MAULANA AZAD
NATIONAL INSTITUTE OF TECHNOLOGY
BHOPAL, INDIA, 462003

HATE SPEECH DETECTION

Minor Project Report

VIth Semester

Submitted by:

Mohit Verma	201112244
Mihir Mall	201112226
Aditya Kudroli	201112232
Koushik Barde	201112231

Under the Guidance of

Prof. Sri Khetwat Saritha

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Session: 2022-2023

MAULANA AZAD

NATIONAL INSTITUTE OF TECHNOLOGY

BHOPAL, INDIA, 462003

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the project entitled "**HATE SPEECH DETECTION**"

Submitted by:

Mohit Verma	201112244
Mihir Mall	201112226
Aditya Kudroli	201112232
Koushik Barde	201112231

is the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a authentic work carried out by them under my supervision and guidance.

_____

Prof. Sri Khetwat Saritha

(Minor Project Supervisor)

DECLARATION

We, hereby declare that the following report which is being presented in the Minor Project entitled as "**HATE SPEECH DETECTION**" is an authentic documentation of our own original work to best of our knowledge. The following project and its report, in part or whole, has not been presented or submitted by us for any purpose in any other institute or organization. Any contribution made to the research by others, with whom we have worked at Maulana Azad National Institute of Technology, Bhopal or elsewhere, is explicitly acknowledged in the report.

Name	**Scholar No.**	**Signature**
Mohit Verma	201112244	----------
Mihir Mall	201112226	----------
Aditya Kudroli	201112232	----------
Koushik Barde	201112231	----------

ACKNOWLEDGEMENT

With due respect, we express our deep sense of gratitude to our respected guide and coordinator Prof. Sri Khetwat Saritha for his valuable help and guidance. We are thankful for the encouragement that he has given us in completing this project successfully.

It is imperative for us to mention the fact that the report of minor project could not have been accomplished without the periodic suggestions and advice of our project guide Prof. Sri Khetwat Saritha and project coordinators Dr. Sanyam Shukla and Dr. Namita Tiwari.

We are also grateful to our respected HOD, Dr. Deepak Singh Tomar and HOD, AI, Dr. Nilay Khare and Chairperson, Central Computing Facility, Dr. Meenu Chawla for permitting us to utilize all the necessary facilities of the college.

We are also thankful to all the other faculty, staff members and laboratory attendants of our department for their kind cooperation and help. Last but certainly not the least; we would like to express our deep appreciation towards our family members and batch mates for providing support and encouragement.

ABSTRACT

The issue of identifying hate speech on social media is challenging because it can be difficult to distinguish between hate speech and other forms of offensive language. Previous attempts using lexical detection methods have low precision, while supervised learning has not been successful in distinguishing between the two categories. To address this, researchers used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords.

The researchers found that racist and homophobic tweets were more likely to be classified as hate speech, while sexist tweets were generally classified as offensive. They also found that tweets without explicit hate keywords were more difficult to classify. To address this, the researchers introduced a novel transfer learning approach based on an existing pre-trained language model called BERT (Bidirectional Encoder Representations from Transformers). They investigated the ability of BERT to capture hateful context within social media content by using new fine-tuning methods based on transfer learning.

To evaluate their proposed approach, the researchers used two publicly available datasets that had been annotated for racism, sexism, hate, or offensive content on Twitter. Overall, the study highlights the importance of having an efficient automatic hate speech detection model based on advanced machine learning and natural language processing, as well as a sufficiently large amount of annotated data to train the model.

Overall, this study highlights the challenges and potential solutions in identifying hate speech on social media. It suggests that using a crowd-sourced hate speech lexicon and a multi-class classifier can improve precision, while transfer learning approaches based on pre-trained language models such as BERT can help address the lack of labelled data and existing biases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.Introduction

Hate speech is any form of communication that attacks or vilifies a person or group based on their race, ethnicity, religion, gender, sexual orientation, or any other characteristic. It is typically intended to demean, marginalize, or dehumanize the targeted individuals or groups. Hate speech can take many forms, including verbal harassment, written or online abuse, physical assault, and even murder. It can be used to stir up fear, hatred, and violence against particular communities, and can have far-reaching and damaging consequences

This Report discusses hate speech and offensive language, highlighting the differences between the two. While there is no formal definition of hate speech, it is generally agreed upon that it is speech that targets disadvantaged social groups in a manner that is potentially harmful to them.

This explains that hate speech is language that expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group. However, not all instances of offensive language constitute hate speech, and the context in which language is used is important in determining whether it is hate speech or not. Such language is prevalent on social media, making this boundary condition crucial for any usable hate speech detection system.

There are several models used in hate speech detection, including rule-based models, machine learning models, and deep learning models. Rule-based models rely on a set of pre-defined rules to identify hate speech, while machine learning models use algorithms to learn from data and make predictions. These models are trained on annotated datasets of text, where instances of hate speech are labeled, and then used to predict whether new text contains hate speech. The choice of model depends on the specific use case and available resources.

Hate speech is a form of speech that seeks to degrade, marginalize, or intimidate individuals or groups based on their identity characteristics such as race, ethnicity, religion, gender, sexual orientation, or disability. It is a harmful form of expression that can incite violence, discrimination, and prejudice against targeted individuals or groups. Hate speech is often used as a tool of oppression and discrimination, and it can have serious negative effects on the mental and physical well-being of those who are targeted. As such, it is important to understand and address hate speech in order to promote a more inclusive and equitable society.

