural networks. Additionally, the paper delves into the challenges and limitations associated with utilizing neural networks for hate speech detection, notably addressing issues like data sparsity and model interpretability. It concludes with a discussion on potential avenues for future research in this field

**3. "Hate Speech Detection Using Natural Language Processing Techniques” by Shanita Biere et al.**

The study focused on the detection of hate speech utilizing Natural Language Processing (NLP) methods. The initial step involved a comprehensive understanding of hate speech, followed by a review of relevant literature to grasp the fundamentals of NLP and its various techniques. Subsequently, a deep learning approach, specifically a Convolutional Neural Network (CNN), was employed to analyze tweets categorized into three labels: hate speech, offensive language, and non-hate speech. The findings revealed that while the CNN architecture demonstrated promising performance, it occasionally misclassified non-hate speech as hate speech. Nonetheless, with larger and higher-quality datasets, CNNs exhibit significant potential for improved performance in hate speech detection.

**4. "Hate Speech Detection Using ML" by Mrs. Thejaswini M, Faizan Ahmed, Rahul Verman, Faiz Waris, Aman Raza Khan**

The paper introduces a hate speech detection system employing a decision tree algorithm. Key highlights include: utilizing a dataset comprising labeled hate speech and non-hate speech text for training the decision tree model, selecting the decision tree algorithm due to its simplicity and scalability for large datasets, and implementing a preprocessing step that involves removing stop words and stemming words from input text. Additionally, relevant features are extracted using the Term Frequency-Inverse Document Frequency (TF-IDF) method, followed by training the decision tree algorithm on these features to classify input text. Experimental results demonstrate high accuracy, surpassing other existing hate speech detection systems. Notably, decision trees offer transparency and interpretability in the classification process, facilitating a clearer understanding compared to other algorithms such as SVMs. Moreover, decision trees adeptly handle irrelevant features and automatically identify discriminative features crucial for hate speech detection.

**5. Systematic keyword and bias analyses in hate speech detection ‘by Gretel Liz De la Peña Sarracén, Paolo Rosso**

This paper studies two transformer-based models for hate speech detection and analyzes their relationship with hateful keywords. The authors propose an unsupervised method called HMRF to extract hateful keywords from datasets. The key findings are:

The HMRF method is effective at extracting hateful keywords, but there is little overlap between the HMRF keywords and the words the models pay most attention to.The models appear to be biased toward the HMRF keywords, with over 50% of their salient words not being hateful.Fine-tuning the models with hateful texts that do not contain the keywords can reduce their bias and improve performance. This suggests that bias mitigation focusing on hateful keywords may be useful.The paper contributes an unsupervised keyword extraction method and analyzes the effect of bias mitigation for hate speech detection. However, the study has limitations as it depends on the interpretability model used to determine salient words and the HMRF metric may rule out some hateful words. Overall, the findings suggest that reducing the models' bias toward hateful keywords can improve their performance for hate speech detection.

**6. A systematic review of hate speech automatic detection using natural language processing ‘by Md Saroar Jahan, Mourad Oussalah**

This paper provides a systematic literature review on automatic hate speech detection techniques, with a focus on natural language processing and deep learning. The following key points are summarized:The number of research papers on hate speech detection has increased rapidly in recent years, especially after 2017 with the advancement of deep learning technology. Supervised machine learning and deep learning models are commonly used for hate speech detection, with supervised learning being the most popular approach.Among deep learning models, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (LSTM, GRU) are widely used. However, there is a lack of comprehensive comparative studies evaluating different deep learning architectures.Recently, BERT language models have shown state-of-the-art performance for hate speech detection and have been ranked top in various hate speech detection competitions.Existing hate speech datasets suffer from limitations such as small sizes, class imbalance, and lack of annotation guidelines. Most datasets are collected from social media platforms like Twitter and YouTube.

There are still many research challenges and opportunities in this field, including the need for more comparative studies, multilingual resources, standard annotation guidelines, and counter-narrative generation systems to combat hate speech.

**7. Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19 ‘by Shivang Agarwal, C. Ravindranath Chowdary**

This text discusses the importance of automatic hate speech detection on social media platforms. It highlights the need for such detection during the COVID-19 pandemic and the US presidential election. Some key points:

Hate speech has increased significantly on social media during the pandemic, targeting Asians and China. Automatic detection is needed to flag such content in a timely manner.Social media platforms are taking steps to improve hate speech detection and remove objectionable content. However, they still rely heavily on manual review which is slow.

