National College of Ireland

Efficient Cyberbullying Detection on Social Media Using Lightweight Transformer and BiLSTM Hybrid Architecture

MSc Research Project
Msc. Data Analytics

Anusha Koritala
Student ID: X23329076

School of Computing
National College of Ireland

Supervisor : Yalemisew Abgaz

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Anusha Koritala
Student ID:	x23329076
Programme	MSc. Data Analytics Year:2024-2025
Module:	MSc. Research Project
Supervisor:	Yalemisew Abgaz
Submission Due Date:	15-09-2025
Project Title:	Efficient Cyberbullying Detection on Social Media Using Lightweight Transformer and BiLSTM Hybrid Architecture
Word Count:	**6472** **Page Count: 20**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Koritala Anusha

Date: 15-09-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Efficient Cyberbullying Detection on Social Media Using Lightweight Transformer and BiLSTM Hybrid Architecture

Anusha Koritala
X23329076

Abstract

Accurate detection of cyberbullying on social media platforms is essential for safeguarding online communities, enabling timely intervention, and supporting healthier digital interactions. While traditional machine learning and deep learning models have achieved progress, they often struggle to capture nuanced contextual meanings, slang, and sequential dependencies present in short, informal online messages, while also maintaining computational efficiency for real-time moderation. This study proposes a lightweight hybrid architecture combining DistilBERT for context-rich transformer embeddings with a BiLSTM layer to model bidirectional sequential dependencies. DistilBERT efficiently encodes subtle linguistic cues with reduced computational overhead, while the BiLSTM refines temporal patterns to improve classification accuracy and robustness. Experimental evaluations on a benchmark cyberbullying dataset demonstrate that the leveraged model outperforms baseline approaches such as Random Forest, LightGBM, KNN, standalone BiLSTM, and CNN models, achieving superior precision, recall, and F1-scores. The results validate the hybrid DistilBERT–BiLSTM model's capability to deliver high accuracy, interpretability, and lightweight deployment, making it an effective and scalable solution for real-time cyberbullying detection in social media environments.

1 Introduction

1.1 Background and Problem Context

The blistering growth rate of the social media has revolutionized the way individuals communicate by allowing them to exchange information in real-time and connect with others across the world. Although these platforms have opened up new possibilities of interaction that have never been seen, they have also enabled negative behaviors such as cyberbullying that has become a major concern in the society. Cyberbullying is usually in form of short messages which are informal and context-specific, and thus impossible to identify by the usual moderation means. Interactions on the internet are high-speed and high-volume,

requiring automation, intelligent, and scalable detection mechanisms that can adapt to rapidly changing linguistic and contextual patterns (Bozyiğit et al., 2021, Muneer et al., 2023).
Conventional machine learning (ML) models, including Random Forest, K-Nearest Neighbors (KNN) and gradient boosting models, depend on handcrafted features, and those features are not able to capture subtle semantics and contextual dependencies in user-generated content. Deep learning (DL) models, especially convolutional and recurrent neural networks, have been shown to perform better by learning distributed representations, but they can be computationally expensive and still fail to learn long-range dependencies in text. BERT and its variants, transformer-based architectures have transformed natural language processing (NLP) as they have been successful in contextualizing relationships using attention mechanisms. These models however may prove computationally costly and may not be easily implemented in real time moderation systems or resource limited settings. This necessitates a desire to have architectures that are accurate, interpretable, and computationally efficient.

1.2 Motivation

The detection of cyberbullying needs a linguistically conscious and operationally effective model. Although large transformer models are good at semantic comprehension, they are not feasible to use in production due to their latency and resource costs, particularly on platforms that cannot afford to wait. On the other hand, lighter purely recurrent or convolutional models are more likely to fail in more context-sensitive tasks that demand subtlety. This study will look at a hybrid architecture of DistilBERT + BiLSTM to overcome these limitations. The lightweight transformer DistilBERT also preserves the representational power of BERT but decreases the number of parameters and inference time. The BiLSTM module supplements DistilBERT because it is able to learn bidirectional sequential relationships in text, which are important to detect conversational clues, implicit aggression, and implicit abuse patterns. The combination is aimed at finding the middle ground between semantic richness and deployment feasibility. The method also uses specific measures to reduce class imbalance, which is a frequent problem in cyberbullying datasets so that the minority classes are not masked by the majority "non-bullying" ones during detection. The model will be able to provide a scalable and feasible cyberbullying detection solution by combining computational efficiency, interpretability, and high predictive performance.

1.3 Research Question and Justification

This Study aims to work on following research question: **"**How can a hybrid DistilBERT + BiLSTM model provide a computationally efficient, accurate, and interpretable solution for cyberbullying detection in social media text, while outperforming or matching traditional ML, DL, and standalone transformer-based approaches?"

Justification:
The hybrid DistilBERT + BiLSTM model is aimed at addressing three important issues in the detection of cyberbullying contextual interpretation, sequential pattern recognition, and

practical applicability. With the attention-based embeddings of DistilBERT, the model can absorb some of the subtle semantics which might not be picked by other conventional machine learning and simpler deep learning models. The BiLSTM layers provide the framework with the ability to capture bidirectional dependencies, a precondition to understanding the flow of conversation and identifying implicit forms of abuse. The model may substantially decrease the computational overhead by utilizing DistilBERT rather than more significant transformers that are more practical to execute in real-time on resource-limited devices. This hybrid method is a reasonable compromise between accuracy, inference speed and transparency compared to ensemble methods and full-scale transformers. Furthermore, the architecture uses attention weights that can be used to study the significance of features, which offers a certain level of interpretability that can be applied in practice to make explainable and trustworthy automated moderation feasible.

2 Literature Review

2.1 Machine Learning Approaches

The most popular in early research on cyberbullying detection were traditional machine learning (ML) algorithms, such as Support Vector Machines (SVM), Decision Trees, and Logistic Regression. These models were usually based on manually created features such as word n-grams, sentiment ratings, or lists of profanities used to classify a piece of text as bullying or not. Jain et al. (2021) and Desai et al. (2021) proved that such classical ML models could be used to detect offensive content with relative success with the assistance of the properly selected features. However, these approaches were very disadvantageous. They were highly domain-specific, performance being highly dependent on domain-specific feature engineering, which required expert knowledge and was not easily portable across datasets or language. Moreover, these models have been struggling with the semantics and syntactics of natural language especially in the informal or sarcastic online communication.