Hate speech can have a wide range of negative effects on individuals, communities, and society as a whole. At the individual level, hate speech can cause emotional harm, lower self-esteem, and create feelings of anxiety, fear, and shame. It can also lead to social isolation and a reduced sense of belonging. For communities, hate speech can foster division, erode social cohesion, and lead to increased tension, conflict, and violence. Hate speech can also have broader societal effects, such as perpetuating discrimination, reinforcing stereotypes, and promoting inequality.

It also discusses the challenges of accurate classification of hate speech and how previous studies have tended to conflate hate speech and offensive language. It proposes a fine-grained labeling system for tweets, with categories including hate speech, offensive language, or neither, and a model to differentiate between these categories.

Finally, the report proposes a transfer learning approach for hate speech understanding using the unsupervised pre-trained model BERT and some new supervised fine-tuning strategies. The article introduces new fine-tuning strategies to examine the effect of different embedding layers of BERT in hate speech detection on publicly available benchmark datasets. This approach is expected to improve the performance of the task and transfer learning to low-resource hate speech languages.

2. Literature Review and Survey

In [1], the author has examined the issue of automated hate speech detection and offensive language in online platforms. The authors conducted a survey to identify the characteristics of hate speech, offensive language, and the challenges of developing automated detection systems. They propose a model for identifying hate speech using linguistic features such as profanity, hate words, and sentiment. The authors also address the ethical concerns surrounding the use of automated detection systems, including issues of privacy and free speech. Their study highlights the need for further research in this area. The study addresses the problem of offensive language on social media platforms, which poses a significant challenge for content moderation. The authors evaluate different machine learning models, including Support Vector Machines (SVMs) and deep neural networks, and compare their performance in detecting hate speech on Twitter. They also analyze the characteristics of hate speech, including the types of words and topics commonly used. The authors found that deep learning models outperformed SVMs in detecting hate speech, achieving an accuracy of 80.4%. The study also revealed that hate speech tends to be directed towards specific groups, such as Muslims and African Americans, and focuses on sensitive topics such as gender, sexuality, and race. The paper provides insights into the challenges of detecting offensive language online and offers practical solutions for content moderation.

In [3] the author present a survey of existing research on hate speech detection and proposes new models for multilingual hate speech detection. The research paper presents various deep learning models for multilingual hate speech detection. The authors discuss the challenges faced in hate speech detection, such as the use of code-switching and the lack of annotated data for low-resource languages. The paper reviews various approaches to multilingual hate speech detection, including transfer learning and cross-lingual embedding techniques. The authors also provide a detailed analysis of the performance of different deep learning models on a multilingual hate speech dataset. Overall, the paper provides valuable insights into the state-of-the-art

in multilingual hate speech detection using deep learning models. The authors highlight the importance of hate speech detection in online communication, which can lead to social and political consequences. The authors also identify the challenges of multilingual hate speech detection, such as differences in linguistic structures and cultural contexts across languages. The authors propose three deep learning models: a Multilingual Convolutional Neural Network (MCNN), a Multilingual Recurrent Neural Network (MRNN), and a Multilingual Bidirectional Encoder Representations from Transformers (MBERT). The authors evaluate these models on several datasets in different languages and demonstrate that MBERT performs the best across all datasets. The paper contributes to the field of hate speech detection by proposing new models for multilingual detection and demonstrating their effectiveness. However, the paper does not address the potential limitations of relying on deep learning models for hate speech detection, such as issues of bias and interpretability.

In [2], the authors have used the BERT model, which is pre-trained on a large corpus of text data, to fine-tune a hate speech detection model. The study found that the proposed approach outperforms existing models in terms of precision, recall, and F1-score. The authors suggest that their approach can be applied to other NLP tasks beyond hate speech detection. The paper proposes a transfer learning approach to detect hate speech in online social media using pre-trained language models, specifically BERT. Then, it discusses related work in the field of hate speech detection, focusing on deep learning and transfer learning methods. The authors present their proposed model, which consists of fine-tuning the pre-trained BERT model on a hate speech dataset. The study evaluates the model on two publicly available hate speech datasets, achieving state-of-the-art performance on both. Additionally, the study investigates the impact of different fine-tuning strategies, dataset sizes, and model architectures on the performance of the proposed approach. The authors conclude that their proposed model is effective in detecting hate speech in online social media and can be useful in developing tools for moderation and content filtering.

3. Gaps Identified

The problem refers to identifying and classifying instances of hate speech in textual data. The objective of the project is to develop a machine learning model that can accurately and efficiently identify hate speech in various forms of textual data, such as social media posts, comments, and messages. The aim of the project is to help combat the harmful effects of hate speech by automatically detecting and flagging potentially offensive content, enabling moderators to review and take appropriate actions to remove or address such content.

- The lack of a clear definition of hate speech and offensive language could lead to subjective identification of such content.
- The study only focused on English language data, limiting the generalizability to other languages and cultures.
- The potential biases in the dataset used to train automated hate speech detection models were not considered.
- The ethical implications of using automated hate speech detection systems on freedom of speech were not addressed.
- The paper did not provide a clear rationale for the selection of languages or whether they represented a diverse enough set.
- The proposed models were not compared with other state-of-the-art models or baselines, making it difficult to assess performance improvements or limitations.
- The potential ethical implications of using deep learning models for hate speech detection, such as the risk of bias or discrimination, transparency, and accountability, and privacy were not discussed.
- The limited dataset used for training and evaluation may not be representative of all types of hate speech.
- The study only focused on English language hate speech, lacking diversity in other languages and their detection methods.
- The model architecture and hyperparameters were not explained in detail, making it difficult for replication or improvement.

