Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

SECURING AI-DRIVEN TEXT CLASSIFICATION AGAINST ADVERSARIAL NLP ATTACKS

Komal Azim[1], Saima Noreen Khosa[2], Saba Tahir[1], Muhammad Altaf Ahmad[1], Wajahat Hussain[1], Urooj Akram[1], Muhammad Faheem Mushtaq[*1]

[1]*Faculty of Computing, The Islamia University of Bahawalpur, 63100, Pakistan*
[2]*Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan*

[1]komalazeem498@gmail.com, [2]saimakhosa@yahoo.com, [1]saba.tahir@iub.edu.pk, [1]muhammadaltaf.ahmad@iub.edu.pk, [1]jamwajahat@gmail.com, [1]urooj.akram@kfueit.edu.pk, [*1]faheem.mushtaq@iub.edu.pk

DOI: https://doi.org/ 10.5281/zenodo.15461905

Keywords
Natural Language Processing, Cyber Security, Text Classification, Adversarial Attacks, Deep Learning, Ensemble Models

Article History
Received on 09 April 2025
Accepted on 09 May 2025
Published on 19 May 2025

Corresponding Author: * Muhammad Faheem Mushtaq

Abstract

The integration of Artificial Intelligence (AI) has revolutionized Natural Language Processing (NLP) enables advanced text classification tasks such as sentiment analysis, spam detection, and news categorization. However, the widespread adoption of AI in NLP has introduced significant cybersecurity risks, as these systems are highly vulnerable to adversarial attacks. These attacks aim to skew predictions and compromise their accuracy and integrity by making minor adjustments to input data, taking advantage of flaws in NLP models. We analyse and assess adversarial assaults on text categorization methods using AG News datasets. We examine how the model's performance might be assessed without human visual notice by implementing relatively straightforward transformation techniques such word substitution, paraphrase, or syntax alterations. These attacks highlight the basic flaws in NLP systems and demonstrate how easily they may be twisted and used maliciously. With up to 97% resilience against hostile attacks, the models proposed ensemble models by integrating the deep learning architectures include Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Recurrent Neural Networks (RNN). CNN performed better at identifying localized features, even though both the LSTM and RNN models showed good sequential processing skills. They significantly increased their resilience by combining complimentary qualities into ensemble frameworks. The highest success rate demonstrates that the ensemble tactics work to reduce adversary manipulation while preserving excellent classification accuracy.

INTRODUCTION

Artificial Intelligence based text classification systems have become vital in the present digital landscape for activities like content moderation, sentiment analysis, and spam detection [1]. These systems achieve phenomenal accuracy and scalability through cutting-edge machine-learning methods like deep neural networks. However, increasing of employability has led to adversarial actions with the aim of misusing the model imputation to cause disruption in operations or pursue malevolent causes [2]. Apart from suggestions for future research in what appears to be a very crucial area, the paper discusses strategies that can be employed for enhancing the robustness of these NLP models. In this way, NLP systems are strengthens in making them more reliable and

trustworthy AI applications in multiple sectors such as healthcare, finance, and security [3]. As such, one will need to consider hostile cyberattacks as an ever-increasing possibility while applying AI to critical domains like security [4], automated driving [5], and health imaging [6].

These attacks are based on fooling AI models by means of slight manipulations of input data that lead to wrong decisions and outcomes. There are several reasons for an in-depth understanding of various adversarial schemes [5]. Adversarial attacks performed on text classifiers usually consist of subtle changes to the input text, like substituting synonyms, introducing errors, or rephrasing terms. These are designed to distract the model while leaving human users completely uninformed. In another instance, a spam classification system may be duped into looking past an email as spam due to the deliberate misspelling of keywords in the email or the use of similar words that are not supported in the spam lexicon. These vulnerabilities threaten organizations and end users alike and call into question the trustworthiness and legitimacy of AI systems [7]. With recent studies emphasizing the profiling of models that are stronger against these kinds of perturbations, the inference has already been made in research that even slight changes of the text input, such as replacing a few words or altering some characters, are enough to severely handicap the performance of the text classifier [8].

The classifier may not be able to discriminate between adversarial modified inputs with similar semantics since current models are not generalized enough. Therefore, providing protection against adversarial-type attacks has become one of the key issues for the safe deployment of AI in solutions for everyday life. This diverse set of skills in AI, which comprises machine learning and deep learning techniques, natural language processing for knowledge representation and reasoning, and knowledge or rule-based expert systems modeling, could be intelligently applied to today's differentiated cybersecurity challenges [9]. Adversarial attacks have far-reaching consequences beyond mere technical errors. The use of adversarial inputs can act in disfavor to propagate content considered harmful or objectionable in applications such as content moderation and bypassing mechanisms of detection [10].

