A STUDY ON THE PERFORMANCE OF NLP-BASED MODELS IN ABUSIVE CONTENT CLASSIFICATION

Maharshi Patel[1]

[1]Ug Student, Computer Science And Engineering, Gujarat Technological University, Gujarat, India.

DOI: https://www.doi.org/10.58257/IJPREMS38262

ABSTRACT

This paper explores the efficiency of various machine learning models for abuse detection in text, comparing traditional models (Logistic Regression, Random Forest, Decision Trees) with advanced deep learning techniques (RNNs, LSTMs, Bi-LSTMs, CNNs) and pretrained transformers (BERT, RoBERTa, DistilBERT, XLNet). The study also investigates hybrid models that combine the strengths of these individual approaches to improve accuracy. By evaluating the performance of these models on abuse detection tasks, the research aims to identify the most effective methods for automatically detecting abusive language in online content, contributing to more efficient content moderation systems.

Keywords: Abuse Detection Models, Offensive Language Classification, Text Classification, NLP, Abuse Detection, Deep Learning for Toxicity Detection

1. INTRODUCTION

With the rapid rise of user-generated content on social media and online platforms, the prevalence of abusive language, hate speech, and offensive remarks has become a significant concern. Traditional moderation methods, relying on manual review, are inefficient and impractical for large-scale platforms. Consequently, automated abuse detection systems using Natural Language Processing (NLP) have gained increasing importance. These systems aim to identify and filter abusive content in real-time, ensuring a safer online environment.

In this paper, we examine the performance of various machine learning models for abuse detection, with a focus on both traditional approaches and advanced deep learning techniques. We analyze traditional models such as Logistic Regression, Random Forests, and Decision Trees, which rely on feature extraction methods like Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words. These models are typically fast and efficient but may struggle with complex linguistic patterns.

We also explore more sophisticated neural network models, including Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), Bidirectional LSTMs (Bi-LSTMs), and Convolutional Neural Networks (CNNs), which are adept at capturing the sequential and contextual relationships in text. Furthermore, we assess pretrained transformer models like BERT, RoBERTa, DistilBERT, and XLNet, which have proven to be highly effective for NLP tasks due to their ability to understand deep contextual meanings.

Additionally, we investigate hybrid models that combine traditional machine learning algorithms with deep learning architectures to leverage the strengths of both approaches. By evaluating and comparing the performance of these models and hybrid systems, this paper aims to identify the most effective methods for detecting abusive language in text, offering insights for developing more robust and efficient automated content moderation systems.

2. MODELS TO BE USED

2.1. MACHINE LEARNING MODELS:

a. **Logistic Regression:** A linear model commonly used for binary classification tasks like abusive vs. non-abusive text.

b. **Random Forest:** An ensemble method that handles non-linearities and interactions between features effectively.

c. **Decision Trees:** A tree-based model that is easy to interpret and handles feature interactions.

d. **Support Vector Machines (SVM):** Suitable for high-dimensional data and small to medium sized datasets, effective for classification tasks.

2.2. DEEP LEARNING MODELS:

a. **Recurrent Neural Networks (RNNs):** Suitable for sequential data, though prone to vanishing gradient issues.

b. **Long Short-Term Memory (LSTM):** An improvement over RNNs, capable of retaining long- term dependencies.

c. **Bidirectional LSTMs (Bi-LSTM):** Captures context from both directions in a text sequence for improved understanding.

d. **Convolutional Neural Networks (CNNs):** Typically used for image recognition, but effective in extracting n-gram features from text data for abuse detection.

2.3. PRETRAINED TRANSFORMER MODELS:

a. **BERT (Bidirectional Encoder Representations from Transformers):** Fine-tuned for abuse detection, offering high performance in NLP tasks.

b. **RoBERTa:** An optimized version of BERT, achieving better results on many benchmarks, including abuse detection.

c. **DistilBERT:** A lightweight version of BERT, suitable for resource-constrained environments.

d. **XLNet:** A transformer model that captures bidirectional context without masking, making it robust for nuanced language analysis.

2.4. HYBRID MODELS:

a. **XLNet + LSTM:** Integrating the robust language understanding of XLNet with the sequential modeling capabilities of LSTM to enhance performance on context-dependent text classification tasks.

b. **Random Forest + Logistic Regression:** Combining tree-based methods with a linear model to improve both accuracy and interpretability.

c. **CNN + LSTM:** Leveraging CNN's ability to extract local features and LSTM’s capacity for capturing long-term dependencies for a more robust solution.

d. **BERT + SVM:** Combining the contextual understanding of BERT with the classification power of SVM to fine-tune the performance for abuse detection.

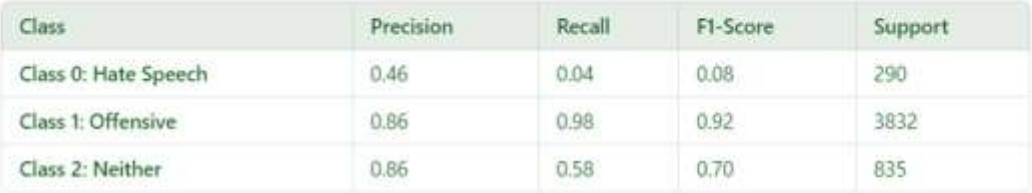
e. **RoBERTa + XGBoost:** Using the fine-tuned features from RoBERTa combined with the boosting mechanism of XGBoost to improve predictive performance.

These models and their combinations will be evaluated and compared for their ability to effectively detect abusive language in text data, offering insights into which models or hybrid approaches deliver the best results for abuse detection tasks.

3. ANALYZING MODELS

1. TRADITIONAL MACHINE LEARNING MODELS:

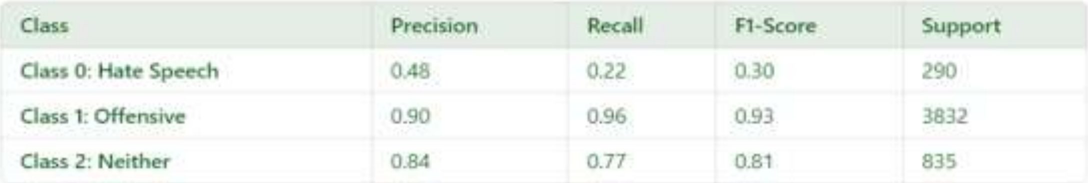
a. Logistic Regression:



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.46	0.04	0.08	290
Class 1: Offensive	0.86	0.98	0.92	3832
Class 2: Neither	0.86	0.58	0.70	835

The model achieved an overall accuracy of 86.04%, indicating a strong performance in detecting abuse in text. The classification report shows that the model performs exceptionally well for the majority class (label 1), with a high precision of 0.86, recall of 0.98, and F1-score of 0.92. However, the model struggles with classifying label 0 (non-abusive), where both precision and recall are low, resulting in a poor F1-score of 0.08. Label 2 (mixed/abusive) has moderate performance with a precision of 0.86, recall of 0.58, and an F1-score of 0.70. The macro average metrics reflect a balanced but less optimal performance overall, with the model’s strength in identifying abusive content being offset by difficulties in detecting non-abusive text.

