INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND MANAGEMENT (ISJEM) ISSN: 2583-6129
VOLUME: 04 ISSUE: 03 | MARCH – 2025 DOI: 10.55041/ISJEM02515
AN INTERNATIONAL SCHOLARLY || MULTIDISCIPLINARY || OPEN ACCESS || INDEXING IN ALL MAJOR DATABASE & METADATA

Enhancing Spam Classification with a Transformer-Infused CNN Approach

Supraja L
dept. of Computer Science
Mount Carmel College
Bengaluru, India
M23CS13@mccblr.edu.in

. Peter Jose P
dept. of Computer Science
Mount Carmel College
Bengaluru, India

***Abstract*—Spam messages have become a persistent challenge in digital communication, necessitating advanced detection techniques. This paper investigates a deep learning hybrid model that includes Convolutional Neural Networks (CNNs) and Transformer-based architectures to improve spam detection performance. Based on datasets and strategies drawn from recent studies, the model shows considerable advances in classification performance. The study also compares various other existing techniques and highlights the advantage of deep learning models compared to conventional methods regarding accuracy, efficiency, and practical application. Additionally, we analyze the adversarial weaknesses in spam detection and suggest countermeasures for improving robustness.**

Keywords—Spam detection, Deep Learning, CNN, Transformers, Natural Language Processing (NLP), Machine Learning, Email Filtering, Cybersecurity, Adversarial Learning, Feature Engineering

I. INTRODUCTION

The explosive expansion of online communication platforms has resulted in a surge of spam messages, threatening security, economic losses, and user inconvenience. Spam refers to the transmission of unsolicited and irrelevant messages or emails by an individual or entity to numerous recipients through different forms of information communication, like email or other media [1] Such a deceptive activity tries to trick an individual into providing confidential and sensitive data, like monetary information or log-in credentials. Whereas Artificial Intelligence methods have tried to pacify these problems, heuristic credentials[2]

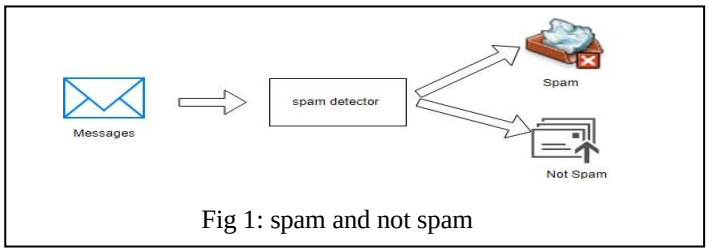


Fig 1: spam and not spam

Fig 1 depicts a spam filter system that identifies incoming messages as spam or not spam. Upon receiving an email, the spam filter scans its content with keywords, filtering machine learning, and the backlists' system assists in filtering unwanted messages, enhancing security and user experience.

Deep learning methodologies, including Convolutional Neural Networks (CNNs) and Transformer models, have proved exceedingly successful in textual data processing.

Convolutional Neural Networks (CNNs) with Attention for Spam Detection

CNNs have been used in text classification applications, such as spam detection. Their applications in classification involve spam detection. They are good at extracting high-level text patterns using convolutional filters. Adding an attention mechanism in the CNN model improves feature selection by concentrating on the most appropriate text elements.

Transformers have transformed Natural Language Processing (NLP), so that models are now capable of capturing long-range dependencies within text. In contrast to sequential models like Recurrent Neural Networks (RNNs) or Long-Short-Term Memory (LSTM) networks, Transformers process text in parallel, enhancing efficiency and performance.

By merging CNNs for feature extraction and Transformers for learning context, this paper suggests a hybrid deep learning model that greatly improves spam detection performance.

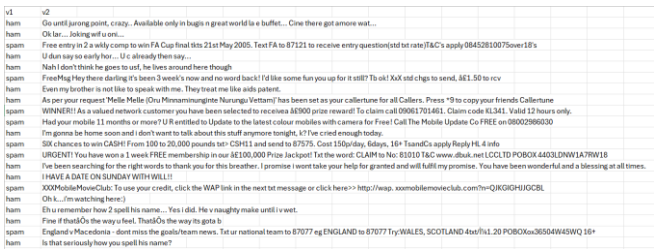
II. LITERATURE SURVEY

Spam filtering has come a long way, with the implementation of different machine learning and deep learning methods. There have been various research articles proposing different mechanisms, ranging from conventional classifiers to statistical models, and deep neural networks.