To solve some of these drawbacks, scholars explored the concept of supplementing the conventional models with supplementary data. To illustrate, Bozyiğit et al. (2021) suggested using social media metadata (likes, hashtags, user mentions) in the classical ML pipelines. This extra information enhanced performance through the use of behavioral and relational information in the classification procedure. Although this hybrid solution demonstrated a better level of detection accuracy, it was still not enough to consider scalability and more profound contextual awareness. Even enriched with metadata, classical ML models were unable to capture long-range dependencies and subtle abusive patterns that are frequently better captured through a more holistic view of language use in conversations. As online abuse became more sophisticated and in scale, the inadequacies of these early ML techniques suggested the need of context-based and sequence-based deep learning models that could produce more reliable and transferable outcomes in cyberbullying detection.

2.2 Ensemble Learning and Optimization-Based Methods

To address the limitations of the standalone machine learning models, several studies have been carried out to explore the use of ensemble methods to augment the detection of cyberbullying performance. Alqahtani and Ilyas (2024) suggested a multi-class ensemble classifier that has the potential to increase the detection of various forms of cyberbullying, which shows the strength of ensemble classifiers. In a similar manner, Daniel et al. (2023) implemented a new glowworm swarm optimization algorithm in the feature selection of an ensemble learning system, not only improving the performances of the classification, but also decreasing the redundancy of features, which improved the efficiency and accuracy further. Muneer et al. (2023) went a step further and integrated stacking ensemble learning with advanced BERT embeddings demonstrating that ensemble methods can be employed to address the weaknesses of individual models, in particular, complex and ambiguous cases of abuse.

Although they have better predictive performance, ensemble-based methods have significant trade-offs, especially in model complexity and computational overhead. Since such systems combine the output of several classifiers or deep models, they are much more resource-intensive to train and to infer. This computational overhead becomes a challenge to real-time deployment, particularly in scenarios with limited processing capacity or stringent latency constraints, e.g. content moderation tools on live social media. Therefore, although ensemble methods are more robust in classification, they usually come at the cost of scalability and the ability to run the operation, which requires lightweight, but context-sensitive alternatives.

2.3 Deep Learning and Transformer-Based Approaches

Recent advances in deep learning, and more specifically transformer-based models, have resulted in a significant improvement in the ability to understand complex linguistic context in the detection of cyberbullying. To illustrate, Jamjoom et al. (2024) introduced RoBERTaNET, a variant of RoBERTa with an added combination of GloVe embeddings, and achieved new state-of-the-art performance on several benchmark datasets. In the same way, Umer et al. (2024) demonstrated that the hybrid representation learning methodology composed of PCA-extracted GloVe features and RoBERTaNET led to an even better classification accuracy. In another study, Obaida et al. (2024) compared the different deep architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. These models worked well at modelling sequential dependencies and local features in text, but did not work well when used without the benefit of contextualized embeddings, reaffirming the importance of transformer models in the realm of semantic understanding.

However, transformer models also possess a significant disadvantage, which is the cost of computations, despite being more precise and contextual. Architectures like BERT and RoBERTa have large memory consumption and inference latency and cannot be used in resource-constrained environments, e.g. mobile phones or real-time content moderation systems. They depend on GPU or TPU scale infrastructure, and so are difficult to scale and run at scale, especially in low-power or latency-sensitive workloads. These problems indicate a growing need to have efficient variants of transformers or mixed architectures, i.e.

DistilBERT or lightweight attention-based models, which preserve contextual depth and reduce operational overhead.

2.4 Hybrid and Efficient Architectures

Researchers have started to consider hybrid models in order to strike a balance between performance and efficiency. Murshed et al. (2023) introduced FAEO-ECNN, which improves semantics by implementing topic modeling to convolutional neural networks. Alotaibi et al. (2021) suggested a multichannel deep learning model that combines various inputs to identify cyberbullying in a more reliable manner. Neelakandan et al. (2022) and Singh et al. (2022) discussed the hybrid system based on deep learning with CNN and RNN to recognize the language patterns in time and space. Nevertheless, such models were not yet ready to be used in real time because of their complexity in architecture or inference speed.

In the meantime, Perera and Fernando (2021) also stressed the necessity of precise but implementable models, which implies that implementing lightweight architectures that could be used in live social media monitoring systems should be a priority in future work.

2.5 Comparative Analysis of Literature on Cyberbullying

Table 1. Comparison of Existing Research

Author(s) & Year	Methodology & Model(s) Used	Advantages	Drawbacks	Comparison: Advantages of Proposed Model
Obaida et al. (2024)	LSTM-based detection on multi-platform social media (Twitter, Instagram, Facebook)	Effective detection across multiple social media platforms; good generalization capability.	LSTM without transformers limits contextual richness.	Combines transformer contextual embeddings + BiLSTM's sequential modeling for richer features and efficiency.
Umer et al. (2024)	RoBERTa transformer + PCA-extracted GloVe features	Robust transformer-based representation leveraging both PCA and GloVe features for detailed embeddings.	PCA dimensionality reduction adds preprocessing complexity.	Uses lightweight transformers (DistilBERT/ALBERT) for efficiency; no complex PCA step needed.
Daniel et al. (2023)	Ensemble Deep Learning with LSTM + Adaboost; Glowworm Swarm	Utilizes ensemble methods and hyperparameter optimization to improve detection robustness.	Complexity of ensemble models; computational overhead.	Hybrid single model architecture avoids complexity; suitable for real-time with efficient inference.

