A Spam Transformer Model for SMS Spam Detection

K Jaya Krishna[1], Dr. D Bujji Babu[2], Yamavarapu Venkata Sai Chadwika[3], Nuthalapati Yagnika[4], Nagumothu Naga Siva Sai Dinesh[5], Vemuri Lakshman[6]

[1]Asst.Professor,[2]Professor & HOD, [3,4,5,6]PG Scholars, Department of MCA,
QIS College of Engineering & Technology (Autonomous) Ongole, AP, India.

ABSTRACT:

In this research, we propose an adaptive Transformers model optimised for identifying SMS spam messages in order to investigate the potential of the Transformers model in this context. We conduct experiments on the SMS Spam Collection v.1 dataset and the UtkMl Twitter Spam Detection Competition dataset, comparing our suggested spam Transformers against a number of well-established machine learning classifiers and cutting-edge SMS spam detection methods. Our SMS spam detection testing demonstrate that the suggested improved spam Transformer provides the best results compared to the other alternatives. The suggested model also performs well on the UtkMl's Twitter dataset, suggesting it might be successfully adapted to other situations with comparable characteristics.

DOI Number: 10.14704/nq.2022.20.11.NQ66268 **NeuroQuantology 2022; 20(11): 2659-2667**

1. INTRODUCTION

With the proliferation of mobile phones and networks, the Short Message Service (SMS) has become an indispensable means of communication in recent years. However, spam sent over SMS is a major problem for many who utilise this communication method. The term "SMS spam," or "drunk message," is used to describe any unwanted communications sent via mobile networks. The proliferation of spam emails may be attributed to a number of factors. First, there are a lot of people with mobile phones in the globe, thus there are a lot of people who may be targeted by a spam message assault. Secondly, the spam attacker may be pleased to learn that spam messages have a minimal cost per send. Last but not least, most mobile phones' inability to accurately and effectively detect spam messages is attributable to the classifier's limited processing resources. One of the most talked-about fields in the previous two decades, machine learning has spawned a plethora of categorization applications across many fields of study. Detecting spam, in particular, is a well-established field of study that employs a number of tried and true techniques. The majority of ML-based classifiers, however, were Wei Xiang oversaw the review process and gave final approval for this article to be published in his capacity as associate editor. reliant on the carefully designed characteristics derived from the training data [2]. Deep learning is a subfield of machine learning that has shown incredible progress in recent years, all because to the exponential increase in computing power over the last several decades. These days, deep learning-based apps are all over the place, simplifying our lives in many ways. Recurrent Neural Network (RNN) and variations like Long Short-Term Memory (LSTM) have been used to spam detection and shown to be quite successful over the last several years, making them one of the most popular and commonly utilised deep learning architectures. The Transformer [3] is a sequence-to-sequence model that was designed specifically for the translation job and has had tremendous success translating between English and German and English and French. Moreover, many enhanced

eISSN 1303-5150

Transformer-based models, including GPT-3 [4] and BERT [5], have been developed lately to deal with various issues in Natural Language Processing (NLP). All that the Transformer and its predecessors have accomplished demonstrates their strength and potential. We want to find out whether the Transformer model can be modified to solve the issue of detecting spam in text messages in this study. As a result, we recommend a customised model based on the original Transformer for spotting SMS spam. We also evaluate and compare the efficacy of our proposed spam Transformer model to that of conventional machine learning classifiers, a deep learning solution based on long short-term memory (LSTM), and human experts in the field of spam identification.

2. LITERATURE SURVEY

2.1 J. P. Singh, and S. Banerjee, ‘‘Deep learning to filter SMS spam,’’ Future Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020

Over the last decade, SMS has steadily increased in use. In terms of productivity, these text messages outperform even emails for commercial use. This is due to the fact that although almost all mobile users read their SMS before day's end, only approximately 20% of emails really get opened. SMS Spam, which is the sending of unsolicited text messages via mobile networks, has become more common as SMS's popularity has grown. Users find them quite frustrating. Until recently, attempts to filter SMS Spam have depended mostly on human detected traits, despite much research into automated methods. This research goes beyond the existing literature by using deep learning to the problem of separating spam from legitimate email. In particular, models based on long short-term memory and Convolutional Neural Networks were used. The suggested models used just text input and generated their own feature set automatically. An impressive 99.44% accuracy was obtained on a benchmark dataset of 747 Spam and 4,827 Not-Spam text messages.

