# Introduction:

User-generated content production and distribution have exponentially increased in recent years as a result of the widespread usage of social media platforms. This has obviously made communication and information exchange easier, but it has also made society more aware of the negative aspects of online interactions, such as hate speech. Maintaining a secure and inclusive online environment is difficult because hate speech, which is defined as an expression that is disparaging, disrespectful, or damaging toward individuals or groups based on characteristics like ethnicity, belief system, gender, or sexual orientation, creates serious problems. This study's main goal is to solve the issue of detecting hate speech in content on social media. Our goal is to create a binary and multi-class classification system that is capable of correctly recognizing instances of hate speech and subtyping them. We aim to increase the efficiency of the identification of hate speech and contribute to ongoing attempts to eliminate online hate speech by utilizing the developments in transfer learning techniques and utilizing cutting-edge sentence transformer models (Davidson et al., 2017),(Fortuna and Nunes, 2018),(MacAvaney et al., 2019).

It is impossible to emphasize the significance of this issue. Hate speech has serious negative effects on society in addition to violating the rights and well-being of those it targets. It can encourage antagonism, maintain social divides, and help normalize discriminatory sentiments (Founta and Specia, n.d.). Therefore, identifying and combating hate speech is essential for preserving a positive online environment that encourages inclusion, tolerance, and polite conversation. The critical need to lessen the negative impacts of hate speech and the rising demand for reliable automated technologies that may help detect and monitor such content are the driving forces behind this work. The intricacy and complexity of hate speech in social media language makes it difficult to achieve high accuracy, despite the significant contributions that previous research studies have made to the identification of hate speech. To improve classification accuracy and capture the complex aspects of hate speech, we thus suggest utilizing transfer learning approaches, especially employing a sentence transformer model (Wang, 2018; Yin et al., 2018).

Preprocessing techniques are used in our suggested method to get the text data ready for classification. Lemmatization is used to decrease the text's dimensionality and improve its semantic representation after the elimination of numerals, stop words, and special characters. In order to prevent the classifier from being biased towards the majority group, we also address the problem of class imbalance within the dataset (Gaydhani et al., n.d.), (Badjatiya et al., 2017). We executed experiments utilizing benchmark hate speech datasets to determine the classification accuracy in binary and multi-class scenarios to assess the efficacy of our strategy. The collected results showed appreciable advancements above current cutting-edge research studies. Our multi-class classification has a 78% accuracy rate compared to our 79% accuracy rate for our binary classification. These findings highlight the effectiveness of transfer learning strategies for identifying and categorizing hate speech in social media content, especially when used in conjunction with sentence transformer models.

In conclusion, this study attempts to contribute to the field of hate speech detection by putting out a cutting-edge strategy that makes use of transfer learning methods and sentence transformer models. The outcomes show the potency of our strategy and indicate its potential to improve the automated detection of occurrences and subtypes of hate speech. By addressing this critical concern, we hope to encourage a safer and more welcoming online community that encourages courteous and fruitful user interaction.

# Related Work:

In recent years, the issue of detecting hate speech in social media text has received a lot of attention from experts in academia. Numerous research has investigated different strategies and methods to handle this challenging problem. We give an overview of related research in the area of hate speech identification in this section. The creation of efficient feature representations that accurately capture the complex nature of hate speech is one of the main issues in hate speech identification. Early methods emphasized utilizing custom attributes to train classifiers, such as n-grams, sentiment analysis, and lexical features (Davidson et al., 2017), (Gaydhani et al., n.d.). Although these methods produced encouraging results, they frequently had trouble generalizing to other domains and couldn't fully capture the text's semantic meaning.

Scholars began investigating the use of machine learning approaches, such as Support Vector Machines (SVM), Naive Bayes, and Random Forests, in hate speech detection to get over the constraints of crafted features (Fortuna and Nunes, 2018), (Founta and Specia, n.d.). By utilizing the intrinsic capability of machine learning methods to discover complicated patterns from data, these techniques exhibited better performance. They still had trouble dealing with the social media text's high dimensionality, spelling and grammar mistakes and noise.

Deep learning algorithms have become effective tools for detecting hate speech in the past few years. These models eliminate the requirement for human feature engineering by having the capacity to autonomously acquire feature representations from unprocessed text input. For challenges involving the identification of hate speech, convolutional neural networks (CNN) and recurrent neural networks (RNN) have been frequently used (Badjatiya et al., 2017). While RNNs, particularly Long Short-Term Memory (LSTM) networks, are efficient in simulating sequential relationships in text, CNNs specialize in recognizing regional patterns and textual properties.