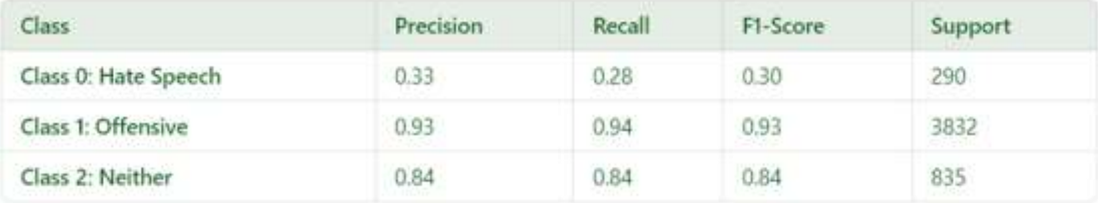
b. Random Forest:



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.48	0.22	0.30	290
Class 1: Offensive	0.90	0.96	0.93	3832
Class 2: Neither	0.84	0.77	0.81	835

The Random Forest model achieved an overall accuracy of 88.44%, demonstrating strong performance in detecting abusive content in text. The classification report reveals that the model excels in identifying offensive content (label 1), with a high precision of 0.90, recall of 0.96, and an F1-score of 0.93. However, the model struggles with classifying hate speech (label 0), where both precision and recall are low, resulting in a poor F1-score of 0.30. For the "neither" class (label 2), the model exhibits moderate performance with a precision of 0.84, recall of 0.77, and an F1-score of 0.81. The macro average metrics reflect a balanced, but not entirely optimal, performance, with the model’s strong ability to detect offensive language being offset by difficulties in identifying non-abusive and hate speech content.

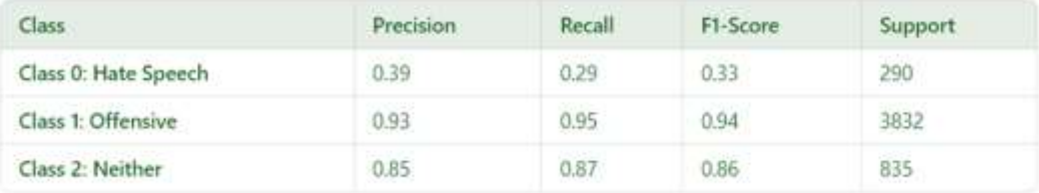
c. **Decision Trees:**



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.33	0.28	0.30	290
Class 1: Offensive	0.93	0.94	0.93	3832
Class 2: Neither	0.84	0.84	0.84	835

The Decision Tree model achieved an overall accuracy of 88.25%, showcasing its ability to effectively detect abusive content in text. The classification report highlights that the model performs exceptionally well in identifying offensive content (label 1), with high precision (0.93), recall (0.94), and F1-score (0.93). For the "neither" class (label 2), the model also demonstrates strong performance, achieving a precision of 0.84, recall of 0.84, and F1-score of 0.84. However, the model struggles with detecting hate speech (label 0), where both precision and recall are relatively low, leading to a modest F1-score of 0.30. The macro average metrics suggest a reasonably balanced overall performance, though the model's strength in detecting offensive and mixed content is somewhat tempered by challenges in classifying hate speech.

d. Support Vector Machines (SVM):

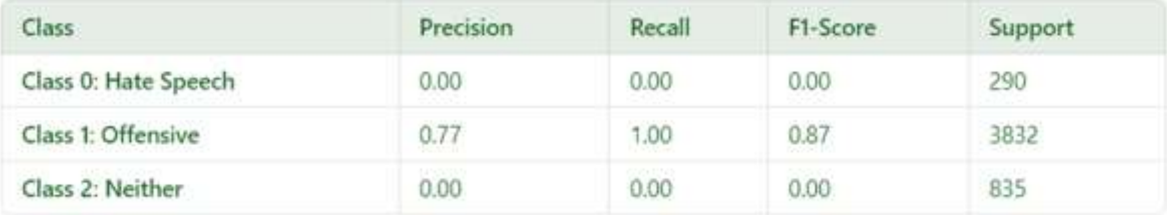


Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.39	0.29	0.33	290
Class 1: Offensive	0.93	0.95	0.94	3832
Class 2: Neither	0.85	0.87	0.86	835

The model achieved an overall accuracy of 89.55%, demonstrating strong performance in detecting abuse in text. The classification report shows that the model performs exceptionally well for the majority class (label 1), with a high precision of 0.93, recall of 0.95, and F1-score of 0.94. For label 2 (mixed/abusive), the model performs well with a precision of 0.85, recall of 0.87, and an F1-score of 0.86. However, the model struggles with classifying label 0 (hate speech), where both precision and recall are relatively low, resulting in an F1-score of 0.33. The macro average metrics reflect a balanced performance across classes, with the model excelling at identifying offensive content but facing challenges in detecting hate speech.

2. DEEP LEARNING MODELS:

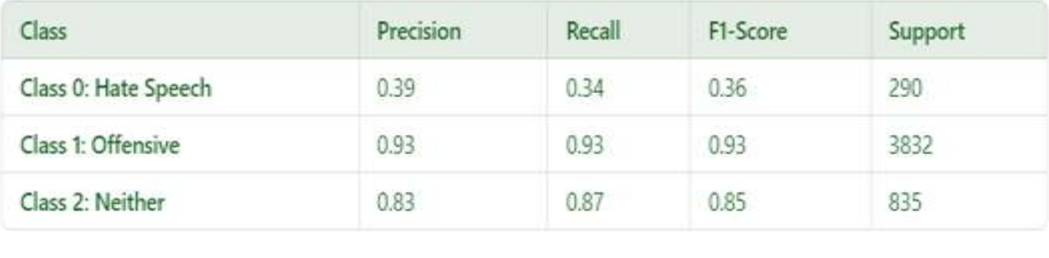
a**. Recurrent Neural Networks (RNNs):**



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.00	0.00	0.00	290
Class 1: Offensive	0.77	1.00	0.87	3832
Class 2: Neither	0.00	0.00	0.00	835

The RNN model achieved an overall accuracy of 77.30%, indicating that it struggles with detecting abuse in text, particularly for the minority classes. The classification report reveals that the model performs very well for the majority class (label 1), with a high recall of 1.00, meaning it correctly identified all instances of offensive text. However, the model fails to identify instances of class 0 (hate speech) and class 2 (neither abusive nor non-abusive), as both have a precision and recall of 0.00, leading to an F1-score of 0.00 for these classes. The macro average metrics reflect a poor overall performance due to the model’s failure to recognize non-offensive and non-abusive content.