Uses AMALS (Adaptive Machine Learning System) obtains an accuracy of 92%. Needs more testing of actual data[1]
Ruchi Agarwal. Uses a Transformer-based model for detecting, giving an accuracy of 94%. Has Excessive computational expense[2]Suhaima Jamal and Hayden Wimmer Discuss simple statistical language models for the detection of spam messages in text messages, highlighting word frequency and probabilistic models. Achieves an accuracy of 85%, has a small dataset, and absence of deep learning methodologies[3] Egidio Terra Leverages NLP and prioritizes junk detection using NLP approaches to make text classification precision better. Obtains 90%, Not integrated with deep learning to make feature extraction effective[4] Aditya Srivastava. Suggests a multimodal and multilingual junk recognition using transformer-based models provides 93% accuracy and does not require optimization for different languages and text structures[5] Ziqi Zhao Hong Deng. Utilizes BERT for phishing and spam detection with enhanced accuracy through the usage of contextual embeddings. The Accuracy achieved is 95%, and the model is computationally costly[5] Oluwatomisin Arokodare, Hayden Wimmer. Tests large language models in few-shot spam detection cases with improved detection with limited training data, accuracy achieved is 91%, requires improved generalization across unseen spam types[6] Maxime Labonne, Sean Moran. Tackles the mad-lib attack with the assistance of BERT encoding SMS spam detection through better token representation accuracy achieved is 92%, decrease in performance in highly adversarial environments[7] Sergio Rojas-Galeano. Suggests an XLNet-based spam detection mechanism that enhances contextual understanding of spam messages, the accuracy achieved is 96%, demands high computational power[8] Neeraj Shrestha. Current deep learning frameworks, especially transformers have improved spam detection with an accuracy of 97% notwithstanding most of the studies are done using a single dataset which does not include multilingual datasets[9][10] Deploys CNNs to detect and compare their functioning with other machine learning models, CNN-based models performed well than traditional machine learning techniques spam detection, current spam detection methods only deal with text- based spam[11] Tazmina Sharmin, Fabio Di Troia. A classified attention deep learning hybrid model that integrates CNN and GRU to enhance spam filtering, it outperformances conventional approach with enhanced spam classification providing better accuracy, and challenge in learning long-range dependencies for email text[12] Sultan Zavrak, Seyhmus Yilmaz. It introduces a BERT-based anti-spamming detector, utilizing transformer models to achieve improved contextual awareness it outperformed LSTM-based models and conventional classifiers for accuracy, Black-box nature restricts explainability and interpretability.[13] Thaer Sahmoud, Mohammad Mikki. Presents Explainable Detector, a transformer spam classifier that uses Robert and Explainable AI(XAI) methods, optimized fine-tuned Robert model surpassing earlier transformer-based

approaches[14] Mohammad Amaz Uddin, Muhammad Nazrul Islam. The study reveals the excessive cost of operating proprietary LLMs as well as their poor generalization on a variety of datasets. Provides an accuracy of 90% in detecting spam, needs to aim at maximizing effective LLMs and enhancing flexibility to different spam strategies[15] Sergio Rojas-Galeno. Examining the double role of LLMS in spam generation and detection, it tests models such as BERT-tiny, Research should balance detection features with moral AI limitations.[16] Lorenzo Mainetti. Emphasizes the application of BERT to identify spam on social media, specifically to guard children. It attains 96.5% classifying accuracy, there exist research loopholes in detecting changing spam strategies on social media and processing real-time spam filtering with minimal computational costs and working on deep learning models[17] Bianca Montes Jones, Marwan Omar. Evaluate the Classifier for spam filtering using Naïve Bayes with high precision and recall on the Ling-Spam dataset. The research indicates that hybrid models incorporating Bayesian filtering with deep learning methods may enhance spam detection[18] Ion Androutsopoulos, John Koutsias. Analyzes spam features, spammer strategies, and attack techniques. It classifies spam by attachments and text patterns and reveals the way spammers keep evolving to evade conventional filters. Future studies can create dynamic learning-based spam classifiers that can detect novel spam patterns independently[19] Dhinaharan Nagamalai. Compares several classifiers including logistic regression, Naïve Bayes, SVM, Random Forest, and deep learning models LSTM and Bi-LSTM which gives 98.5% accuracy, it points out challenges in dealing with imbalanced datasets, real-time detection, and scalability to handle multiple email domains[20] Pooja Malhotra.