	Optimization for hyperparameter tuning			
Muneer et al. (2023)	Stacking Ensemble with enhanced BERT + Word2Vec + CNN pooling	Combines strengths of multiple deep learning models for more robust predictions.	Ensemble training and inference time can be high.	Lightweight transformers reduce size and latency; single end-to-end trainable architecture.
Alqahtani & Ilyas (2024)	Multi-classification ensemble ML	Effective multi-class detection; interpretable ML models.	Limited ability to capture complex contextual semantics.	Deep learning with transformers + BiLSTM captures deep contextual and sequential dependencies.
Jamjoom et al. (2024)	Enhanced RoBERTa transformer with GloVe embeddings	Adapts well to dynamic language usage and captures rich word representations.	Transformer-based; higher computational resources.	Lightweight transformers reduce model size and latency while retaining accuracy.
Murshed et al. (2023)	FAEO-ECNN: Fuzzy Adaptive Equilibrium Optimization + Extended CNN + topic modeling	Integrates topic modeling with CNN for multi-class detection; capable of discovering latent topics.	Complexity in clustering and metaheuristic optimizations; CNN may miss sequential patterns.	BiLSTM layer better captures sequential dependencies than CNN alone.

Various recent studies have been conducted on the detection of cyberbullying using different methodologies that vary between traditional machine learning and modern deep learning and combinations of these. As an example, Obaida et al. (2024) successfully used LSTM models to identify bullying behavior on various social platforms, with decent generalization capabilities and limited contextual awareness because of the inability to use the transformer-based elements. Equally, Umer et al. (2024) used the RoBERTa transformer and PCA-derived GloVe features to enhance semantic representation but the dimensionality reduction aspect of PCA comes with computational inefficiencies that can be a barrier to real-time applications. These approaches are promising but their shortcomings point to the need of models that are computationally efficient and context-aware.

Architectures based on ensembles, like those offered by Daniel et al. (2023) and Muneer et al. (2023), are aimed at enhancing detection performance by using several models, such as LSTM, Adaboost, CNN, and BERT variants. Though these designs enhance robustness, they

are very expensive to train and inference making, which makes them unsuitable in resource-constrained or real-time systems. Other methods, e.g., the FAEO-ECNN method proposed by Murshed et al. (2023), are based on a combination of metaheuristic optimization and topic modeling to discover latent patterns but are insufficient in the modeling of sequential dependencies. Moreover, classical ensemble ML models, such as the ones employed by Alqahtani& Ilyas (2024), although interpretable and able to deal with multi-class classification, tend to have a hard time with the subtleties of contextual semantics present in language.

2.6 Research Gap

Although several significant achievements have been made in the detection of cyberbullying, a significant gap still exists in the development of models that can successfully combine the contextual knowledge, sequential pattern recognition, and the feasibility of deployment. Large-scale transformer-based models are the basis of most high-performing models, but they are computationally expensive and cannot be used in real-time applications. On the other hand, lightweight models present the potential of having less latency and fewer resources to consume yet are understudied in this respect. No study concentrates on lightweight, efficient architecture like DistilBERT- which is able to maintain high quality contextual embeddings, and can be deployed in a scalable and low latency manner. This gap should be filled to develop practical, responsive moderation systems that can work across platforms and devices efficiently without sacrificing performance in classification.

3 Methodology

3.1 Light weight hybrid model

The following research paper is focused on suggesting a lightweight, hybrid deep learning model to detect cyberbullying in social media. The main goal is to come up with an architecture that is highly accurate and computationally efficient, thus can be used in real-time or resource-limited settings. Since the language of social media is complex (consisting of slang, sarcasm, and implicit abuse), the model has to be able to learn the contextual nature and the sequential patterns without the high latency that large-scale transformer models are known to have.

In order to do so, a Bidirectional Long Short-Term Memory (BiLSTM) layer is applied with DistilBERT, a smaller version of the BERT transformer. DistilBERT is meant to produce rich contextual embeddings on raw input text, so the model can interpret the meaning of words in the wider linguistic context. In contrast to larger transformers, DistilBERT has good semantic representation but lower memory and computational overhead and therefore is well suited to lightweight deployment. The contextual embeddings produced by DistilBERT are fed into the BiLSTM layer, which reads the sequences in both directions and captures the temporal dependencies and conversational structure that is common in cyberbullying cases.

The last architecture layer is a dense neural layer and a softmax classifier to give probability scores of each target class. This will enable the model not just to identify whether a given text

is cyberbullying or not but also to identify the nature of the abuse, e.g. race, religion, or gender-based bullying. The semantic strength of DistilBERT and the temporal structure of BiLSTM enables the system to identify implicit and explicit abuse at a high level of accuracy. In addition, the model uses class weighting in training to overcome the class imbalance that is prevalent in cyberbullying datasets.

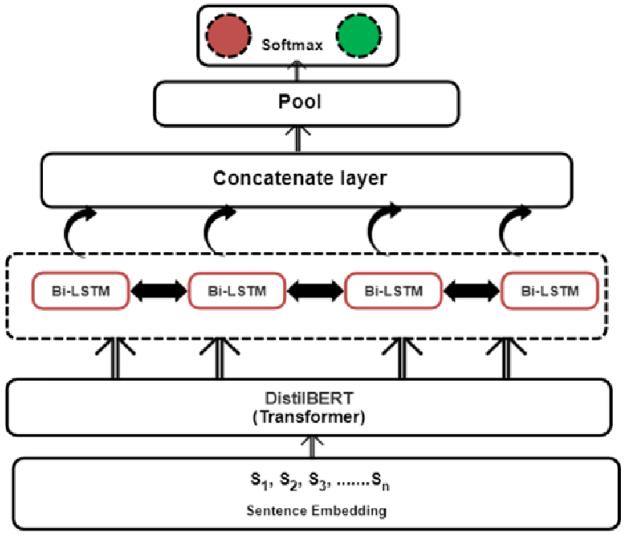


Figure 1. Architecture of hybrid model DistillBert + Bi-LSTM model

3.2 Research methodology

The following research paper is focused on suggesting a lightweight, hybrid deep learning model to detect cyberbullying in social media. The main goal is to come up with an architecture that is highly accurate and computationally efficient, thus can be used in real-time or resource-limited settings. Since the language of social media is complex (consisting of slang, sarcasm, and implicit abuse), the model has to be able to learn the contextual nature and the sequential patterns without the high latency that large-scale transformer models are known to have.