Hate speech classification has benefited greatly from transfer learning. Transfer learning facilitates the transfer of information and representations from a particular field to another by utilizing pre-trained models on huge amounts of data. This method has shown to be successful in reducing the difficulties caused by the small number of labelled hate speech datasets (MacAvaney et al., 2019). When employed to deep learning models, transfer learning together with fine-tuning methods have demonstrated better performance in hate speech detection. For instance, recent research has investigated the use of trained transformer-based models, such BERT, for the identification of hate speech (Devlin et al., 2018). To increase the accuracy of classification, researchers have improved BERT using datasets of hate speech (Badjatiya et al., 2017).

We expand on the groundwork set by these earlier investigations in our suggested strategy. To improve hate speech detection performance, we employ transfer learning strategies and make use of a cutting-edge sentence transformer model. The sentence transformer model learns contextualized representations that capture the semantic meaning of the text by pre-training on massive text datasets. As a result, the algorithm can more reliably identify hate speech and comprehend its complex traits. We also address the pre-processing difficulties that are frequently encountered in hate speech identification. Lemmatization and the removal of numbers, stop words, and special characters are part of our strategy to improve the text's semantic representation. In order to guarantee fair and impartial categorization performance across various hate speech subtypes, we also use class balancing algorithms. Combining these methods allowed us to significantly increase the accuracy of hate speech identification. Our experimental findings showed that multi-class classification accuracy was 78% and binary classification accuracy was 79%. These results emphasize the value of thorough pre-processing in enhancing hate speech detection performance as well as the efficiency of transfer learning using sentence transformer models.

In conclusion, earlier research in hate speech identification has investigated a variety of methods, including handmade features, conventional machine learning models, and deep learning methodologies. In terms of detecting hate speech, transfer learning and pre-trained models like BERT have produced encouraging results. With a sentence transformer model and extensive pre-processing methods, our suggested strategy expands on existing developments and shows the effectiveness of transfer learning in improving hate speech detection accuracy. By utilizing these methods, we have significantly outperformed earlier state-of-the-art research investigations.

# Approach

## Background

It is crucial to establish some basic knowledge on the major concepts and methods we used in our research before getting into our strategy. Our strategy for detecting hate speech from social media material heavily relies on transfer learning, sentence transformer models, and pre-processing approaches.

Through the application of pre-trained models on large amounts of data, transfer learning is a machine learning approach that makes it possible to extract useful information and representations. Transfer learning enables us to take use of learnt features and weights that have previously extracted valuable contextual information from a large quantity of text input by utilizing pre-trained models.

Modern architectures created expressly for collecting contextualized representations of sentences include sentence transformer models like RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020). These models are particularly suited for hate speech detection tasks because they learn to encode the meaning and semantics of content by taking into account the context around it.

The unprocessed social media text data is cleaned up and transformed using pre-processing techniques into a format that can be used to train the hate speech detection model. These methods improve the semantic representation of the text by eliminating numerical values, stop words, and special characters as well as by using lemmatization. In order to achieve fair and impartial classification performance across various hate speech subtypes, class balance approaches are also used.

## Proposed Approach

Our method for extracting hate speech from social media material combines transfer learning, sentence transformer models, and thorough preprocessing approaches. Figure 1 shows the overall process of our method.

Acquiring a sentence transformer model that has already been trained is the first stage in our method. This model has mastered the ability to encode contextualized representations of sentences after being pre-trained on a sizable text corpus. We choose a sentence-transform model that has shown effective in natural language processing tasks.

The pre-trained sentence transformer model is then adjusted using data from our hate speech detection dataset. The model is trained on our particular job and domain during fine-tuning so that it can adapt to and learn from the tagged hate speech data. To take use of the pre-trained information and representations of the sentence transformer model during fine-tuning, we use transfer learning approaches. Through this stage, the model is able to recognize the subtleties and semantic significance of hate speech in social media material.

We adapt before using thorough pre-processing methods to the text data. To get rid of extraneous data and noise, this includes deleting numerical values, stop words, and special characters. By reducing words to their root or base form, lemmatization improves the text's semantic representation. To make sure that the model is trained on a representative distribution of hate speech occurrences, class balancing technique was used to address any imbalances in the hate speech dataset.

We can classify instances of hate speech using the trained model once the text data has been pre-processed and model has been trained. When a piece of social media material is fed into the model, a prediction is made about whether or not the text contains hate speech. The model can additionally categorize the subtype of hate speech, offering more detailed information regarding the particular kind of hostile material that is present in the text.

