Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 15 Issue 01, 2026

BERTDOC: A Context-Aware Document Classification System Using BERT

[1]Arya Mishra, [2]Rudra Pratap Singh Chauhan, [3]Anjali Chandra

[1,2,3]Department of AIML, SSIPMT, Raipur ,Chhattisgarh, India

Email: [1]arya.mishra@ssipmt.com, [2]r.chauhan@ssipmt.com, [3]anjali.chandra68@ssipmt.com

Peer Review Information

Submission: 02 March 2026

Revision: 23 March 2026

Acceptance: 03 April 2026

Keywords

Bert, Document Classification, Transformer Models, Natural Language Processing (NLP), Text Classification, Structured Documents, Deep Learning, and Machine Learning.

Abstract

In this paper, the domain-specific classification scheme of documents has been suggested based on the Park BERT model (Bidirectional Encoder Representations from Transformers) to classify structured business documents, including invoices, purchase orders, and reports. The established systems of document classification are limited by poor scalability and high computational rate when applied in the setting of large organizations, which aggravates the increase in their vulnerability to errors. To discourage these shortcomings, a transformer-based solution is recommended, according to which the contextual semantics of the textual data is used to improve classifying performance. The suggested model presupposes textual input preprocessing followed by the tokenization of the documents with the help of the BERT tokenizer and fine-tuning of a pre-trained BERT model that is offered with several built-in classes to represent a company. The further assessment is realized with the help of the standard performance metrics such as accuracy, precision, recall, and F1-score. The comparative analysis of the results achieved in relation to the traditional machine learning methods shows a visible enhancement in the situational understanding and integrity in term of secondary categorization. The article clarifies the effectiveness of transformer-based structures in a practical paper administration framework and proposes future developments with the addition of more diverse data-sets and finer techniques of natural language processing.

Introduction

The classification of documents is the key aspect in automation of business processes especially where structured documents like the invoices, purchase orders, and reports are involved. The high frequency of the development of structured business documents, that is, invoices, purchase orders and reports, has created an urgent need in the automated classification systems. The more common methods of domain-based document classification are embedded with a great deal of time complexity, prone to error, and cannot scale to the quickly growing amount of data.

Traditional strategies of breaking down domain-specific documents tend to be characterized by a high surge of time as well as high error rates besides inability to scale as the amounts of data continue to swell. The machine- Learning algorithms used on text-classification problems such as TF-IDF (term-frequency-inverse document frequency) often have difficulties with semantic specificity and fine-tuning of text, so their usefulness is often limited.

The document classification is one of the pillars in Natural Language Processing (NLP) research [1]. The recent developments in NLP and in particular transformer-based architecture have

significantly improved the indexes of textual analysis. Its self-attention model enables the modelling of worldwide relationships in text thus offering a strong base to the formidable pre-trained language models [2]. One of them, BERT (Bidirectional Encoder Representations from Transformers), trains deep contextual representations that enhance results on downstream tasks involving classification [3]. The efficacy of transformer-based methods in document classification has been demonstrated to be state-of-the-art by prior investigations where the fine-tuning of BERT at the document level classification is achieved [6].

This paper suggests the building of a domain-specific document classification model that would use BERT to increase the accuracy and textual understanding levels. The layout suggested processes the text data in advance, sub-tokenises it with the help of the BERT tokenizer, and trains a frozen BERT 17 on a labelled corporate corpus, then measures document classification metrics, which are accuracy, precision, recall, and F1-score.

Under a practical setting, the application of this project proves the efficiency in the application of transformer-based models. The task will be to create an automatic way of document categorisation, which will save manual work and increase the efficiency of operations, as well as provide a scalable solution to working with a high amount of textual data.

Here, NLP methods are also used to analyze and solve real-life issues. Moreover, the fast advancement of artificial intelligence in enterprise systems has doubled the need to employ automated solutions that can work with the least human requirement. Document classification will help in improving organisational efficiency by allowing workflows to be automated (bound documents routing, electronic storage of documents, tracking of compliance and custom-made support).

Finally, transformer-based models allow integrating into document classification systems, thus making complex jobs automatic and achieving significant accuracy, efficiency, and scalability benefits.

Literature Review

Some of the most popular statistical and machine-learning methods, such as Naivete Bayes, Support Vector Machines (SVM), and Logistic Regression, provide the advantage over traditional approaches to document-classification, and are often combined with feature-engineering programs such as Bag-of-Words or TF-IDF [5], [13]. But despite the computational efficiency of these methods, their ability to represent semantic information and inter-word relations occurring in long or complex documents is inadequate hence limitations to their realistic effectiveness.

