IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Performance Comparison of a Transformer Model with Sparse and Standard Attention in Fake News Classification

Adhavan S. Viswapurna

British School Muscat, PO Box 1907, Ruwi 112, Sultanate of Oman; adhavansenviswa@gmail.com

ABSTRACT: The recent increased prevalence of false information across numerous internet platforms has made distinguishing facts from misinformation difficult among the general public. This concern is further exacerbated by the advancement of AI. The main purpose of this research is to propose a transformer model that aims to resolve the issue outlined above. With the intention of assessing the model performance of classifying labelled data with both sparse attention (SA) and the standard multi-head attention (MHA), this research aims to ascertain the relative effectiveness of the sparse attention model compared to standard transformer models, utilizing multi-head self-attention, when classifying fake news. The described models adopt an encoder architecture, where the encoder will aid the model in learning the contextual data. The models are trained with data of fake or biased information, provided by the LIAR dataset, to maximize the performance of the model. Furthermore, the sparse attention mechanism in the model is based on the BigBird model to maximize analysis of large text. The SA transformer model presented in this research achieved 20% higher accuracy results as well as a 4.5% increase in F1 score when compared to a baseline transformer model.

KEYWORDS: Computer Science, Transformers, Fake News Detection, Multi-Head Attention, Sparse Attention.

Introduction

In recent times, online news and social media platforms have become the primary medium of acquisition and retention of information. Consequently, the predicament of fabricated narratives has become an increasingly pervasive issue,[1] making them indistinguishable from authentic facts. Moreover, advancements in Artificial Intelligence (AI) have allowed superficially authentic sources to circulate information rapidly.[2] Generative AI can create near-authentic images and videos, which further complicates information across the internet rapidly.[3] This circumstance was particularly present during the COVID-19 pandemic,[4] under which social media facilitated the distribution of malicious content. A multitude of alleged cures for the virus was exacerbated on social media, which resulted in the term infodemic, which can be defined as a significant volume of misinformation that advances quickly across the internet.

Deep learning models have proven to be effective in text classification and AI-generated news. Therefore, this research article presents a transformer model with a sparse attention mechanism to classify fake or biased information.[5] The transformer model, which was first presented in 2017, was a revolutionary model for the sequential analysis of text. The architecture of the model consists of an encoder-decoder structure entailing an attention mechanism.[6] The purpose of the encoder is to process the input statement that is given, and the role of the decoder is to process the input and give the desired output.[6] The model mentioned above contains elements such as positional encodings and layers of normalization that proved to be more effective than Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN).

The fundamental equation (1) that is employed in the model is the attention equation, where q represents the query, k represents the key matrix, and v represents the value matrix. Dk is the dimension of the key vectors, which is utilized for scaling. $Q * K\hat{}T$ is used to measure the amount of similarity for each query and key. It is scaled to prevent large dot products once the dk is substantial. The activation function of softmax is to convert the scores into probabilities, with a higher score equating to closer attention. It is then multiplied by the value matrix to use the attention weights to compute a weighted sum of the value vectors.[6]

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

However, the model in this research article presents a modified version of this architecture. It utilizes a sparse attention (SA) mechanism instead of a multihead attention (MHA), theorizing whether or not the model's accuracy improves following this modification. Specifically, the model will be derived from a BigBird attention mechanism.[7] This hybrid mechanism is dependent on three parts: Local Attention, Global tokens, and Random Attention. To ensure that no bias or any other outstanding factors affect the performance of the model, it will be trained under a shuffled dataset, and a train-test split will be applied. This research aims to investigate the model's performance and how accurate the model becomes in fake news classification after modifying the attention mechanism. The results will be compared to other standard transformer models, which utilize the standard multi-head attention mechanism. The null hypothesis of this paper is that replacing the standard multi-head attention mechanism in a Transformer model

DOI: 10.36838/IJHSR89.39

with a sparse attention mechanism will not lead to measurable changes in performance or efficiency

This research article consists of 5 sections. Section 1 entails a literature review, analyzing methods to detect fake news and their relevance. Section 2 will elaborate on the datasets that are used to train the model, as well as the architecture that is utilized. The third section will be a comparison and review of the results of the empirical study. Section 4 will highlight the significance of the model and evaluate the conclusions drawn from the model. The final section will conclude this paper by summarizing the central hypothesis, key findings, and outlining potential directions of future research to advance this field, as well as discussing limitations of the study.