Adversarial manipulations can distort sentiment analysis metrics of public opinion, thereby shaping consumer behavior and ad campaigns. The banking sector is most vulnerable regarding adversarial attacks agitating fraud detection systems that cause cash loss and compromise user data's confidentiality [11]. By investigating the workings of adversarial NLP attacks and suggesting practical defenses, this study aims to address these issues. In this research, expertise model built on the foundational work in adversarial machine learning. In this sense, the proposed study focuses on techniques created especially for text classification, such as adversarial training, robust model designs, and anomaly detection systems. The main objective is to provide a holistic approach to enhancing the security of AI-powered text classifiers, gathering insights from recent research and experimentations [12].

Here, the proposed model shows the current interactions between adversarial attacks and defenses in NLP. The rapid growth of artificial intelligence indicates a fast change in the nature of hostile threats, thereby demanding constant evolution and interdisciplinary approaches to protect AI-based text classification applications. The major contributions of this work are as follows:

1. The vulnerability of text categorization systems to different adversarial NLP attack techniques, such as word substitution, paraphrase, and syntax changes, is methodically investigated in this paper. The study offers a comprehensive evaluation of attack efficacy and model weaknesses by utilizing benchmark datasets like AG News.
2. The research introduces an ensemble model framework by combining LSTM, RNN, and CNN architectures. This framework achieves up to 97% resilience against adversarial attacks by leveraging the complementary strengths of sequential processing and feature extraction.
3. The proposed model is evaluated using the evaluation parameter such as accuracy, precision, recall, F1 score and test loss.
4. Strong sequential processing abilities are displayed by LSTM and RNN models, which are able to capture textual contextual linkages. Localized characteristics and patterns, like phrases or particular word arrangements, are highly detectable by CNN models.

The remainder of this paper is structured as follows: Section 2 emphasize the related work that are relevant to the propose study. Section 3 explore the proposed methodology, such as preprocessing techniques, deep learning algorithms and in detail about dataset. Section 4 presents results and discussion to assess the effectiveness of these approaches. Finally, Section 5 outlines conclusion and future research directions, emphasizing the need for adaptive defenses and ethical considerations.

2. Related Work

It is well known that AI-driven text classification systems are vulnerable to adversarial attacks. Adversarial instances were first studied by Kurakin, that show the small changes in input data can result in large misclassifications in machine learning models [12]. After that, the necessity of model robustness is highlighted during training and suggested strong optimization strategies to thwart adversarial attacks [2]. HotFlip technique was presented for creating white-box adversarial examples especially for text classification problems [7]. This technique demonstrated how vulnerable NLP models are to character-level anomalies that are typically invisible to human observers. Text Bugger further emphasizes the efficacy of any such attack by generating adversarial texts with a view toward real-world applications [13]. Local Interpretable Model-Agnostic Explanations (LIME) and the techniques proposed for improving model transparency both draw attention to the importance of interpretability in countering adversarial attacks. The work was extended to the study of adversarial instances in natural language processing, where the resulting impact of semantic ambiguity and context-dependent interpretations were discussed [14]. Adversarial training has been gaining traction as a method of defense against these threats. The proposal involves including adversarial examples during the training dataset so that model robustness may be enhanced [12].

This study also realized that this usually results in overfitting and therefore would recommend exploring some other avenues like ensemble learning and regularization techniques [15]. Anomaly detection algorithms will provide some positive proactive means of finding suspicious inputs and hence prop up others-the defense mechanisms. Conventional machine learning had encouraging results, whereas deep learning techniques surpassed machine learning in accuracy and loss scores as evaluation metrics [16]. In this regard, the present study aims to improve text classification performance on the AG News dataset using tuning of word-level Convolutional Neural Network (CNN) model hyperparameters, Bidirectional Long-Short Term Memory (BiLSTM) model, and ultimately Bidirectional Encoder Representations from Transformers (BERT) architecture. The results show that the encoder-based BERT transformer architecture performs exceptionally well on news classification [17].

Many crucial elements for news categorization are not included in current news datasets since they only concentrate on textual features and infrequently use image features. In this work, we present a novel dataset, N24News, which is derived from the New York Times and comprises textual and visual data for every news item. In this study, use a multimodal approach to multitasking. The technique and experimental results show that multimodal news classification works better than text-only news categorization. It is possible to increase the classification accuracy by up to 8.11%, depending on how long the text [18]. In this paper, we have designed an IoT-assisted using an advanced machine learning approach. This strategy will offer intelligent cyber protection. system that will use blockchain technology to assist in identifying network security concerns. Despite these advancements, generalizing defensive strategies to a range of NLP tasks and languages remains challenging [19].

The need for adaptive systems that can adapt to new opponent tactics is emphasized by research [20-21]. Furthermore, ethical issues emphasize how important it is to balance security protocols with user privacy and transparency [22]. Digital infrastructure is protected by cybersecurity, yet there are more and more potential cyberattacks every day. They are impervious to common algorithmic defenses. The defensive actions of experts are ineffective. As a result, networks are being better protected through the application of artificial intelligence (AI). To prevent hostile AI, guarantee AI security, and protect cooperative learning, AI models require specific network safety measures and assurance technologies. Based on these

two perspectives, we investigate the relationship between AI and digital security [23].