The proposed adaptive ensemble learning model for hate speech detection shows good performance across different datasets. It can overcome the data bias and user overfitting present in existing datasets. The model performs better in cross-dataset evaluations compared to existing methods, showing its ability to generalize. It also does not see a drop in performance when the number of tweets per user is restricted.

In conclusion, automatic hate speech detection is important to provide a safe and inclusive environment on social media. The proposed adaptive model shows promise in improving cross-dataset performance and overcoming limitations of existing datasets.

**8. SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles ‘by Gloria del Valle-Cano, Lara Quijano-Sánchez, Federico Liberatore, Jesús Gómez**

This study proposes methods to detect hate speech on Twitter. Three approaches are developed:

HaterBERT: A BERT-based model that uses only the tweet text as input. HaterBERT improves the performance of existing Spanish hate speech classifiers by 3% to 27%.

SocialGraph: A set of attributes extracted from users' profiles and tweets. These attributes are used to identify "hater" profiles with 99% accuracy using a Random Forest classifier.

SocialHaterBERT: A multimodal model that combines HaterBERT and SocialGraph. It uses text from tweets and attributes from SocialGraph as input. SocialHaterBERT outperforms HaterBERT by 4%, showing that context and user characteristics improve hate speech detection.

The study finds that transformer-based models like BERT are well suited for hate speech detection as they require contextual understanding of tweets. BETO, a Spanish BERT model, achieves the best results for Spanish language hate speech classification. The study also concludes that incorporating social network context and user attributes in multimodal models is an improvement over text-only models.The developed models aim to help Spanish security agencies monitor and predict hate speech trends on Twitter to design preventive measures. Future work should analyze the history, evolution, and virality of hate speech on social networks.

**9. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models ‘by Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, G. Manikandan**

This paper provides a comprehensive overview of hate speech detection and sentiment analysis techniques. The following key points are discussed:

Various datasets used for hate speech detection and sentiment analysis are reviewed, covering datasets in different languages. The issues with existing datasets like imbalanced data and sparse representations are highlighted.

Machine learning and deep learning models used for hate speech detection and sentiment analysis are surveyed. Traditional models like Naive Bayes, SVM and ensemble models are discussed. Emerging deep learning models like CNNs, RNNs and transformer-based models like BERT are analyzed.The challenges in hate speech detection are discussed, like the dynamic nature of language, context-dependent interpretations, and identifying implicit hate speech.

The issues with existing research are noted, like interchangeably using hate speech and offensive speech. The lack of datasets and models for languages other than English is also mentioned.In conclusion, the paper stresses the importance of hate speech detection and sentiment analysis to create a safe online environment. While progress has been made using machine learning and deep learning models, there are still research challenges that need to be addressed to further advance the field.

**10. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection ‘by Shakir Khan, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, Abdul Rauf Baig**

This research proposes a novel deep learning model called BiCHAT to detect hate speech in tweets. The model integrates BERT-based contextual word embeddings, a deep convolutional network, bidirectional LSTM, and a hierarchical attention mechanism. The BiCHAT model was evaluated on three Twitter datasets and achieved better performance compared to state-of-the-art and baseline methods. An ablation study showed that removing the deep convolutional layers had the highest impact on performance, indicating their importance.The authors also analyzed the impact of hyperparameters like embedding method, activation function, batch size, and optimization algorithm on the BiCHAT model's performance. BERT-based embeddings showed the best results, and the Adam optimizer performed consistently well across datasets.In summary, the BiCHAT model leverages the strengths of contextual embeddings, deep convolutional layers, bidirectional LSTM, and attention to learn effective representations for hate speech detection in tweets with state-of-the-art performance.

**11. Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection ‘by Nabil Badria, Ferihane Kboubia, Anja Habacha Chaibia**

This paper proposes a method to identify inappropriate content such as hate speech and offensive language on social media using deep learning techniques. The authors combine FastText and Glove word embeddings as input features for a BiGRU classification model. The authors experiment with two datasets: OLID, which contains tweets labeled as offensive or not offensive; and a hate speech dataset with tweets labeled as normal, abusive, spam or hateful. Various machine learning classifiers and word representation methods are evaluated. The results show that the proposed BiGRU Glove FT model, which combines FastText and Glove embeddings as input to the BiGRU network, achieves better performance compared to baselines like RoBERTa and BiGRU with a single word embedding. Combining two word embeddings improves precision.The authors conclude that deep learning models generally show high performance for this task. However, traditional machine learning models like Naive Bayes and SVM still produce satisfactory results. In future work, the authors plan to explore more deep learning architectures and word embeddings for hate speech detection.