Related Work:

Having outlined the initial context of the problem, this section addresses the most relevant work in relation to this paper. Lu & Yao presented a deep learning model that utilized multi-head attention mechanisms and Residual Networks to effectively work against the vanishing gradient issue. The research paper utilized a hybrid CNN architecture to enhance the performance of the Natural Language Processing (NLP) tasks that it is capable of. The residual attention mechanism consists of 3 core components, which make it beneficial to mitigate the vanishing gradient problem. The performance of the model when presented with datasets such as Liar, FakeNewsNet, and Weibo was impactful. It obtained an accuracy score of 0.997 by increasing the sample size to 3000, consequently surpassing the accuracy of other common models. Nevertheless, the model presented was unable to have specific classifications of subjects in identifying misinformation. As a result of the model having not encountered any sample as such during the trial phases, the generalizability of the model is limited. However, the overall performance of the model exceeded Bidirectional Encoder Representations from Transformers (BERT).[8,9] It also outperformed the Robustly Optimized BERT Pretraining Approach (RoBERTA) model.[10]

Rout *et al.* outlined an optimized transformer model specifically adapted for intricate language arrangements. The model consisted of an enhanced MHA with a robust positional encoding mechanism and a computationally effective classification head. However, the architecture of the model shares many attributes with a normal transformer model, with the layers of tokenization, embedding, encoding, and a Feed Forward Neural Network. The model managed to attain a mean accuracy mark of 79.85% and 79.84% in F1. Regarding the dataset of FakeNewsNet, the model obtained scores of 80%, 100%, and 59.56%. Nonetheless, these scores should not be weighed into deep learning news classification as the model did not have an extensive volume of real-world labeled data.[11]

Liu proposed a dual-component system to detect fake news, which included a system of Content-Driven Encoder (CoDE) and the Domain-Informed Detection Adapter (DiDA). CoDE utilized a bold attention mechanism with a hierarchical semantic encoding and domain-aware feature modulation in order to allow adaptive representation learning across unimodal and multimodal inputs. DiDA-infused domain priors through conservative learning and retrieval, improving inference robustness beneath concentrated domain shifts. When faced with the dataset FakeNewsNet, the model scored an accuracy of 95% and an F1 score of 92%.[12] However, DiDA relied on prior accurate domains for contrastive learning, which might limit performance in scenarios where domain information is unreliable.

Methods

Model Architecture:

Building on these prior findings, this section describes the methodology that is utilized in this paper. The model that is used in this research to classify fake news is a transformer model with a BigBird sparse attention mechanism. The model has many similarities to a standard transformer model, but is modified for our specific task. The model has been developed using the coding language Python and the libraries of Torch, pandas, re, collections, and Sklearn. To prepare the model, the dataset that is used has been loaded and classified into two values: 1, for Fake news, and 0, for factual information. After this, the two csv files have been combined into a single dataframe, and the order of the information is randomized to maximize the training of the model. The dataframe is then cleaned using the function clean_txt and tokenized to develop a standard of the most common words. The model then begins with an embedding layer, converting the IDs of the token into vector values, 64 as per our d_model, which is then followed by a positional encoding layer to preserve the sequence of the tokens. Figure 1 is a diagram of the standard transformer model, consisting of an encoder-decoder structure.

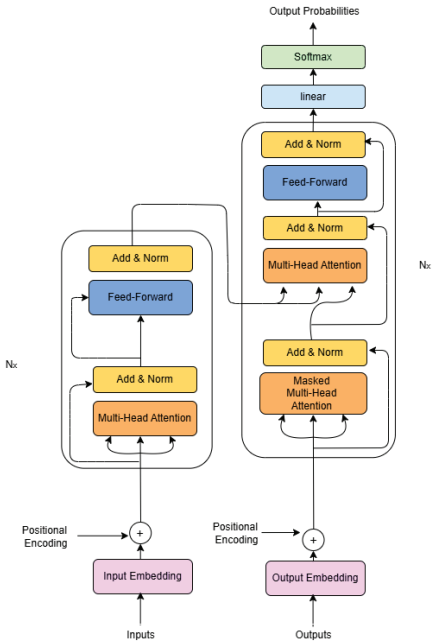

Figure 1: The diagram presents a Standard Transformer Model, which utilizes an encoder-decoder architecture to process inputs and give desired outputs. The left side conveys the encoder, while the right side conveys the decoder. The multi-head attention layers are depicted between the inputs and the outputs, as well as positional encodings added into the input embeddings to store the order of the sequence. The arrows display the flow of information through attention and feed-forward layers. This figure provides the baseline transformer model used for comparison with the proposed sparse attention model.

The next layer is the sparse attention layer, based on the BigBird Model. This is the section of the model that differs from a standard transformer. The sparse attention mechanism consists of a sparse mask to limit the computation of attention each token conducts to every other one. It generates a binary matrix to define which token attends to which. This is divided into three attention masks: local, global, and random. Every transformer block entails a sparse self-attention layer, a layer of normalization, an FNN, another layer of normalization, and a residual connection. Subsequently, a classification head follows the transformer blocks, with the purpose of implementing mean pooling layers across the length of the sequence. The model is then trained with the dataframe established at the start. In this case, the model is trained in loops over batches of data, resetting the gradients, computing the training loss, backpropagating, and updating the weights. Finally, the training loss and validation accuracy will be presented after running each epoch. Figure 2 highlights the architecture of this research paper's model.

