Using BERT to Generate Contextualised Textual Images for Sentiment Analysis

Harpreet Singh[0000−0001−6039−6834], Second Author[0000−0001−6687−0306], Roderick Adams, and Yi Sun

University of Hertfordshire, Department of computer science, UK. {h.singh23, n.helian, r.g.adams, y.2.sun}@herts.ac.uk

Abstract. Sentiment Analysis could be performed on textual data and indicates the general 'tone' or emotional state of the writing. It is important in business, for instance in marketing, to determine customer opinions and trends, and in analysing social media to help weed out inappropriate or discriminatory language. Recently improved performance has been obtained by first converting the text to a grayscale image and then using a BLSTM and deep CNN, specifically ResNet, to classify the data. This paper investigates the addition of more context to the original text using a pre-trained BERT model to produce contextualised textual images. This produces a marked improvement over the previous results. The proposed BERT-BLSTM-ResNet model outperforms the BERT model on smaller datasets and above a threshold data size, the BERT performance is comparable.

Keywords: Contextualised Textual Images · BERT in NLP · sentiment Analysis · Deep 2D CNN on Text

1 Introduction

Sentiment analysis is a popular text classification task where a text document can be classified into positive, negative, and neutral classes, which has attracted several applications. For example, the sentiment of customers towards a product is crucial to improving services and user satisfaction. Many sentiment analysis models have surfaced which incorporate supervised learning to predict the sentiment of the text. Instead of using high sparsity vectors, which neglect the order of the words in a document, a more compact and efficient vector called Word Embeddings was adopted, which restores the word order in the sentence [11]. Deep learning emphasises a layered structure with different modules in different layers. Feed-forward, Recurrent, and Convolution-based layers are among the most popular for extracting features from the input vectors. The two-dimensional convolutional layer has proved beneficial in computer vision for Image classifying tasks, so, based on this success, a tensor equivalent to images was made from textual data, and the resultant images are called textual images [17]. The textual images are formed by transforming the word embedding matrix into a greyscale image with the help of the recurrent neural network (RNN) block. This representation of the text helped to improve information extraction using deep 2D Convolutional Neural Network (CNN) [13, 6].

The attention layer addition to textual data processing in combination with self-supervised training has proven beneficial in boosting neural network performance [2] and in natural language processing (NLP) [15]. The Bidirectional Encoder Representation from Transformer (BERT) [3] is one such pre-trained model which can be fine-tuned for any text-based task. The fine-tuned models can be utilised to generate contextualised textual images and it has been shown to improve performance. Very recently, pre-trained Large language models like ChatGPT have shown how effectively these models can be deployed in real-world applications [14].

Textual image generation concerns converting text to represent them as images and for that we will be using a pre-trained BERT model together with an existing model [13]. This new approach using BERT-BLSTM-ResNet for creating Images is explored in this paper since it incorporates more context awareness into the textual images and reaps the benefit of the pre-trained BERT model. Experiments were also conducted to analyse the impact of the dataset size of this textual image classification.

The research aim addressed in this paper is: **To investigate the impact of adding contextualised embeddings to textual image Classification.**

The paper is organised into five sections. Section 2 covers related work. Section 3 outlines our proposed model and the detailed implementation of the BERT-BLSTM-ResNet model. Section 4 describes the experiment and results in detail and finally, Section 5 concludes the paper.

2 Related Work

In the last five years, transformer-based neural networks have outperformed RNN and CNN in various text classification tasks. The transformer model was first introduced in [15], which is solely based on attention mechanisms and has encoder-decoder blocks. Transformers were introduced to deal with sequence-to-sequence modelling tasks, but they can be used in many text-based NLP classification problems by transforming the final layer. The BERT [3] model uses only the encoder part of the Transformer. BERT authors tried to overcome the shortcomings of other Transformer models, which learned only unidirectional representation, to have bidirectional self-attention. It incorporates two different self-supervised pre-training tasks which are Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) (refer [3] for details). The pre-trained model can then be fine-tuned to perform various NLP downstream tasks competitively.