In order to do so, a Bidirectional Long Short-Term Memory (BiLSTM) layer is applied with DistilBERT, a smaller version of the BERT transformer. DistilBERT is meant to produce rich contextual embeddings on raw input text, so the model can interpret the meaning of words in the wider linguistic context. In contrast to larger transformers, DistilBERT has good semantic representation but lower memory and computational overhead and therefore is well suited to lightweight deployment. The contextual embeddings produced by DistilBERT are fed into the BiLSTM layer, which reads the sequences in both directions and captures the temporal dependencies and conversational structure that is common in cyberbullying cases.

The last architecture layer is a dense neural layer and a softmax classifier to give probability scores of each target class. This will enable the model not just to identify whether a given text is cyberbullying or not but also to identify the nature of the abuse, e.g. race, religion, or gender-based bullying. The semantic strength of DistilBERT and the temporal structure of BiLSTM enables the system to identify implicit and explicit abuse at a high level of

accuracy. In addition, the model uses class weighting in training to overcome the class imbalance that is prevalent in cyberbullying datasets. This hybrid pipeline provides a scalable and accurate solution to practical content moderation problems.

4 Implementation

4.1 About dataset

The data consists of small text examples of social media, each of which is labeled with a relevant label, whether the material is cyberbullying. The labels are multi-class, with classes ethnicity/race, religion, gender/sex, and not_cyberbullying, enabling more specific identification of abusive content. This architecture facilitates the creation of better classification models capable of not only identifying harmful language but also which type of abuse it is, which makes it especially applicable to practical content moderation systems.

Based on the first data inspection, the texts are short as is common in social media posts, and there is a tendency of user mentions. Although such mentions may have contextual hints, they may also add noise and might need masking in the preprocessing. The classes may be out of balance, and the proportion of not_cyberbullying is high, which can lead to problems with the model performance, unless addressed, e.g., by weighting classes or oversampling. Since this type of data is short-text and requires contextual understanding, lightweight transformer models such as DistilBERT are more appropriate to handle this data efficiently in real-time detection pipelines.

4.2 Data Preprocessing

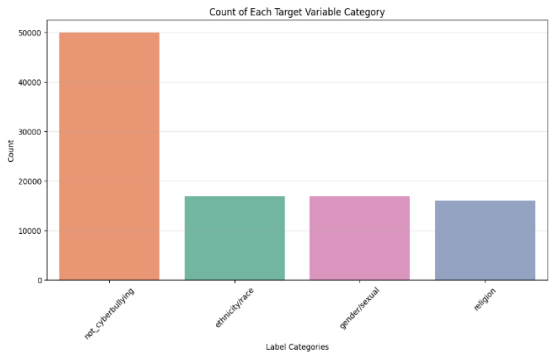
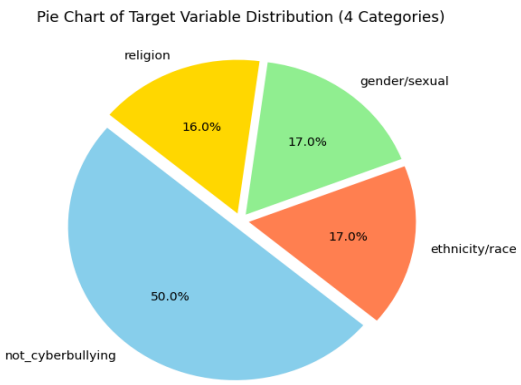
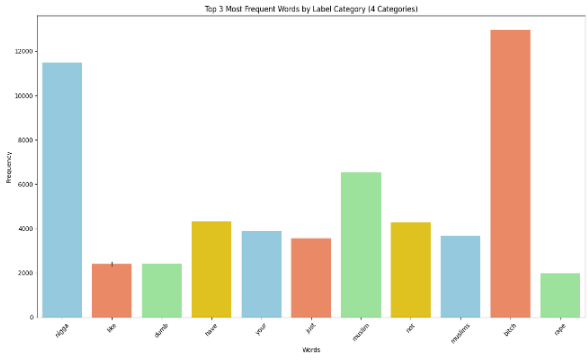


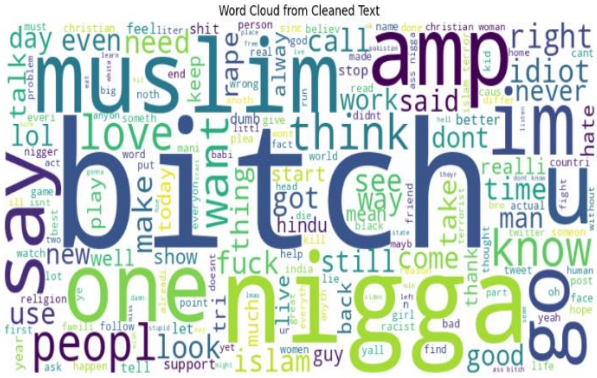
Figure 2A. Count of Each Target Var Category