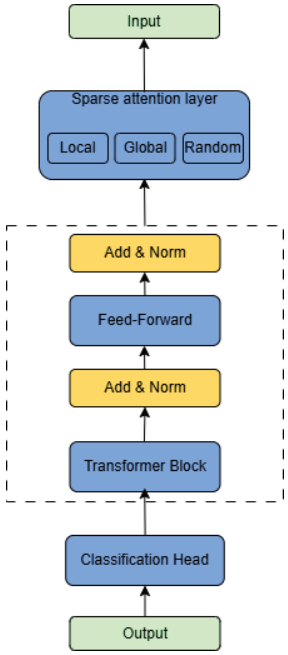

Figure 2: The image displays an encoder-only model with BigBird architecture. It differs from a standard transformer due to the addition of sparse attention. The model takes an input, which then passes through the sparse attention layer, separated into local, global, and random attention. The output then passes through a transformer block, including a feed-forward layer with residual "Add & Norm" connections. Then the information is passed to a classification head to continue to the output. This architecture reduces attention complexity and enables efficient processing of long text sequences compared to standard transformers.

BigBird Architecture:

Now that the BigBird model has been established, this section outlines the model and the attention mechanisms' function. TheBigBird model has 3 types of attention: Local, global, and random. Local attention is used as each takes only to attend to a standard window of nearby tokens. This captures the short-range dependencies much like a CNN model, with each word only tending to ones nearby. When a model utilizes full attention, each token computes attention scores with every other token using the formula (1). This essentially states that a mask is applied to the layer to restrict the number of calculations from a full matrix to a diagonal band. Figure 3 presents the diagonal window of tokens that each token can attend to. For example, the window size is limited to 2, the token (i) can attend to i-2,i-1, i, i+1, i+2. Then the attention score formula is applied:

$$M_{ij}^{local} = \begin{cases} 1 & \text{if } |i-j| \leqslant w \\ 0 & otherwise \end{cases} \quad (2)$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

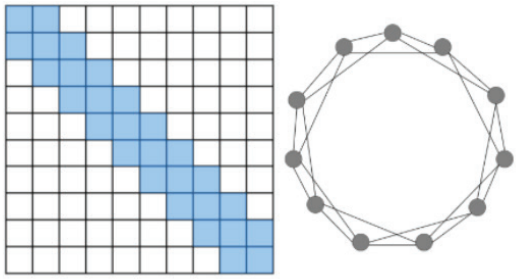

Figure 3: TThe image illustrates the local attention mechanism. It conveys the connections between tokens and how local attention limits the window that each token can attend to any other one. The token in the image has no more than 4 connections and only to its closest tokens. This results in reduced computational expenses. By restricting attention to neighbouring tokens, this mechanism lowers the computational cost while maintaining local contextual information.

The use of global attention here is vital for text classification as any global dependencies are not essential and can be subsided with global tokens. Global attention selects wider tokens, such as CLS or section headers, that can pay attention to all tokens. This prevents the network from being trapped in local windows and allows any information to be transmitted globally. Figure 4 conveys a visual representation of the connections between tokens. The formula (3) below refers to a global attention mask with $Mi,j\ global$ determining whether token i is able to attend to j^{**}, and G is the set of global tokens. $i \in G$ states that the token is global and can attend to every other token, while $i \in G$ states that the token is globally visible and can be seen by any other token. If any of these two conditions are met, attention can be applied.

$$M_{ij}^{global} = \begin{cases} 1 & \text{if } i \in G \text{ or } j \in G \\ 0 & otherwise \end{cases} \quad (3)$$

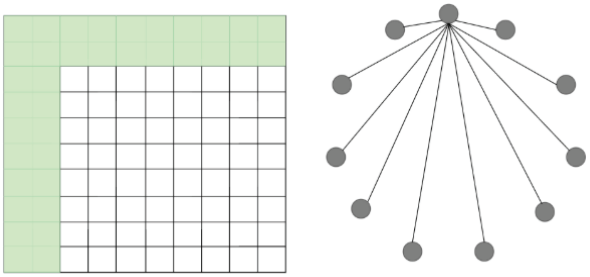

Figure 4: The global attention mechanism allows certain tokens, such as the CLS tokens, to attend to all tokens that are in that sequence. This ensures that important summaries or classification tokens have full access to the context. This avoids the need for every token to compute attention to each other. Global attention ensures that classification retains access to full-sequence context, which is critical for higher accuracy in fake news detection.