BERT has two popular model configurations BERT$_{Base}$ (L=12, H=768, and A=12 with 110 million parameters) and BERT$_{large}$ (L=24, H=1024 and A=16 with 340 million parameters) where L is the number of transformer blocks, H is hidden units, A is the number of self-attention units. The former is less computationally demanding than the latter. At the time of publishing, BERT$_{large}$

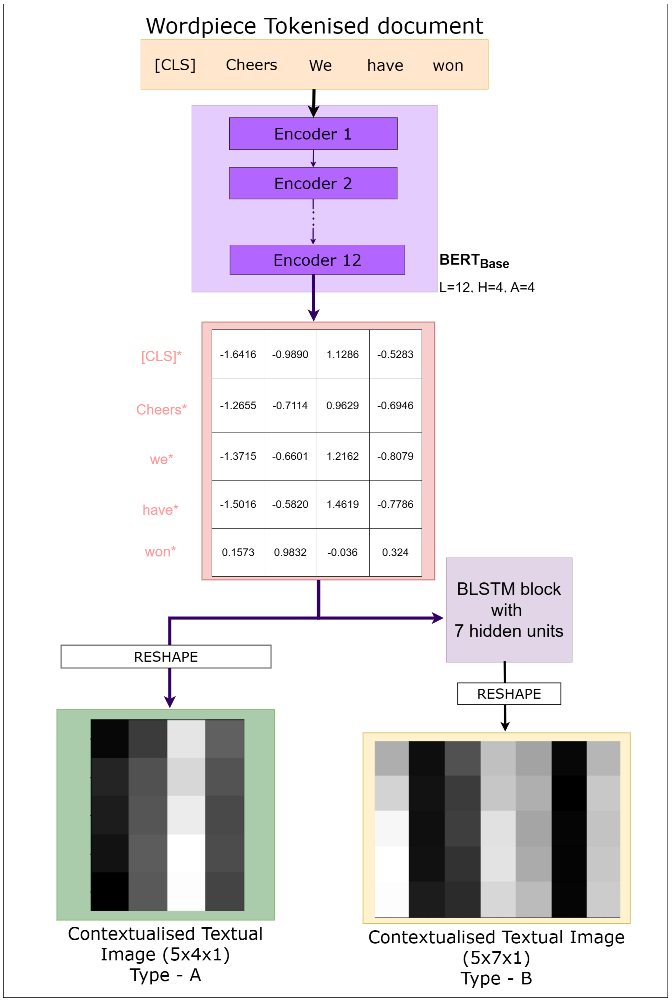
outperformed eleven NLP tasks including a sentiment analysis dataset. Many models have since been proposed that use the BERT architecture [10, 16, 12, 8] with different pre-training strategies, optimisation approaches, parameter tuning, model distillation techniques, and a lighter variant, respectively.

For textual image classification, CNN is the foundation of modern computer vision systems. Image data is three-dimensional and can be analysed using 2D convolution windows, which slide over the image to perform the convolution on the underlying pixel values. The modern architecture of a multi-layer CNN is first used in LeNet [9] for handwritten character recognition, which made auto sorting of mail possible based on the handwritten address on it. 2012 was the landmark period that made deep CNN the industry favourite in solving image classification tasks with the advent of AlexNet [7], which performed best in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-2012 competition. After AlexNet, networks with even deeper layers started performing better in the ImageNet competition every year, for example, ZF in 2013, VGGNet in 2014, GoogleNet in 2014, and ResNet in 2015, Inception-NetV4 in 2016 etc [1]. ResNet is short for Residual Network, and it is a deep 2D CNN architecture that solved the inherent problems of going deeper in a CNN, i.e., the degradation problem (refer [5] for details). ResNet uses a skip connection, which makes the back propagation of the error more effective during training. This makes it possible to train a very deep neural network without increasing the error propagation, which earlier restricted the maximum depth of the network. The brilliant feature of ResNet is Scaling up capabilities limited by the Image Resolution or Hardware available to run the network. Based on the resource availability, different depths of the ResNet can be achieved, for example, from ResNet-18 to ResNet-1202 or even deeper, where 18 and 1202 represent the count of convolution layers in the network.

3 Proposed Model

Contextualized textual Images are greyscale images like the textual images explored in [13]. Contextualised embeddings are generated in this paper by using the BERT$_{Base}$ neural model. The BERT$_{Base}$ model contains 12 encoder blocks of the Transformer model and can process a maximum of 512-word tokens. This model is used as a feature extractor where the tokenised text document is input to the model to produce the corresponding word embeddings. The Transformer encoder model uses the attention mechanism to generate the context in the embeddings, and usually multiple attention layers are stacked, and in this case, 12 stacked layers are used. The output from the last encoder block is used as the final contextualised embedding for each word token.