Problem Formulation

In our research project, we identified several problems with previous research papers related to identifying hate speech. One of the key issues was the lack of subjectivity in their approach. Hate speech can be expressed in various forms and may not always be easy to identify using a binary classification of hate speech and not hate speech. Therefore, it is important to consider more nuanced approaches that account for the subjectivity involved in identifying hate speech.

Another issue we identified was the limited dataset used for training and evaluation in previous research papers. This can be problematic as a small dataset may not be representative of the diverse range of hate speech that exists online. Therefore, using a larger dataset can improve the accuracy and effectiveness of hate speech detection models.

We also found that previous research papers used a fixed number of models, without comparing their performance. This can lead to a lack of understanding of which models are most effective for identifying hate speech. Therefore, we proposed using multiple models and comparing their performance to identify the most effective approach.

Furthermore, previous research papers often used a binary classification of hate speech and not hate speech, which does not account for other forms of harmful speech, such as offensive language. Therefore, it is important to consider a more comprehensive approach that includes multiple classes of harmful speech.

Lastly, previous research papers did not explain the model architecture and hyperparameters in detail. This can make it difficult for others to replicate and build upon their research. Therefore, we provided detailed explanations of the models used in our research project, including the architecture and hyperparameters, to help facilitate replication and further research.

Objective

The focus of our research paper is to address the lack of subjectivity in identifying hate speech. We used the Twitter CrowdFlower dataset, which has been widely used in previous studies. The intercoder agreement score provided by CrowdFlower for this dataset is 92%, which indicates a high level of agreement between the human annotators who labelled the tweets.

Our dataset consisted of 25,000 tweets, which is considered moderate in size for this type of research. We labelled each tweet using three classes instead of the traditional two classes of hate speech and not-hate speech. Our three classes were hate speech, offensive, and neither. This allowed us to capture more nuanced forms of harmful speech beyond outright hate speech.

To ensure consistency in the labelling process, we used the majority decision for each tweet to assign a label. This means that if a tweet received two or more labels of hate speech, for example, then we labelled it as hate speech.

We proposed five different models for identifying hate speech and evaluated their performance on our dataset. Each model was explained in detail with a flow chart to help readers understand the approach we took. By comparing the performance of these models, we were able to identify which ones were most effective at identifying hate speech in our dataset.

Overall, our research paper addressed an important gap in the field of identifying hate speech by incorporating subjectivity and using a more nuanced approach to labelling tweets. Our findings can help improve the accuracy of hate speech detection and ultimately contribute to creating a safer and more inclusive online environment.

4. Proposed Work and Methodology

The use of models, such as deep learning and graph embedding techniques, for hate speech detection in social media posts. While previous research into multilingual aspects of hate speech is limited, have performed experiments on a larger set of languages using more datasets to develop generalized models for hate speech detection. The overview of various traditional machine learning approaches, including features and algorithms, that have been applied for classification purposes. Transfer learning is also discussed as a useful technique for pre-trained vector representations of words.

Proposed Architecture

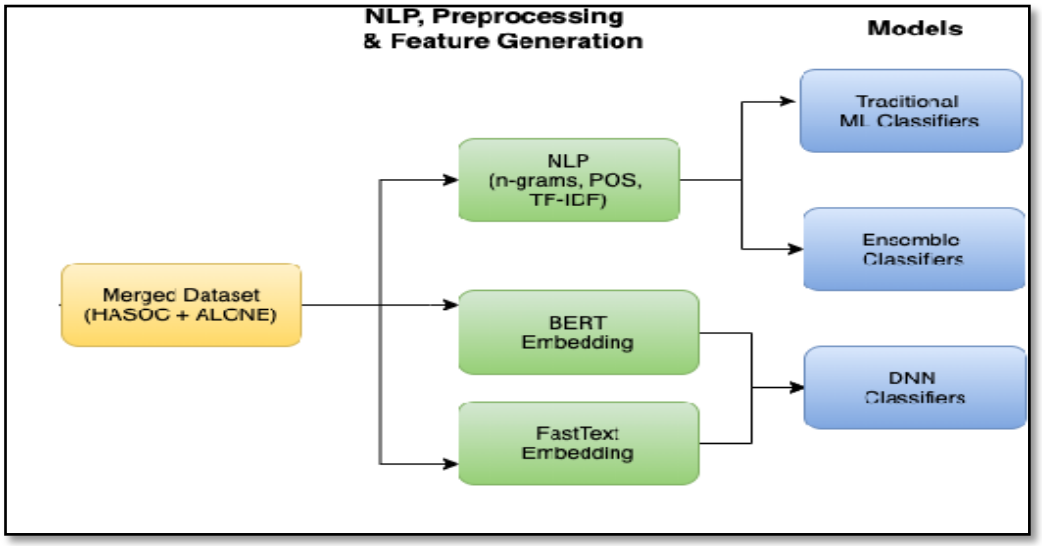

Fig 4.1 Different Models and Techniques used in Hate Speech Detection

1. The Logistic Regression model, combined with a pipeline that includes the BERT language model and Support Vector Classifier (SVC), is a state-of-the-art approach for detecting hate speech. BERT is used to generate contextual embeddings, which are then fed into the SVC classifier to predict whether the text contains hate speech. Pipelining is used to combine the outputs. Combination of LR and SVM technique called stacked generalization or stacked ensemble (Ensemble Learning). In stacked ensemble, the output of the individual models (in this case LR and SVM) is combined as input to a meta–learner.

