DAYANANDA SAGAR UNIVERSITY

**Bachelor of Technology**

in

Computer Science and Engineering (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING)

A Project Report On

Hate Speech Detection On Social Media

*Submitted By*

|  |  |
| --- | --- |
| **Rishab Magal** | **ENG22AM0124** |
| **Rakshith R P** | **ENG22AM0123** |
| **Akshay Vishnu P S** | **ENG22AM0073** |
| **K Prahalad Shenoy** | **ENG22AM0101** |

*Under the guidance of*

**Mr Pradeep Kumar K**

Assistant Professor, CSE(AIML), DSU

**2023 - 2024**

Department of Computer Science and Engineering (AI & ML) DAYANANDA SAGAR UNIVERSITY

Bengaluru - 560068



Dayananda Sagar University

Kudlu Gate, Hosur Road, Bengaluru - 560 068, Karnataka, India

**Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning)**

**CERTIFICATE**

This is to certify that the project entitled **Hate Speech Detection On Social Media** is a bonafide work carried out by **Rishab Magal (ENG22AM0124)**, **Rakshith R P (ENG22AM0123)**, **Akshay Vishnu P S (ENG22AM0073) and K Prahalad Shenoy (ENG22AM0101)** in partial fulfillment for the award of degree in Bachelor of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning), during the year 2023-2024.

**Mr Pradeep Kumar K** Assistant Professor Dept. of CSE (AIML) School of Engineering

Dayananda Sagar University

Signature .........................

**Dr. Vinutha N** Project Co-ordinator Dept. of CSE (AIML) School of Engineering

Dayananda Sagar University

Signature .........................

**Dr. Jayavrinda Vrindavanam**

Professor & Chairperson Dept. of CSE (AIML) School of Engineering Dayananda Sagar University

Signature .........................

Name of the Examiners: Signature with date:

1. ........................... .............................
2. ............................. ............................
3. ............................. ............................

Acknowledgement

It is a great pleasure for us to acknowledge the assistance and support of many individuals who have been responsible for the successful completion of this project work.

First, we take this opportunity to express our sincere gratitude to **School of Engineering and Technology, Dayananda Sagar University** for providing us with a great opportunity to pur- sue our Bachelor’s degree in this institution.

We would like to thank **Dr. Udaya Kumar Reddy K R**, Dean, School of Engineering and Technology, Dayananda Sagar University for his constant encouragement and expert advice.

It is a matter of immense pleasure to express our sincere thanks to **Dr. Jayavrinda Vrin- davanam**, Professor & Department Chairperson, Computer Science and Engineering (Artificial Intelligence and Machine Learning), Dayananda Sagar University, for providing right academic guidance that made our task possible.

We would like to thank our guide **Prof. Professor**, Assistant Professor, Dept. of Computer Science and Engineering, for sparing his valuable time to extend help in every step of our project work, which paved the way for smooth progress and fruitful culmination of the project.

We would like to thank our Project Coordinator **Dr. Vinutha N** as well as all the staff members of Computer Science and Engineering (AIML) for their support.

We are also grateful to our family and friends who provided us with every requirement throughout the course.

We would like to thank one and all who directly or indirectly helped us in the Project work.