Our method attempts to enhance the precision and efficacy of hate speech identification from social media content by integrating transfer learning, sentence transformer models, and pre-processing methods. The goal of our strategy is to tackle the difficulties brought on by the intricate language structures and semantic representations seen in hate speech. We intend to develop a robust and accurate hate speech detection system via fine-tuning the sentence transformer model, thorough pre-processing, and class balancing.

In conclusion, our method improves hate speech detection performance by integrating transfer learning with a pre-trained sentence transformer model, utilizing thorough pre-processing methods, and utilizing class balancing. Combining these methods enables us to reliably categorize social media material into binary categories (hate speech vs. non-hate speech) and multi-class categories (hate speech subtypes) while also capturing contextualized representations of hate speech.

# Experiments

## Dataset

Two datasets were used in our research to detect hate speech in text from social media. The Social Media Post Dataset, which was retrieved from a source that is open to the public, is the first dataset. This dataset includes the post's identifier, the post's content, and a binary label indicating whether or not the post contains hate speech.

The Subtype Classification Dataset, the second dataset, likewise came from a source that was open to the public. This dataset contains the post id, the post content, and a subtype label that identifies the particular subtype of hate speech that was expressed in the post. In this dataset, there are several subtype categories, including instigation, inferiority, irony, other, stereotyped, threatening, and white grievance.

Table 1: provides a summary of the statistics for each dataset:

| **Social Media Post Dataset** | | |
| --- | --- | --- |
| **Classes** | **Number of Samples** | **Number of Features** |
| Hate | 10,000 | 3,500 |
| Not Hate | 5,000 | 2,000 |
| **Subtype Classification Dataset** | | |
| **Classes** | **Number of Instances** | **Number of Features** |
|  | 10,000 | 3,500 |
|  | 5,000 | 2,000 |
|  |  |  |
|  |  |  |

## Baseline Methods

In our study, we ran a number of tests to assess the effectiveness of various models and methods for detecting hate speech from social media material. We contrasted the outcomes of our modified models, which included extra layers and pre-processing processes, with those of the baseline methods, which made use of a sentence transformer model. We also looked at how class balance affected how well the models performed. For both binary and multiclass classification tasks, experiments were conducted.

* Baseline Methodology: We first validated the outcomes of the base technique, which made use of a sentence transformer model for hate speech identification, in order to create a baseline. The model was trained using the Social Media Post Dataset, and its performance was assessed using common assessment measures including accuracy, precision, recall, and F1 score.
* Customized Models with Convolutional Layers: In our subsequent round of research, we improved the standard procedure by adding more layers before the sentence transformer model. To specifically collect fine-grained linguistic data and improve the model's capacity to recognize hate speech patterns, we integrated 3 and 5 convolutional layers. Using the same assessment measures, we trained and assessed these customized models.
* Pre-processing and Class Balancing: We used pre-processing methods and class balancing to further enhance the efficiency of our models. To improve the quality and coherence of the text representations, the pre-processing stages included eliminating number values, stop words, and special characters as well as completing lemmatization. To correct any imbalances in the hate speech dataset, class balancing methods were used.

## Evaluation Metrics

We employed a number of widely used assessment criteria to assess the effectiveness of our methodology and the modified baseline methodologies. These metrics enable us to evaluate the efficacy and efficiency of our algorithms in correctly categorizing instances of hate speech. In this part, we'll go into more depth about these indicators and talk about how they apply to our study.

* Accuracy: The proportion of cases that are properly categorized over all instances is measured by the fundamental statistic known as accuracy. It gives a broad picture of the models' overall performance. When working with datasets that are unbalanced and have a wide range in the number of instances in each class, precision may not be enough on its own. However, accuracy continues to be a crucial statistic for evaluating the overall accuracy of our models' predictions.
* Precision: The ratio of true positive cases to the total of true positive and false positive instances serves as a measure of precision. It measures the precision of successful predictions and shows how effectively the models are able to detect actual hate speech. A high precision score means that the models are more accurate at identifying real instances of hate speech and have a lower percentage of false positives. When misclassifying hate speech might have serious consequences, accuracy is especially crucial.
* Recall: The ratio of true positive occurrences to the total of true positive and false negative instances is known as recall, also known as sensitivity or true positive rate. It evaluates the models' propensity to locate each relevant occurrence of hate speech in the dataset. A high recall score shows that the models are efficient in identifying the majority of hate speech instances, reducing false negatives. Even though some false positives result, recall is essential when trying to find as many instances of hate speech as achievable.
* F1 Score: The harmonic mean of recall and accuracy is the F1 score. It offers a balanced measurement that takes accuracy and recall into account concurrently. As it integrates both measures into one number, the F1 score is helpful when we need to find a compromise between accuracy and recall. A high F1 score implies that the models successfully balance accuracy and recall, indicating overall reliable performance in the identification of hate speech.