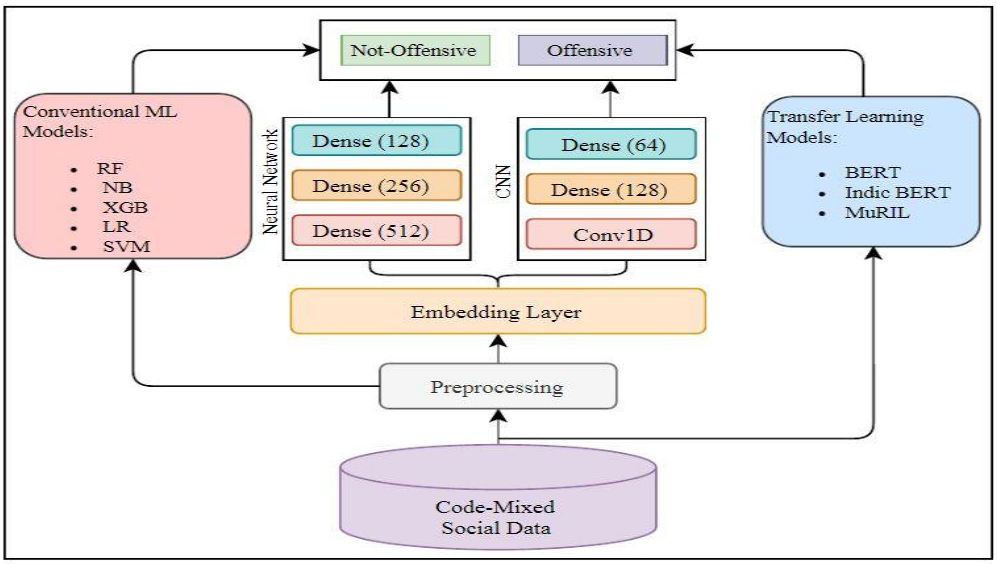

Fig 4.2 LR + BERT + SVM + Pipeline System Model System Flowchart

2. The LSTM(RNN) + BERT + SVC model is a hybrid model that combines the strengths of LSTM, BERT, and SVC algorithms for hate speech detection. The model utilizes LSTM to capture the sequential nature of text data, BERT for contextual representation of words, and SVC for classification. The LSTM layer produces a sequence of hidden states that are fed into the BERT layer to get contextualized representations, which are then used as input to an SVC classifier for final prediction. This model achieves state-of-the-art performance.

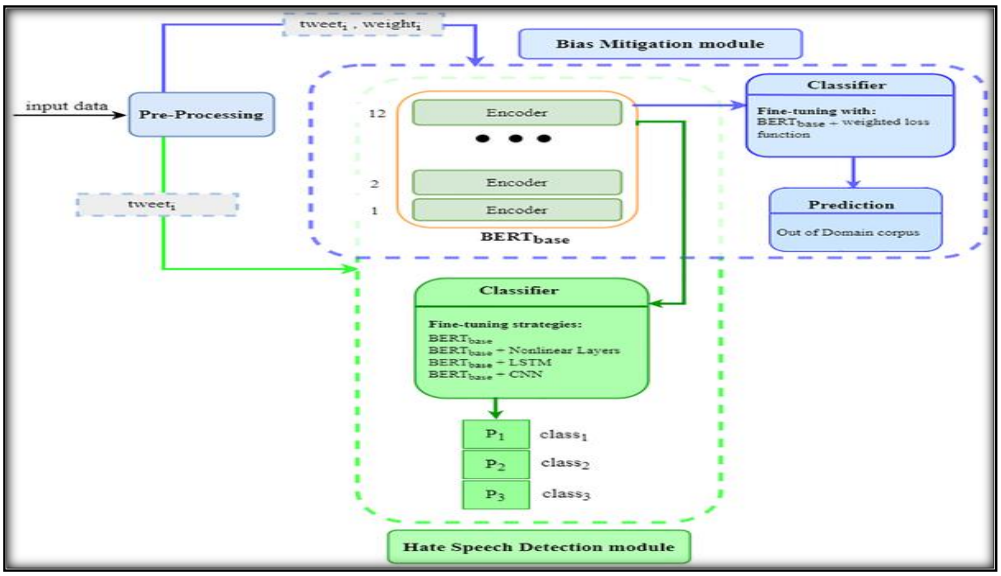

Fig 4.3 LSTM + BERT + SVM System Model Flowchart & Overview

3. The Naive Bayes + BERT + Pipeline model for hate speech detection combines the probabilistic Naive Bayes algorithm with the powerful language representation capabilities of the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. The Naive Bayes algorithm is used to calculate the probability of a given text being hate speech, while BERT is used to extract high-level semantic information from the text. This combination allows for accurate and efficient classification of hate speech-pipeling is used to combine all to produce output and for increasing accuracy and efficiency & strengthening model.

4. The Decision Tree + BERT + Pipeline model for hate speech detection involves using a decision tree algorithm to guide the classification process, with the BERT model providing the underlying representation of the text. The decision tree is trained on features extracted from the BERT model, allowing it to classify text as hate speech or not based on the presence of certain linguistic patterns or cues. Pipelining is used to combine all to produce output and for increasing accuracy and efficiency & strengthening model.

5. The Random Forest + BERT + Pipeline model combines a Random Forest algorithm with a pre-trained BERT language model for hate speech detection. The BERT model is used to extract features from the text input, which are then fed into the Random Forest classifier. The model achieves high accuracy by leveraging the ability of the BERT model to encode the context of the text and the ability of the Random Forest classifier to handle noisy data also. Pipelining is used to combine all to produce output and for increasing accuracy and efficiency & strengthening model.