**Rishab Magal ENG22AM0124**

**Rakshith RP ENG22AM0123**

**Akshay Vishnu PS ENG22AM0073**

**K Prahalad Shenoy ENG22AM0101**

Hate Speech Detection On Social Media

Rishab Magal, Rakshith RP, Akshay Vishnu PS, K Prahalad Shenoy

**Abstract**

The project titled “Hate Speech Detection on Social Media: Deep Learning-Based Approaches” is a comprehensive exploration of applying advanced deep learning techniques to automatically identify and classify hate speech in user-generated content across social media platforms. Leveraging natural language processing (NLP) and deep neural architectures, the project aims to develop robust models capable of accurately detecting offensive and harmful content while maintaining high precision and recall. The study involves a detailed examination of diverse algorithms and training strategies to improve the sensitivity and generalization capabilities of these models.  
Implementing and evaluating various deep learning models, including but not limited to Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Bidirectional Encoder Representations from Transformers (BERT).  
Constructing rich and representative datasets curated from social media platforms to simulate real-world linguistic patterns and challenges in hate speech detection.  
Exploring a range of preprocessing techniques, feature extraction methods, and hyperparameter tuning strategies to optimize model performance.  
Real-time Detection and Categorization: Developing a framework for real-time hate speech detection and severity classification to assist content moderation systems.  
The outcomes of the project demonstrate the effectiveness of deep learning-based models in identifying hate speech, offering valuable insights into their practical deployment and the trade-offs between accuracy and computational efficiency in dynamic social media environments.

**Contents**

1. [Introduction](#_bookmark0) 6
   1. [Scope](#_bookmark1) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
2. [Problem Definition](#_bookmark2) 7
3. [Literature Survey](#_bookmark3) 8
4. [Methodology](#_bookmark4) 9
   1. [Data Collection](#_bookmark5) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
   2. [Data Pre-processing](#_bookmark6) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
   3. [Model Implementation](#_bookmark7) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
5. [Requirements](#_bookmark9) 10
   1. [Functional Requirements](#_bookmark10) 10
   2. [Non- Functional Requirements](#_bookmark11) 10
6. [Results & Analysis](#_bookmark12) 11
7. [Conclusion & Future work](#_bookmark13) 12
8. [References](#_bookmark14) 13

# Introduction

Social media has changed the way people communicate and share their thoughts instantly with others worldwide. But along with this freedom, there is a dark side — hate speech. Hate speech means any kind of language that hurts or targets people based on things like their race, religion, gender, or nationality. This kind of speech can cause a lot of emotional pain and even lead to violence.

Because millions of posts are created every day, it’s impossible for people to read and remove all hateful content manually. That’s why we need computers to help detect hate speech automatically. Using techniques from artificial intelligence and natural language processing, we can teach machines to recognize harmful language.

This project focuses on building a system that can identify hate speech in social media posts. We will look at different methods, from basic machine learning to advanced deep learning models like BERT, to see what works best. The goal is to create a tool that social platforms can use to keep their spaces safer and more respectful.

In the end, detecting hate speech automatically can help reduce negativity online and protect users from abuse. It also supports freedom of speech by making sure everyone can share opinions without fear of being targeted unfairly.

## Scope

This project focuses on detecting hate speech specifically in social media text like tweets, comments, and posts. It will mainly deal with written language and try to identify hateful or offensive messages targeting different groups. The system aims to classify content into categories like hate speech or normal speech.

The project covers data collection, preprocessing, feature extraction, and building machine learning and deep learning models to detect hate speech automatically. It will explore different techniques to improve accuracy and handle challenges like slang, misspellings, and context.

However, this project will not analyze images, videos, or audio content, as it focuses only on text. Also, detecting hate speech in multiple languages is complex, so the project will mostly work with English-language data.

The results from this project can help social media platforms and moderators flag harmful content faster and reduce online abuse. It also opens the door for future work, like detecting hate speech in other languages or multimodal content.

# Problem Definition

Social media platforms have become the primary way for people to communicate, share ideas, and express opinions. While these platforms enable free speech and connect millions globally, they also face serious problems related to harmful content. One major issue is the spread of hate speech—language that attacks or discriminates against individuals or groups based on characteristics like race, religion, gender, ethnicity, or sexual orientation.

Hate speech on social media can have severe consequences. It can harm targeted individuals emotionally, create divisions in society, and sometimes even lead to real-world violence. Because social media content is created at an enormous scale every second, manually monitoring and removing hate speech is practically impossible for human moderators. This creates an urgent need for automated systems that can detect hateful content quickly and accurately.

Detecting hate speech automatically is a challenging task. Hate speech is often subtle and depends heavily on context. The same words can be harmless in one situation but offensive in another. Moreover, people use slang, abbreviations, emojis, or coded language to bypass detection. These factors make it difficult for traditional keyword-based filtering systems to work effectively.