A machine learning-based system for identifying distributed denial of service (DDoS)/DoS assaults is proposed in this study. A sizable dataset comprising the application layer's network traffic is used for this purpose. A new method to achieve better performance, a multi-feature strategy combining principal component analysis (PCA) and regular value decomposition (SVD) features is suggested [24]. According to the findings, 56% of the discovered AI-driven cyberattack technique was exhibited during the access and penetration phase, 12% during the exploitation phase, and 12% during the command-and-control phase. 9% was shown during the delivery phase of the study, and 11% during the reconnaissance phase kill chain for cybersecurity. The results of this study demonstrate that current cyber defense systems will not be able to handle the growing complexity and speed of AI-driven attacks. Consequently, to mitigate these new risks, companies need to start investing in AI cybersecurity infrastructure [25].

3. Proposed Methodology

The suggested method employs an ensemble approach that incorporates RNNs, LSTMs, and CNNs to fortify text categorization models against adversarial attacks. The robustness is assured against adversarial perturbations, which efficiently capture local and global textual characteristics. The proposed ensemble model is illustrated in Figure 1.

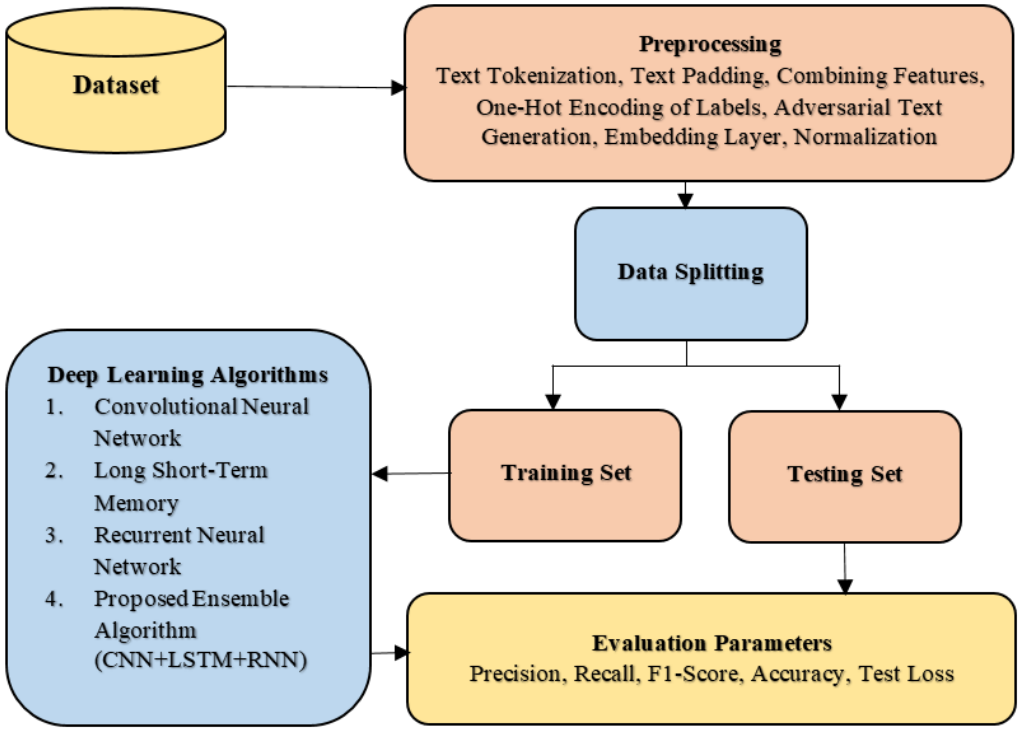

Figure 1. Proposed Methodology

3.1 Dataset Selection

The AG News dataset is one of the most popular benchmarks in text classification that was used for evaluating the proposed method. With more than 120,000 training instances and 7,600 test cases, the AG News dataset is divided among four classes: World, Sports, Business, and Science/Technology. Each sample in the dataset is a short news story on a wide array of subjects within various domains, with a typical text length of between 50 and 500 characters. The diverse nature of the materials allows

comprehensive testing of the resilience of text categorization models to adversarial perturbations. In order to present more generalized evaluation for model resilience under real attack scenarios, augment the dataset with some generated adversarial instances through the proposed adversarial text generation approach. The adversarial modified examples incorporated with real-world news data provide solid ground for evaluating the performance of the proposed technique. The AG news dataset word cloud is shown in Figure 2.