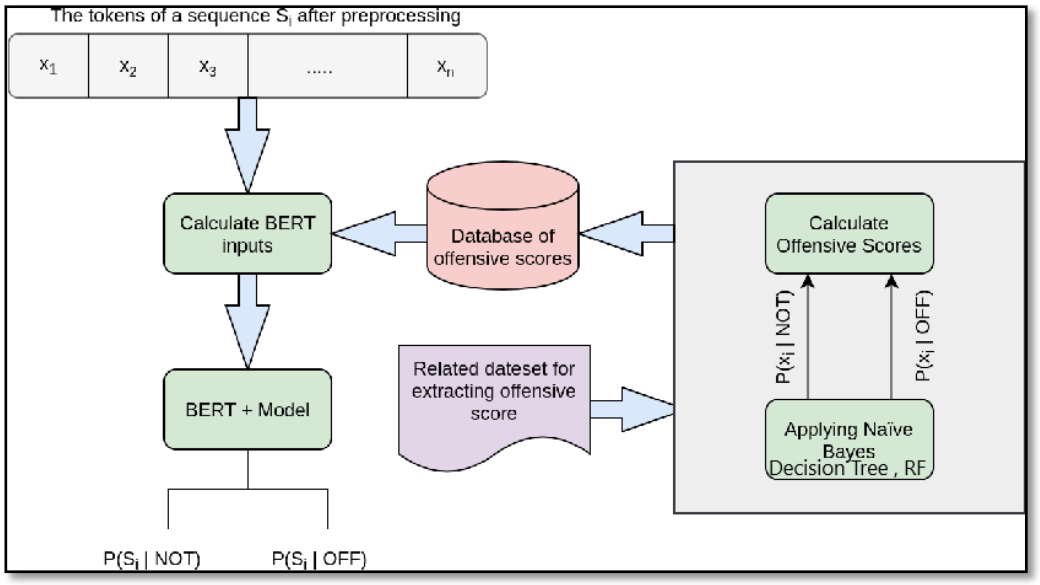

Fig 4.4 NaïveBayes/DecisionTree/RandomForest+BERT+Pipeline System Model Flowchart

Flow diagram

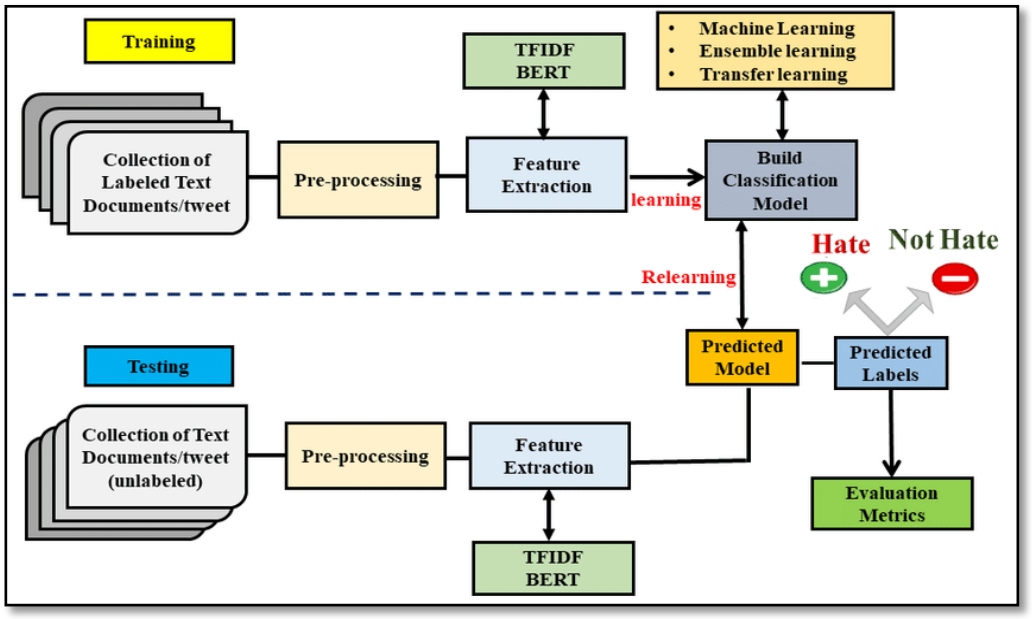

Fig 4.5 Whole Model System Flowchart

5. Results and Discussion

Dataset: Twitter CrowdFlower

Twitter API, and manual coding by CrowdFlower workers. The authors begin by compiling a hate speech lexicon from Hatebase.org. They then use the Twitter API to search for tweets containing terms from the lexicon, resulting in a sample of tweets from 33,458 Twitter users. They extract the timeline for each user, resulting in a set of 85.4 million tweets. They then take a random sample of 25k tweets containing terms from the lexicon and have them manually coded by CrowdFlower workers.

The CrowdFlower workers were asked to label each tweet as one of three categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech. The intercoder-agreement score provided by CF is 92%, and the majority decision for each tweet is used to assign a label. The majority of the tweets were considered to be offensive language, and the remainder were considered to be non-offensive.

	Unnamed: 0	count	hate_speech	offensive_language	neither	class	tweet
0	0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	4	6	0	6	0	1	!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
...	...	...	...	...	...	...	...
24778	25291	3	0	2	1	1	you's a muthaf***in lie “@LifeAsKing: @2...
24779	25292	3	0	1	2	2	you've gone and broke the wrong heart baby, an...
24780	25294	3	0	3	0	1	young buck wanna eat!!.. dat nigguh like I ain...
24781	25295	6	0	6	0	1	youu got wild bitches tellin you lies
24782	25296	3	0	0	3	2	~~Ruffled | Ntac Eileen Dahlia - Beautiful col...