DOI: 10.36838/IJHSR89.39

Random attention selects in no sequential order any few tokens to attend to them. It allows for a model to not depend on tokens to be attended globally for long-range dependencies, along with connections in different parts of the sequence to pass information. It avoids calculating all n2 token pairs, limiting to r number of connections. The visual representation of these processes is presented in Figure 5. The formula (4) below refers to a random attention mask with $Mi,jrand$ as an entry in the random attention matrix, Ri as a randomly selected subset of positions that $I\ can$ attend to. If the position of j is included within the random subset, then attention can be applied. Otherwise, attention is not paid to the value. This allows the model to utilize full attention with fewer computations.

$$M_{ij}^{rand} = \begin{cases} 1 & \text{if } j \in R_i \\ 0 & otherwise \end{cases} \quad (4)$$

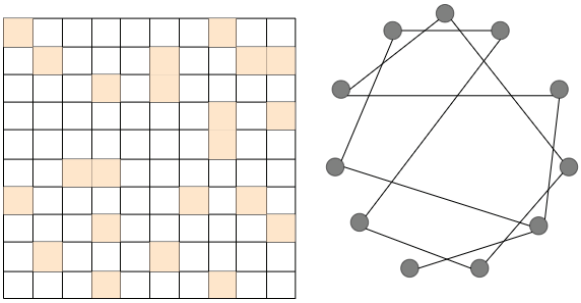

Figure 5: Diagram of the random attention mechanism in the BigBird architecture. The mechanism selects a random subset of tokens to attend to. The highlighted connections in the figure convey how information can be transferred across long-range positions in the initial input sequence, which supports the classification of larger statements. This ensures long-range dependencies are caught whilst not wasting computational energy. Random attention enables the model to capture long-range dependencies without the use of full attention computation.

Feed-Forward Neural Network:

In addition to the BigBird architecture, a feed-forward neural network (FNN), which is one of the simpler neural networks, is utilized. It applies linear transformation, non-linear activation functions, and layered abstraction to learn patterns from data. They consist of a layered structure of an input layer, a hidden layer, and an output layer, with each neuron fully connected to the others. The purpose of an FNN is to transform data from one representation to another one, introduce non-linearity, and build layered abstractions. In the model presented in this paper, it works on each token's embeddings individually to transform the token vector after attention has changed its context. After the attention mechanism, the FNN expands the dimensionality of the token, then makes use of the ReLU function to introduce the nonlinearity, and finally compresses the token back to its original size. This is to ensure that they can be used in stacking, and a residual connection can be applied.[13] Sharma *et al.* provide a comprehensive review of feedforward neural networks. A diagram of this process is present in Figure 6.

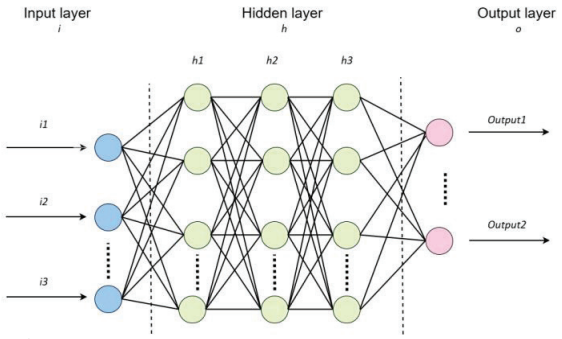

Figure 6: The illustration conveys a Feed-Forward Neural Network with an input layer, a hidden layer, and an output layer. Every node in a layer is connected to every node in the subsequent layer by weighted connections. The network processes input features via these layers, applies activation functions, and then presents output predictions. This network refines token representations after attention to improve feature abstraction.

Residual connections are streamlined methods that bypass a layer and add the input of that layer directly to its output. This assists the model to learn modifications in addition to the original input, as opposed to learning from basics. This is used in order to fight against the vanishing gradient problem. Deep networks may struggle to overcome gradients backward past many layers: residuals can provide gradients to flow directly back to previous layers, facilitating a more stable training. The model uses residual connections twice, once in the sparse attention and the second in the FNN. Each residual connection wraps around the sub-layer and is then followed by a layer of normalization.

Functions:

In this model, the two main types of activation functions that have been utilized are ReLU and Softmax. The activation function of ReLU in the model is used in the FNN, and it presents non-linearity to allow the model to understand complex patterns and sequences. It is preferred here, as unlike other activation functions, such as Sigmoid and tanh, it helps prevent the vanishing gradient issue. The function ReLU is an elementwise function: it will convert all negative values to 0 and leave all positive values unchanged, which is represented in the graph of Figure 7. This may result in a limited number of neurons being completely shut off, creating a gated non-linearity.

$$ReLU(x) = max(0, x) \quad (5)$$

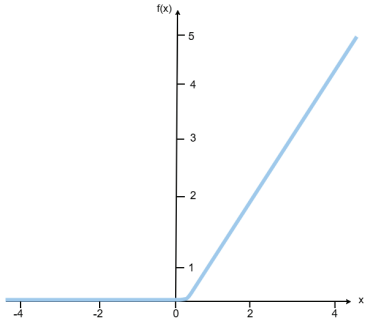

Figure 7: The graph shows the ReLU function, where every negative input number is plotted to zero, and positive numbers remain unchanged. The ReLU function is very computationally efficient, which provides faster training. This is due to the function adding non-linearity to avoid the vanishing gradient issue. The ReLU function improves training stability and speed by preventing the vanishing gradient issue.