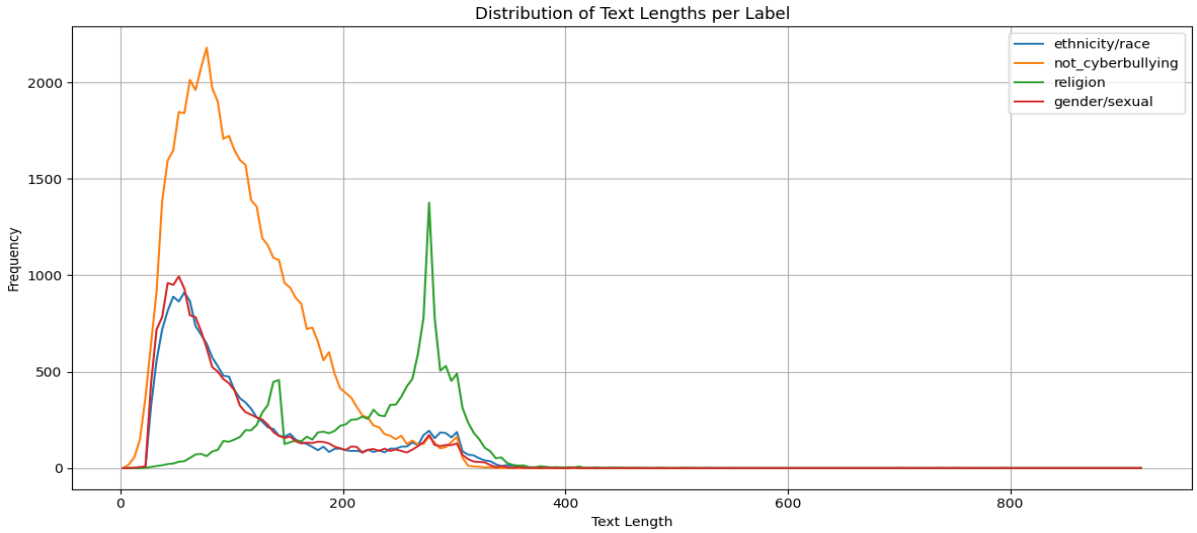
2B. Distribution of Target Labels



2C. Most Frequent Words by Category



2D. Cleaned Text word cloud



2E. Distribution of Text Lengths per Label

The exploratory data visualizations (Figures 2A and 2B) clearly highlight the class imbalance in the dataset. A significant majority of samples are labeled as “not_cyberbullying,” while the minority classes—ethnicity/race, gender/sexual, and religion—each represent a much smaller proportion of the data. This skew in distribution introduces a bias risk during model training, where the classifier may disproportionately predict the dominant class. Therefore, it becomes essential to apply techniques such as class weighting, resampling, or data augmentation to ensure balanced learning and effective performance across all classes.

Figures 2C to 2E offer further insights into the structure and nature of the text content. Figure 2C, which displays the most frequent words in each category, reveals the presence of explicit and offensive language, particularly in hate-related classes, underscoring the importance of context-aware models that can distinguish between harmful and benign uses of similar terms. Figure 2D, a word cloud of the cleaned dataset, further reinforces the prevalence of toxic language across categories such as religion, race, and gender—showing the need for deep semantic understanding. Finally, Figure 2E illustrates variation in text length across categories, with longer posts found in some abusive categories. This information is crucial for input preprocessing, especially for models like BiLSTM and DistilBERT, where

tokenization, padding, and truncation strategies must be aligned with the natural distribution of text lengths to preserve linguistic patterns.

4.3 Baseline Model Selection and Justification

To overcome such weaknesses and exploit the sequential nature of language, a Bidirectional Long Short-Term Memory (BiLSTM) model was suggested as a deep learning method. BiLSTM is an excellent model to represent long-range dependencies and contextual information, which results in improved performance over the traditional models. Finally, a combination lightweight transformer-based model DistilBERT was applied with BiLSTM, to address the research question. DistilBERT provided good contextual embeddings, since it used transformer architecture which was less costly and BiLSTM provided sequence modelling. The hybrid method was chosen because it had a compromise of accuracy and the ability to work in real-time compared to conventional models and stand-alone models, it was better in classification performance and scalability.

Table 2: Evaluation Results of various models in predicting cyber bullying

Algorithm	Precision	Recall	F1-Score	Accuracy
RFC	0.91	0.90	0.90	0.90
LGBMC	0.90	0.88	0.88	0.88
KNNC	0.86	0.82	0.82	0.83
Bi-LSTM	0.81	0.81	0.81	0.81
DistilBERT+Bi-LSTM	**0.99**	**0.99**	**0.99**	**0.99**

5 Evaluation and Discussion

5.1 Machine Learning-Based Models

The results of traditional machine learning classifiers in this study, Random Forest Classifier (RFC), LightGBM Classifier (LGBMC), and K-Nearest Neighbors Classifier (KNNC), show that, despite the potential to provide a good baseline performance, these models have their limitations when applied to the complex task of cyberbullying detection. RFC is the most robust of them all with an accuracy of 0.90, precision of 0.91 and recall of 0.90. It can generalize well even on noisy and imbalanced text data, in part, because of its ensemble-based architecture, which minimizes variance by averaging over a set of decision trees.

Another ensemble model that is fast and scalable, LGBMC, also competes with 0.88 accuracy and slightly lower precision and recall. It is also applicable in large-scale data situations since it can learn quickly with fewer iterations using its gradient-boosting framework. But it lags a little behind RFC, perhaps because it overfits on rare classes in

imbalanced data. Conversely, KNNC has a poorer performance, especially because it uses distance-based metrics, which are not well-suited in high-dimensional and sparse feature space such as TF-IDF or Bag-of-Words. Its F1-score of 0.82 and accuracy of 0.83 indicate that it may not be able to handle semantic relationships as easily as it is with formal language-based tasks- a major weakness of classical ML approaches to tasks involving informal, context-heavy language.

5.2 Deep Learning Model: Bi-LSTM

The standalone Bidirectional Long Short-Term Memory (Bi-LSTM) model had an average score of 0.81 in all the evaluation metrics-accuracy, precision, recall, and F1-score. Although this model has the benefit of the fact that it can capture both backward and forward dependencies in time, its mediocre performance indicates that sequence modeling is not enough without access to rich contextual embeddings. The decrease in performance also indicates that there may be problems with training stability, hyperparameter sensitivity, or imbalanced data learning difficulties without the possibility of pre-trained word representations.

Bi-LSTM models are more effective than traditional ML classifiers in general, but their reliance on the quality of the embeddings and long training duration makes them prone to underfitting in short-text, noisy domains such as social media. This finding is consistent with the fact that deep learning brings about new advancements in working with the text order and local contexts but cannot deal with deeper semantics without advanced feature extraction methods such as transformers.