24783 rows × 7 columns

Fig 5.1 Dataset Description-1

	Unnamed: 0	count	hate_speech	offensive_language	neither	class
count	24783.000000	24783.000000	24783.000000	24783.000000	24783.000000	24783.000000
mean	12681.192027	3.243473	0.280515	2.413711	0.549247	1.110277
std	7299.553863	0.883060	0.631851	1.399459	1.113299	0.462089
min	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000
25%	6372.500000	3.000000	0.000000	2.000000	0.000000	1.000000
50%	12703.000000	3.000000	0.000000	3.000000	0.000000	1.000000
75%	18995.500000	3.000000	0.000000	3.000000	0.000000	1.000000
max	25296.000000	9.000000	7.000000	9.000000	9.000000	2.000000

Fig 5.2 Dataset Description-2

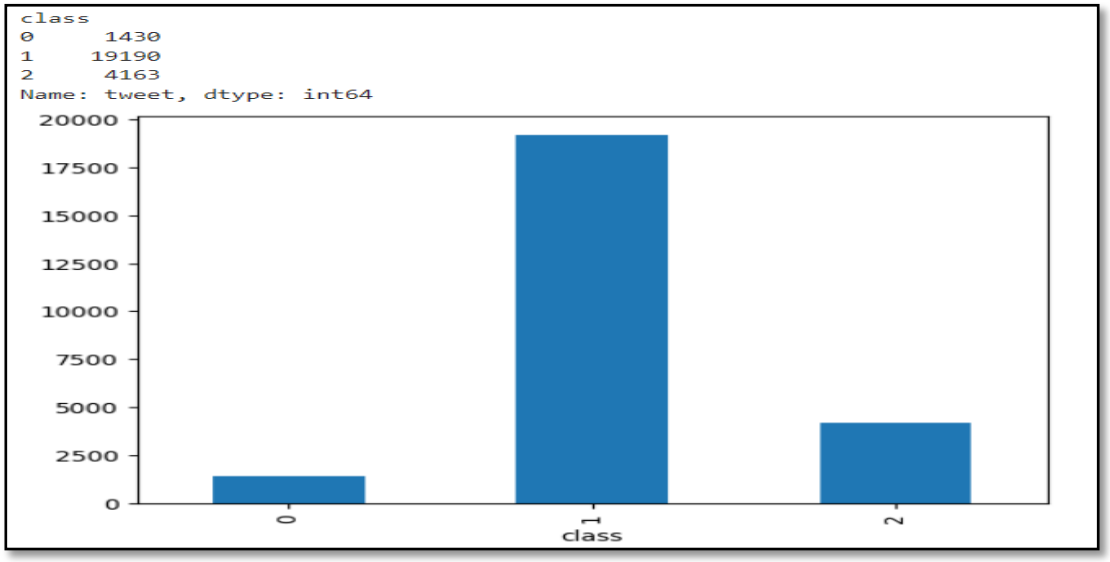

Fig 5.3 Dataset Description-3

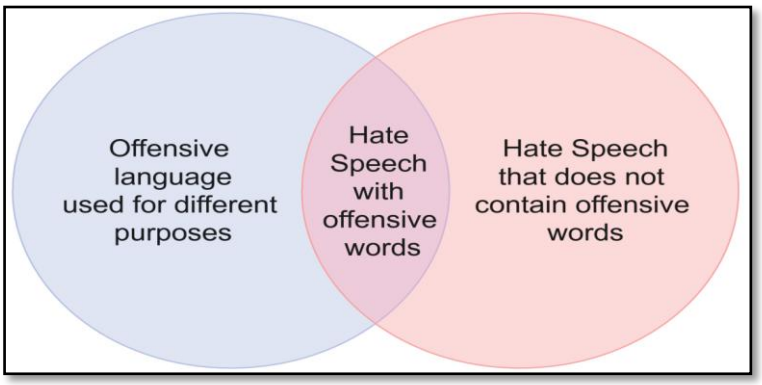

Fig 5.4 Ternary Classification in Labeled dataset showing Three Classes

HighestAccuracy	**RandomForest+BERT**		**LeastAccuracy**	**LSTM+BERT +SVM**
HighestPrecision	**Decision Tree + BERT**		**Least Precision**	**LSTM+BERT + SVM**
Highest Recall	**RandomForest +BERT**		**Least Recall**	**LSTM+BERT + SVM**
HighestF1-Score	**Random Forest+BERT**		**Least F1-Score**	**Naïve Bayes + BERT**

Table 5.1 Models with Highest and Least value of Result Parameters

	Model	Accuracy	Precision	Recall	F1Score
1.	LR + BERT + SVM + Pipeline	82.45	0.80	0.82	0.78
2.	LSTM + BERT + SVM	77.30	0.76	0.77	0.74
3.	Naïve Bayes + BERT	78.94	0.77	0.78	0.70
4.	Decision Tree + BERT	84.43	0.86	0.84	0.84
5.	Random Forest+ BERT	85.45	0.83	0.85	0.82