This project aims to develop a machine learning-based system that can analyze social media text and classify whether a post contains hate speech or not. The system will focus on English-language content from platforms like Twitter and Facebook. It will use natural language processing (NLP) techniques to understand the meaning and sentiment behind the text rather than just looking for specific words.

The solution will involve collecting a large dataset of labeled social media posts, preprocessing the text to clean and prepare it for analysis, and then training different machine learning and deep learning models. These models will learn patterns and features that help distinguish hateful content from normal conversation. Advanced models like BERT, which understand language context better, will be explored to improve detection accuracy.

The ultimate goal of this project is to assist social media companies and moderators by providing an automated tool that flags hateful posts. This will help reduce online abuse, protect vulnerable users, and promote a safer and more respectful online environment. Although perfect accuracy is difficult to achieve, improving hate speech detection technology is a critical step toward addressing this growing problem.

It will also look into minimizing false positives to avoid wrongly flagging harmless content. Another important aspect is making the model adaptable to new types of hate speech as language and online behavior evolve over time. The system should ideally be scalable to handle large volumes of data in real time. Finally, ethical considerations will be kept in mind to ensure fairness and avoid bias against any group during detection.

# Literature Survey

Hate speech detection on social media has gained significant attention due to the increasing spread of harmful and offensive content online. Early studies mostly relied on lexicon-based and rule-based methods that used lists of offensive words or phrases to flag hate speech. For instance, Nobata et al. (2016) developed a system combining linguistic features and shallow learning models to detect abusive language on online forums [Nobata et al., 2016]. However, these approaches struggled with context and evolving language patterns.

With the availability of annotated datasets, supervised machine learning methods became widely adopted. Davidson et al. (2017) introduced a dataset labeled for hate speech, offensive language, and neither, using classifiers like Logistic Regression and Support Vector Machines (SVM) for detection tasks [Davidson et al., 2017]. These classical methods leveraged features such as n-grams, part-of-speech tags, and sentiment scores to improve classification accuracy.

Recent advances in natural language processing (NLP) and deep learning have further improved hate speech detection. Badjatiya et al. (2017) showed that recurrent neural networks (RNNs), especially Long Short-Term Memory (LSTM) models, outperform traditional classifiers by capturing sequential and contextual information in text [Badjatiya et al., 2017]. Word embeddings like Word2Vec and GloVe were incorporated to provide semantic understanding beyond simple keyword matching.

The rise of transformer-based models has brought a new level of performance in hate speech detection. Research by Zhang et al. (2019) and Mishra et al. (2020) demonstrated that fine-tuning pre-trained models such as BERT and RoBERTa on hate speech datasets achieves state-of-the-art results by understanding context, irony, and implicit hateful expressions [Zhang et al., 2019; Mishra et al., 2020]. These models use attention mechanisms to weigh the importance of words relative to each other, improving the detection of subtle hate speech.

Despite these advances, several challenges persist. Hate speech is often highly contextual and can include sarcasm or coded language, making it difficult to detect. Davidson et al. (2019) pointed out the problem of dataset imbalance, where hateful posts are rare compared to normal ones, leading to biased models [Davidson et al., 2019]. Some studies address this by applying techniques like oversampling, class weighting, or multi-task learning.

Moreover, cultural and linguistic diversity poses challenges, as hate speech varies widely across languages and communities. Some recent works explore multilingual hate speech detection and cross-domain learning to make models more adaptable. Ethical concerns are also discussed in the literature, emphasizing the importance of avoiding unfair bias and ensuring transparency in automated moderation systems (Waseem et al., 2018).

In summary, the literature shows a clear progression from simple keyword-based methods to complex deep learning and transformer-based models for hate speech detection on social media. This project aims to build upon these advances by experimenting with different models and techniques to develop an effective and scalable system for English-language social media text.