2.2 J. Cendrowska, "PRISM: An method for inducing modular rules"

One of Quinlan's ID3 algorithm's significant flaws is the decision tree output. It's not only hard to understand and fiddle with, but it also often requires extraneous data to be given when used in expert systems. According to the data shown here, the issue is inherent to the induction technique itself, and the only solution is to completely overhaul the underlying methodology. The paper details a novel algorithm called PRISM that, although based on ID3, employs a unique induction technique to induce modular rules, sidestepping many of the issues that plague decision trees in the process. To begin with, an introductory In recent years, several academics have focused on improving expert systems' methods of knowledge acquisition, with a particular emphasis on rule induction algorithms. A lot of focus has been placed on Ross Quinlan's ID3 algorithm (Quinlan, 1979a, 1979b, 1983a), which proved effective in the field of chess end-games and was quickly put into use in a variety of commercial settings. Despite its apparent effectiveness, the ID3 method has been shown to have significant drawbacks that render its application inappropriate in many fields (Bundy, Silver, & Plummer, 1984; Cendrowska, 1984; Hart, 1985; O'Rorke, 1982). Concern has been indicated regarding the method in which the results of the induction process are presented, however this is an active field of study (A-Razzak, Hassan &Pettipher, 1985; Hart, 1985; Lavrac et al., 1986; Michie, 1983; Quinlan, 1983b), as is the algorithm's incapacity to cope with noisy input data. Here we focus on the second of those two restrictions; its discussion fills the bulk of this

eISSN 1303-5150

study. The decision tree format that ID3 uses to display its findings might be difficult to understand, edit, and explain (by computers for humans). It is suggested that the present efforts to improve ID3's decision tree output are misguided, and that the decision tree output is a flaw in the algorithm that can only be fixed by completely revamping.

2.3"SmiDCA: An anti-Smishing model using machine learning technique," by G. Sonowal and K. S. Kuppusamy 2.4 August 2018 issue of the Computer Journal. Phishing has evolved into a major threat to online safety, and it is now being transmitted over several channels (email, SMS, etc.) in an effort to get sensitive personal data from unsuspecting victims. Despite the fact that several cutting-edge anti-phishing measures have been created, the problem of phishing continues to be unsolved. Smishing is a kind of phishing that uses text messages sent to a mobile phone using the Short Messaging Service (SMS) protocol in order to steal the user's login information. In this article, we introduce the 'SmiDCA' anti-phishing model (SMIshing Detection based on Correlation Algorithm). Initially, 39 characteristics were retrieved from the collected smishing messages using the suggested model. Feature reduction (BFSA) and non-reduction (AFSA) were tested in machine learning-based studies using the SmiDCA model. The model's validity was tested on both the English and non-English datasets, with positive results (96.40% accuracy for the English dataset and 90.33% accuracy for the non-English dataset, respectively). When almost half of the characteristics were removed, the model still performed at 96.16% accuracy.

3. **SYSTEM ANALYSIS**

3.1 EXISTING SYSYEM:

Over the last several decades, many categorization application strategies based on machine learning have been suggested. Nearly all of these methods for identifying SMS spam rely on time-tested machine learning strategies including Logistic Regression (LR), Random Forest (RF) [10], Support Vector Machine (SVM) [11], Nave Bayes (NB), and Decision Trees (DT). Many new approaches have been created to combat SMS spam in recent years, many of which rely on deep learning-based solutions like Convolutional Neural Networks (CNN).

DISADVANTAGES:

- Low Accuracy.
- However, most ML-based classifiers relied on characteristics manually retrieved from the ML-based classifier's training data.