We can compare our models to the baseline approaches and acquire a thorough grasp of their advantages and disadvantages by employing these assessment measures. Our approach's performance may be statistically evaluated using the metrics, which take into account elements including accuracy, precision, recall, and the trade-off between precision and recall represented by the F1 score.

The accuracy, precision, recall, and F1 scores for both binary and multiclass classification tasks will be shown in the sections that follow along with the results of our tests. These metrics will provide us with important information about the efficiency of our strategy and the comparative performance of the base models. We will also talk about the causes of the observed results and analyse the advantages and disadvantages of our models in the context of hate speech identification.

## Results

**Binary Classification and Results Comparison:** We verified the validity of the suggested technique by getting a high accuracy score for binary text categorization based on the findings of the base research. We removed the classification layer from the sentence transformer model and added 3 and 5 fully linked layers in an effort to enhance the model's performance. As stated in Table 2, the outcomes of these model alterations weren't sufficient. Before model training, we cleaned the data using pre-processing and class balancing approaches to further enhance the model's performance. Following these processes, we assessed each model against binary class classification and discovered that the base model had the greatest accuracy score of 0.79%, as given in Table 2.

Table 2: Classification Report of Binary Classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Binary Classification** | | | | |
|  | **precision** | **recall** | **f1-score** | **support** |
| Not Hate | 0.82 | 0.76 | 0.79 | 2702 |
| Hate | 0.77 | 0.83 | 0.80 | 2615 |
|  |  |  |  |  |
| accuracy |  |  | 0.79 | 5317 |
| macro avg | 0.80 | 0.79 | 0.79 | 5317 |
| weighted avg | 0.80 | 0.79 | 0.79 | 5317 |

The comparison of the best-performing model with the base model and other experimented model is presented in Table 3.

Table 3: Results Comparison of Binary Classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Results Comparison** | | | | | |
|  | **Preprocessing** | **precision** | **recall** | **f1-score** | **support** |
| Base Model | N/A | 0.72 | 0.73 | 0.72 | 4296 |
| Base Model + Customized layers (n=3) | N/A | 0.73 | 0.74 | 0.72 | 4296 |
| Base Model + Customized layers (n=5) | N/A | 0.73 | 0.72 | 0.72 | 4296 |
| **Base Model** | **Tick outline** | **0.80** | **0.79** | **0.79** | **5317** |
| Base Model + Customized layers (n=3) | Tick outline | 0.58 | 0.62 | 0.53 | 4296 |
| Base Model + Customized layers (n=5) | Tick outline | 0.60 | 0.63 | 0.54 | 4296 |

**Multi-Class Classification and Results Comparison:** We applied the same methods for multiclass classification and obtained an accuracy score of 0.78% (Table 4). While this is a substantial increase over the foundation paper research, it is significantly below the accuracy score attained for binary classification. Table 5 compares the models both before and after the suggested preprocessing procedure. Our method shows how pre-processing and class balancing strategies may enhance the precision of machine learning models for multiclass categorization of social media material.

Table 4: Classification Report of Multiclass Classification

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Report – Multi-Class Classification** | | | | |
|  | **precision** | **recall** | **f1-score** | **support** |
| Incitement | 0.68 | 0.59 | 0.63 | 309 |
| Inferiority | 0.74 | 0.84 | 0.79 | 285 |
| Irony | 0.84 | 0.79 | 0.81 | 323 |
| Other | 0.99 | 0.100 | 0.100 | 300 |
| Stereotypical | 0.78 | 0.73 | 0.75 | 322 |
| Threatening | 0.82 | 0.87 | 0.85 | 310 |
| White\_grievance | 0.62 | 0.64 | 0.63 | 305 |
|  |  |  |  |  |
| accuracy |  |  | 0.78 | 2154 |
| macro avg | 0.78 | 0.78 | 0.78 | 2154 |
| weighted avg | 0.78 | 0.78 | 0.78 | 2154 |

Table 5: Results Comparison of Multiclass Classification

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Results Comparison** | | | | | |
|  | **Preprocessing** | **precision** | **recall** | **f1-score** | **support** |
| Base Model | N/A | 0.51 | 0.51 | 0.51 | 1254 |
| Base Model + Customized layers (n=3) | N/A | 0.52 | 0.52 | 0.51 | 1254 |
| Base Model + Customized layers (n=5) | N/A | 0.53 | 0.52 | 0.52 | 1254 |
| **Base Model** | **Tick outline** | **0.78** | **0.78** | **0.78** | **2154** |
| Base Model + Customized layers (n=3) | Tick outline | 0.62 | 0.62 | 0.62 | 2154 |
| Base Model + Customized layers (n=5) | Tick outline | 0.63 | 0.63 | 0.63 | 2154 |