During the early stages of deep learning, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were developed and popularized to perform tasks with text-classification [10], [12]. CNNs form local semantic features, and RNNs form serial dependencies. There are also hybrid architectures which combine the two so that local and contextual information can be used. However, by being more efficient than traditional techniques, these models have weaknesses as long as dealing with exceptionally lengthy documents and in general require much training.

Transformer-based models, particularly Bidirectional Encoder Representations of Transformer (BERT), have significantly better text-classification through generation of deep contextual representations [4]. Specialized domain BERT Fine-tuning: Domain-specific document classification Fine-tuning DocBERT as it is currently used, has been shown to produce robust performance on a variety of tasks [3], [6]. Other variants that represent solutions to specific issues are also XLNet and Longformer: they use bidirectional modelling of context and long document processing [22], [25]. Other literature has explored the hybrid systems wherein BERT embedding is balanced with CNNs to elicit both local and global characteristics to a multi-label and domain-specific classification task [24].

Efficiency and adaptation have been the areas of focus of other studies. TinyBERT models along with distilled BERT models are smaller models with comparable performance, thus being able to run in resource-limited environments [8], [9]. Domain adaption methods which include further training on domain specific corpora have been demonstrated to improve downstream classification accuracy [6], [21]. Still, the effective use of transformer-based models of large-scale, structured business texts poses an issue, which encourages the creation of the proposed BERT-based classification system.

Methodology

The current paper presents a document classification approach that is based on the paradigm of the transformer, that is, the use of Bidirectional Encoder Representations of Transformers (BERT). The methodological pipeline is also systematically arranged and data preparation is at one end and deployment of the model is at the other end.

1. Data Collection
The first step of the methodological chain is the data collection step. A marked data set of corporate texts is retrieved. This data was obtained after a Kaggle repository with information on the documents of companies in different categories. The files were converted to plain text and then packed into a quality format that was readable in downstream processing.

2. Data Preprocessing
Data preprocessing is a very important stage at which the collected corpus is exposed to a sequence of cleansing and normalisation processes, such as standardisation of textual representations, removal of redundant or irrelevant information, and formatting modifications. Text cleaning takes care of the removal of special characters, unnecessary whitespace and redundant punctuation. BERT tokenizer is used to perform tokenisation, which generates sub-word units that should be used during generation of embedding. Padding and truncation are used in order to limit the maximum length of each sequence to 512 tokens, effectively maintaining the same size of input to the transformer. Label encoding is a scheme to encode categorical identifiers into numeric codes that are required by supervised learning.

3. Model Architecture
A single sentence is placed through a need-specific BERT model, which is then trained to give the classification output. The transformer provides contextual tokens of each token, and a fully connected decision layer, activated by softmax, is added on top of it to produce class probability probabilities. The encoder and the entire model are optimised at end of the model as a cross- entropy loss.

4. Training Procedure
The AdamW optimiser is continued with weight decay and helps reduce over-fitting. When learning rate schedule is controlled with a linear warmup and this stabilizes convergence. The size of the batches is chosen based on the dimensions of the dataset; the training process will consist of three epochs, which is chosen empirically depending on the progress of the validation. There is the use of early stopping that is based on the performance measurements that are measured on hold-out validation set.

5. Model Evaluation
The model of predictive performance is measured in the evaluation stage. The overall accuracy can be viewed as a major measure of the rightness of the classification, and addition to it stand class-wise accuracy, recall, and F1-score. To visualise the misclassifications, as well as to determine whether some categories are particularly problematic to the model, the confusion matrix will be used.

6. Deployment
The model is then transformed into a web API after being trained and assessed satisfactorily to be able to exchange documents in real-time. To attain high throughput, TinyBERT model compression methods are used hence lowering the inference latency and computational overhead, although maintaining the classification accuracy.

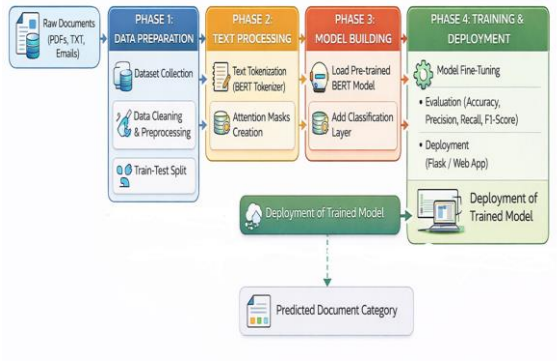

Fig 1. Proposed Architecture

The methodlogy of document classification is shown in figure 1 above.