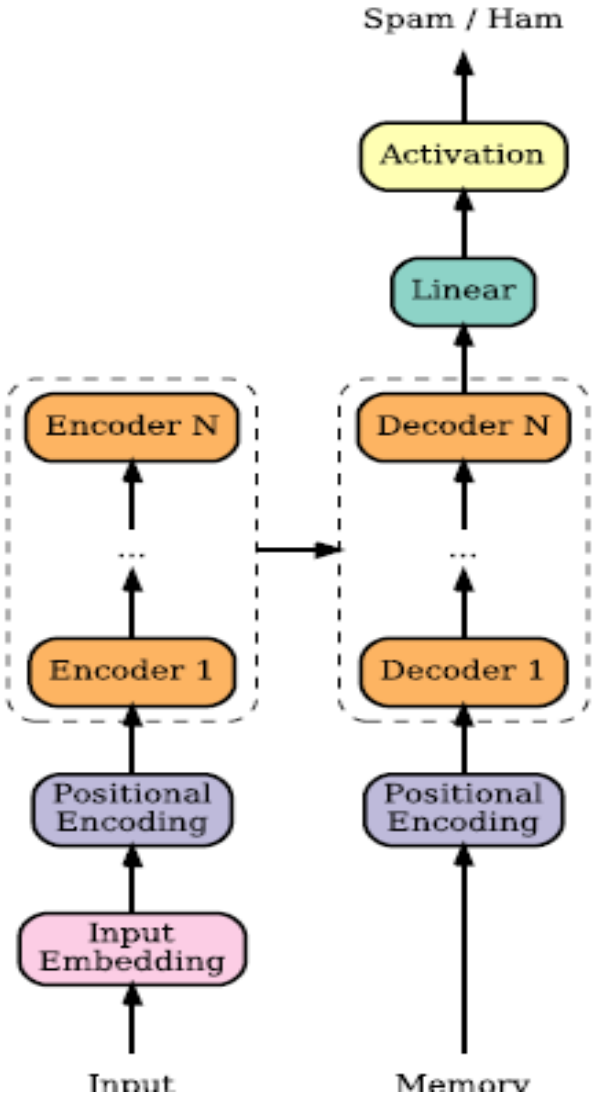
3.2 PROPSED SYSTEM:

A sequence-to-sequence model, as in the Transformer (Seq2Seq). Transformer's primary novelty is that, unlike earlier Seq2Seq models, it relies only on the attention mechanism to effectively learn from the most important bits. While it has been shown that LSTM and other RNN variations work well as encoders and decoders in Seq2Seq based models, the high training consumption of recurrent models becomes a key limitation. This sequential structure of processing precludes parallelization, computational efficiency, and RNN variant training.

ADVANTAGES:

- A great degree of accuracy is achieved.
- RNN models' ability to remember and apply data from the past is a huge benefit when it comes to natural language processing (NLP) issues, where the context is often crucial to a correct interpretation of a phrase.

4. METHODOLOGY



The SMS spam detection task mostly benefits from more RAM. Since the SMS spam detection task does not have an output sequence (target sequence), we substituted a set of trainable parameters dubbed "memory" for output sequence embedding. The amount of RAM may be changed as a hyper-con gurable. Each memory byte is a vector of dimension dmodel so that it may be readily adapted to the Transformer model. We store our memories in a matrix with the dimensions lenmemorydmodel. To save space, the output embedding layer from the original Transformer model has been removed since that there are no longer any target sequence texts to be transformed to numeric vectors. Similar to how the output sequence in the basic Transformer model works, the positional information is injected into memory at the positional encoding layer before being passed into decoders. The memory matrix and its accompanying parameters are taught to classify messages as spam or not. Thus, the updated spam Transformer model's decoders may use memory to help in locating the crucial part of the output sequence of the encoder stack that summarised the message, thus facilitating the categorization of spam SMS messages using an attention mechanism. The _nal activation function is the _cation's second functional analogue. T dmodel, where T is the target sequence length and dmodel is the model size, describes the dimensions of the outputs from the decoder layers in the

eISSN 1303-5150

foundational Transformer (number of features). To find the closest candidate word in the dictionary, it makes sense to use the linear layers to convert the output to a vector of the same dimension as the number of words in the dictionary, and then to apply a softmax function to the vector. SMS spam detection is a binary classification challenge, which is a major setback. Therefore, the subsequent linear layers are also modied to turn the output from the decoder stacks of dimension dmodel into a singular probability that the message is spam. Instead of using a vector map to the decoder stack's output, a single neuron makes up the linear layer in the updated Transformer model for SMS spam detection. This means that the probabilities sent by the decoder stack's outputs may be expressed as a single value. In addition, the _nal activation function should be replaced with another function that can transform the result to a binary outcome.

5. CONCLUSION

In this study, we improved upon an existing Transformer model to detect spam SMS. We evaluated our spam Transformer model against many other approaches to SMS spam detection using the SMS Spam Collection v.1 dataset and the UtkMl Twitter dataset. Experiments show that compared to state-of-the-art algorithms like Logistic Regression, Naive Bayes, Random Forests, Support Vector Machine, Long Short-Term Memory, and CNN-LSTM [22], our proposed spam Transformer model performs better. Our spam Transformer has better accuracy, recall, and F1-Score than any other classifier used to the SMS Spam Collection v.1 dataset. Particularly, we enhanced our prior spam Transformer approach, resulting in a state-of-the-art F1-Score. In addition, our improved spam Transformer model beats the prior approaches mentioned in this study on the UtkMl Twitter dataset in all four metrics. Our spam Transformer, for instance, has an excellent F1-Score because of its amazing recall capability.