The second activation function, which is represented in Figure 8, that this model employs is the Softmax function in the attention layers. This function normalises attention scores to convert them into probabilities. The scores are scaled and masked, then the function is applied to convert them into attention weights. This is vital for the model as it allows the model to focus on the most important tokens, the ones that are deemed to be weighted the most, and allows them to be classified. The weights are calculated with the formula:

$$Softmax(s_i) = \frac{e^{s_i}}{\sum_{j \in allowed(i)} e^{s_j}} \quad (6)$$

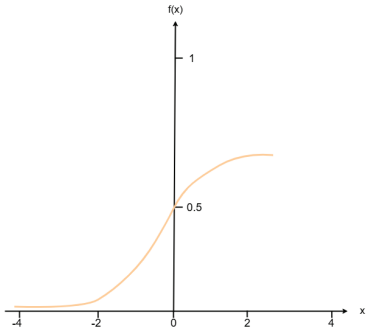

Figure 8: The graph conveys the softmax function. It converts a vector of scores into probabilities that are equal to 1. Each curve depicts how increasing one input value increases its output probability whilst decreasing other values. It is vital in neural networks to provide a clear measure of the model's predicted outcomes. Softmax allows the model to assign higher importance to informative tokens by converting attention scores into probabilities.

Dataset:

Having established the model architecture, the next step is to describe the dataset that is utilized to train and evaluate the model. The dataset that will be utilized in this paper is the LIAR dataset. The dataset consists of a range of factual and fake news, presented as short-labeled statements. These were collected through PolitiFact.com API. This website is suitable for fake news classification due to its being an accuracy assessment of each dataset. As opposed to other available datasets, LIAR provides a larger selection of short statements of genuine and false information. It contains over 12,800 labeled statements. The dataset labels each statement with one of the following: Pants on fire, extremely inaccurate, False, Half-True, and True. In order to allow models to train effectively, there is a healthy amount of data in each labeled category. LIAR is unique due to its range of political speakers, including Republicans, Democrats, and other independent figures. Furthermore, the statements which have been collected vary from a range of topics, all including Radio, Interviews, Speeches, Twitter posts, Political debates, and Facebook posts. In addition to the sources, the dataset guarantees a broad coverage of most topics, with the top ten most pertinent being: Taxes, economy, healthcare, Federal budget, Jobs, education, state budget, immigration, candidate biographies, and elections. The dataset presents how labeled statements are conveyed for training and evaluation.[14]

Results

After training the model with the LIAR dataset, the following section outlines the performance of the sparse attention model. To evaluate the accuracy and efficiency of the model, the model was presented with test values from the LIAR dataset in order to analyze whether the model's performance is better than existing models and whether it is efficient at fake news classification. This research collects the data from many other transformer models utilizing standard attention mechanisms to compare the effectiveness of our sparse attention model. The model uses the base transformer model and the multi-head attention transformer-based model to compare our results.[15,16] The model uses 4 layers and sets a maximum number of epochs at 15 with a maximum batch size of 48. The model utilized 8 attention heads to allow for more precise classification and increase performance. The dimension size is 256 to have a balanced size for token embeddings. The hyperparameters used for our model are established in Table 1 below:

Table 1: The table of contents presents the chosen hyperparameters to optimize training and validation scores. These hyperparameters were selected to balance model capacity, training stability, and computational efficiency.

Batch size	32
Epochs	15
Number of layers	2
Dimensions	256
Number of Heads	12

Training and Validation Loss:

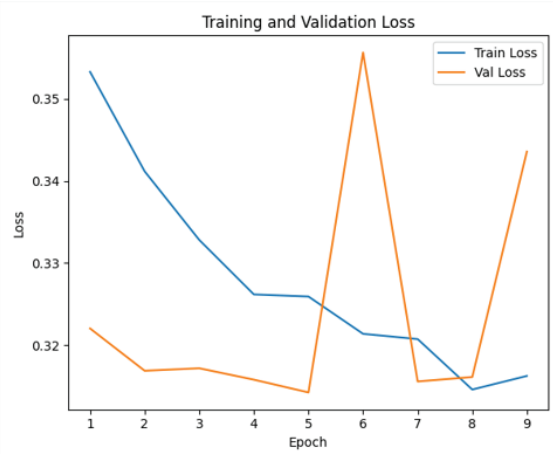

Figure 9: The graph presents the training and validation loss that the model achieves using the AdamW optimizer. Overall, there is a decrease in loss over the epochs, with only a few spikes in the validation loss. Gradually, the two lines of training and validation do reduce. The overall downward trend in both curves indicates effective model training, despite minor validation instability at specific points.