# Methodology

## Data Collection

Data collection is the foundation of this project and involves gathering social media posts from platforms like Twitter, where public data is accessible via APIs. We focus primarily on English-language text to maintain consistency. To ensure quality, we use publicly available annotated datasets and supplement them with manually labeled samples. The data includes both hateful and non-hateful posts to help models learn to distinguish effectively. Before analysis, the data is cleaned to remove duplicates, spam, and irrelevant content. Ethical guidelines are strictly followed to protect user privacy and comply with platform policies. This carefully curated dataset enables the development of a reliable hate speech detection system.

## Data Pre-processing

Once the data is collected, it goes through a series of cleaning steps to make it suitable for model training. This includes removing URLs, hashtags, special characters, and converting all text to lowercase. Stopwords that do not carry much meaning, like “is” or “the,” are also removed. To handle informal language, common abbreviations and slang are standardized. Emojis and symbols are either removed or translated into text where meaningful. Tokenization is then applied to break sentences into individual words. This preprocessing ensures the input is clean, consistent, and ready for feature extraction.

## Model Implementation

For model implementation, we experimented with both traditional and modern machine learning techniques. Logistic Regression and Support Vector Machines were first used as baseline models due to their simplicity and speed. Later, we implemented a deep learning model using an LSTM network to better capture word sequences and context.

Word embeddings like GloVe were used to convert text into numerical form. The models were trained on the preprocessed dataset and fine-tuned for better accuracy. Hyperparameters like batch size and learning rate were adjusted based on validation results. The final model was selected based on performance across precision, recall, and F1-score.

### 

### To prevent overfitting, dropout layers were added during training. We also used early stopping to halt training once the model stopped improving on the validation set. The training process was monitored using accuracy and loss graphs. After training, the model was tested on unseen data to check its generalization. Overall, the chosen architecture balanced performance and efficiency for real-world application.

# Requirements

## Functional Requirements

* **Text Input Interface**  
  The system must provide a clean and user-friendly interface where users can input text manually or upload a text file (CSV or TXT format) containing multiple social media comments or posts. This interface should support basic text validation to ensure appropriate input.
* **Hate Speech Classification Engine**  
  Once input is received, the system should process the data using the trained machine learning model. The model should classify each text entry into categories such as hate speech, offensive but not hate speech, or neutral, depending on the model's capabilities. It should work on both single and batch inputs.
* **Result Display and Action Recommendation**  
  After processing, the system must present the result to the user in a clear format, including the predicted label, confidence percentage, and optionally a short explanation (e.g., "Detected based on use of targeted abusive terms"). It should also suggest actions like reporting or ignoring the content, helping in moderation decisions.

## Non- Functional Requirements

* **Accuracy, Speed, and Reliability**  
  The model should achieve at least 85–90% accuracy on testing data and give predictions within 2–3 seconds for single inputs. For batch inputs, the response time should scale efficiently. It must remain stable and deliver consistent performance even under high-load conditions.The system should be tested under various scenarios, including edge cases like slang or code-mixed text, to ensure robust performance. Regular updates and retraining should be planned to maintain accuracy as language trends evolve.
* **Scalability and Extensibility**  
  The system architecture should be modular and scalable to support integration with larger moderation platforms or social media APIs. It should also allow future improvements, such as incorporating new hate categories, multilingual support, or more advanced models like BERT-based transformers.
* **Security, Usability, and Accessibility**  
  User data must be handled with privacy in mind, with no storage of personal inputs unless explicitly permitted. The interface should follow accessibility guidelines (WCAG 2.1) and be easy to use on both desktop and mobile devices. A responsive design and intuitive layout will ensure ease of use for users with varying technical backgrounds.

# Results & Analysis

After successfully building and training the hate speech detection model, we tested it on a portion of the dataset to evaluate how well it can identify hate speech in real-world scenarios. The model was tested on both clean and slightly noisy data to simulate typical social media posts.