Once the contextualised embeddings are generated using BERT, they can either be reshaped in the form of an image (Type A) or can be transformed into images using a BLSTM block (Type B), see Fig. 1. Images obtained from both these methods are called Contextualised textual Images. The image dimension, in a Type A image, is (document length)$\times$(BERT hidden units)$\times$1. In compar-


Wordpiece Tokenised document
[CLS]
Cheers
We
have
won
Encoder 1
Encoder 2
Encoder 12
BERT$_{Base}$
L=12. H=4. A=4
BLSTM block with 7 hidden units
RESHAPE
RESHAPE
Contextualised Textual Image (5x4x1) Type - A
Contextualised Textual Image (5x7x1) Type - B

[CLS]*	-1.6416	-0.9890	1.1286	-0.5283
Cheers*	-1.2655	-0.7114	0.9629	-0.6946
we*	-1.3715	-0.6601	1.2162	-0.8079
have*	-1.5016	-0.5820	1.4619	-0.7786
won*	0.1573	0.9832	-0.036	0.324

Fig. 1. A tokenised text document is shown with five tokens, where [CLS] is a special token signifying classification task. The contextualised embeddings are generated using BERT$_{Base}$ as a feature extractor and the output of the last encoder layer is used. Two different methods for creating contextualised textual images are depicted. BERT$_{Base}$ output can directly be reshaped into a grayscale image, or it can be given to the BLSTM block and then reshaped into an image.

ison, Type B has dimensions of (document length)×(BLSTM hidden units)×1. The formed contextualised textual images will be classified using a standard 2D ResNet model.

Based on the contextualised textual image used as input, various models have been created.

1. BERT$_{Base}$-BLSTM-ResNet18 (BBR18)

2. BERT_{Base}-BLSTM-ResNet34 (BBR34)
3. BERT_{Base}-BLSTM-ResNet50 (BBR50)

The suffix BERT_{Base} indicates that the contextualised embeddings are used, and the tag BLSTM in models indicates using the BLSTM block for image generation. The ResNet followed by a number represents a ResNet model with the number of convolution layers in the ResNet block (more details in [5, 13]).

4 Experiments and Results

4.1 Dataset

The dataset used to test out the above architectures for text classification is Sentiment140 [4], which consists of 1.6 million tweets. The dataset has two classes, one for positive and one for negative tweets. It is a balanced dataset with each class containing 0.8 million data points.

4.2 Training Setup

BERT_{Base} will be only referred to as BERT further in the section 4. A pre-trained BERT model is fine-tuned on our training set, and the output of the last transformer block is used to generate the actual contextualised embeddings. This model is then used as the feature extraction step in our BERT-BLSTM-ResNet models. A Training/Test set split ratio of 9:1 is used for every subset of Sentiment140, which are stratified samples of the entire dataset with random seed 42. For example, for the 20,000 subset, 18,000 samples will be used for training and the rest for the testing set. The training set is used to fine-tune the weights of the pre-trained BERT and the proposed models.

The maximum token length of the Sentiment140 [4] after data cleaning and word piece tokenisation was just below 60; therefore 60 is used as the maximum document length and padding is applied to documents with smaller lengths. The hidden units chosen for the BLSTM model are 300 based on [13, 17]. BLSTM-ResNet12 is used as the baseline model as it uses non-contextualised Glove embeddings and is the best-performing model in [13]. Also, for standardisation, we have used Standard ResNet model configurations proposed in [5] instead of custom ResNet blocks explored in [13].

The optimiser is Adam, the Batch size is 32 (as are commonly used), and the maximum epoch is kept to 20. The early stopping training method is adopted during training, and the best model weights are saved. The saved model weights are then used to record the performance of the model on the test set. During Experimentation, several balanced subsets of the Sentiment140 dataset are used, and Accuracy is used for comparison.

4.3 The Experiment

It has already been found that textual images improved performance [17, 13] of sentiment analysis. In this paper, we investigated if there was any further improvement possible when adding contextualise embeddings before the formation of the images. The contextualised embeddings were supplied by BERT [3], an already trained and exceptionally well-performing model. So to investigate if an attention mechanism would also improve the textual image classification task, the finetuned BERT, is used to generate contextualised word embeddings.