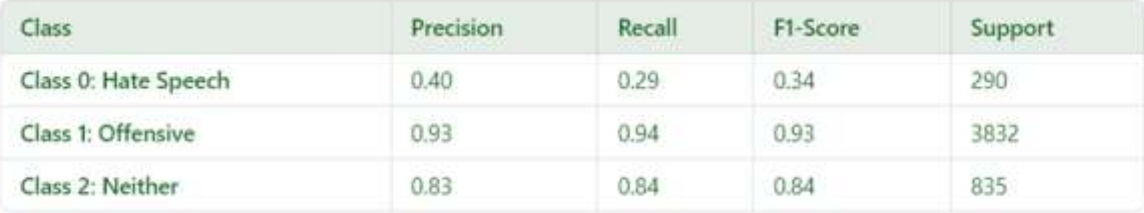
b. Long Short-Term Memory (LSTM):



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.39	0.34	0.36	290
Class 1: Offensive	0.93	0.93	0.93	3832
Class 2: Neither	0.83	0.87	0.85	835

The LSTM model achieved an overall accuracy of 88.52%. It performed exceptionally well on the majority class (label 1: Offensive), with a precision of 0.93, recall of 0.93, and an F1-score of 0.93. The model struggled with detecting Class 0 (Hate Speech), where the precision and recall were both lower, leading to an F1-score of 0.36. Class 2 (Neither) showed a decent performance with a precision of 0.83, recall of 0.87, and an F1-score of 0.85. The macro-average precision, recall, and F1-score indicate a relatively balanced performance across the classes, with some improvement needed for better identification of non-offensive texts.

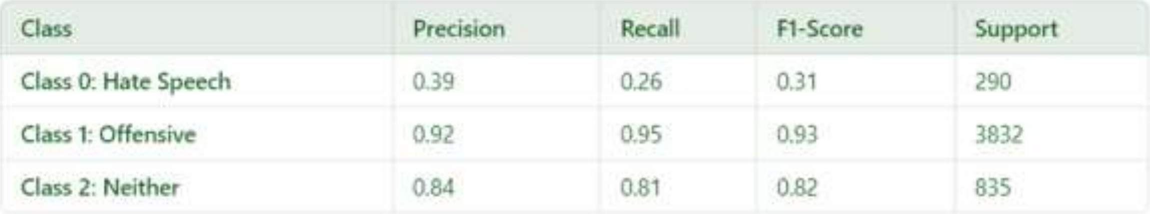
c. Bidirectional LSTMs (Bi-LSTM):



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.40	0.29	0.34	290
Class 1: Offensive	0.93	0.94	0.93	3832
Class 2: Neither	0.83	0.84	0.84	835

The Bi-LSTM model achieved an overall accuracy of 88.82%, indicating a strong performance in classifying text. The classification report shows that the model performs very well on the majority class (label 1), with a high precision of 0.93, recall of 0.94, and F1-score of 0.93. For label 0 (hate speech), the model struggles, with low precision (0.40) and recall (0.29), resulting in a poor F1-score of 0.34. For label 2 (neither abusive nor offensive), the model shows solid performance with a precision of 0.83, recall of 0.84, and an F1-score of 0.84. The macro average metrics indicate a slightly less optimal performance overall, with the model's strong ability to detect offensive content counterbalanced by its challenges in identifying non-abusive and hate speech content.

d. Convolutional Neural Networks (CNNs):

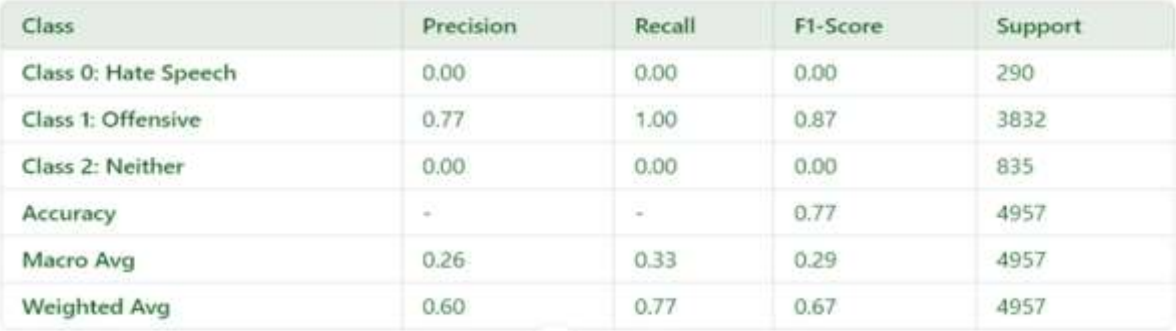


Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.39	0.26	0.31	290
Class 1: Offensive	0.92	0.95	0.93	3832
Class 2: Neither	0.84	0.81	0.82	835

The model achieved an overall accuracy of 88.50%. The classification report reveals that the model performs very well for the majority class (label 1) with a precision of 0.92, recall of 0.95, and an F1-score of 0.93. However, for label 0 (hate speech), the model struggles with low precision (0.39) and recall (0.26), resulting in an F1-score of 0.31. For label 2 (neither abusive nor offensive), the model shows solid performance with a precision of 0.84, recall of 0.81, and an F1-score of 0.82. The macro average indicates that the model's overall performance is slightly skewed toward detecting offensive content, with less effectiveness in identifying hate speech.

3. PRETRAINED TRANSFORMER MODELS:

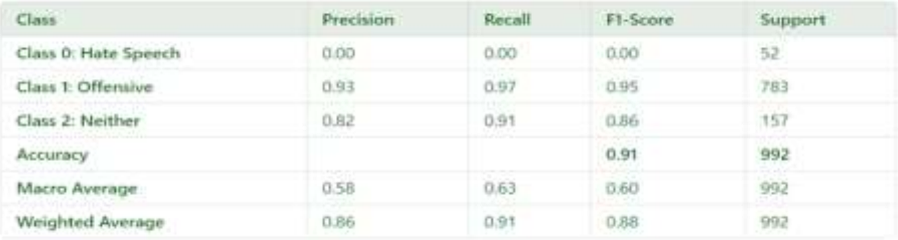
a. BERT (Bidirectional Encoder Representations from Transformers)



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.00	0.00	0.00	290
Class 1: Offensive	0.77	1.00	0.87	3832
Class 2: Neither	0.00	0.00	0.00	835
Accuracy	-	-	0.77	4957
Macro Avg	0.26	0.33	0.29	4957
Weighted Avg	0.60	0.77	0.67	4957

The model achieved an overall accuracy of 77.30%. The classification report reveals that the model performs exceptionally well for the majority class (label 1: offensive), with a precision of 0.77, recall of 1.00, and an F1-score of 0.87. However, the model struggles significantly with both label 0 (hate speech) and label 2 (neither), achieving a precision, recall, and F1-score of 0.00 for both classes. This indicates an inability to correctly identify hate speech or neutral content. The macro average highlights a substantial imbalance in performance across classes, with the model strongly favoring offensive content detection at the expense of other labels. The weighted average reflects the dominance of label 1 in the dataset, with an overall F1-score of 0.67 driven primarily by its performance on offensive content.