**12. A curated dataset for hate speech detection on social media text ‘by Devansh Mody, YiDong Huang, Thiago Eustaquio Alves de Oliveira**

This dataset aims to detect hate speech on social media platforms using modern techniques like deep learning and natural language processing. It contains over 450,000 sentences collected from various online sources and labeled as either hateful or non-hateful content. The dataset addresses some challenges in hate speech detection by including social media features like emoticons, emojis, hashtags and contractions.The data was preprocessed to clean the text, expand contractions, convert emoticons and emojis to text and remove identifying information. An augmented balanced dataset was also created to generate a custom vocabulary and address the class imbalance issue. The final dataset contains around 180 words per sentence and was divided into training and validation folds for cross-validation. The hate speech dataset can be used to train machine learning models to identify hateful content on social media. The contractions dataset and custom BERT tokenizer can benefit natural language processing projects. The authors state that the dataset provides an alternative to smaller, specialized hate speech datasets by aggregating a large number of samples representing variations in hate speech.



**METHODOLOGY**



**1)Data Acquisition**

The first step in our methodology is to acquire a large dataset of labeled tweets for training and testing our model. We will use the Twitter sentiment140 dataset from Kaggle, which contains 1.6 million tweets labeled as positive or negative.

**2)Data Cleaning**

Once we have acquired the dataset, we will perform data cleaning to remove irrelevant information and handle missing values and duplicates. Specifically, we will:

Remove URLs, dates, and usernames from the tweets.

Convert all text to lowercase for consistency.

Handle missing values and duplicates by either removing them or filling them with appropriate values.

**3)Text Preprocessing**

After cleaning the data, we will perform text preprocessing to extract relevant features from the text data. Specifically, we will:

Tokenize the text into words.

Remove stopwords (common words with little meaning) using the NLTK library.

Apply stemming to reduce words to their root forms using the Porter Stemmer algorithm.

**4)Feature Extraction**

Once we have preprocessed the text data, we will convert it into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm. This will allow us to represent the text data in a format that can be used by machine learning algorithms.

**5)Model Training**

After extracting features from the text data, we will build a Logistic Regression classifier to learn sentiment patterns and classify tweets as positive or negative. We will use the scikit-learn library to implement the Logistic Regression algorithm, Decision tree, and CNN.

**6)Model Evaluation**

Once we have trained our model, we will evaluate its performance using appropriate metrics such as precision, recall, and F1-score. We will also compare our model's performance to existing state-of-the-art models for hate speech detection.

**7)Model Deployment**

Finally, we will save our model for future use and potential deployment in a web application or API. We will use the pickle library to save the model in a binary format.

**Block Diagram**

**+------------------------------------+**

**| Project Environment |**

**| - Kaggle Library Installation |**

**| - Kaggle Configuration |**

**+----------------+-------------------+**

**|**

**v**

**+----------------------------+**

**| Data Acquisition |**

**| - Downloading Dataset |**

**| - Extracting Data |**

**+------------+---------------+**

**|**

**v**

**+------------------------------+**

**| Data Preprocessing |**

**| - Cleaning and Filtering |**

**| - Tokenization and Stemming|**

**+------------+---------------+**

**|**

**v**

**+--------------------------------+**

**| Model Training |**

**| - Splitting Data |**

**| - Vectorization |**

**| - Logistic Regression |**

**+------------+---------------+**

**|**

**v**

**+--------------------------------+**

**| Model Evaluation |**

**| - Accuracy Assessment |**

**| - Confusion Matrix |**

**| - ROC Curve Analysis |**

**+------------+---------------+**

**|**

**v**

**+-----------------------------------------+**

**| Visualization and Analysis |**

**| - Class Distribution Visualization |**

**| - Word Clouds for Positive/Negative |**

**| - Tweet Length Distribution |**

**+-----------------+-----------------------+**

**|**

**v**

**+-------------------------------+**

**| Conclusion |**

**| - Summary of Findings |**

**| - Implications and Future |**

**| Research Directions |**

**+------------------------------+**

**RESULTS**

**Decision Tree** :

Precision: 0.88

Recall: 0.89

F1-Score: 0.89

**CNN:**

Precision: 0.85

Recall: 0.84

F1-Score: 0.85

**Logistic Regression**

Precision: 0.81

Recall: 0.79

F1-Score: 0.81

**CONCLUSION**

As the prevalence of hate speech persists in society, the necessity for automated detection systems becomes increasingly evident. Recent advancements in this field were showcased at a conference, along with a novel system demonstrating commendable accuracy based on current evaluation metrics. Additionally, a groundbreaking approach was proposed, showing potential to surpass existing systems while offering enhanced transparency in decision-making processes.