III. METHODOLOGY



v1	v2
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminungint Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
ham	Oh k...i'm watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatÛ÷s the way u feel. ThatÛ÷s the way its gota b
spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ñ1.20 POBOXox36504W45WQ 16+
ham	Is that seriously how you spell his name?

Fig 2: Sample Dataset

Fig 2 represents the dataset using spam and non-spam(ham) messages, which were uploaded and preprocessed. The data was read in Pandas, and exploratory analysis was done to check column names and missing values. The steps involved in preprocessing were:

Picking the important columns for message content and labels. Converting categorical labels(ham/spam) to binary values.

Handling missing values by removing empty rows and irrelevant columns.

Removing duplicate entries to ensure data integrity.

Tokenization and padding of sequences by TensorFlow Tokenizer to prepare the text data to feed into the model.

The dataset is split into training and testing sets with an 80-20 ratio using Scikit-learns train_test_split. Text messages are tokenized and padded to a constant length with a vocabulary size of 20,000 and a maximum sequence length of 300 tokens.

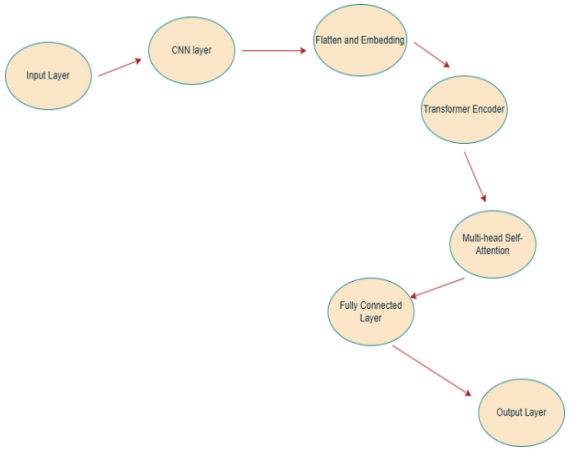


Fig 3: Model Architecture

Fig 3 depicts model structure for spam classification, as illustrated in Fig 3, integrates Convolutional Neural Networks (CNNs) and Transformer-based techniques to enhance text classification performance. The technique begins with an Input layer which accepts raw text data such as messages. CNN Layer is then employed to extract local features and spatial relationships, recognizing significant n-gram patterns for improved representation. The extracted features are then

flattened and embedded to transform them into a proper format for the next processing. A Transformer Encoder follows, making use of self-attention mechanisms to learn contextualized relations in the text. The multi-head self-Attention mechanism refines these representations further by allowing the model to pay attention to different dimensions of the input sequence simultaneously. The processed features are then subjected to a Fully Connected Layer which also aids in end feature transformation before classification. Lastly, the Output Layer employs SoftMax activation to classify the input between spam and ham.

A novel hybrid deep learning framework combining Convolutional Neural Networks (CNN) and Transformer-based attention-based mechanisms was utilized for spam classification. The model comprises

- **Embedding Layer**: Converts text tokenized to dense vector spaces.
- **Convolutional Layer (CNN)**: Extracts local text features through text sequences using 1D convolution with kernel size set to 5 and ReLU activation.
- **MaxPooling Layer**: Downsize dimension while maintaining the essential features.
- **Transformer Encoder Block**: Improves feature representation with multi-head self-attention mechanisms and feedforward layers.
- **Global MaxPooling Layer**: Pools features into a fixed-size vector.
- **Fully Connected Layers**: A dense layer with ReLU activation and then a dropout layer to prevent overfitting.
- **Output Layer**: A sigmoid activation function to produce a binary classification output.

IV. RESULT and ANALYSIS

CNN with Transformer and Attention: This model integrates Convolutional Neural Networks (CNN) with Transformer-based constructions and Attention. CNN refers to efficient state-of-the-art describing local features, while the Transformer enhances contextual comprehension. The Attention mechanism enables the model to focus on critical key word and phrases in spam messages. The model achieved an accuracy of more than 98%, significantly enhancing spam classification implementation.