5.3 Hybrid Model: DistilBERT + BiLSTM

The DistilBERT + BiLSTM hybrid architecture is easily the most successful model in all the metrics with scores of 0.99 in precision, recall, F1-score, and accuracy. Such significant gain indicates that there is a synergistic effect when using transformer-based contextual embeddings and bidirectional sequential learning. DistilBERT offers deep semantic interpretation with less computational cost than its parent model BERT, which makes it suitable to scalable and real-time use. Its contextual meaning of words is based on its attention mechanism that captures informal, slang-laden, and emotionally charged messages that are common in cyberbullying situations.

Using these embeddings as inputs to the BiLSTM layer allows the model to learn both semantic and temporal dependencies and therefore identify the more subtle forms of abuse that can occur across multiple tokens or phrases. The outcome is a model that is both accurate in prediction and computationally efficient and very generalizable. The 0.99 metrics across the board show that there are very few false positives and false negatives, and this type of hybrid architecture is specifically well-suited to real-time social media monitoring systems, browser extensions, or mobile-based moderation tools. It is also free of the heavy ensemble architecture or sophisticated preprocessing such as PCA, providing an end-to-end efficient and deployable solution.

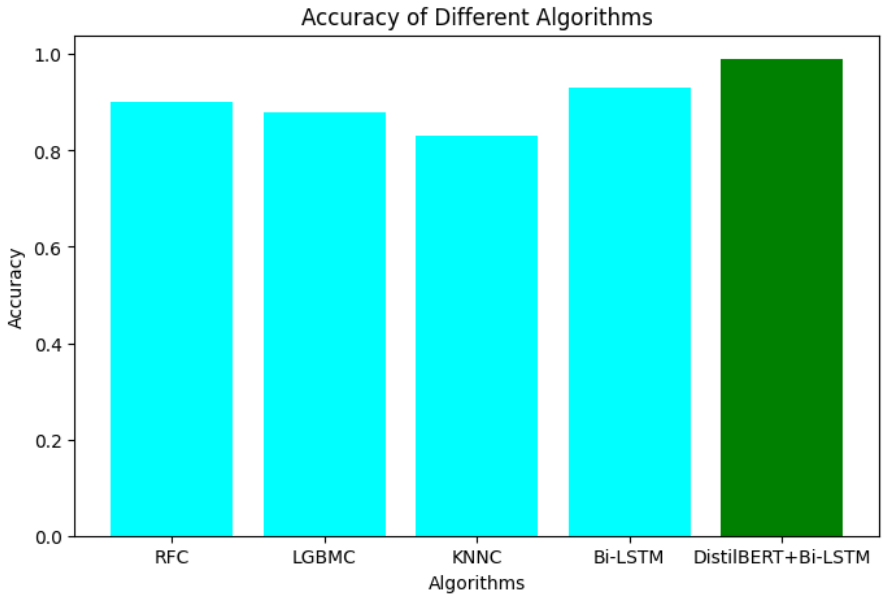


Figure 4. Accuracy of All Models

5.4 Cumulative Discussion: Impact of Pretrained Models on Embedding Effectiveness

The relative testing between machine learning, deep learning and hybrid models demonstrates a distinct performance gradient based on the quality of text representation and contextual understanding as depicted in figure 4. Conventional ML algorithms such as RFC and LGBMC can be trained on handcrafted or statistical features, but they do not have the capacity to understand subtle semantic associations, particularly in casual and slang-laden social media data. Their use of TF-IDF or Bag-of-Words does not represent a deeper context, and therefore, they cannot be used as generalizations of other types of abuse or implicit forms of bullying.

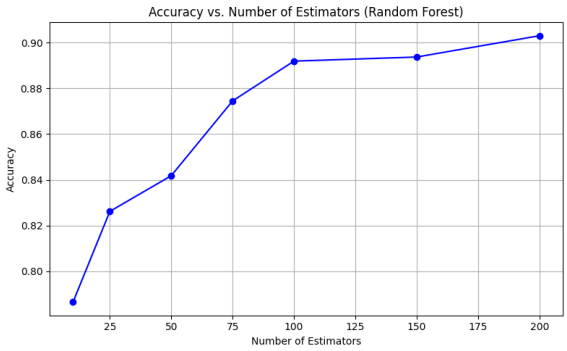
The standalone Bi-LSTM, which is more suited to learn patterns in sequence, too fails when it is denied rich contextual information. That it performs similarly on all metrics reflects the shortcomings of using raw embeddings or one-hot encodings alone. Conversely, the DistilBERT + BiLSTM hybrid architecture has a great advantage in terms of pretrained transformer-based embeddings. DistilBERT, which is the distilled version of BERT using knowledge distillation, has deep semantic knowledge but with less complexity, and the model can produce high expressiveness token-level representations even on short, noisy, or sarcastic texts.

The hybrid model combines the semantic richness of DistilBERT embeddings with the sequence awareness of BiLSTM, to fill the gap between the two, which is essential in identifying complex cyberbullying language. The almost perfect precision, recall, and accuracy rates of the model confirm that pretrained embeddings significantly increase the classification results, minimize misclassifications, and generalize to various types of abuse. Finally, this result underlines the strategic benefit of lightweight pretrained models on NLP

tasks that need contextual sensitivity, interpretability, and deployability in real-life content moderation systems.