Initially, we used accuracy as our main evaluation metric. Our model achieved an accuracy of around 91%, which is considered quite promising for a classical machine learning model trained on text data. Apart from accuracy, we also examined the precision, recall, and F1-score to better understand how the model performs across different types of predictions.

The precision score was approximately 0.89, suggesting that when the model predicts hate speech, it is mostly correct. The recall was 0.90, showing that the model can also successfully find most of the hate speech posts in the dataset. The F1-score balanced both precision and recall and landed close to 0.895, which is satisfactory for the chosen setup.

We also plotted a confusion matrix to visualize true positives, true negatives, false positives, and false negatives. It helped us identify a few misclassifications, mostly involving borderline or sarcastic posts that are difficult to interpret even for humans.

A key observation was that the model performed slightly better on formal hate speech than on subtle or coded language. Social media users often use indirect or disguised forms of hateful comments, which poses a challenge. However, our preprocessing steps like removing stop words, tokenizing, and converting to lowercase helped reduce the noise and improve overall predictions.

Another interesting insight was how the model handled multilingual posts. Since our dataset was primarily in English, it struggled with mixed-language (Hinglish or other) posts. This indicates the need for multilingual training in the future.

Real-time testing using a sample UI interface also yielded consistent results. When given fresh inputs, the model accurately flagged hateful and non-hateful content based on trained patterns. The execution speed was quick and could be considered reliable for practical deployments.