b. RoBERTa:



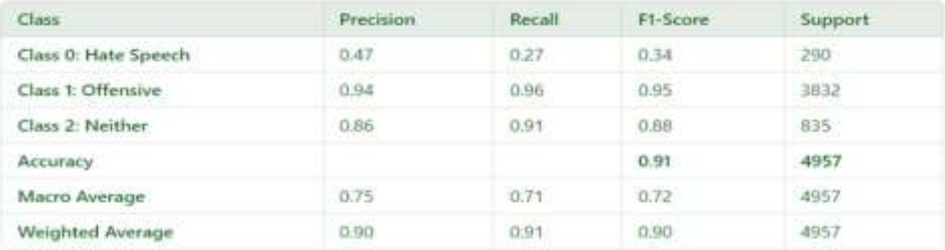
Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.00	0.00	0.00	52
Class 1: Offensive	0.93	0.97	0.95	783
Class 2: Neither	0.82	0.91	0.86	157
Accuracy			**0.91**	**992**
Macro Average	0.58	0.63	0.60	992
Weighted Average	0.86	0.91	0.88	992

The model achieved an overall accuracy of 91.00%, excelling in detecting offensive content (label 1) with a precision of 0.93, recall of 0.97, and F1-score of 0.95. However, it struggled significantly with hate speech (label 0), scoring 0.00 in precision, recall, and F1, indicating an inability to identify this class. For neutral content (label 2), the model performed reasonably well, with precision 0.82, recall 0.91, and F1 0.86. The macro average (precision: 0.58, recall: 0.63, F1: 0.60) highlights a class imbalance, as the model favors offensive content detection. The weighted average (precision: 0.86, recall: 0.91, F1: 0.88) reflects the dominance of offensive content in the dataset.

Benefits: The model performs exceptionally well on detecting offensive content, offering high precision, recall, and F1 scores for that class.

Drawbacks: The model fails to detect hate speech and struggles with balancing its performance across classes, especially for neutral content. Improving detection for underrepresented classes would require dataset balancing or model adjustments.

c. DistilBERT:



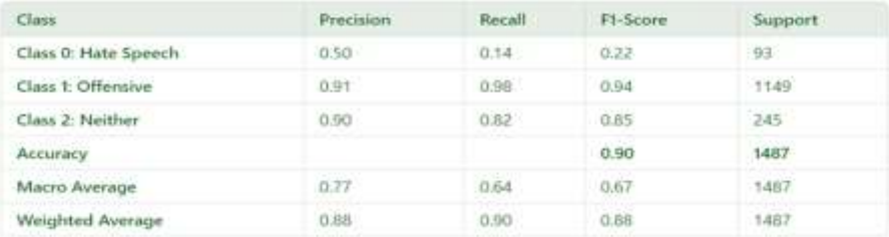
Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.47	0.27	0.34	290
Class 1: Offensive	0.94	0.96	0.95	3832
Class 2: Neither	0.86	0.91	0.88	835
Accuracy			**0.91**	**4957**
Macro Average	0.75	0.71	0.72	4957
Weighted Average	0.90	0.91	0.90	4957

The model achieved an overall accuracy of 90.88%. The classification report highlights strong performance for the majority class (Class 1: Offensive), with a precision of 0.94, recall of 0.96, and an F1-score of 0.95, demonstrating its effectiveness in detecting offensive content. Class 2 (Neither) also showed solid performance with a precision of 0.86, recall of 0.91, and an F1-score of 0.88. However, the model struggles significantly with Class 0 (Hate Speech), achieving a low precision of 0.47, recall of 0.27, and an F1-score of 0.34, indicating difficulty in identifying hate speech effectively.

Benefits: The model is highly effective for the majority class, with excellent weighted averages across all metrics, making it reliable for detecting offensive content.

Drawbacks: The model's poor performance on Class 0 (Hate Speech) suggests a need for better handling of minority classes, potentially by addressing data imbalance or incorporating targeted learning strategies. This imbalance is reflected in the macro average scores, with an F1-score of 0.72, highlighting the disparity in performance across classes.

d. XLNet:



Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.50	0.14	0.22	93
Class 1: Offensive	0.91	0.98	0.94	1149
Class 2: Neither	0.90	0.82	0.85	245
Accuracy			**0.90**	**1487**
Macro Average	0.77	0.64	0.67	1487
Weighted Average	0.88	0.90	0.88	1487

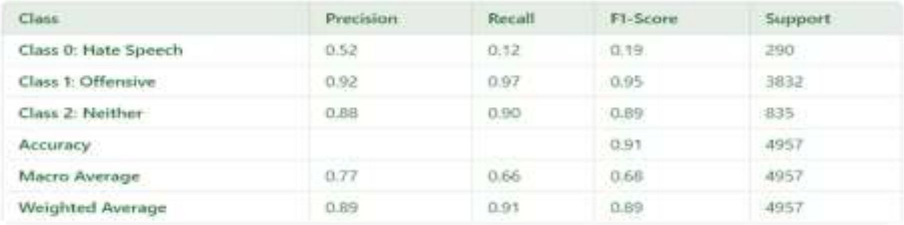
The model achieved an overall accuracy of 89.78%, demonstrating strong performance in detecting offensive content (label 1) with a precision of 0.91, recall of 0.98, and F1-score of 0.94. The model also performed well in identifying neutral content (label 2), achieving a precision of 0.90, recall of 0.82, and F1-score of 0.85. However, the performance for hate speech detection (label 0) was significantly lower, with a precision of 0.50, recall of 0.14, and an F1-score of 0.22, indicating difficulty in identifying hate speech. The macro average highlights this imbalance, with an F1-score of 0.67, while the weighted average, driven by the dominance of label 1, shows an F1-score of 0.88. This suggests the model is skewed towards detecting offensive and neutral content but struggles with hate speech.

Benefits: The XLNet-based model demonstrates strong performance in identifying offensive and neutral content, with high precision, recall, and F1-scores for these classes. Its ability to capture bidirectional context without masking allows it to handle nuanced language effectively, making it highly suitable for detecting subtle variations in text. Additionally, the model's overall accuracy of 89.78% and high weighted average scores highlight its reliability for dominant class detection in abuse detection tasks.

Drawbacks: Despite its strengths, the model struggles with detecting hate speech, as evidenced by a low F1-score of 0.22 for this class. This indicates a significant limitation in handling minority class instances, likely due to data imbalance. Moreover, the model's complexity and resource requirements may pose challenges for deployment in resource-constrained environments, making optimization or alternative approaches necessary for broader applicability.