Table 1. Accuracy in the percentage of various models on Sentiment140 subsets

Dataset size Model	20,000	200,000	400,000	800,000	1,600,000
BLSTM-ResNet 12 (Baseline)	72.85	78.82	79.89	80.68	81.74
BBR18	**83.40**	**83.50**	84.61	85.54	84.50
BBR34	82.20	83.10	**84.65**	85.68	84.66
BBR50	82.55	83.00	84.58	**85.72**	**84.73**
BERT	81.00	83.24	**85.60**	**86.10**	**85.07**

Subsets of size 20,000, 200,000, 400,000, 800,000, and 1,600,000 of the Sentiment140 dataset are selected. The models are trained on these subsets, and the corresponding accuracy is shown in the first four rows of Table 1. This table and Fig. 2 show that adding BERT to the front significantly improved the accuracy compared to the Baseline (best previous) model. This demonstrates the significance of using context-aware embeddings in textual image classification.

Since 20,000 is a small dataset size, the ResNet18 variant performed better than other BERT-BLSTM-ResNet models. As the size of the dataset gets larger, the additional depth of the models can be made use of so that first, the ResNet34 variant is best, and then the ResNet50 variant takes over. These findings are in line with the findings of the [13].

Since BERT produced a significant improvement, we then wondered what would happen if we just used BERT and did not use the rest of the model, which, curiously, most other authors do not seem to do. Interestingly, as the blue results in the last row of Table 1 and Fig. 3 show, BERT on its own either outperforms all the other models or gives close results. Significantly, it seems that the dataset size is the deciding factor. BERT alone is better for datasets of a size of about 200,000 and larger but is less good for smaller datasets.

To further investigate this result and justify our claim, we performed a more computationally intensive task of using 5-fold cross-validation on the results of several smaller subsets. This is shown in Fig. 4 and Table 2. This gives a value for the margin of error and shows a point around 60,000 where the two models are very similar and is a possible crossover point. Sentiment140 smaller subset sizes 1250, 2500, 5,000, 10,000, 20,000, 40,000, 80,000 and 160,000 were used. This is an important result as it establishes that BERT requires more data to optimise

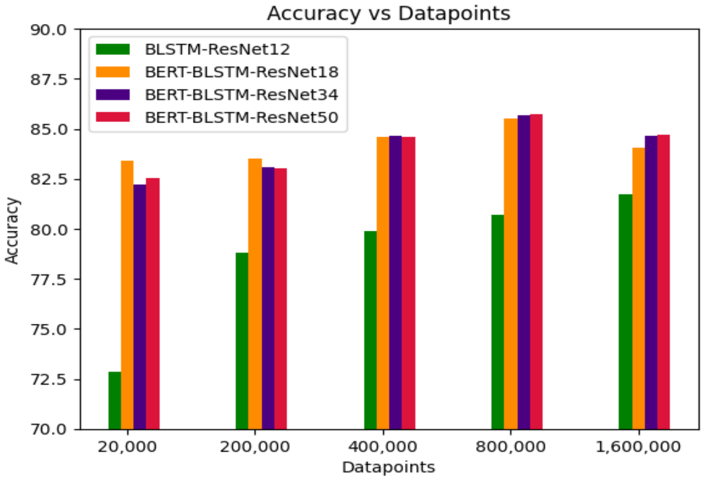


Fig. 2. Accuracy vs Dataset size for subset 20,000 to 1,600,000. Comparison between BERT-BLSTM-ResNet models with the Baseline model BLSTM-ResNet-12

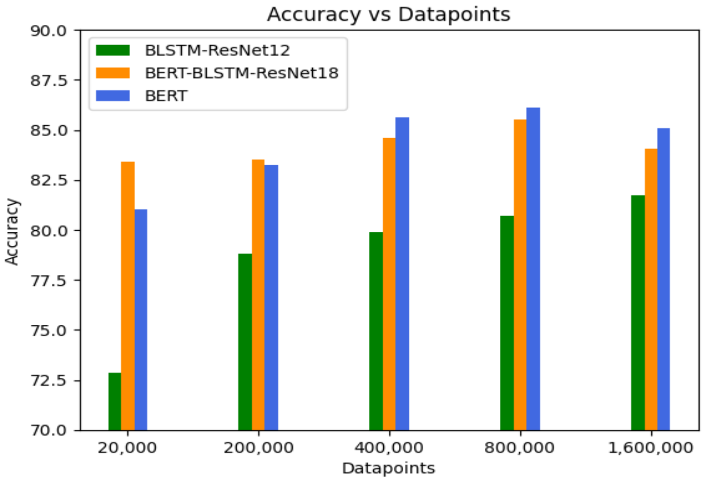


Fig. 3. Accuracy vs Dataset size for subset 20,000 to 1,600,000. Comparison between BERT-BLSTM-ResNet models and Baseline BLSTM-ResNet-12 with BERT