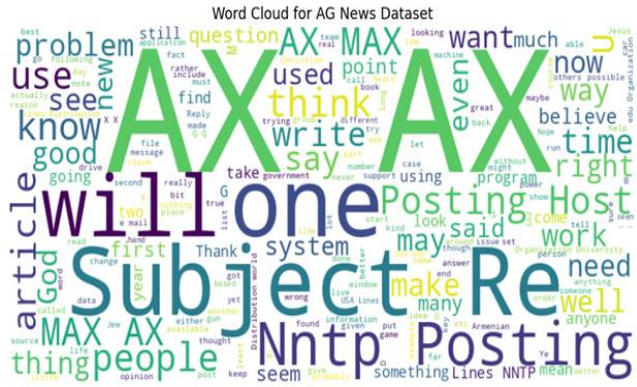

Figure 2. AG News Dataset Word Cloud

3.2 Preprocessing

In NLP models, text tokenization is the first preprocessing step quite important as it directly influences model accuracy and security. Conventional tokenization techniques usually constitute some form of word-level splitting. Such techniques are prone to being manipulated in adversarial ways. We propose hybridized tokenization methods that include both sub word-level and word-level tokenization's to bring a balance between vocabulary size and model resistance against adversarial attacks. Another vital preprocessing step is padding to keep the input size fixed for deep learning models, although this technique can be exploited by attackers to insert noise or to change the structure of text. The dynamic padding techniques we study to counter this insertion make the manipulations difficult for the adversaries since they change the padding length based on the complexity of their input. Besides, classical models usually utilize one feature extraction method such as bag-of-words or TF-IDF [26]. To ensure that the resulting representations are more robust and can withstand adversarial perturbations, we propose the integration of multiple methods like n-grams, word embedding, and syntactic features.

3.3 Data Splitting

Data splitting is about dividing the entire dataset available into smaller meaningful segments such that the model being tested would be put through rigorous testing and training. Usually, the data can be split into two sections: training and testing. The training set is for teaching the model to predict, while the testing set is for evaluating model performance on unseen data and thereby determining the generalization capacity. For this reason, another variation of data splitting is required to ensure the model's steady efficacy under adversarial attacks. This often requires the addition of a second validation set to the training data during hyperparameter tuning. Instances of adversarial nature engineered to mislead the model can be added purposely to this second validation set so that during training the model learns to detect and address such hostile perturbations.

3.4 Deep Learning Algorithms

The dataset's text categorization task performance was improved through the application of deep learning techniques [27-28]. The deep learning algorithms listed below are used to improve accuracy.

3.4.1 Convolutional Neural Network (CNN)

CNNs apply convolutional filters on word sequences in order to identify local patterns [29]. For classification tasks like sentiment analysis or spam detection, this enables the network to recognize critical features like n-grams or significant words. CNNs are capable of extracting hierarchical characteristics by using many layers of convolution and pooling, which enables them to recognize both simple and intricate textual patterns.

3.4.2 Long Short-Term Memory (LSTM)

LSTM are fitted with memory cells that can store information over long sequences in order to capture temporal dependencies across greater textual spans. In NLP applications where context and word order are essential, such as machine translation, named entity

recognition, and sentiment analysis, LSTM models are particularly well-suited. By using gates to control the information flow, LSTMs are able to selectively remember or forget information. According to [30], Such processes would enhance the model's chances of dealing with complicate relationships in sequential data.

3.4.3 Recurrent Neural Network (RNN)

An RNN's output at every time instance depends on the previous hidden state and the current input. RNNs process input sequences one element at a time. This means that many of the applications in NLP require the ability to say something about temporal relationships in the text. Unfortunately, the ability of the standard RNNs to learn long-range dependencies is hindered due to problems such as vanishing gradients. Nonetheless, RNNs continue to be a prevailing architecture in many NLP applications, especially where the understanding of text is dependent on the temporal sequence and flow of information [31].

3.4.4 Proposed Ensemble Model

This paper proposes an ensemble method to enhance text classification model performance uniting the strengths of CNNs, LSTMs, and RNNs. On the one hand, CNNs excel at detecting local patterns and n-grams; on the other hand, LSTMs are superior in modeling long-term dependencies. RNNs, however, provide strong methods for endowing models with sequential-handling capabilities. The proposed approach synthesizes these three models into one ensemble and thus benefits from their complementary characteristics to enhance learning of local features and global context alike. While the CNN captures noteworthy features in small text windows, the LSTM and RNN maintain appropriate temporal context and sequential dependencies. For more advanced NLP tasks, understanding not just the local but also the long-range interactions involved is important. This combined version can increase text classification systems' accuracy and robustness widely. In our perspective, such an ensemble would help in finding a dependable and effective solution in solving real-life problems in text classification.

4. Results and Discussion

The findings of the proposed ensemble model as well as the outcomes of various deep learning models, including CNNs, LSTMs, and RNNs, are covered in detail in this section.

4.1 Model Performance under Adversarial Conditions

The experimental evaluations using the AG News datasets were aimed at investigating the robustness of CNN, LSTM, RNN, and an ensemble of these architectures. Word replacements, paraphrasing, and perturbation in syntax formed the core of the evaluation strategies that were adopted in the testing of the models. The various epoch performances are summarized in Table 1. Several of the significant assessment parameters are also employed for measuring the performance of text classification models, especially against adversarial NLP attacks. These metrics describe how well the model performs in general and its defence capabilities under different scenarios. The models are evaluated using metrics such as accuracy, precision, recall, f1-score, and test loss.