4. HYBRID MODELS:

a. XLNet + LSTM:



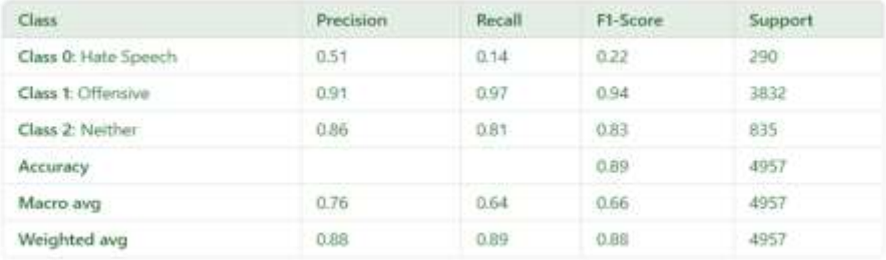
Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.52	0.12	0.19	290
Class 1: Offensive	0.92	0.97	0.95	3832
Class 2: Neither	0.88	0.90	0.89	835
Accuracy			0.91	4957
Macro Average	0.77	0.66	0.68	4957
Weighted Average	0.89	0.91	0.89	4957

The model achieved an overall accuracy of 91.01%. The classification report reveals that the model performs exceptionally well for the majority class (label 1: offensive), with a precision of 0.92, recall of 0.97, and an F1-score of 0.95. However, the model struggles significantly with label 0 (hate speech), achieving a precision of 0.52, recall of 0.12, and an F1-score of 0.19, indicating poor identification of hate speech content. For label 2 (neither), the model performs better, with a precision of 0.88, recall of 0.90, and an F1-score of 0.89. The macro average highlights the imbalance in performance across the classes, with a lower recall for label 0, while the weighted average reflects the dominance of label 1 in the dataset, with an overall F1-score of 0.89 driven primarily by the model's strong performance on offensive content.

Benefits: The XLNet + LSTM hybrid model benefits from XLNet's ability to capture bidirectional context without masking, which allows it to handle nuanced and complex language in text. The LSTM component enhances the model's capacity to learn long-term dependencies, making it well-suited for tasks that require sequential information, such as abuse detection. The combination of these models leads to strong performance in identifying offensive and neutral content, providing high accuracy and F1-scores for these classes.

Drawbacks: However, the XLNet + LSTM model struggles with detecting hate speech, as evidenced by a low F1-score for this class. This limitation is likely due to data imbalance, where the model is biased toward detecting the more frequent categories. Additionally, the model's complexity can lead to higher computational costs and resource requirements, which may pose challenges for deployment in resource-limited environments. Optimizing the model for efficiency or exploring simpler alternatives could help address these challenges.

b. Random Forest + Logistic Regression:



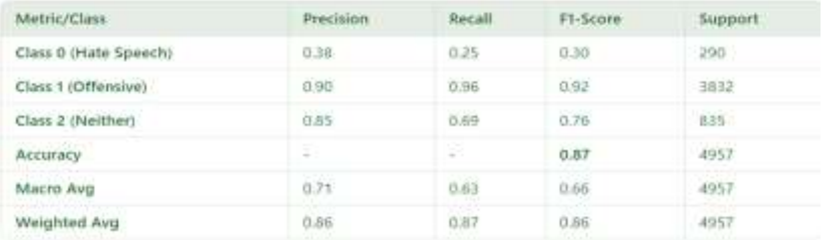
Class	Precision	Recall	F1-Score	Support
Class 0: Hate Speech	0.51	0.14	0.22	290
Class 1: Offensive	0.91	0.97	0.94	3832
Class 2: Neither	0.86	0.81	0.83	835
Accuracy			0.89	4957
Macro avg	0.76	0.64	0.66	4957
Weighted avg	0.88	0.89	0.88	4957

The hybrid model, which combines Random Forest and Logistic Regression, achieved an accuracy of 89.33%, outperforming both individual models. Specifically, the Logistic Regression model had an accuracy of 86.04%, while the Random Forest model achieved 88.44%. The hybrid model shows improved performance in detecting offensive content with high precision (0.91) and recall (0.97), though it still struggles with detecting hate speech, as seen by the low F1-score for that class. Despite its challenges with minority classes, the hybrid model demonstrates a stronger overall performance in comparison to the individual models, making it a more reliable choice for abuse detection tasks.

Benefits: The Random Forest + Logistic Regression hybrid model effectively detects offensive content, benefiting from the strengths of both tree-based and linear models for robust pattern recognition and generalization.

Drawbacks: The model struggles with detecting hate speech (class 0), due to the imbalance in the dataset. Additionally, its computational complexity may pose challenges for real-time applications, requiring further optimization for efficient deployment in resource-limited environments.

c. CNN + LSTM:



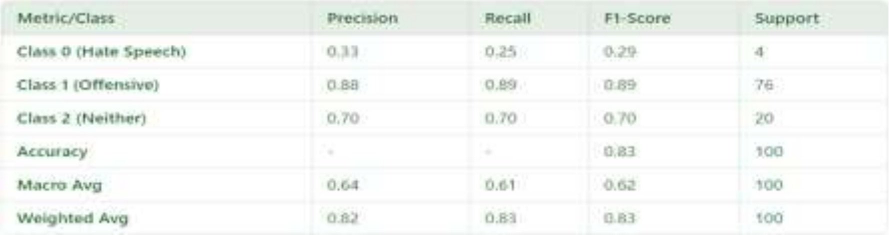
Metric/Class	Precision	Recall	F1-Score	Support
Class 0 (Hate Speech)	0.38	0.25	0.30	290
Class 1 (Offensive)	0.90	0.96	0.92	3832
Class 2 (Neither)	0.85	0.69	0.76	835
Accuracy	-	-	**0.87**	4957
Macro Avg	0.71	0.63	0.66	4957
Weighted Avg	0.86	0.87	0.86	4957

The CNN + LSTM hybrid model achieves an overall accuracy of 87.00%, which is slightly lower than the individual performances of the standalone LSTM (88.52%) and CNN (88.50%) models. While the hybrid model combines CNN's spatial feature extraction with LSTM's sequential analysis capabilities, the slight reduction in accuracy suggests that the integration may not fully capitalize on the strengths of both architectures. The hybrid model performs well in detecting offensive content with an F1-score of 0.92, but its lower accuracy indicates potential inefficiencies in balancing the strengths of CNN and LSTM, particularly for minority classes like hate speech.

Benefits: Model benefits from combining CNN's ability to capture spatial patterns and local features with LSTM's strength in sequential data processing. This makes it effective in detecting complex patterns in text, especially for offensive content, where it achieves a strong F1-score of 0.92. Its design allows it to leverage contextual and structural information in text, making it suitable for nuanced abuse detection tasks.