Sentiment140 data fully and needs a certain minimum dataset size to release its full potential. Cross-validation is performed only on BERT and BBR18 because of the lack of available computation resources.

To visualise all the model performances from a smaller to a bigger dataset size, a single BERT is trained on various subsets. Then, the model weights were used to train all BERT-BLSTM-ResNet variants, see Table 3. The values are

Table 2. BERT and BERT-BLSTM-ResNet18 (BBR18) using 5-fold cross validation. Average Accuracy % (A) and with standard deviation (Std) on various Sentiment140 subsets. The best results are shown in bold.

dataset Model		1,250	2,500	5,000	10,000	20,000	40,000	80,000	160,000
BERT	A	71.98	80.56	78.05	80.22	81.03	82.21	**83.53**	**84.21**
	Std	2.02	1.46	0.42	0.58	0.53	0.22	0.17	0.09
BBR18	A	**82.80**	**87.40**	**89.70**	**85.53**	**86.37**	**83.78**	82.15	82.98
	Std	4.60	3.64	5.77	1.48	3.39	6.01	0.99	1.16

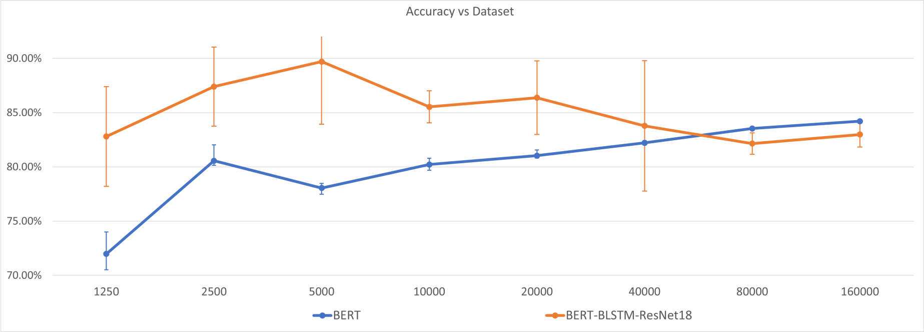


Fig. 4. 5-fold cross validation Accuracy vs Dataset size plot for the smaller subsets of Sentiment140 from 1250 - 160,000. Comparison between BERT-BLSTM-ResNet 18 model with BERT

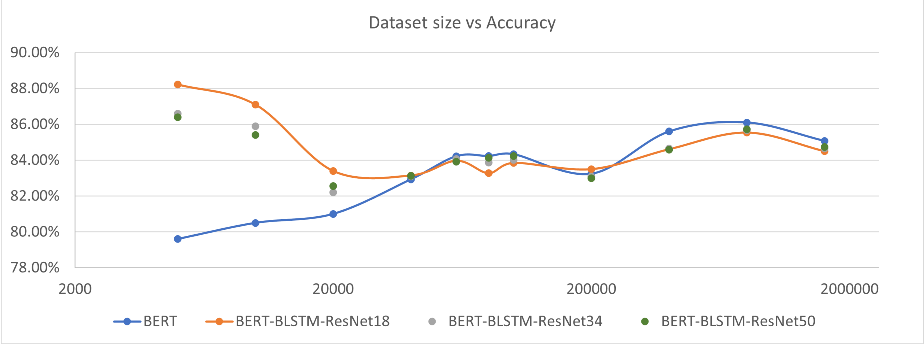


Fig. 5. Log plot of Accuracy vs Dataset size over various datasets. Comparison between BERT-BLSTM-ResNet models with BERT. For clarity, only the BERT and ResNet18 variant are shown with a line to illustrate trends.

then plotted in logarithmic scale, see Fig. 5. This plot captures the entire trend of the model performances. BERT underperforms initially and then overtakes as the dataset size increases. BBR18 performs best till 40,000 subsets and is then superseded by another model. BERT seems to be performing pretty well on its own, but for a smaller dataset, it can benefit directly from having spatial-temporal features extracted using the BLSTM-ResNet model.