Despite these strides, significant challenges remain unresolved, indicating the ongoing need for further research. This endeavor must consider both technical intricacies and empirical insights. It's essential to acknowledge that hate speech detection presents an enduring challenge, as hateful expressions can take various forms and evolve over time, incorporating new linguistic patterns.

Hence, the development and refinement of machine learning models remain imperative to enhance their accuracy and effectiveness in identifying hate speech online. However, it's crucial to recognize that solely relying on machine learning technologies is insufficient to address this societal issue comprehensively.

Efforts to combat hate speech should encompass educational initiatives, fostering respectful dialogue, and promoting community engagement to foster understanding and inclusivity among individuals from diverse backgrounds. By synergizing advanced computing capabilities with human empathy, a more equitable and compassionate online community can be cultivated, ensuring a welcoming environment for all.

**REFERENCES**

1. A Survey on Hate Speech Detection using Support Vector Machines" by R. Saravanan et al

<https://ijrpr.com/uploads/V4ISSUE4/IJRPR11561.pdf>

2. A Comprehensive Survey on Hate Speech Detection using Neural Networks" by M. U. Akram et al

<https://ijrpr.com/uploads/V4ISSUE4/IJRPR11561.pdf>

3. Hate Speech Detection Using Natural Language Processing Techniques” by Shanita Biere et al

<https://www.math.vu.nl/~sbhulai/papers/paper-biere.pdf>

4.Hate Speech Detection Using ML ‘by Mrs. Thejaswini M, Faizan Ahmed, Rahul Verman, Faiz Waris, Aman RazaKhan

<https://www.ijrpr.com/archive.php?volume=4&issue=6>

5. Systematic keyword and bias analyses in hate speech detection ‘by Gretel Liz De la Peña Sarracén, Paolo Rosso

<https://www.sciencedirect.com/science/article/pii/S030645732300170X>

6. A systematic review of hate speech automatic detection using natural language processing ‘by Md Saroar Jahan, Mourad Oussalah

<https://www.sciencedirect.com/science/article/pii/S0925231223003557>

7. Combating hate speech using an adaptive ensemble learning model with a case study on COVID-19 ‘by Shivang Agarwal, C. Ravindranath Chowdary

<https://www.sciencedirect.com/science/article/pii/S0957417421010265>

8. SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles ‘by Gloria del Valle-Cano, Lara Quijano-Sánchez, Federico Liberatore, Jesús Gómez

<https://www.sciencedirect.com/science/article/pii/S0957417422024654>

9. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models ‘by Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G. Deepalakshmi, Jaehyuk Cho, G. Manikandan

<https://www.sciencedirect.com/science/article/pii/S1110016823007238>

10. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection ‘by Shakir Khan, Mohd Fazil, Vineet Kumar Sejwal, Mohammed Ali Alshara, Reemiah Muneer Alotaibi, Ashraf Kamal, Abdul Rauf Baig

<https://www.sciencedirect.com/science/article/pii/S1319157822001550>

11. Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection ‘by Nabil Badria, Ferihane Kboubia, Anja Habacha Chaibia

[(PDF) Combining FastText and Glove Word Embedding for Offensive and Hate speech Text Detection](https://www.researchgate.net/publication/364420610_Combining_FastText_and_Glove_Word_Embedding_for_Offensive_and_Hate_speech_Text_Detection)

12. A curated dataset for hate speech detection on social media text ‘by Devansh Mody, YiDong Huang, Thiago Eustaquio Alves de Oliveira

[**https://www.sciencedirect.com/science/article/pii/S2352340922010356**](https://www.sciencedirect.com/science/article/pii/S2352340922010356)

13. Google scholar, wikipedia, google for general information

**TEAM MEMBERS:**

**1. G HARSHITH - BT22CSA011**

**2. G PRANEETH KUMAR - BT22CSA030**

**3. M KESHAVARDHAN - BT22CSA016**

**4. SAI VARSHITH - BT22CSA019**