A SPAM TRANSFORMER MODEL FOR SMS SPAM DETECTION

S.SARIKA[1], P RAHUL[2], G ANJANI PRASAD[3], N HARI TEJA[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF CSE, CMR INSTITUTE OF TECHNOLOGY, KANDLAKOYA VILLAGE, MEDCHAL RD, HYDERABAD, TELANGANA 501401

ABSTRACT- This paper presents a modified Transformer model designed specifically for detecting spam Short Message Service (SMS) messages. The primary goal is to explore the effectiveness of the Transformer architecture in the context of SMS spam detection. The model is evaluated using two distinct datasets: the SMS Spam Collection v.1 dataset and the UtkMl's Twitter Spam Detection Competition dataset. To benchmark the performance of the proposed model, it is compared against multiple established machine learning classifiers and other state-of-the-art SMS spam detection approaches. The experimental results indicate that the modified Transformer model achieves exceptional performance across several evaluation metrics, including an accuracy of 98.92%, a recall of 0.9451, and an F1-Score of 0.9613 on the SMS dataset. This performance surpasses that of all other comparison models, demonstrating the Transformer's strong potential for handling SMS spam classification tasks. Additionally, the model's effectiveness is not limited to the SMS dataset; it also delivers strong results on the UtkMl's Twitter dataset, further showcasing its versatility and adaptability to other similar problems. The success of the Transformer model in these tests highlights its ability to accurately identify spam messages and offers promising opportunities for its application in other domains, such as social media spam detection. This research emphasizes the Transformer model's potential as a robust solution to tackle the increasing challenge of spam messages in various forms of communication.

INDEX TERMS-SMS spam detection, Transformer model, machine learning classifiers, F1-Score, accuracy, recall, UtkMl's Twitter Spam Detection, state-of-the-art approaches.

I. INTRODUCTION

Short Message Service (SMS) has become a widely used communication tool, providing quick and efficient ways for people to stay in touch. However, as its usage grew, so did the prevalence of SMS spam. SMS spam refers to unsolicited and irrelevant messages, often sent in bulk, that disrupt user experience and create various inconveniences. The primary reasons behind the rise of spam messages are the large number of mobile phone users worldwide, the low cost of sending these messages, and the limited computational capabilities of most mobile phones to detect and filter spam effectively. These challenges highlight the need for robust and efficient spam detection methods that can operate in resource-constrained environments. Although there are existing spam filtering methods, many of them rely on traditional machine learning algorithms, which often require handcrafted features to be manually extracted from data, limiting their scalability and performance.

In recent years, deep learning techniques, particularly Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, have shown significant promise in handling spam detection due to their ability to learn and capture complex patterns in sequential data. Despite this, the introduction of Transformer models, initially designed for tasks such as machine translation, has revolutionized the field of Natural Language Processing (NLP) by leveraging self-attention mechanisms to process sequences of data more efficiently. Transformer-based models like BERT and GPT have demonstrated remarkable performance across a wide range of NLP tasks. This paper aims to explore whether the Transformer model, with its attention mechanism, can be adapted for SMS spam detection. We propose a modified version of the Transformer specifically tailored for SMS spam classification and evaluate its performance against traditional machine learning models and deep learning approaches like LSTM. Our objective is to determine if the Transformer's architecture can offer significant improvements in the accuracy and efficiency of SMS spam detection.

II.LITERATURE SURVEY

A) P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," Future Generation Computer Systems, vol. 102, pp. 524–533, 2020.

In this paper, the authors propose a deep learning-based approach to filter SMS spam using neural network architectures. As the volume of SMS spam increases globally, efficient filtering mechanisms are needed to protect users from unwanted and potentially harmful messages. Traditional methods of spam detection rely on keyword-based filtering or manual rule-based systems, which are often ineffective in handling large-scale, dynamic datasets. The paper explores various deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for their potential in classifying SMS messages as spam or non-spam. The authors specifically focus on improving the accuracy and scalability of spam detection systems by leveraging deep learning's ability to automatically learn relevant features from data without the need for manual feature extraction. They evaluate their proposed model on publicly available SMS datasets and compare the results with traditional machine learning approaches. The experimental results show that deep learning models outperform conventional methods in terms of accuracy, precision, and recall, highlighting the effectiveness of these models in handling the complexities of SMS spam classification. The proposed approach demonstrates the potential of deep learning in improving SMS spam filtering systems, making them more efficient and adaptable to changing patterns of spam messages.