Table 3. BERT and BBR Models performance for subsets from a size of 5k up to the entire dataset of Sentiment140. in this table, 'k' represents 1000.

Model	5k	10k	20k	40k	60k	80k	100k	200k	400k	800k	1600k
BERT	79.60	80.50	81.00	82.93	**84.22**	**84.23**	**84.33**	83.24	**85.60**	**86.10**	**85.07**
BBR18	**88.22**	**87.10**	**83.40**	**83.15**	83.97	83.27	83.85	**83.50**	84.61	85.54	84.50
BBR34	86.60	85.90	82.20	83.09	84.06	83.86	84.01	83.10	84.65	85.68	84.66
BBR50	86.40	85.40	82.55	83.13	83.92	84.14	84.24	83.00	84.58	85.72	84.73

5 Conclusion

This paper introduces a new attention-based neural network model for sentiment analysis to generate and classify textual images called BERT-BLSTM-ResNet. The model incorporates a fine-tuned BERT model to extract a contextualised embedding, which was then used to create Contextualised Textual images and a standard ResNet network is used to classify them. The experimental results demonstrate the BERT-BLSTM-ResNet network outperformed the previous best textual image classifier. It is also noted that the BERT model on its own performed well; however, on smaller datasets, it can benefit from the use of the full BLSTM-ResNet model. In the future, we would like to incorporate new state-of-the-art computer vision models and methods to further enrich the contextualised word embeddings.

References

1. ImageNet — image-net.org. https://image-net.org/challenges/LSVRC/, [Accessed 01-11-2023]
2. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR (sep 2015), https://arxiv.org/abs/1409.0473v7
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**, 4171–4186 (oct 2018), https://arxiv.org/abs/1810.04805v2
4. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. Ph.D. thesis, CS224N Project Report, Stanford (2009), http://tinyurl.com/cvvg9a

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2016-Decem, pp. 770–778. IEEE Computer Society (dec 2016). https://doi.org/10.1109/CVPR.2016.90, http://image-net.org/challenges/LSVRC/2015/
6. Jiang, D., He, J.: Text Semantic Classification of Long Discourses Based on Neural Networks with Improved Focal Loss. Computational Intelligence and Neuroscience **2021** (2021). https://doi.org/10.1155/2021/8845362
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (jun 2017). https://doi.org/10.1145/3065386, http://code.google.com/p/cuda-convnet/.
8. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R., Research, G.: ALBERT: A Lite BERT for Self-supervised Learning of Language Representations (sep 2019). https://doi.org/10.48550/arxiv.1909.11942, https://arxiv.org/abs/1909.11942v6
9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2323 (1998). https://doi.org/10.1109/5.726791
10. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., Allen, P.G.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (jul 2019). https://doi.org/10.48550/arxiv.1907.11692, https://arxiv.org/abs/1907.11692v1
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. International Conference on Learning Representations, ICLR (jan 2013), http://ronan.collobert.com/senna/
12. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (oct 2019). https://doi.org/10.48550/arxiv.1910.01108, https://arxiv.org/abs/1910.01108v4
13. Singh, H., Helian, N., Adams, R., Sun, Y.: Sentiment Analysis using BLSTM-ResNet on Textual Images. International Joint Conference on Neural network (IJCNN) pp. 1–8 (sep 2022). https://doi.org/10.1109/IJCNN55064.2022.9892883
14. Singh, H., Tayarani-Najaran, M.H., Yaqoob, M.: Exploring computer science students perception of chatgpt in higher education: A descriptive and correlation study. Education Sciences **13**(9) (2023). https://doi.org/10.3390/educsci13090924, https://www.mdpi.com/2227-7102/13/9/924
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. Advances in Neural Information Processing Systems **2017-Decem**, 5999–6009 (jun 2017), https://arxiv.org/abs/1706.03762v5
16. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Information Processing Systems **32** (jun 2019). https://doi.org/10.48550/arxiv.1906.08237, https://arxiv.org/abs/1906.08237v2
17. Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B.: Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers pp. 3485–3495 (nov 2016). https://doi.org/10.48550/arxiv.1611.06639, https://arxiv.org/abs/1611.06639v1