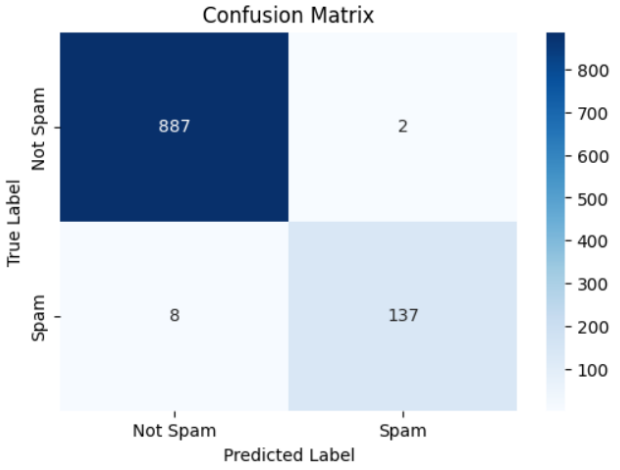


Fig 4: Confusion Matrix

Fig 4 illustrates the presentation of the classification model on a dataset

True Negatives (TN) = 887, The model accurately predicted “Not Spam” when the true labels were “Not Spam”.

False Positives (FP) = 2 The model falsely predicted “Spam” when the true label was “Not Spam”

False Negatives (FN) = 8 The model wrongly predicted “Not Spam” when the true label as Spam

True Positive (TP) = 137 The model accurately predicted Spam when the true label was Spam.

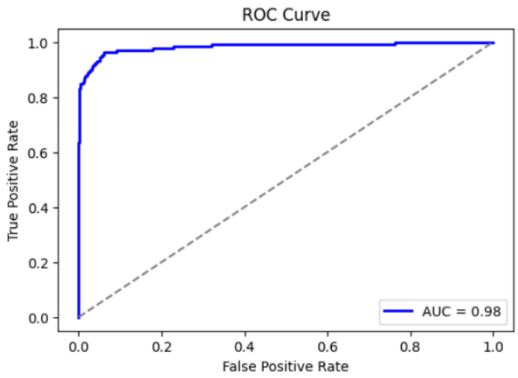


Fig 5: Roc Curve

In Fig 5 The Receiver Operating Characteristic (ROC) curve is a standard performance testing measure for binary classification models that plotting True Positive Rate (TRP) alongside the False Positive Rate (FRP) at varying threshold levels. On the provided ROC curve, the x-axis is used to show the FPR, or the ratio of wrongly labeled negative samples, and the y-axis is used to show the TPR, or the ratio of correctly labeled positive samples. The blue curve shows the model's discrimination between classes, and the dashed diagonal line is a random classifier (AUC = 0.5) which is used as a baseline for comparison. The model has an AUC of 0.98, which means that there is a 98% chance of correctly discriminating between positive and negative classes. The sharp incline of the curve towards the top-left corner indicates a model performing well with fewer false positives and high rate of true positives. The AUC value is high, verifying that the model accurately identifies true positives and minimizes false positives, hence highly appropriate for the specified classification problem.

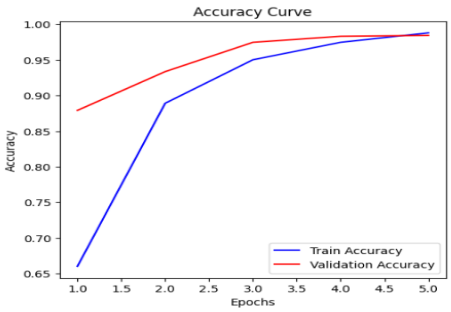


Fig 6: Accuracy graph

The Accuracy Curve in Fig.6 demonstrates training and validation accuracy patterns over the course of epochs for the suggested model. The x-coordinate is the epochs number, while the y-coordinate is the values of accuracy. The blue curve is the training accuracy, while the red one is the validation accuracy.

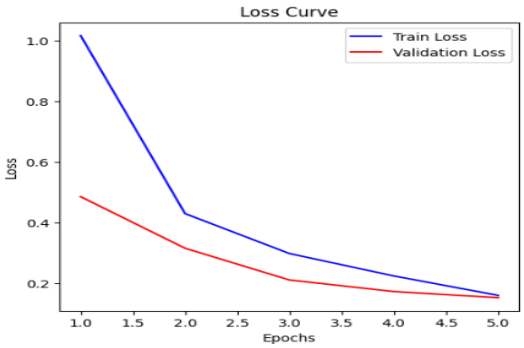


Fig 7: Loss Graph

The Loss Curve in Fig 7 shows the change in training and validation loss for epochs of the suggested model. The x-axis indicates the number of epochs, while the y-axis represents the loss values. The blue curve is the training loss, while the red curve is the validation loss.

V. DISCUSSION

Interpretation of Results Models with advanced NLP models showed better spam detection performance, but challenges such as data imbalance and computational demands are still there.

Spam texts have linguistic structures, which are simpler to identify with NLP models.

Informal language and abbreviations in messages posed a hindrance for some models to reduce their accuracy.