B) G. Jain, M. Sharma, and B. Agarwal, "Optimizing semantic LSTM for spam detection," International Journal of Information Technology (Singapore), vol. 11, no. 2, pp. 239–250, 2019

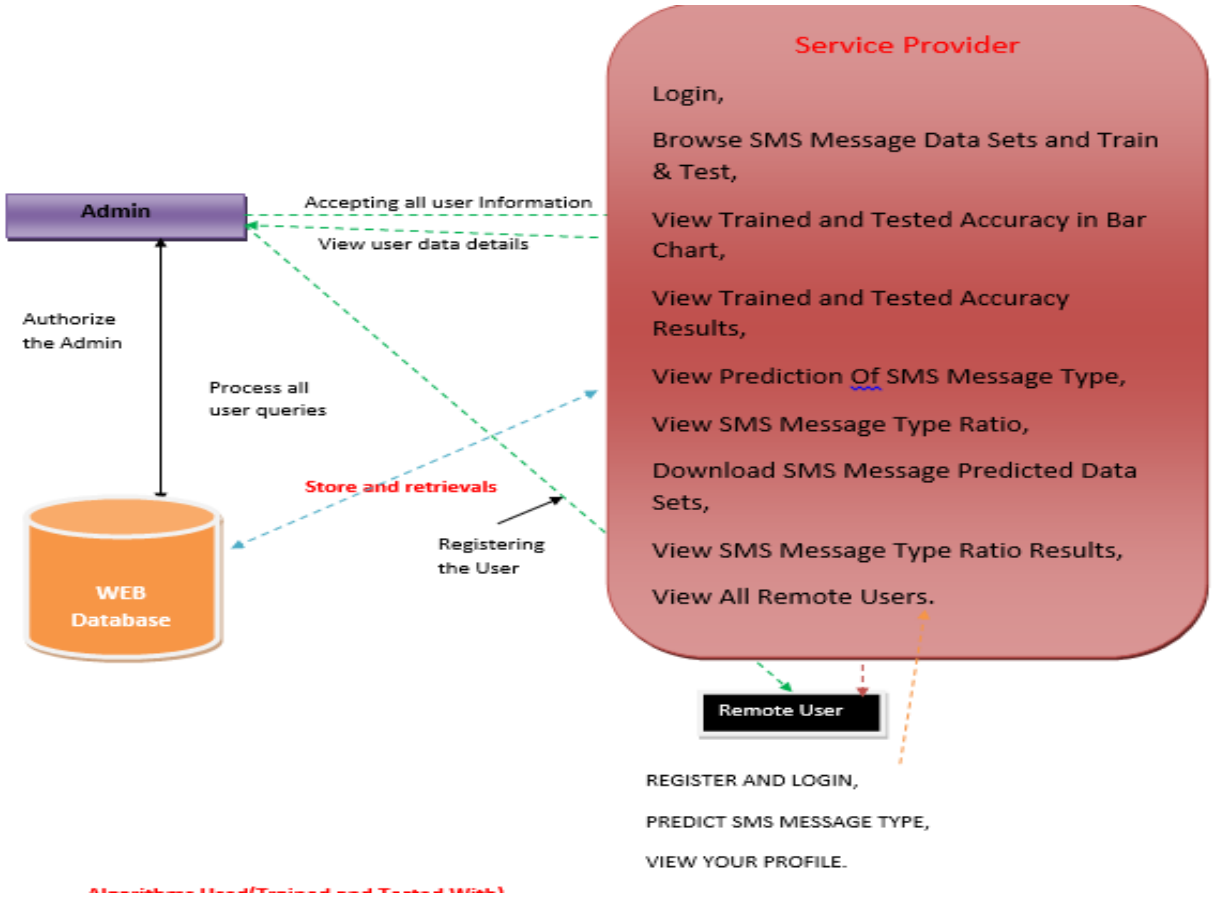
In this paper, the authors propose an optimized version of the Semantic Long Short-Term Memory (LSTM) model for spam detection in text messages. The LSTM architecture is known for its ability to capture long-range dependencies in sequential data, making it an effective tool for processing and classifying text. The proposed method aims to

improve the performance of traditional spam detection systems by enhancing the semantic understanding of the text. Instead of relying solely on surface-level features or keywords, the Semantic LSTM model incorporates semantic information, allowing it to better understand the context and meaning of the messages. The authors also optimize the model by tuning the hyperparameters and incorporating techniques like word embeddings to improve its accuracy and efficiency. The model is evaluated on several datasets, and the results show that the optimized Semantic LSTM outperforms conventional spam detection methods, achieving higher accuracy and better generalization. The findings suggest that incorporating semantic understanding into spam detection can lead to more robust and reliable systems capable of handling various forms of spam, including those with subtle or misleading content.

C) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5999–6009

In the paper *"Attention is All You Need"*, Vaswani et al. introduced the Transformer model, a novel architecture that revolutionized the field of Natural Language Processing (NLP) by replacing traditional recurrent and convolutional layers with an attention-based mechanism. This attention mechanism, specifically self-attention, allows the model to consider the entire input sequence at once, enabling it to capture long-range dependencies more effectively. Unlike previous models like RNNs and LSTMs, which process data sequentially and are limited by computational inefficiencies, the Transformer processes input data in parallel, making it faster and more scalable. The architecture is divided into two components: the encoder and decoder, both of which leverage multi-head attention to learn relationships between words, regardless of their position in the sequence. This innovation dramatically improves computational efficiency and accuracy in NLP tasks. The Transformer model achieved state-of-the-art results in machine translation, outperforming earlier models like LSTMs in both speed and translation quality. Its effectiveness and efficiency led to the development of more advanced models such as BERT, GPT, and T5, which are widely used in various NLP applications, including text classification, sentiment analysis, text summarization, and question answering. The success of the Transformer has established it as the foundation for most modern NLP models, and its ability to handle large-scale datasets with high accuracy continues to influence research and practical applications in the field.

III. PROPOSED SOLUTION



Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Antifraud Model for Internet Loan Prediction, Find Internet Loan Prediction Type Ratio, View Primary Stage Diabetic Prediction Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT PRIMARY STAGE DIABETIC STATUS, VIEW YOUR PROFILE.

CONCLUSION

In this paper, we introduced a modified Transformer model aimed at identifying SMS spam, and evaluated its performance against several other SMS spam detection methods using two datasets: the SMS Spam Collection v.1 and UtkMl's Twitter dataset. The experimental results demonstrated that our proposed spam Transformer model outperformed other classifiers, including Logistic Regression, Naïve Bayes, Random Forests, Support Vector Machine, Long Short-Term Memory (LSTM), and CNN-LSTM. Specifically, on the SMS Spam Collection v.1 dataset, our spam Transformer achieved superior results in terms of accuracy, recall, and F1-Score compared to the other models. The modified spam Transformer not only surpassed the other models in overall performance but also achieved an exceptional F1-Score, indicating its ability to correctly classify both spam and non-spam messages with a high degree of precision. Furthermore, when tested on the UtkMl's Twitter dataset, the modified spam Transformer also showed improved performance across all evaluation metrics, including accuracy, precision, recall, and F1-Score, outperforming the alternatives. Notably, the spam Transformer excelled in recall, which further contributed to a higher F1-Score, confirming the model's robustness and efficiency in detecting spam messages in both SMS and social media contexts. These promising results highlight the potential of Transformer-based models for spam detection tasks and suggest their applicability in real-world spam filtering systems.

REFERENCES

[1] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," Future Generation Computer Systems, vol. 102, pp. 524–533, 2020.

[2] G. Jain, M. Sharma, and B. Agarwal, “Optimizing semantic LSTM for spam detection,” International Journal of Information Technology (Singapore), vol. 11, no. 2, pp. 239–250, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, 2017, pp. 5999–6009.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models are Few-Shot Learners,” arXiv, may 2020. [Online]. Available: http://arxiv.org/abs/2005.14165

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), vol. 1, pp. 4171–4186, oct 2019.

[6] G. Sonowal and K. S. Kuppusamy, “SmiDCA: An AntiSmishing Model with Machine Learning Approach,” The Computer Journal, vol. 61, no. 8, pp. 1143–1157, aug 2018.

[7] J. W. Joo, S. Y. Moon, S. Singh, and J. H. Park, “SDetector: an enhanced security model for detecting Smishing attack for mobile computing,” Telecommunication Systems, vol. 66, no. 1, pp. 29–38, sep 2017.

[8] S. Mishra and D. Soni, “Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis,” Future Generation Computer Systems, vol. 108, pp. 803–815, jul 2020.

[9]T. K. Ho, “Random decision forests,” in Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, vol. 1, 1995, pp. 278–282.

[10] C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[11] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, “A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers,” in 2018 11th International Conference on Contemporary Computing (IC3), 2018.

[12] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, “Contributions to the study of sms spam filtering: New collection and results,” in Proceedings of the 11th ACM Symposium on Document Engineering, 2011, p. 259–262.

[13] A. K. Jain and B. B. Gupta, “Rule-Based Framework for Detection of Smishing Messages in Mobile Environment,” in Procedia Computer Science, vol. 125, 2018, pp. 617–623.

[14] W. W. Cohen, “Fast Effective Rule Induction,” in Machine Learning Proceedings, 1995, pp. 115–123.

[15] J. Cendrowska, “PRISM: An algorithm for inducing modular rules,” International Journal of Man-Machine Studies, vol. 27, no. 4, pp. 349–370, 1987.