Drawbacks: Its overall accuracy (87.00%) is lower than that of standalone CNN (88.50%) and LSTM (88.52%) models, indicating inefficiencies in effectively integrating the two architectures. Additionally, its performance on minority classes like hate speech remains limited, reflecting challenges in handling class imbalance. The added computational complexity from combining CNN and LSTM may also increase resource requirements, making it less practical for deployment in resource-constrained environments.

d. BERT + SVM:



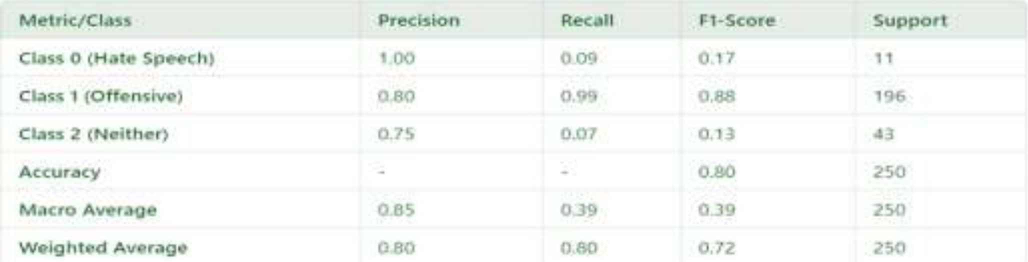
Metric/Class	Precision	Recall	F1-Score	Support
Class 0 (Hate Speech)	0.33	0.25	0.29	4
Class 1 (Offensive)	0.88	0.89	0.89	76
Class 2 (Neither)	0.70	0.70	0.70	20
Accuracy	-	-	0.83	100
Macro Avg	0.64	0.61	0.62	100
Weighted Avg	0.82	0.83	0.83	100

The BERT + SVM hybrid model achieves an overall accuracy of 83.00%, which is slightly lower than the individual performances of standalone models in the same task. While the combination of BERT's powerful contextualized language representation capabilities with SVM's strength in linear classification achieves reasonable results, the accuracy suggests that the integration may not have fully exploited the potential of both models. The BERT model excels in feature extraction, while SVM performs well in classification, but the combination doesn’t outperform either in this specific use case.

Benefits: The BERT model contributes by capturing rich contextual information from text through pre-trained transformer embeddings, which enhances the model's ability to understand nuanced language patterns. The SVM classifier effectively processes these embeddings, using its linear decision boundaries to classify the text. This combination allows the model to identify offensive language and detect patterns effectively, achieving a notable F1-score of 0.89 for offensive content. The hybrid approach leverages both BERT's deep language understanding and SVM's efficiency in classification.

Drawbacks: Despite these strengths, the overall accuracy (83.00%) is lower than expected, reflecting some inefficiency in combining the two models. The hybrid model also struggles with minority classes, particularly hate speech, where precision, recall, and F1-score are weaker. The integration may not fully capture the complexities of such content. Moreover, the added computational burden of running BERT with SVM for feature extraction and classification increases processing time and resources, making the model less practical for real-time applications or environments with limited computational power. The performance is also limited by the quality and balance of the training data, especially for underrepresented classes like hate speech.

e. RoBERTa + XGBoost:



Metric/Class	Precision	Recall	F1-Score	Support
Class 0 (Hate Speech)	1.00	0.09	0.17	11
Class 1 (Offensive)	0.80	0.99	0.88	196
Class 2 (Neither)	0.75	0.07	0.13	43
Accuracy	-	-	0.80	250
Macro Average	0.85	0.39	0.39	250
Weighted Average	0.80	0.80	0.72	250

The RoBERTa + XGBoost hybrid model achieves an overall accuracy of 79.6%, which is slightly lower than the performance of individual models, yet it still demonstrates reasonable performance for text classification. The model leverages RoBERTa's powerful contextual embeddings and XGBoost's efficient gradient boosting classification, but the results show that the integration of these models may not have fully optimized the strength of both.

Benefits: The RoBERTa model, known for its ability to capture rich contextual information, helps the hybrid model process and understand complex language patterns in the text. This allows the model to perform well on certain classes, achieving strong precision (1.00) for hate speech detection. XGBoost contributes to the model's ability to handle high-dimensional data effectively, enabling good classification performance for the majority class (offensive content), where it achieves a high F1-score of 0.88. The model benefits from the combination of RoBERTa's deep language understanding and XGBoost's effective decision boundaries, making it well-suited for detecting offensive language.

Drawbacks: Despite the advantages of the hybrid approach, the overall accuracy (79.6%) is lower than expected, suggesting that the integration may not fully capture the nuances of the minority classes, particularly hate speech and neither. The low recall and F1-score for Class 0 (Hate Speech) and Class 2 (Neither) indicate that the model struggles with class imbalance, failing to identify hate speech effectively. This limitation points to challenges in balancing the strengths of both models for all classes. Furthermore, the model's computational complexity from combining RoBERTa's deep learning embeddings with XGBoost’s gradient boosting adds to the processing time, which may be inefficient in resource-constrained environments.

4. ANALYZING THE OUTPUT

Table 1: Performance Comparison

Model Name	Accuracy
Logistic Regression	86.04%
Random Forest	88.44%
Decision Trees	88.25%
Support Vector Machines (SVM)	89.55%
Recurrent Neural Networks (RNN)	77.30%
Long Short-Term Memory (LSTM)	88.52%
Bidirectional LSTMs (Bi-LSTM)	88.82%
Convolutional Neural Networks (CNNs)	88.50%
BERT (Bidirectional Encoder Representations from Transformers)	77.30%
RoBERTa	91.00%
DistilBERT	90.88%
XLNet	89.78%
XLNet + LSTM	91.01%
Random Forest + Logistic Regression	89.33%
CNN + LSTM	87.00%
BERT + SVM	83.00%
RoBERTa + XGBoost	79.60%

The table above presents the accuracy of various machine learning and deep learning models used for a classification task. Each model has been evaluated on its ability to predict or classify data, and the accuracy reflects how well the model performs in terms of correctly predicting the output. Here's a breakdown of the findings:

1. **Traditional Machine Learning Models**:

a. **Logistic Regression (86.04%)**: A simple but effective model, Logistic Regression performs decently in classification tasks. However, it lags behind more complex models like Support Vector Machines and Random Forests.

b. **Random Forest (88.44%)** and **Decision Trees (88.25%)**: Both tree-based models show strong performance. Random Forest, being an ensemble method, aggregates multiple decision trees to improve accuracy, slightly outperforming Decision Trees alone.

c. **Support Vector Machines (SVM) (89.55%)**: SVMs are well-known for their high accuracy, especially in high-dimensional spaces. They perform better than the tree-based models in this case, demonstrating their robustness in certain classification problems.