In Figure 9, the training loss and validation loss decrease steadily across each epoch. As well as the overall low loss vowels being at around 0.31-0.36, suggesting the model fits adequately well. However, the sharp fluctuations at epochs 6 and 9 may suggest instability or sensitivity to the validation

DOI: 10.36838/IJHSR89.39

set. The absence of a downward trend may suggest noisy validation, and at some point, the loss increases whilst training, which could indicate overfitting.

Training and Validation Accuracy:

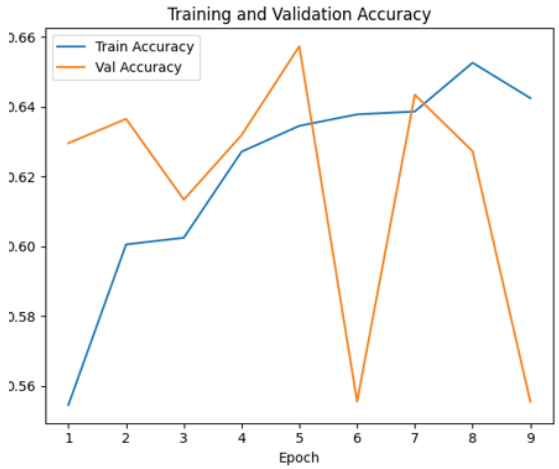

Figure 10: The graph displays the train and validation accuracy over each epoch. Improvement of training and validation accuracy over time, with some declines at certain epochs. The training and validation lines exhibit a moderate degree of linearity for the overall performance. The small gap between training and validation accuracy suggests limited overfitting and reasonable generalization performance.

The training accuracy steadily improves over each epoch, reaching above 0.64 at epoch 5 in Figure 10. This indicates consistent learning from the model. The gaps between the train and validation are small at most points, which means that the overfitting is reduced in those moments. However, there is still some fluctuation of the accuracy score, indicating the model is unstable and is susceptible to overfitting. This can be common within models that are trained on large amounts of the same data.

Training and Validation F1:

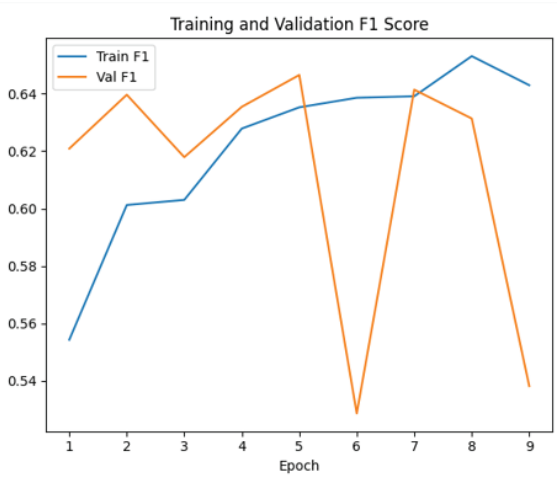

Figure 11: The graph displays the train and validation F1 over each epoch. There is an improvement in training and validation F1 over time, with some declines in performance at certain epochs. The line graphs are slightly unstable. Fluctuations in the F1 score reflect sensitivity to class imbalance within the dataset, while the overall trends indicate improving classification performance.

Figure 11. indicates certain epochs where the model performs well. However, there is instability, as seen before with accuracy, which further suggests sensitivity to certain keywords within the dataset. Furthermore, the model was able to achieve an F1 score of 0.6431 with a precision score of 0.6480 and a recall of 0.6383. This suggests the model favours precision over recall to gain high-performance classification of fake news. The most prevalent issue may be the class imbalance and the differentiation ability between certain labels.

Table 2: This table presents the metric values attained by the sparse attention model in validation of the LIAR dataset. These results show that the sparse attention model achieves balanced precision and recall on the LIAR dataset.

Metric	Score
Accuracy	0.6543
F1	0.6431
Recall	0.6383
Precision	0.6480

Table 3: This table presents the comparison of this paper's transformer model to a Triple-BERT and Base transformer model. The models in comparison were trained by the same LIAR. The sparse attention model outperforms baseline transformer models in accuracy, demonstrating the effectiveness of the sparse attention mechanism.

Model	Accuracy	F1
Sparse Attention Model	0.6543	0.6431
Base transformer model	0.4055	0.4292
Triple-BERT	0.371	0.768

Discussion:

The null hypothesis initially proposed is disproved with changes in marginal accuracy and F1 scores. In comparison to a base transformer model of Qazi, the accuracy and F1 score of this research paper's model outperform this standard attention-based transformer model. The model presented by Qazi *et al.* is a base transformer model with an attention mechanism and an encoder-decoder architecture.[15] The standard transformer model is trained on the LIAR dataset, and when tested, scores an accuracy value of 0.4055 and a F1 score of 0.4292. As presented in Table 2, the accuracy score of 0.6543 and F1 score of 0.6431 are obtained with the BigBird architecture. The sparse attention model's ability to analyze certain tokens with the mechanism of local, global, and random attention helps the model to work at a faster rate with greater accuracy. Moreover, the analysis of the Triple BERT model for fake news classification has also been significantly outperformed,[17] with the transformer model achieving the accuracy score of 0.371. The Triple-Bert model is trained and tested on the LIAR dataset and utilizes a transformer-based model with minor adjustments to the architecture. However, the Triple-Bert model does obtain a higher F1 score but has achieved low accuracy results. The model proposed within this research significantly outperforms the model in accuracy but lacks in F1 compared to a BERT model.