REFERENCES

[1] P. K. Roy, J. P. Singh, and S. Banerjee, ‘‘Deep learning to filter SMS spam,’’ Future Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020.

[2] G. Jain, M. Sharma, and B. Agarwal, ‘‘Optimizing semantic LSTM for spam detection,’’ Int. J. Inf. Technol., vol. 11, no. 2, pp. 239–250, Jun. 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5999–6009.

[4] T. B. Brown et al., ‘‘Language models are few-shot learners,’’ 2020, arXiv:2005.14165. [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,’’ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, Jun. 2019, pp. 4171–4186.

[6] G. Sonowal and K. S. Kuppusamy, ‘‘SmiDCA: An anti-Smishing model with machine learning approach,’’ Comput. J., vol. 61, no. 8, pp. 1143–1157, Aug. 2018.

[7] J. W. Joo, S. Y. Moon, S. Singh, and J. H. Park, ‘‘S-detector: An enhanced security model for detecting Smishing attack for mobile computing,’’ Telecommun. Syst., 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 Gener. Comput. Syst., vol. 108, pp. 803–815, Jul. 2020.

[9] C. Li, L. Hou, B. Y. Sharma, H. Li, C. Chen, Y. Li, X. Zhao, H. Huang, Z. Cai, and H. Chen,

eISSN 1303-5150

''Developing a new intelligent system for the diagnosis of tuberculous pleural effusion,'' Comput. Methods Programs Biomed., vol. 153, pp. 211–225, Jan. 2018.

[10] T. K. Ho, ''Random decision forests,'' in Proc. Int. Conf. Document Anal. Recognit. (ICDAR), vol. 1, 1995, pp. 278–282.

[11] C. Cortes and V. Vapnik, ''Support-vector networks,'' Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[12] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, ''A comparative study of spam SMS detection using machine learning classifiers,'' in Proc. 11th Int. Conf. Contemp. Comput. (IC3), Aug. 2018, pp. 1–7.

[13] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, ''Contributions to the study of SMS spam filtering: New collection and results,'' in Proc. 11th ACM Symp. Document Eng., Sep. 2011, pp. 259–262.

[14] A. K. Jain and B. B. Gupta, ''Rule-based framework for detection of Smishing messages in mobile environment,'' ProcediaComput. Sci., vol. 125, pp. 617–623, 2018.

[15] W. W. Cohen, ''Fast effective rule induction,'' in Machine Learning Proceedings, 1995, pp. 115–123. [16] J. Cendrowska, ''PRISM: An algorithm for inducing modular rules,'' Int. J. Man-Machine Stud., vol. 27, no. 4, pp. 349–370, Oct. 1987.

[17] J. H. Friedman, ''Greedy function approximation: A gradient boosting machine,'' Ann. Statist., vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[18] L. Bottou, ''Large-scale machine learning with stochastic gradient descent,'' in Proc. COMPSTAT.Physica-Verlag, 2010, pp. 177–186. [19] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' in Proc. Int. Conf. Learn.Represent., 2013.

[19] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean,

"Efficient estimation of word representations in vector
space," in International Conference on Learning Repre-
sentations, 2013.

[20]G. A. Miller, "WordNet: A Lexical Database for En-
glish," Communications of the ACM, vol. 38, no. 11,
pp. 39–41, 1995.

[21]H. Liu and P. Singh, "ConceptNet - a practical com-
monsense reasoning tool-kit," BT Technology Journal,
vol. 22, no. 4, pp. 211–226, 2004.

[22]A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A
hybrid CNN-LSTM model for SMS spam detection in
arabic and english messages," Future Internet, vol. 12,
no. 9, p. 156, 2020.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams,
"Learning representations by back-propagating errors,"
Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[24]Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-
Term Dependencies with Gradient Descent is Difficult,"
IEEE Transactions on Neural Networks, vol. 5, no. 2,
pp. 157–166, 1994.

[25]R. Pascanu, T. Mikolov, and Y. Bengio, "On the diffi-
culty of training Recurrent Neural Networks," 30th In-
ternational Conference on Machine Learning (ICML),
no. PART 3, pp. 2347–2355, 2013.

[26]S. Hochreiter and J. Schmidhuber, "Long Short-Term
Memory," Neural Computation, vol. 9, no. 8, pp. 1735–
1780, 1997.