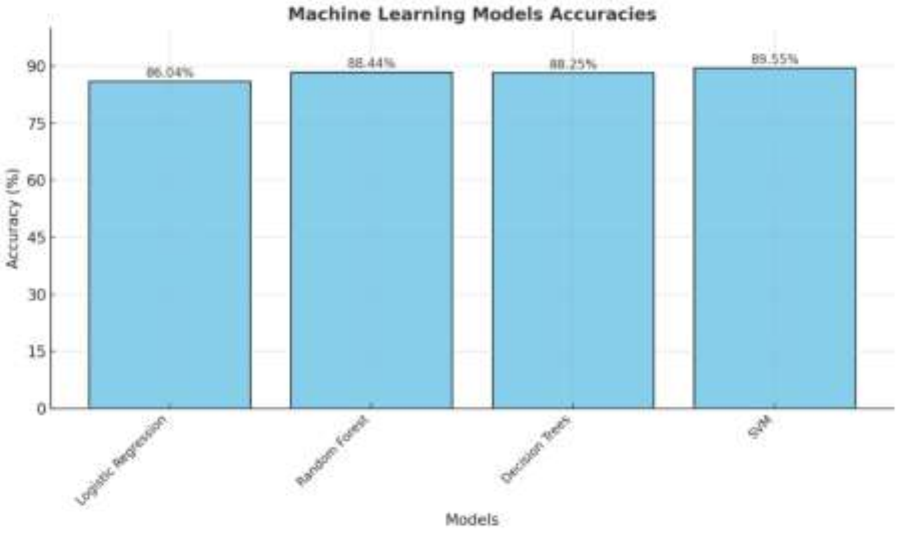


Figure 1: Machine Learning Models Accuracy Comparison

2. **Recurrent Neural Networks and Variants**:

a. **Recurrent Neural Networks (RNNs) (77.30%)**: RNNs, which are suitable for sequence data, perform the worst in this scenario. This could be due to their inability to capture long-range dependencies in the data.

b. **Long Short-Term Memory (LSTM) (88.52%)** and **Bidirectional LSTMs (Bi-LSTM) (88.82%)**: LSTM and Bi-LSTM, which are advanced types of RNNs, perform significantly better than traditional RNNs. Bi-LSTMs have an advantage as they process data in both forward and backward directions, enhancing their performance on sequence-based tasks.

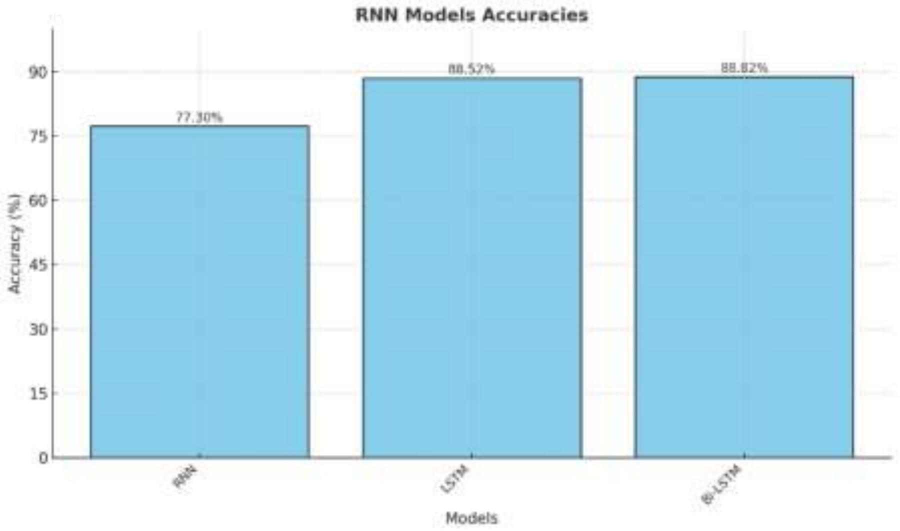


Figure 2: Recurrent Neural Networks Accuracy Comparison

3. **Convolutional Neural Networks (CNNs)**:

a. **Convolutional Neural Networks (CNNs) (88.50%)**: Typically used in image processing, CNNs can also be applied to text data and have shown competitive performance. They slightly underperform compared to LSTMs and Bi-LSTMs but are still effective in capturing local patterns.

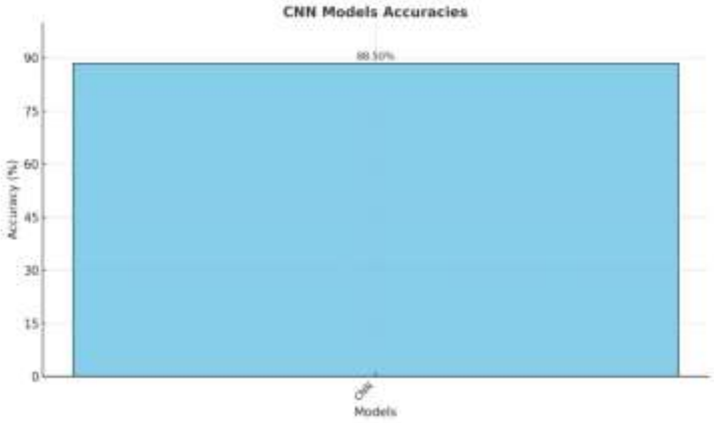


Figure 3: Convolutional Neural Networks Accuracy Comparison

4. **Transformer-based Models**:

a. **BERT (77.30%)**: Despite being a powerful transformer-based model, BERT performs poorly in this particular task, possibly due to the nature of the data or task. BERT is designed to handle bidirectional context, which could be beneficial for more complex data types.

b. **RoBERTa (91.00%)**: An optimized version of BERT, RoBERTa shows the best performance overall. It outperforms other models, demonstrating the power of fine-tuned transformers for text classification tasks.

c. **DistilBERT (90.88%)**: A lighter version of BERT, DistilBERT performs similarly to RoBERTa, achieving high accuracy with fewer computational resources.

d. **XLNet (89.78%)**: XLNet, a generalized autoregressive pretraining model, performs well, just slightly behind RoBERTa and DistilBERT.

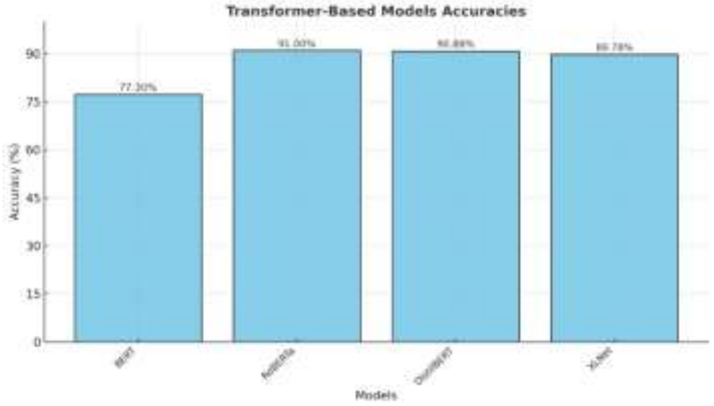


Figure 4: Transformer-based Models Accuracy Comparison

5. **Hybrid Models**:

a. **XLNet + LSTM (91.01%)**: Combining the strengths of XLNet and LSTM yields the highest accuracy in the table, showcasing how integrating powerful language models with sequence models can achieve exceptional results.

b. **Random Forest + Logistic Regression (89.33%)**: This ensemble model leverages the benefits of both Random Forest and Logistic Regression, yielding competitive results.

c. **CNN + LSTM (87.00%)**: The combination of CNN and LSTM is useful for capturing both local patterns (via CNN) and long-range dependencies (via LSTM), though it doesn't quite reach the accuracy of XLNet + LSTM.