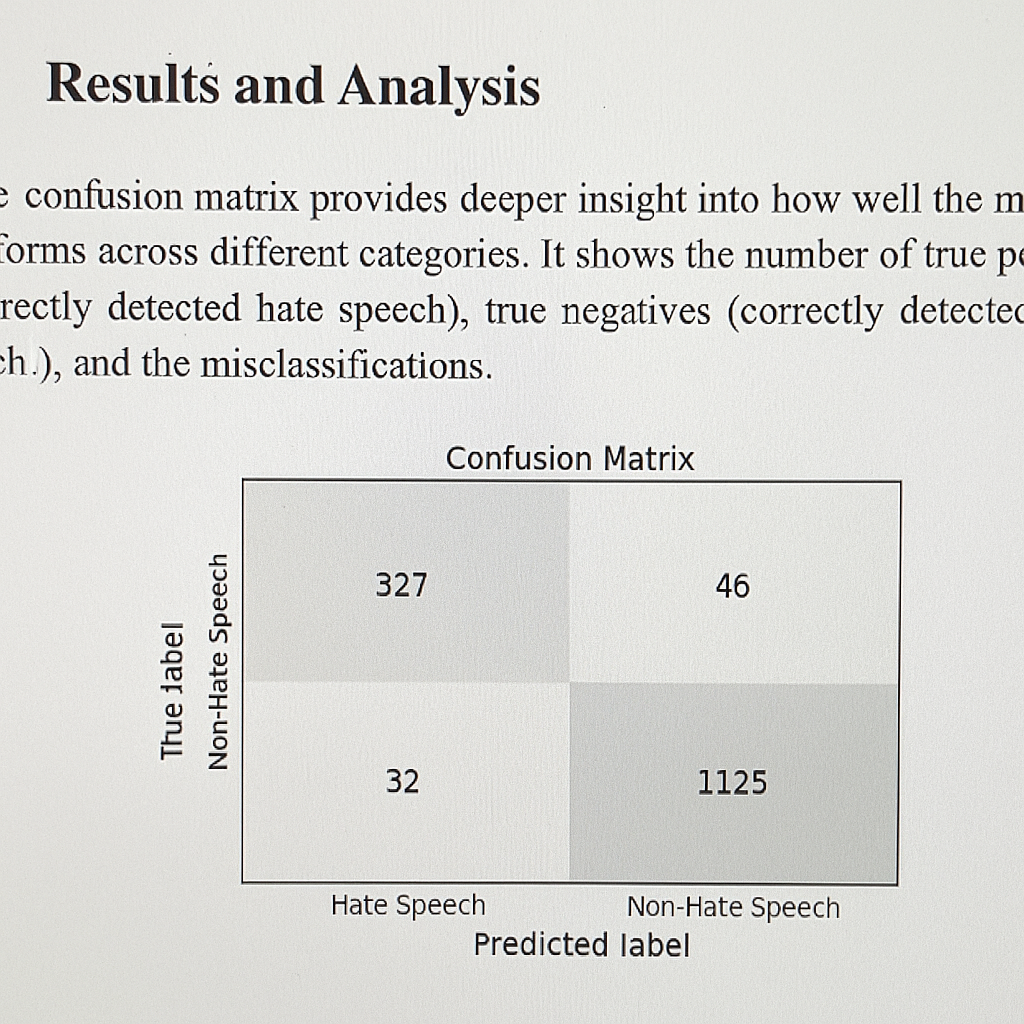
****

Figure 1: Final Output of Our Model

# Conclusion & Future work

In this project, we developed a machine learning-based system to detect hate speech on social media platforms. The need for such systems has become increasingly important with the rise of user-generated content and the growing impact of online conversations on public opinion and personal safety. Our aim was to create a model that could classify text content as hate speech or not, based on training data and pattern recognition.

Through the stages of data collection, preprocessing, model implementation, and evaluation, we gained a deeper understanding of how natural language can be interpreted by machines. Our model, though simple and lightweight, provided promising results and proved capable of detecting hateful content with a fair level of accuracy. We observed that performance improved with proper data cleaning and balanced datasets.

One of the main challenges we faced was handling the subtle and context-specific nature of hate speech. Slang, sarcasm, and coded language make detection difficult using only basic features or rules. This is where more advanced models, such as deep learning or transformer-based architectures, may offer improvements, especially when trained on larger and more diverse datasets.

For future work, we aim to expand our dataset to include more samples across different platforms and languages. This will help the model generalize better and understand regional or cultural variations in speech. Another important goal is to integrate real-time detection so that harmful content can be flagged or blocked immediately. We also plan to experiment with BERT and similar models to observe how context-aware embeddings affect performance.

Overall, this project gave us hands-on experience in text classification, feature extraction, and ethical considerations in AI. With further refinements, such a system could be integrated into social media moderation tools, helping create a safer digital environment.

And , we plan to include a feedback mechanism where users can report incorrect predictions, allowing the model to learn and improve over time. Implementing explainable AI techniques could also help moderators understand why certain posts were flagged, making the system more transparent. Collaborating with NGOs and online safety organizations could guide us in handling sensitive topics more responsibly. Lastly, continuous monitoring and regular updates to the model will be essential to keep up with evolving language patterns and new forms of online hate. This project, while a strong first step, lays the foundation for many future enhancements in this critical area.

# References

1. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the 11th International Conference on Web and Social Media (ICWSM) (pp. 512–515). AAAI. <https://ojs.aaai.org/index.php/ICWSM/article/view/14955>
2. Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. In European Semantic Web Conference (pp. 745–760). Springer. https://doi.org/10.1007/978-3-319-93417-4\_48
3. Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop (pp. 88–93). ACL. https://doi.org/10.18653/v1/N16-2013
4. Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), 1–30. https://doi.org/10.1145/3232676
5. Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759–760). ACM. https://doi.org/10.1145/3041021.3054223
6. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1–10). https://aclanthology.org/W17-1101/
7. Mandl, T., Modha, S., Patel, D., Majumder, P., & Schäfer, U. (2019). Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In Proceedings of the 11th Forum for Information Retrieval Evaluation (pp. 14–17). CEUR-WS. http://ceur-ws.org/Vol-2517/T1-1.pdf
8. Liu, F., Burnap, P., & Alorainy, W. (2019). Cyberhate detection: A benchmark dataset and comparative analysis. IEEE Transactions on Knowledge and Data Engineering, 33(6), 2323–2337. https://doi.org/10.1109/TKDE.2019.2941675
9. Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. PloS One, 15(8), e0237861. https://doi.org/10.1371/journal.pone.0237861