[27] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[28]J. Koutník, K. Greff, F. Gomez, and J. Schmidhuber, "A Clockwork RNN," 31st International Conference on Machine Learning (ICML), vol. 5, pp. 3881–3889, 2014.

[29]C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," arXiv:1511.08630, 2015.

[30]I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems, vol. 4, no. January, pp. 3104–3112, 2014.

[31]R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in Proc. Interspeech 2017, 2017, pp. 939–943.

[32]S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," in IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4534–4542.

[33] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in 3rd International Conference on Learning Representations (ICLR), 2015.

[19] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in International Conference on Learning Representations, 2013.

[20]G. A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[21]H. Liu and P. Singh, "ConceptNet - a practical commonsense reasoning tool-kit," BT Technology Journal, vol. 22, no. 4, pp. 211–226, 2004.

[22]A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, "A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages," Future Internet, vol. 12, no. 9, p. 156, 2020.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[24]Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[25]R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training Recurrent Neural Networks," 30th In-

ternational Conference on Machine Learning (ICML),
no. PART 3, pp. 2347–2355, 2013.
[26]S. Hochreiter and J. Schmidhuber, “Long Short-Term
Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–
1780, 1997.
[27] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bah-
danau, F. Bougares, H. Schwenk, and Y. Bengio,
“Learning phrase representations using RNN encoder-
decoder for statistical machine translation,” in Con-
ference on Empirical Methods in Natural Language
Processing (EMNLP), 2014, pp. 1724–1734.
[28]J. Koutník, K. Greff, F. Gomez, and J. Schmidhuber,
“A Clockwork RNN,” 31st International Conference
on Machine Learning (ICML), vol. 5, pp. 3881–3889,
2014.
[29]C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, “A C-LSTM Neural Network for Text Classification,”
arXiv:1511.08630, 2015.
[30]I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to
Sequence Learning with Neural Networks,” Advances
in Neural Information Processing Systems, vol. 4, no.
January, pp. 3104–3112, 2014.
[31]R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. John-
son, and N. Jaitly, “A comparison of sequence-to-
sequence models for speech recognition,” in Proc. In-
terspeech 2017, 2017, pp. 939–943.
[32]S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney,
T. Darrell, and K. Saenko, “Sequence to sequence –
video to text,” in IEEE International Conference on
Computer Vision (ICCV), 2015, pp. 4534–4542.
[33] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural
machine translation by jointly learning to align and
translate,” in 3rd International Conference on Learning
Representations (ICLR), 2015.

[19] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in International Conference on Learning Representations, 2013.

[20] G. A. Miller, “WordNet: A Lexical Database for English,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[21] H. Liu and P. Singh, “ConceptNet - a practical commonsense reasoning tool-kit,” BT Technology Journal, vol. 22, no. 4, pp. 211–226, 2004.

[22] A. Ghourabi, M. A. Mahmood, and Q. M. Alzubi, “A hybrid CNN-LSTM model for SMS spam detection in arabic and english messages,” Future Internet, vol. 12, no. 9, p. 156, 2020.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[24] Y. Bengio, P. Simard, and P. Frasconi, “Learning LongTerm Dependencies with Gradient Descent is Difficult,” IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[25] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” 30th International Conference on Machine Learning (ICML), no. PART 3, pp. 2347–2355, 2013.

[26] S. Hochreiter and J. Schmidhuber, “Long

Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735– 1780, 1997.

[27] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoderdecoder for statistical machine translation," in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.

[28] J. Koutník, K. Greff, F. Gomez, and J. Schmidhuber, "A Clockwork RNN," 31st International Conference on Machine Learning (ICML), vol. 5, pp. 3881–3889, 2014.

[29] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, "A C-LSTM Neural Network for Text Classification," arXiv:1511.08630, 2015.

[30] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems, vol. 4, no. January, pp. 3104–3112, 2014.

[31] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-tosequence models for speech recognition," in Proc. Interspeech 2017, 2017, pp. 939–943.

[32] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence – video to text," in IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4534–4542.

[33] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in 3rd International Conference on Learning Representations (ICLR), 2015.

[34] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in 32nd International Conference on Machine Learning (ICML), vol. 3, 2015, pp. 2048– 2057.

[35] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1412–1421.

[36] E. S. D. Reis, C. A. D. Costa, D. E. D. Silveira, R. S. Bavaresco, R. D. R. Righi, J. L. V. Barbosa, R. S. Antunes, M. M. Gomes, and G. Federizzi, "Transformers aftermath," Communications of the ACM, vol. 64, no. 4, pp. 154–163, apr 2021.

[37] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv:1207.0580, 2012.