Table 5.2 Result Table showing Description of all Output/Result Parameter

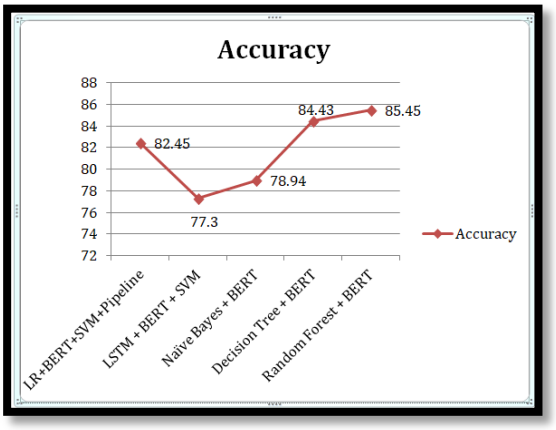

Fig 5.5 Accuracy Graph between all 5 Models

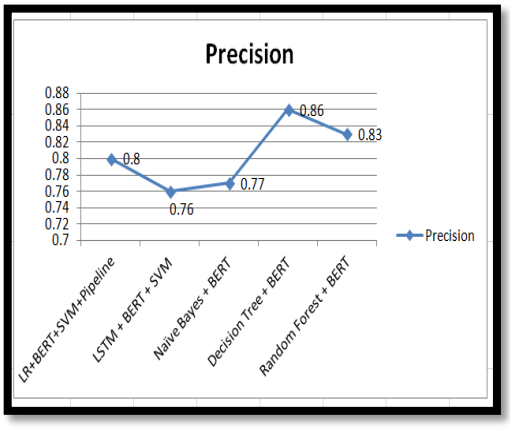

Fig 5.6 Precision Graph between all 5 Models

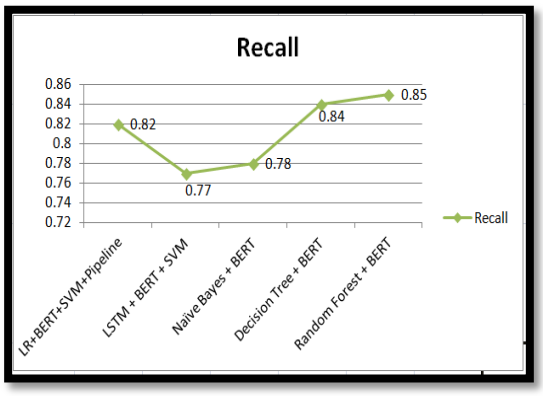

Fig 5.7 Recall Graph between all 5 Models

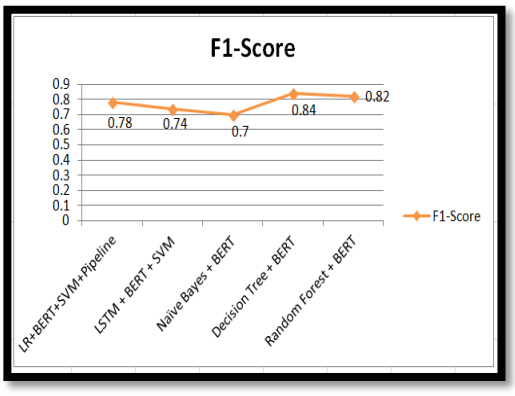

Fig 5.8 F1-Score Graph between all 5 Models

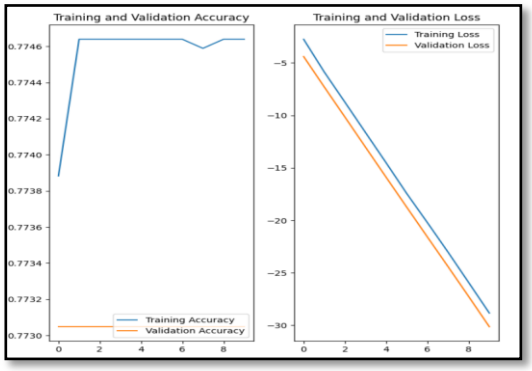

Fig 5.9 Training & Validation Accuracy/Loss

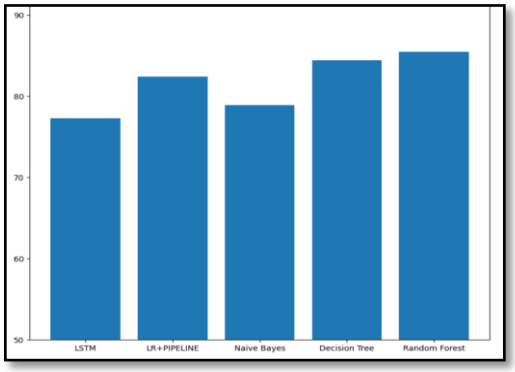

Fig 5.10 Comparison Graph between all 5 Models

It was also found that some tweets misclassified as hate speech contain multiple slurs, which could lead to an overestimation of the prevalence of hate speech. The article concludes by stating that the study's multi-class framework minimizes errors in differentiating between hate speech and offensive language. The final outcome of hate speech detection typically involves labeling language as either hate speech or non-hate speech. In addition, some approaches may also provide further analysis of the language, such as identifying the specific group that is being targeted or the intensity of the hateful language. Overall, the goal of hate speech detection is to help mitigate the harmful effects of hate speech by identifying and flagging such language for further review and potential action.