4.2 Performance of Convolutional Neural Network

We see the CNN model continuously showing good scores of precision, recall, and F1 during all epochs. During epoch 5, it stood a balanced 97.67% using all the given metrics and reasonably distinguished between positives and negatives. Starting from epoch 10 through epoch 20, these measures remained stable, with slight variations in the range of 97.56-97.75. Notably, during epoch 10, the precision, recall, and F1-score of the model were measured at 97.58%, 97.56%, and 97.56%, and the scores for both epochs 15 and 20 rather settled around 97.74%, which just indicated the model was capable of discriminating between classes with very little bias throughout. In our case, some signs of overfitting have also been noticed, indicated by sluggish test loss increment over time; nevertheless, this goes to suggest that the CNN model has done quite well for itself in terms of being accurate and robust concerning the high precision, recall, and F1-scores along with low test losses during the later epochs. Table 1 and Figure 3 present the performance of the CNN model.

Table 1. Results of Convolutional Neural Network

Epoch	Testing accuracy	Training accuracy	Test Loss	Precision	Recall	F1 score
5	0.9134	0.9767	0.4015	0.9767	0.9767	0.9767
10	0.9018	0.9756	2.0083	0.9758	0.9756	0.9756
15	0.9034	0.9774	1.8422	0.9775	0.9774	0.9774
20	0.903	0.9774	1.5535	0.9774	0.9774	0.9774

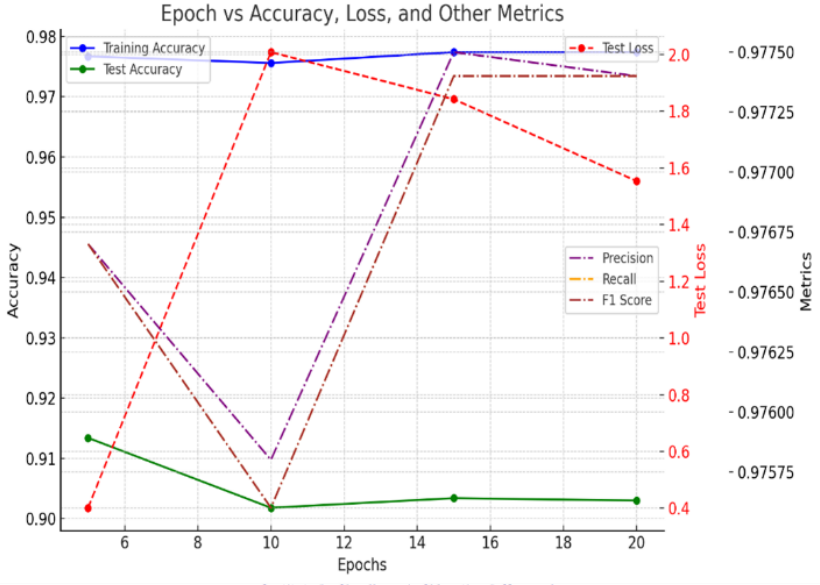

Figure 3. Performance of CNN model

4.3 Performance of Long Short-Term Memory

Contrarily, the outcomes of the LSTM model show consistent attainment of strong performances on various assessment measures, with appreciable gains with increasing epochs. An epoch 5 test loss of 0.3586 and training accuracy of 96.25% were recorded for the model, with a lowest test accuracy of 90.59% across all epochs, which is indicative of good performance but leaves much room for improvement. The increase in test loss from 0.3586 at epoch 5 to 0.8068 at epoch 10, though, did not prevent an increase in the training accuracy (97.38%) and test accuracy (90.08%), respectively, which suggests that the trend of increasing training accuracy and stationary test accuracy is being maintained ahead, marking the process of advancement towards overfitting. This period of training continued through epoch 15. The model's training accuracy was 97.47%, but its test accuracy fell to 89.32%, and its test loss grew to 1.0301. Training accuracy held consistent at 97.47% at epoch 20. By epoch 20, the overfitting pattern was further supported by training accuracy staying constant at 97.47% and test accuracy marginally improving to 89.68%, while the test loss hit 1.3326. The performance of the LSTM model is displayed in Table 2 and Figure 4.