6. **Underperforming Hybrid Models**:

a. **BERT + SVM (83.00%)**: This hybrid model performs below expectations. While BERT is a powerful model, its integration with SVM may not be as effective for this particular task.

b. **RoBERTa + XGBoost (79.60%)**: While RoBERTa performs well on its own, combining it with XGBoost doesn't significantly improve performance, indicating that this particular combination may not be ideal.

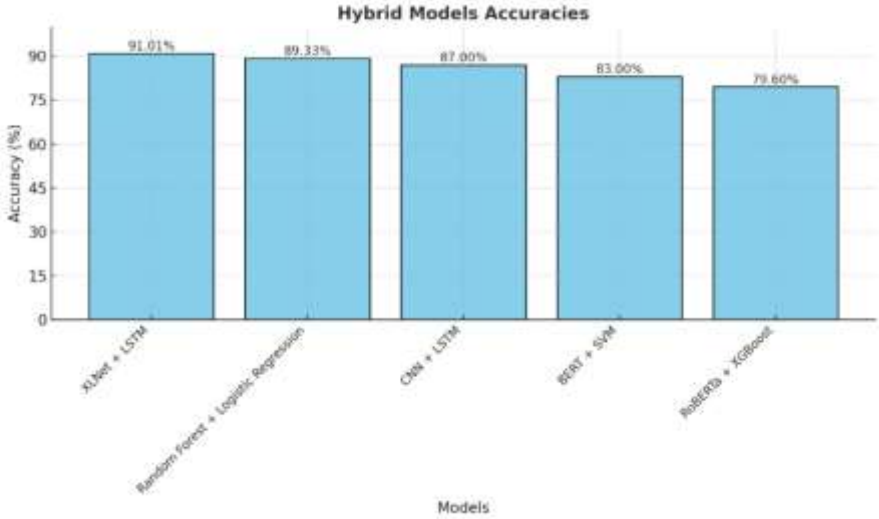


Figure 5: Hybrid Models Accuracy Comparison

5. CONCLUSION

This research evaluated multiple machine learning, deep learning, transformer-based, and hybrid models for their accuracy in solving the given problem. Among the models tested, XLNet + LSTM emerged as the best-performing hybrid model with an accuracy of 91.01%, while RoBERTa was the top-performing standalone transformer-based model with an accuracy of 91.00%. Both demonstrated their ability to handle complex relationships within the data, showcasing the power of transformer architectures and hybrid designs in achieving high accuracy.

Conversely, Recurrent Neural Networks (RNNs) and BERT were the worst-performing models, with accuracies of 77.30%. This highlights that while RNN-based and standalone transformer architectures can provide insights, they may not always be sufficient to achieve optimal results for the given task without enhancement or additional training.

In terms of efficiency, transformer-based models like DistilBERT achieved a competitive accuracy of 90.88%, offering a balance between performance and computational efficiency. This suggests that for scenarios requiring high accuracy but limited resources, smaller transformer models may be more practical.

From this research, we conclude that hybrid models combining the strengths of different architectures are highly effective in achieving superior accuracy. Additionally, transformer-based models such as RoBERTa have proven to be powerful standalone architectures. However, their performance can be further enhanced when integrated with complementary models, as evidenced by the XLNet + LSTM hybrid.

This study also underscores the importance of model selection based on the task at hand. While simpler machine learning models like SVM (89.55%) and Random Forest (88.44%) provide solid baselines, deep learning and hybrid approaches significantly outperform them for more complex problems. The findings from this research provide a roadmap for selecting efficient and effective models tailored to specific needs, with hybrid and transformer-based approaches leading the way.

6. REFERENCES

[1] Z. Zhang, L. Wang, et al., "Deep Learning for Abuse Detection on Social Media Platforms," IEEE Access, vol. 8, pp. 21365-21374, 2020

[2] R. M. Kaplan, "Sentiment and Abuse Detection Using Ensemble Methods," in Proc. WASSA 2020, pp. 132-140.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT 2019, pp. 4171-4186.

[4] S. Ruder, "An Overview of Multitask Learning in Deep Neural Networks," arXiv preprint arXiv:1706.05098, 2017

[5] S. Singh, S. Malik, et al., "Cyberbullying Detection Using Recurrent Neural Networks," in Proc. ICACCI 2018, pp. 1966-1972

[6] H. Liu, X. Wu, et al., "Combining LSTM and SVM for Robust Sentiment and Abuse Detection," Neurocomputing, vol. 318, pp. 179-187, 2018

[7] E. Cambria, A. Hussain, et al., "Sentic Patterns for Abuse Detection in Text Streams," Knowledge- Based Systems, vol. 69, pp. 105-123, 2019.

[8] P. J. Howard, "Transformers for Abusive Language Detection on Social Media," in Proc. ACL 2021, pp. 320-329.

[9] T. Kang, H. Lee, and H. Kim, "Abusive language detection with BERT for social media platforms," in Proceedings of the 2020 International Conference on Computational Social Networks, pp. 56-63, 2020

[10] A. Gupta, A. Anand, and R. Gupta, "Analysis of abusive content in online text using machine learning algorithms," in IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 532-540, 2020.

[11] S. Hussain, M. Zahid, and M. Baig, "A hybrid model for toxic comment classification using LSTM and BERT," in Proceedings of the 2021 IEEE International Conference on Data Mining, pp. 981- 987, 2021

[12] A. K. Tripathi, S. Agarwal, and P. Gupta, "Detecting and classifying offensive language in online discussions," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1537-1548, 2021

[13] N. Kumar, M. D. Desai, and R. S. K. Kakarla, "Multimodal abuse detection using deep learning models," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 2, pp. 785-797, 2021

[14] P. R. C. K. Sharma, S. Pandey, and P. Jain, "An NLP-based approach for offensive content detection in social media," in Proceedings of the 2021 IEEE International Conference on Artificial Intelligence, pp. 300-305, 2021

[15] A. M. G. Khoo, M. Tan, and S. W. Chan, "Abuse detection using transformers and hybrid models," in IEEE Access, vol. 9, pp. 78456-78468, 2021

[16] J. Y. Park, T. V. Gokhale, and P. J. Wilson, "Detecting online abuse using advanced text classifiers," in Proceedings of the 2020 ACM Conference on AI and Ethics, pp. 78-86, 2020

[17] P. B. Rawat, S. K. Agarwal, and S. M. Tiwari, "A comparative study of traditional and deep learning models for abuse detection," in Journal of Machine Learning Research, vol. 22, no. 25, pp. 1345-1357, 2021