The reasons that the sparse attention model was able to outperform the standard model stem from the attention mechanism. Sparse attention limits each token to only attend to the most relevant positions. This structure combines well with the manner in which fake news signals tend to appear: key cues often in short, high-density spans or unverifiable claims. When the model is prevented from spreading attention across the entire sequence, spare attention is more efficient in the high-signal regions. However, full attention diffuses weight across many, if not all, tokens. This results in standard transformer models being susceptible to overfitting or lexical frequency patterns instead of focusing on deceptive devices in text. The sparse attention model's higher F1 score suggests that it has a better ability to balance precision and recall, identifying unknown clues without having high rates of false positives. The Triple-BERT model utilizes multiple BERT encoders to theoretically deepen semantic representation. However, in the absence of an appropriate mechanism for managing the redundancy between each encoder, the model seems to overlap features. This increases noise, most observably on this dataset, where cues are subtle.

These issues are then further amplified by the baseline full attention mechanism that is already prevalent in the model. A conclusion is that the sparse attention mechanism interprets subtle linguistic cues and structure of misinformation at a higher quality than a standard attention mechanism model.

However, the proposed model can be improved. Despite the high peak scores of 0.65, the sparse attention model can be uncertain and lower at certain validation statements. The model may be unable to classify some vague statements, decreasing its accuracy. The model does tend to marginally overfit when tested on the LIAR dataset. The process of overfitting usually occurs when a model comprehends certain text patterns that are hyperspecific to the data it was trained on. As a result, we can see that the model may perform at a lower standard for new statements. In addition to this, whilst the model does steadily increase in accuracy measures, there are some periods of decline. This conveys that the model may be somewhat insatiable. The use of local, window, and global tokens may result in the model weighting certain tokens differently from others. This could lead to more inconsistency within the model's performance. Another possibility could be that the model is highly sensitive to the hyperparameters set in place, such as the rate of learning, warm-up steps, and the sparsity ratios. The shift in attention patterns can be caused by the smallest changes in these values, leading to inconsistent validation performance. In addition to this, the size of the LIAR dataset could be a factor in how there are periods of decline due to its smaller size of data capacity. This is due to the classes having fewer examples for the model to train on. This makes the model receive uneven gradient signals, creating sudden changes in loss and momentary declines in accuracy. Furthermore, the recall and precision scores reflect that the model may be risk-averse and prioritize being correct, which could lead to many false positives and the slight declines in F1, which Figure 11. presents. Despite this, the model still achieves an average score of 0.6 for accuracy. This could be due to the model being highly sensitive to certain features or keywords, and when they are present, the model could predict with higher confidence.[18] To further improve the model, it can be trained with a larger threshold of epochs to guarantee the model learns, as opposed to memorizing certain statements.

■ Limitations and Further Work

The main limitations of this study are: the time available to train the model as well as test it, the dataset's capacity, and the limited amount of additional linguistic features. This study tests the model's accuracy and F1 over 9 epochs. This may result in less time for the model to train on data, as long text sequence models require more training cycles, resulting in there being restricted conclusions. Enhancing the linguistic complexity of the model could significantly improve the ability of the model to analyze subtle sentiment cues. Furthermore, the dataset remains slightly narrow. This means that the model may struggle with unfamiliar text or ambiguous phrasing found in more real-life cases of fake news. Future work may be focused on extending the time allocated for more BigBird model variations to be trained, allowing a greater understanding of the strengths of the architecture. Additional time may also be used to integrate more linguistic features for the model and improve its accuracy. This can be further improved with features such as the linguistic complexity of a sentence added to the architecture, enabling a better sentiment analysis of a statement. Furthermore, data could actively be generated for the model to train off in order to help the model learn to handle ambiguity and provide a larger sample of examples to train from, as well as increasing the diversity of the training set. Overall, this is an advancing field, and the model presented in this research paper highlights the importance of using sparse attention for long sequence text analysis, such as fake news classification, rather than standard attention.