Table 2. Results of Long Short-Term Memory

Epoch	Testing accuracy	Training accuracy	Test Loss	Precision	Recall	F1 score

5	0.9059	0.9625	0.3586	0.9629	0.9625	0.9626
10	0.9025	0.9738	0.8068	0.9739	0.9738	0.9737
15	0.8932	0.9747	1.0301	0.9747	0.9747	0.9747
20	0.8968	0.9728	0.7777	0.9748	0.9747	0.9747

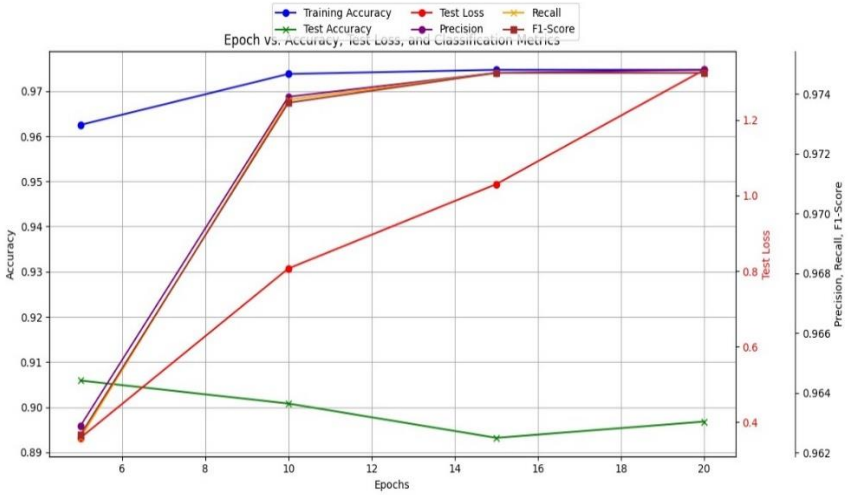

Figure 4. Performance of LSTM model

4.4 Performance of Recurrent Neural Network

The Recurrent Neural Network (RNN) model showed good performance across all epochs, with training accuracy consistently around 97%. At epoch 5, test accuracy was 90.86%, and test loss was 0.3317. However, as training progressed, test accuracy gradually decreased, reaching 89.16% by epoch 20, while test loss increased to 0.7777, indicating some overfitting. Despite this, precision, recall, and F1-score remained high throughout, with values close to 97%, reflecting the model's strong ability to accurately classify both positive and negative instances. Table 3 and Figure 5 shows the performance of RNN model.

Table 3. Results of Recurrent Neural Network

Epoch	Testing accuracy	Training accuracy	Test Loss	Precision	Recall	F1 score
5	0.9086	0.9719	0.3317	0.9719	0.9719	0.9719
10	0.9024	0.975	0.5289	0.9752	0.975	0.975
15	0.897	0.9724	0.6576	0.9724	0.9724	0.9724
20	0.8916	0.9728	0.7777	0.9731	0.9728	0.9728

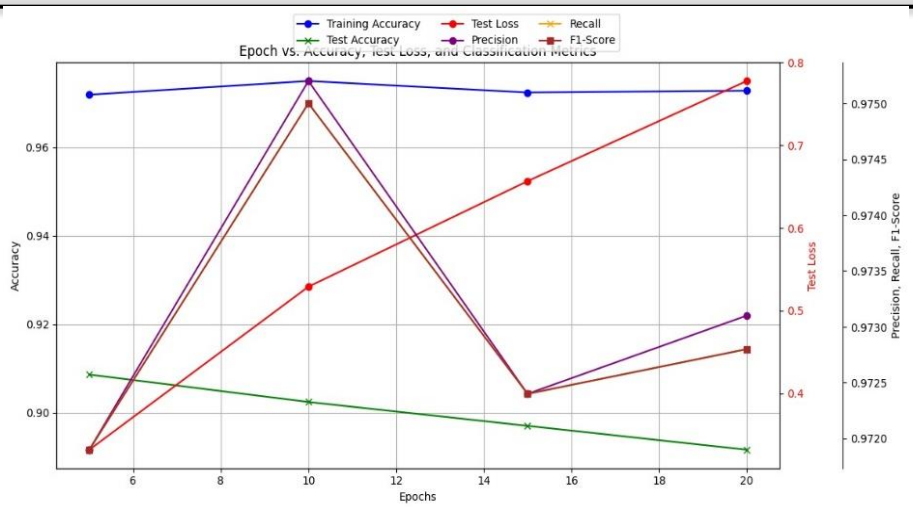

Figure 5. Performance of RNN model

4.4 Proposed Ensemble Model

The results from the proposed ensemble model combining RNN, LSTM, and CNN demonstrate strong performance with consistent high training accuracy across all epochs. At epoch 5, the model achieved a training accuracy of 96.68% and a test accuracy of 90.79%, with a test loss of 0.395. As the model continued training, it showed steady improvement in both training and test accuracy. By epoch 10, training accuracy increased to 97.44%, but test accuracy slightly decreased to 89.75%, with the test loss rising to 0.7121, indicating the model was beginning to overfit. The trend of high training accuracy and increasing test loss continued with training accuracy reaching 97.46% at epoch 15, and test accuracy decreasing slightly to 89.34%, with a test loss of 0.9294. By epoch 20, however, the training accuracy spiked to 99.87%, but the test accuracy was still 89.62%, with test loss rising to 1.21, reflecting a strong overfitting tendency.