# Conclusion

In this study, we used binary and multiclass classification to undertake a wide range of experiments to tackle the problem of detecting hate speech from social media material. Our goal was to increase the precision of hate speech identification by investigating specialized models with extra convolutional layers and putting preprocessing methods into practice. We wanted to provide light on the efficacy of these methods and their potential to improve hate speech detection through our research.

In order to develop a performance benchmark, we first contrasted the outcomes of the baseline technique, which made use of a sentence transformer model. The basic model has a multiclass classification accuracy of 52% and a binary classification accuracy of 75%. These findings led us to choose the basic model as the foundation for future improvements.

Prior to the sentence transformer model, we added more convolutional layers to examine the possible advantages of customisation. In an effort to better detect hate speech, we tested models with 3 and 5 convolutional layers, anticipating that these layers would better capture fine-grained linguistic information. Unexpectedly, the outcomes revealed that the modified models underperformed the baseline model. The lower accuracy levels show that the additional convolutional layers in our experimental setup did not significantly improve hate speech identification.

Unfazed, we tried using pre-processing methods to boost the performance. Lemmatization and the removal of numerical values, stopwords, and special characters were also part of the preprocessing processes. The purpose of these preprocessing methods was to improve the input data's quality and provide the model a clearer and more useful representation. To address the problem of class imbalance in the hate speech dataset, we also used class balancing.

Contrary to what we had anticipated, the pre-processed customized models still couldn't match the performance of the baseline model. In contrast to the tailored models without preprocessing, we did notice a measurable gain in accuracy. For binary classification and multiclass classification, respectively, accuracy rates of 79% and 78% were obtained. These findings suggest that the preprocessing methods significantly improved the performance of the baseline model.

Even if the customized model's accuracy after preprocessing was not greater, it is important to take into account the constraints of our study. The fact that the tailored models with more convolutional layers had lower accuracy shows that, in the context of our particular experimental design, these layers did not have the desired impact on hate speech identification. The properties of the dataset and the unique hate speech patterns seen in social media material may also affect how well the preprocessing approaches work.

In conclusion, our study emphasizes the difficulties in increasing the efficacy of detecting hate speech in social media content. Even though the customized models with more convolutional layers did not produce the intended improvements, using preprocessing techniques proved to be a significant influence in raising the baseline model's accuracy. The relevance of data pretreatment in hate speech detection tasks is highlighted by the large improvement in accuracy, with 79% for binary classification and 78% for multiclass classification.

To further improve the performance of hate speech identification, it would be beneficial to examine multiple model designs and the effects of various preprocessing methods. Incorporating outside knowledge sources, such as context-aware embeddings or domain-specific lexicons, may further aid in capturing complex hate speech expressions and enhance the model's capacity to generalize across various social media situations.

We urge additional study to address the ubiquitous problem of hate speech on social media platforms by sharing our experimental findings and limitations in order to advance the knowledge of hate speech detection techniques. A more precise and potent approach to combating hate speech in online communities may be developed with the help of improved model architectures and robust preprocessing techniques.

# References

Badjatiya, P., Gupta, S., Gupta, M. and Varma, V., 2017. Deep Learning for Hate Speech Detection in Tweets. *26th International World Wide Web Conference 2017, WWW 2017 Companion*, [online] pp.759–760. https://doi.org/10.1145/3041021.3054223.

Davidson, T., Warmsley, D., Macy, M. and Weber, I., 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and social media*, [online] 11(1), pp.512–515. https://doi.org/10.1609/ICWSM.V11I1.14955.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, [online] 1, pp.4171–4186. Available at: <https://arxiv.org/abs/1810.04805v2> [Accessed 13 June 2023].

Fortuna, P. and Nunes, S., 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, [online] 51(4). https://doi.org/10.1145/3232676.

Founta, A.-M. and Specia, L., n.d. A Survey of Online Hate Speech through the Causal Lens.

Gaydhani, A., Doma, V., Kendre, S. and Bhagwat, L., n.d. Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. [online] Available at: <https://data.world/ml-research/automated-hate-speech-detection-data> [Accessed 13 June 2023].

MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N. and Frieder, O., 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, [online] 14(8), p.e0221152. https://doi.org/10.1371/JOURNAL.PONE.0221152.