■ Conclusion

This research paper presents a transformer model with a BigBird-style sparse attention mechanism and an encoder-only architecture. This research aimed to replace the standard multi-head attention in a transformer model to enhance its performance in classifying fake and biased information. The model focused on reducing the computational thinking and expenses that arose from a normal transformer classifying fake news due to sequential token analysis. The BigBird model was utilized to analyze only certain tokens with the local, global, and random attention mechanisms. The model was trained and tested on the LIAR dataset due to its balance across all its labeled data. The model achieved accuracy, F1, recall, and precision scores of 0.6543, 0.6431, 0.6383, and 0.6480 outperforming standard transformer models utilizing standard multi-head attention mechanisms. However, the model can be improved with a more extensive training in order to reduce the unstable values during validation. The model may be improved with more adjustments and tuning to global tokens to increase pattern analysis, as well as using more varied data with less noise, so the model doesn't become overfitted, a limiting

DOI: 10.36838/IJHSR89.39

factor that has affected this study. Therefore, future work will be required for further development of this transformer model for classifying fake news. It is important to recognize that this study is limited to the analysis of a standard transformer model with a multi-head attention mechanism and does not compare the presented model with alternative hybrid models.

Acknowledgement

I am immensely thankful to my mentor, Dr. Eric Sakk, for his guidance and support. I would like to extend my gratitude to my parents and family for their unwavering support.

References

1. Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; Zittrain, J. L. The Science of Fake News. *Science* **2018**, *359* (6380), 1094–1096. https://doi.org/10.1126/science.aao2998.
2. Nazar, S.; Bustam, M. R. Artificial Intelligence and New Level of Fake News. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *879* (1), 012006. https://doi.org/10.1088/1757-899x/879/1/012006.
3. Vayadande, K.; Bohri, M.; Chawala, M.; Kulkarni, A. M.; Mursal, A. The Rise of AI-Generated News Videos. In *How Machine Learning is Innovating Today's World*; John Wiley & Sons, Ltd, 2024; pp 423–451. https://doi.org/10.1002/9781394214167.ch25.
4. Zarocostas, J. How to Fight an Infodemic. *The Lancet* **2020**, *395* (10225), 676. https://doi.org/10.1016/s0140-6736(20)30461-x.
5. Maktabdar Oghaz, M.; Babu Saheer, L.; Dhame, K.; Singaram, G. Detection and Classification of ChatGPT-Generated Content Using Deep Transformer Models. *Front. Artif. Intell.* **2025**, *8*. https://doi.org/10.3389/frai.2025.1458707.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. arXiv August 2, 2023. https://doi.org/10.48550/arXiv.1706.03762.
7. Zaheer, M.; Guruganesh, G.; Dubey, A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; Ahmed, A. BigBird: Transformers for Longer Sequences. arXiv January 8, 2021. https://doi.org/10.48550/arXiv.2007.14062.
8. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805.
9. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv July 26, 2019. https://doi.org/10.48550/arXiv.1907.11692.
10. Lu, Y.; Yao, N. A Fake News Detection Model Using the Integration of Multimodal Attention Mechanism and Residual Convolutional Network. *Sci. Rep.* **2025**, *15* (1). https://doi.org/10.1038/s41598-025-05702-w.
11. Rout, J.; Mishra, M.; Saikia, M. J. Towards Reliable Fake News Detection: Enhanced Attention-Based Transformer Model. *J. Cybersecurity Priv.* **2025**, *5* (3), 43. https://doi.org/10.3390/jcp5030043.
12. Liu, B. Research on Natural Language Misleading Content Detection Method Based on Attention Mechanism. *IEEE Access* **2025**, *13*, 132410–132425. https://doi.org/10.1109/ACCESS.2025.3591455.
13. Sharma, P.; Malik, N.; Akhtar, N. FEEDFORWARD NEURAL NETWORK: A Review.
14. Wang, W. Y. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. arXiv May 1, 2017. https://doi.org/10.48550/arXiv.1705.00648.
15. Qazi, M.; Khan, M. U. S.; Ali, M. Detection of Fake News Using Transformer Model. In *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*,
16. Wang, Y.; Han, H.; Ding, Y.; Wang, X.; Liao, Q. Learning Contextual Features with Multi-Head Self-Attention for Fake News Detection. In *Cognitive Computing – ICCC 2019*; Xu, R., Wang, J., Zhang, L.-J., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2019; Vol. 11518, pp 132–142. https://doi.org/10.1007/978-3-030-23407-2_11.
17. Mehta, D.; Dwivedi, A.; Patra, A.; Anand Kumar, M. A Transformer-Based Architecture for Fake News Classification. *Soc. Netw. Anal. Min.* **2021**, *11* (1), 39. https://doi.org/10.1007/s13278-021-00738-y.
18. Gupta, S.; Gupta, A. Dealing with Noise Problem in Machine Learning Data-Sets: A Systematic Review. *Procedia Comput. Sci.* **2019**, *161*, 466–474. https://doi.org/10.1016/j.procs.2019.11.146.

Author

Adhavan Senguttuvan Viswapurna is a GCSE student at British School Muscat. He aspires to pursue a career as a computer science engineer, specifically in cybersecurity and machine learning. His strong passion for algorithmic thinking and AI machine learning facilitated his progress in programming.