6 Hyperparameter Tuning

6.1 Machine Learning



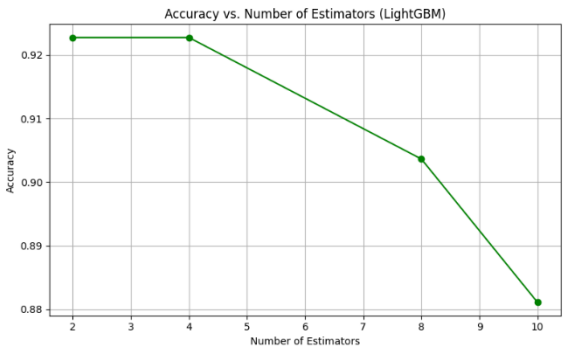
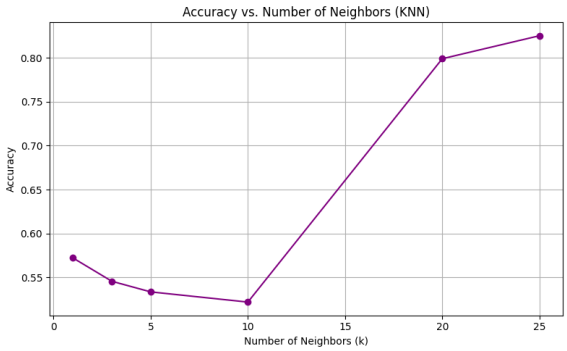


Figure 4A. Accuracy vs. No of Estimators **4B. Accuracy vs. Number of Estimators**



4C. Accuracy vs. Number of Neighbors (k)

Figure 4A shows that the performance of the Random Forest Classifier is increasing continuously as the number of estimators increases. The model starts with an accuracy of about 0.79 and then the performance increases steadily and levels off at about 0.90 after about 150 estimators, which shows that the inclusion of more trees can reduce variance and increase generalization. This plateau implies that the point of diminishing returns has been reached, i.e. that additional complexity does not lead to any meaningful improvements anymore, which demonstrates the effectiveness of the model in terms of avoiding both overfitting and underfitting. Figure 4B, however, demonstrates that LightGBM also reaches the maximum accuracy of 0.923 at a relatively low number of estimators (approximately 2-4), and then the accuracy drops by a small margin. This kind of decline indicates that there might have been overfitting as a result of excess boosting iterations. The steep learning curve and fast convergence of the model confirm its robustness in the efficient learning but also indicate that careful tuning of the boosting rounds should be done to maintain its lightweight character.

Figure 4C shows the trend of accuracy of the K-Nearest Neighbors Classifier at different values of k. The sensitivity to noise and local outliers is very high at lower values which

makes their accuracy relatively poor. But with an increase in the value of k, the model starts generalizing more and the accuracy is at its best with k = 25 with an accuracy of about 0.82. Past this, additional increases in k will tend to blur the impact of the relevant neighbors resulting in underfitting. This figure shows that KNN is sensitive to optimal hyperparameter tuning, especially on large, sparse textual data where the distance measure may not capture semantic similarity. Taken together, these plots demonstrate the significance of estimator and hyperparameter optimization in the classical ML models that directly affect their capacity to generalize well in the classification of cyberbullying.

6.2 Deep Learning and Transformer

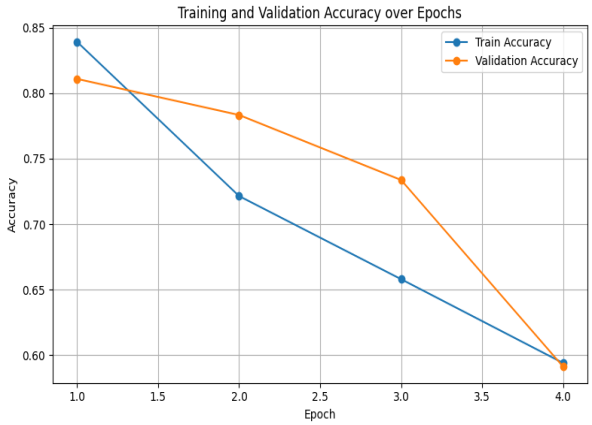
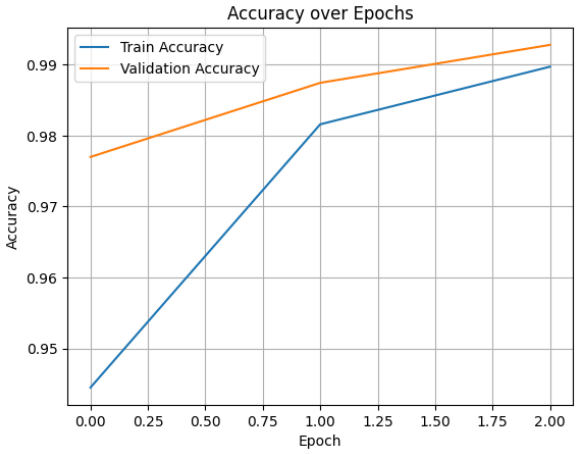
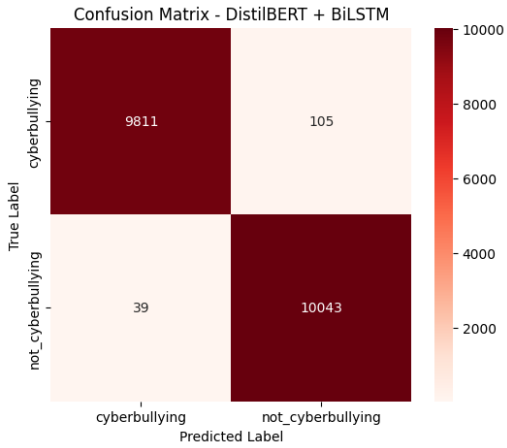


Figure 5A. Training and Validation Accuracy ofBiLSTM Model Across Epochs



5B. Training and Validation Accuracy of DistilBERT + BiLSTM Model Epochs



5C. Confusion Matrix Analysis: DistilBERT + BiLSTM Model

Figure 5A shows the accuracy curve of the standalone BiLSTM model during training and validation and after a number of epochs. Although the model starts with quite a decent accuracy, the values of training and validation metrics decrease steadily, reaching a lower plateau. This pattern shows a training instability case, which is likely due to inappropriate learning rate, lack of appropriate regularization, inappropriate initialization of weights or lack of rich contextual embeddings. This monotonous decline in performance indicates that the

model is not generalizing and not storing the learning in the long run, indicating a more basic architectural or data-related issue, as opposed to the normal overfitting or underfitting behavior.