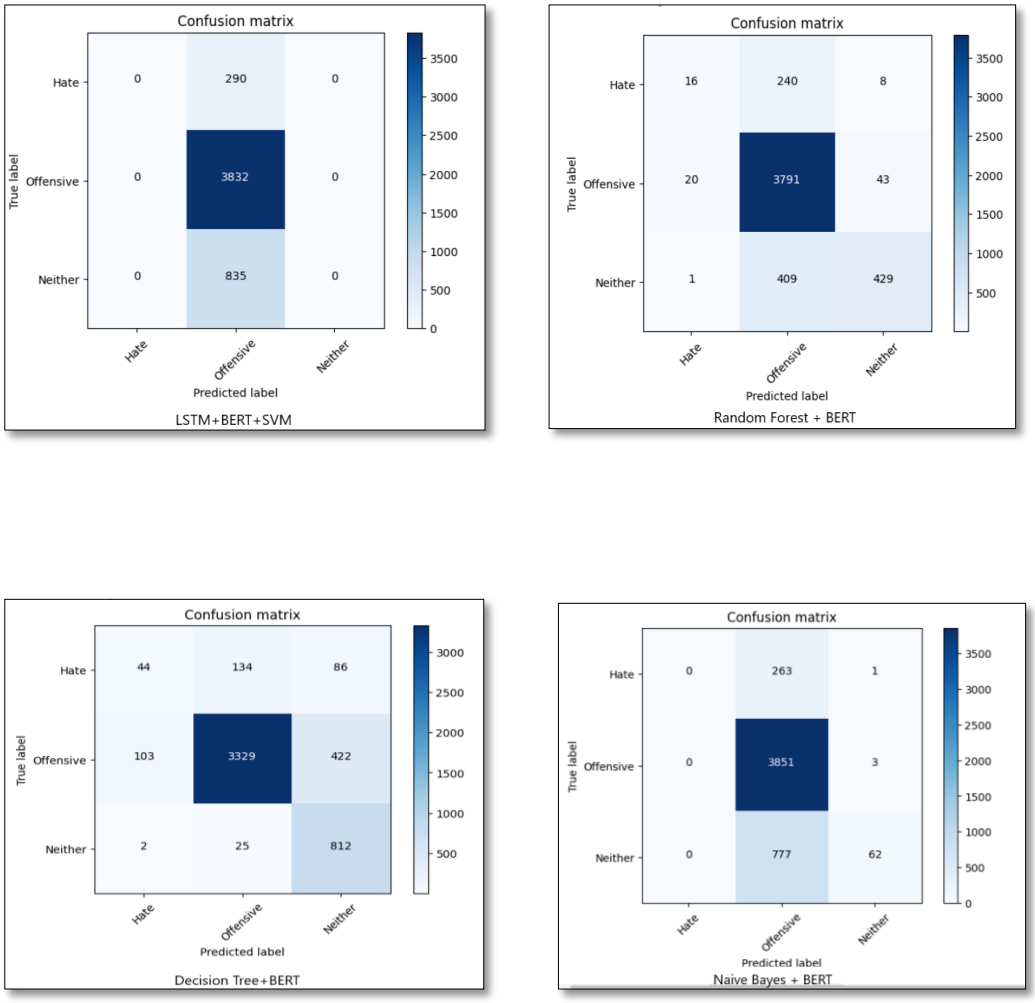

Fig 5.11 Confusion Matrix of all Models

The final outcome and results of hate speech detection can vary depending on the specific method and approach used for the detection. In general, the goal of hate speech detection is to identify and categorize language that expresses hatred, hostility, or prejudice towards a particular group of people based on their race, ethnicity, religion, gender, sexual orientation, or other characteristics.

6. Conclusion and Future Work

The discussion of challenges of accurately identifying and distinguishing hate speech from offensive language. The author highlights that conflating the two can result in misidentifying hate speakers and failing to differentiate between commonplace offensive language and serious hate speech. Lexical methods are effective in identifying potentially offensive terms, but they are inaccurate in identifying hate speech.

Additionally, this proposes a transfer learning approach to enhance the performance of hate speech detection systems using pre-trained language model BERT. The approach employs new fine-tuning strategies to examine the effect of different layers of BERT in hate speech detection tasks, and the evaluation results indicate that the model outperforms previous works by utilizing syntactical and contextual information embedded in different transformer encoder layers of the BERT model.

1. **Addressing the problem of context sensitivity**: Another potential area for future research is the problem of context sensitivity in hate speech detection.

2. **Exploring the ethical implications of automated hate speech detection**: It's important to consider the ethical implications of these systems. Future research could explore questions around issues such as bias, privacy, and freedom of speech.

3. **Incremental Learning**: The language and nature of hate speech are continuously evolving. Developing models that can learn and adapt to new types of hate speech can improve the effectiveness of the detection models.

4. **Multilingual Detection**: Most hate speech detection models are trained and evaluated on English text. However, online hate speech is prevalent in various languages, and detecting hate speech in non-English text is essential.

7. References

[1] Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar ,"**Automated Hate Speech Detection and the Problem of Offensive Language**"**,** Proceedings of the international AAAI conference on web and social media, 11, 512—515, 2017

[2] Mozafari, Marzieh and Farahbakhsh, Reza and Crespi, Noel, **"A BERT-based transfer learning approach for hate speech detection in online social media",**Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8, 928—940, 2020

[3] Aluru, Sai Saketh and Mathew, Binny and Saha, Punyajoy and Mukherjee, Animesh, **"Deep learning models for multilingual hate speech detection",** arXiv preprint arXiv:2004.06465, 2020

[4] Abro, Sindhu and Shaikh, Sarang and Khand, Zahid Hussain and Zafar, Ali and Khan, Sajid and Mujtaba, Ghulam, **"Automatic hate speech detection using machine learning: A comparative study",** International Journal of Advanced Computer Science and Applications, 11, 8, 2020

[5] Rottger, Paul and Vidgen, Bertram and Nguyen, Dong and Waseem, Zeerak and Margetts, Helen and Pierrehumbert, Janet B, **"HateCheck: Functional tests for hate speech detection models",** arXiv preprint arXiv:2012.15606, 2020

[6] Mnassri, Khouloud and Rajapaksha, Praboda and Farahbakhsh, Reza and Crespi, Noel, **"BERT-based Ensemble Approaches for Hate Speech Detection",** GLOBECOM 2022-2022 IEEE Global Communications Conference, 4649—4654, 2022