In terms of classification performance, precision, recall, and F1-score remained impressively high throughout, hovering around 97%. During epoch 5, precision was 97.51%, recall was 97.50%, and the F1-score was 97.50%. These metrics remained stable during the 10th and 15th epochs with slight changes: 97.44% (precision), 97.44% (recall), and 97.43% (F1-score) for epoch 10; while all three metrics recorded a value of 97.46% at epoch 15. In epoch 20, slight drops to 97.38%, 97.37%, and 97.37% were observed for precision, recall, and F1-score, respectively. The overall high precision, recall, and F1-score of the ensemble model thus indicate its ability to maintain robust classification performance on the training and test datasets, notwithstanding the overfitting demonstrated by the steady rise of test loss and slight descent of test accuracy. Some measures against overfitting should be considered, but the ensemble consisting of CNN, LSTM, and RNN proves useful for many text classification tasks. The proposed ensemble model's performance is depicted in Table 4 and Figure 6.

Table 4. Proposed Ensemble Model Results

Epoch	Testing accuracy	Training accuracy	Test Loss	Precision	Recall	F1 score
5	0.9079	0.9668	0.395	0.9751	0.975	0.975
10	0.8975	0.9744	0.7121	0.9744	0.9744	0.9743
15	0.8934	0.9746	0.9294	0.9746	0.9746	0.9746
20	0.8962	0.9987	1.21	0.9738	0.9737	0.9737

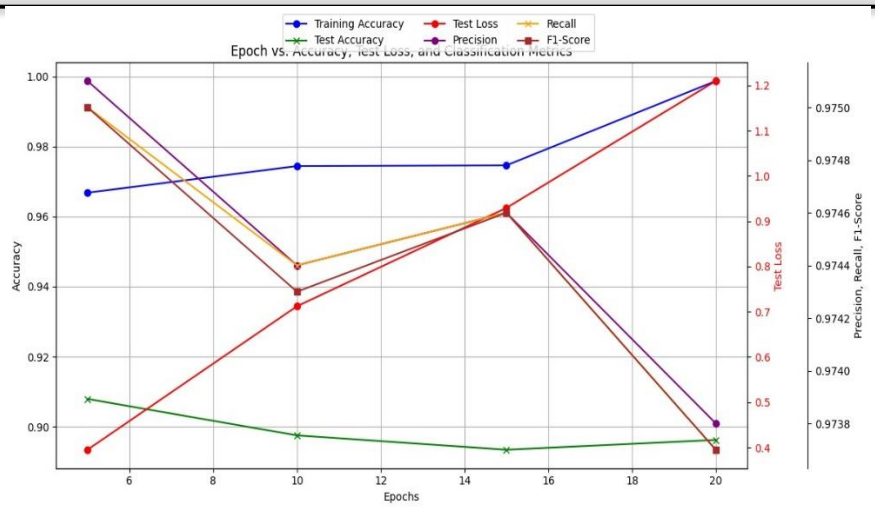

Figure 6. Performance of the Proposed Ensemble Model

4.5 Performance Analysis of All Models

The performance of deep learning architectures including CNN, LSTM, RNN is assessed and contrasted, with each showing distinct advantages in various facets of pattern recognition. Long Short-Term Memory (LSTM) networks solve the vanishing gradient problem when learning long-range dependencies, Recurrent Neural Networks (RNNs) efficiently capture sequential relationships in temporal data, and Convolutional Neural Networks (CNNs) are excellent at extracting spatial features from structured data. The ensemble model achieved the highest accuracy by integrates the temporal memory retention of LSTM, the sequential modeling of RNN, and the spatial feature extraction of CNN. Table 5 shows the comparative performance of all models.

Table 5. Performance Comparison of All Models

Epoch	Model	Testing accuracy	Training accuracy	Test Loss	Precision	Recall	F1 score
20	CNN	0.903	0.9774	1.5535	0.9774	0.9774	0.9774
15	LSTM	0.8932	0.9747	1.0301	0.9747	0.9747	0.9747
10	RNN	0.9024	0.975	0.5289	0.9752	0.975	0.975
20	Proposed Model	0.8962	0.9987	1.21	0.9738	0.9737	0.9737

5. Conclusion and Future Work

This study presents a hybrid ensemble text classification model that integrates CNN, LSTM, and RNN architectures, achieving strong performance across key evaluation metrics, including F1-score, recall, accuracy, and precision. The model was continually training successfully with high accuracy, demonstrating that it could accurately learn situations, whether positive or negative, as evidenced by its approximately 97% precision, recall, and F1-score. Yet, while the test accuracy remained plateauing and test loss was increasing across epochs, since accuracy practically kept increasing during training. For generalization and overfitting mitigation, more regularization methods or data augmentation approaches can be needed. Although the ensemble approach seems promise in protecting text

categorization models. Although, the ensemble approach seems promise in protecting text categorization models to optimize its resilience and applicability in a variety of real-world NLP applications.

REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6572

[2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.06083

[3] T. S. L. Prasad, K. B. Manikandan, and J. Vinoj, "Shielding NLP Systems: An In-depth Survey on Advanced AI Techniques for Adversarial Attack Detection in Cyber Security," in 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS), 2024, pp. 1059–1067.

[4] V. Kovtun, I. Izonin, and M. Gregus, "Reliability model of the security subsystem countering to the impact of typed cyber-physical attacks," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-17254-4.

[5] W. Z. Khan, M. Y. Aalsalem, M. K. Khan, and Q. Arshad, "When social objects collaborate: Concepts, processing elements, attacks and challenges," Computers and Electrical Engineering, vol. 58, pp. 397–411, Feb. 2017, doi: 10.1016/j.compeleceng.2016.11.014.

[6] Y. Ruan and A. Durresi, "A survey of trust management systems for online social communities – Trust modeling, trust inference and attacks," *Knowl Based Syst*, vol. 106, pp. 150–163, 2016, doi: https://doi.org/10.1016/j.knosys.2016.05.042.

[7] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "HotFlip: White-Box Adversarial Examples for Text Classification," Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.06751

[8] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," Jul. 2016, [Online]. Available: http://arxiv.org/abs/1607.01759

[9] I. H. Sarker, M. H. Furhad, and R. Nowrozy, "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions," *SN Comput Sci*, vol. 2, no. 3, p. 173, 2021.

[10] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A New Benchmark for Natural Language Understanding." [Online]. Available: https://parl.ai/.

[11] S. Garg and G. Ramakrishnan, "BAE: BERT-based Adversarial Examples for Text Classification," Apr. 2020, doi: 10.18653/v1/2020.emnlp-main.498.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.01236

[13] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "TEXTBUGGER: Generating Adversarial Text Against Real-world Applications," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019*, The Internet Society, 2019. doi: 10.14722/ndss.2019.23138.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[15] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment," Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11932

[16] Mushtaq, M. F., et al., A Survey on the Cryptographic Encryption Algorithms. *International Journal of Advanced Computer Science and Applications*, 2017. 8(11): pp. 333-344.

[17] S. Ozdemir, "News Classification with State-of-the-Art Deep Learning Methods," in *8th International Artificial Intelligence and Data Processing Symposium, IDAP 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/IDAP64064.2024.10710921.

[18] Z. Wang, X. Shan, X. Zhang, and J. Yang, "N24News: A New Dataset for Multimodal News Classification," Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.13327

[19] R. Majeed, N. A. Abdullah, and M. F. Mushtaq, "IoT-based Cyber-security of Drones using the Naïve Bayes Algorithm." [Online]. Available: www.ijacsa.thesai.org
[20] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," May 2019, [Online]. Available: http://arxiv.org/abs/1905.05583
[21] R. Xu, J. Wu, Y. Cheng, Z. Liu, Y. Lin, and Y. Xie, "Dynamic security exchange scheduling model for business workflow based on queuing theory in cloud computing," *Security and Communication Networks*, vol. 2020, 2020, doi: 10.1155/2020/8886640.
[22] M. F. Mushtaq et al., "Key schedule algorithm using 3-dimensional hybrid cubes for block cipher," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, 2019.
[23] S. Minhajul Hassan and D. Wasim, "Journal Of Aeronautical Materials Study Of Artificial Intelligence In Cyber Security And The Emerging Threat Of Ai-Driven Cyber Attacks And Challenges," vol. 43, pp. 1557–1570, 2023, [Online]. Available: https://ssrn.com/abstract=4652028
[24] F. Rustam, M. F. Mushtaq, A. Hamza, M. S. Farooq, A. D. Jurcut, and I. Ashraf, "Denial of Service Attack Classification Using Machine Learning with Multi-Features," *Electronics (Switzerland)*, vol. 11, no. 22, Nov. 2022, doi: 10.3390/electronics11223817.
[25] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, and V. Pospelova, "The Emerging Threat of Ai-driven Cyber Attacks: A," 2022.
[26] M. A. Abid, M. F. Mushtaq, U. Akram, M. A. Abbasi, and F. Rustam, "Comparative analysis of TF-IDF and loglikelihood method for keywords extraction of twitter data," *Mehran University Research Journal of Engineering and Technology*, vol. 42, no. 1, p. 88, Jan. 2023, doi: 10.22581/muet1982.2301.09.
[27] Ali, Mudasir, et al. "Hybrid machine learning model for efficient botnet attack detection in iot environment" *IEEE Access* (2024).
[28] Akram, Urooj, et al. "IoTTPS: Ensemble RKSVM model-based Internet of Things threat protection system." *Sensors* 23.14 (2023): 6379.
[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.
[30] Ishaq, Abid, et al. "Extensive hotel reviews classification using long short term memory." *Journal of Ambient Intelligence and Humanized Computing* 12 (2021): 9375-9385.
[31] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Physica D*, vol. 404, p. 132306, 2020, doi: https://doi.org/10.1016/j.physd.2019.132306.