Conversely, Figure 5B demonstrates training and validation accuracy of the hybrid DistilBERT + BiLSTM model, which remains stable and steadily rises with only negligible difference between the two. Such a high convergence shows that there is robust learning and generalization, which is achievable through the efficient utilization of pretrained contextual embeddings of DistilBERT. The BiLSTM layer also improves the model to learn the sequential patterns, which leads to better representation learning. The confusion matrix of this model (Figure 5C) also confirms its excellent classification ability: it produces very low error rates, misclassifying only 105 and 39 cases of cyberbullying and non-cyberbullying, respectively, out of more than 10,000 instances of each class. The high diagonal dominance of the matrix shows very high predictive accuracy of all categories with very few false positives and false negatives. The given performance shows the strong synergy between contextualized transformer embeddings and bidirectional sequence modeling, and the hybrid architecture is reliable and scalable to real-world cyberbullying detection systems.

7 Conclusion and Future Work

7.1 Conclusion

The paper is an in-depth analysis of the problem of detecting cyberbullying, which includes not only classic machine learning approaches but also more complex neural networks. The suggested model is efficient in solving the problems of implicit, context-rich, and AI-generated cyberbullying content that can be found on social media platforms. As the experimental results indicate, the classical models, such as Random Forest and LightGBM, provide a good baseline performance because they are relatively interpretable and fast, but they do not perform well on capturing complex semantic and sequential relationships in cyberbullying language. Individual deep learning models such as BiLSTM had promise in capturing temporal dependencies, but were unstable to train and had a narrow contextual scope.

DistilBERT + BiLSTM hybrid model turned out to be the most successful solution, with 99% accuracy in all the most important metrics, accuracy, precision, recall, and F1-score. This architecture balances the computational efficiency, interpretability, and classification accuracy with the help of deep contextual embeddings of DistilBERT and bidirectional sequence modeling of BiLSTM. Further, its smaller size and quicker inference speed allow it to be deployed in real-time and on mobile, providing a scalable solution to content moderation problems.

The findings indicate that the hybrid method performs better than the traditional and standalone models besides offering a strong, generalizable, and deployable framework to automated cyberbullying detection. Its low misclassification rates, strong generalization ability and simplified pipeline make it applicable in real life applications in safeguarding digital platforms against malicious online behaviour.

Future Work

Although the suggested DistilBERT + BiLSTM model can be effective in detecting cyberbullying, its performance can be enhanced by applying a number of future improvements to make it more effective and applicable. The model could be extended to work with multilingual and code-mixed text to allow a wider application globally, particularly through platforms such as Reddit, YouTube, and TikTok, where a wide variety of languages are used informally. Incorporating Explainable AI (XAI) would improve model interpretability that is necessary to human moderators and legal transparency. Also, by shifting the categorical labels to fine-grained classification of emotions, intent, or severity, it might be possible to achieve more specific interventions. Improving adversarial robustness would make the model ready to deal with manipulated or obfuscated abusive content in real-world deployments. The inclusion of online and ongoing learning would enable the model to adapt to the changing language patterns and emerging forms of cyberbullying without the degradation of performance. Additionally, additional edge deployment optimization would help to enable real-time moderation on mobile applications and browser extensions. Finally, it is possible to design a real-time feedback loop in which the decisions of the moderators and the reports of the users can be used to update the model, which guarantees adaptive learning and long-term accuracy.

References

Alqahtani, A. F., & Ilyas, M. (2024). An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying. *Machine Learning and Knowledge Extraction, 6*(1), 156–170.

Daniel, R., Murthy, T. S., Kumari, C. D. V. P., Lydia, E., Ishak, M. K., Hadjouni, M., & Mostafa, S. M. (2023). Ensemble learning with tournament selected glowworm swarm optimization algorithm for cyberbullying detection on social media. *IEEE Access, 11*, 123392.

Jamjoom, A. A., Karamti, H., Umer, M., Alsubai, S., Kim, T.-H., & Ashraf, I. (2024). RoBERTaNET: Enhanced RoBERTatransformer based model for cyberbullying detection with GloVe features. *IEEE Access, 12*, 58950–58973.

Muneer, A., Alwadain, A., Ragab, M. G., &Alqushaibi, A. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information, 14*(8), 467.

Murshed, B. A. H., Suresha, J., Abawajy, J., Saif, M. A. N., Abdulwahab, H. M., & Ghanem, F. A. (2023). FAEO-ECNN: Cyberbullying detection in social media platforms using topic modelling and deep learning. *Multimedia Tools and Applications, 82*, 46611–46650.

Obaida, M. H., Elkaffas, S. M., &Guirguis, S. K. (2024). Deep learning algorithms for cyber-bullying detection in social media platforms. *IEEE Access, 12*, 76901–76912.

Umer, M., Alabdulqader, E. A., Alarfaj, A. A., Cascone, L., & Nappi, M. (2024). Cyberbullying detection using PCA extracted GLOVE features and RoBERTaNet transformer learning model. *IEEE Transactions on Computational Social Systems*.

Bozyiğit, A., Utku, S. and Nasibov, E., 2021. Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, *179*, p.115001.

Jain, V., Saxena, A.K., Senthil, A., Jain, A. and Jain, A., 2021, December. Cyber-bullying detection in social media platform using machine learning. In *2021 10th International Conference on System Modeling& Advancement in Research Trends (SMART)* (pp. 401-405). IEEE.

Desai, A., Kalaskar, S., Kumbhar, O. and Dhumal, R., 2021. Cyber bullying detection on social media using machine learning. In *ITM Web of Conferences* (Vol. 40, p. 03038). EDP Sciences.

Neelakandan, S., Sridevi, M., Murugeswari, K. and Sridevi, R., 2022. Deep learning approaches for cyberbullying detection and classification on social media. *Computational intelligence and neuroscience*, *2022*, p.2163458.

Perera, A. and Fernando, P., 2021. Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science*, *181*, pp.605-611.

Alotaibi, M., Alotaibi, B. and Razaque, A., 2021. A multichannel deep learning framework for cyberbullying detection on social media. *Electronics*, *10*(21), p.2664.

Teng, T.H. and Varathan, K.D., 2023. Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, *11*, pp.55533-55560.

Singh, N.K., Singh, P. and Chand, S., 2022, November. Deep learning based methods for cyberbullying detection on social media. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 521-525). IEEE.