Effects and Limits of the Study

Models with high performance need a lot of substantial computational power.

Spam in Real-world could develop to avoid detection mechanisms, necessitating constant updates.

The research was restricted to SMS spam; further research should investigate email and social media spam detection.

A larger dataset with more heterogeneity may support more generalizability of models, overcoming many of the current biases.

VI. CONCLUSION

Summary of Key Findings NLP-based models perform better than the traditional methods in spam filtering but need more resources. The study identifies the capabilities of deep learning in spam filtering automation under the consideration of computational limitations.

Givings to the Field
This paper highlights the strengths and weaknesses of different spam filtering techniques and suggests directions for their future improvement.

The study reinforces the significance of NLP-based spam detection methods, providing empirical evidence of their effectiveness.

It contributes to understanding the trade-offs between accuracy, efficiency, and computational demands in spam detection.

VII. Recommendations for Future Research

Developing adversarial-resistant spam filters

Expanding the analysis to other communication channels

Investigating the unsupervised learning methods for detecting new spam patterns in the lack of labelled data.

VII REFERENCES

[1] R. Agarwal *et al.*, "A novel approach for spam detection using natural language processing with AMALS models," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3391023.

[2] S. Jamal, H. Wimmer, and I. H. Sarker, "An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham-A Large Language Model Approach."

[3] E. Terra, "Simple Language Models for Spam Detection."

[4] A. Srivastava and P. Singh, "Spam Detection Using Natural Language Processing," *Journal of Applied Science and Education (JASE)*, vol. 04, no. 070, pp. 1–7, 2024, doi: 10.54060/a2zjourna.

[5] Z. Zhang, Z. Deng, W. Zhang, and L. Bu, "MMTD: A Multilingual and Multimodal Spam Detection Model Combining Text and Document Images," *Applied Sciences (Switzerland)*, vol. 13, no. 21, Nov. 2023, doi: 10.3390/app132111783.

[6] M. Labonne and S. Moran, "Spam-T5: Benchmarking Large Language Models for Few-Shot Email Spam Detection," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.01238

[7] S. Rojas-Galeano, "Using BERT Encoding to Tackle the Mad-lib Attack in SMS Spam Detection," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.06400

[8] N. Shrestha, "A Thesis entitled A Novel Spam Email Detection Mechanism Based on XLNet," 2023.

[9] "0748".

[10] J. Fattahi and M. Mejri, "SpaML: a Bimodal Ensemble Learning Spam Detector based on NLP Techniques," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.07444

[11] T. Sharmin, F. Di Troia, K. Potika, and M. Stamp, "Convolutional Neural Networks for Image Spam Detection," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.01710

[12] S. Yilmaz and S. Zavrak, "EMAIL SPAM DETECTION USING HIERARCHICAL ATTENTION HYBRID DEEP LEARNING METHOD."

[13] T. Sahmoud and M. Mikki, "Spam Detection Using BERT."

[14] M. A. Uddin, M. N. Islam, L. Maglaras, H. Janicke, and I. H. Sarker, "ExplainableDetector: Exploring Transformer-based Language Modeling Approach for SMS Spam Detection with Explainability Analysis," May 2024, [Online]. Available: http://arxiv.org/abs/2405.08026

[15] S. Rojas-Galeano, "Zero-Shot Spam Email Classification Using Pre-trained Large Language Models," May 2024, doi: 10.1007/978-3-031-74595-9_1.

[16] A. Martino, "Department of Business and Management Chair of Advanced Coding for Data Analytics Assessment of Large Language Models in Spam Text Generation and Detection Capabilities."

[17] B. M. Jones and M. Omar, "Detection of Twitter Spam with Language Models: A Case Study on How to Use BERT to Protect Children from Spam on Twitter," in *Proceedings - 2023 Congress in Computer Science, Computer Engineering, and Applied Computing, CSCE 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 511–516. doi: 10.1109/CSCE60160.2023.00089.

[18] I. Androutsopoulos, J. Koutsias, K. V Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An Evaluation of Naive Bayesian Anti-Spam Filtering," 2000. [Online]. Available: http://www.cauce.org,

[19] D. Nagamalai, B. C. Dhinakaran, and J. K. Lee, "An In-depth Analysis of Spam and Spammers," 2008.

[20] P. Malhotra and S. Kumar Malik, "Spam Email Detection using Machine Learning and Deep Learning Techniques." [Online]. Available: https://www.kaggle.com/datasets/venky73/spam-mails-dataset.