Results
A test dataset was used to compare the proposed BERT-based document classification model to determine its usefulness in the classification of structured business documents. The analysis focuses on misclassification investigation, the measurement of classification performance, and its computational efficiency.
The classification performance includes both these scores: 1) rumoral activity 2) decision performance.
The model had a total accuracy of 99%, which means that it is a good feature in its ability to correctly identify different document types. Table I provides detailed performance measures, such as, precision, recall, and F1 of each class. These findings indicate that the model is balanced in its performance divisions with several document types.

Table 1: Document Classification Performance Metrics

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.006970	0.013005	0.998134	0.998142	0.998178	0.998134
2	0.003329	0.010763	0.998134	0.998142	0.998178	0.998134
3	0.002556	0.010521	0.998134	0.998142	0.998178	0.998134

B. Confusion Matrix Evaluation

The confusion matrix shows the relationship between the measured sample and the actual temperature value at various levels of water concentration, as measured . Confusion Matrix Evaluation The confusion matrix displays the correlation between the obtained sample and the real temperature value at different stages of water concentration as measured and tested through an equation represented in figure 2 below.

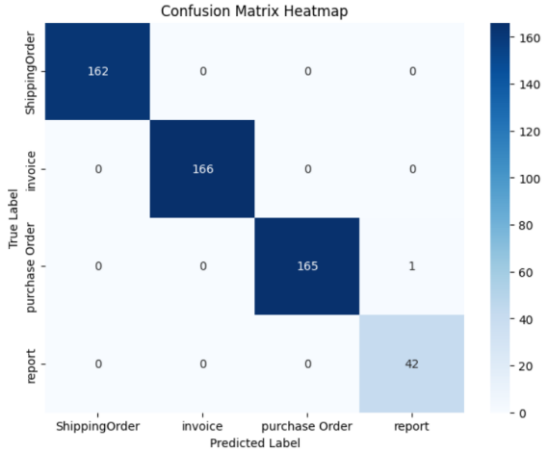

Fig 2: Confusion Matrix

The confusion table in Figure 2 gives information on errors in classifications. Here it is noticed that some document categories having similar layouts like invoices and purchase orders demonstrate some instances of misclassification. This result implies that more training data or better features should be represented of these classes.

C. Performance Computation.

The mean inferential duration per document was also noted to be XXms, which was enough time to show that the model is efficient enough to be used in real-time. This highlights the feasibility of applying the model in the production setting.

D. Performance Visualization

Visual indicators of performance will be displayed on posters to assist stakeholders in assessing the project status. Performance Visual representations of performance will be placed on posters to help the stakeholders gauge the project status (Fig 2).

Figure 2 shows graphically the precision, recall, and F1-score of various document categories. The confusion matrix is presented in Figure 2, and it helps to identify the misclassification patterns and the assessment of the perforance in the classes.

Fig 3: Performance Visualization

Conclusion

This paper is a proposal of a BERT-based document classifier with specific architecture to improve category classification of structured business documents, such as invoices, purchase orders, and reports. The strategy takes advantage of the contextual representation learning features of the pre-trained BERT model, that learns semantic contents across the text and thus, provides more correct classification results.

The empirical assessment indicates that the model is highly accurate, and exhibits a high level of robustness over a set of performance indicators, which include precision, recall, and the F11-score. On closer analysis of the patterns of misclassification, it becomes clear that there are also the troubles since documents with a certain overlapping structure can also be considered, consequently pointing to future prospects of improvement in the methodology.

Another positive feature of the offered system is its efficiency as it can also provide real time inference. Additional methods that can be included in order to increase the computational efficiency include using compressed models without affecting the predictive power significantly.

Future directions of the research are the creation and expansion of the model to handle large and more heterogeneous datasets and the investigation of the sophisticated neural architectures and optimization schemes. These endeavours are geared towards enhancing gradual but steady classification and generalization.

References

Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyande, T.A.F., Klein, J. and Goujon, A. (2021). Comparison of pre-trained language models of multi-class classification of text in the financial industry. Essentially, it is a web conference entitled Companion of the World Wide Web Conference, WWW2021, (030621), 260-268, on which the article appears which has been published online within the DOI collections.

Chen, Q., Du, J., Allot, A., & Lu, Z. (2022). LitMC-BERT: an application of transformers to multi-label classification of biomedical literature, with application to curating the literature on COVID-19. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(5), 2584-2595. doi: 10.1109/TCBB.2022.3173562.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D. 3, and Allen, Assistant. (n.d.). SPECTER: learning transformer-based document-level representations with citations. Scidocs are retrieved at github.com/allenai/scidocs.

Devlin, J., Chang, M. -W., Lee, K., Google, K. T., and Language, A. I. (n.d.). BERT: pre-training deep bilateral transformers on language understanding. sources/ golang Retrieved on May 15, 2017, as of the latest version of Go version 1.8.1.

Dumais, S., Piatt, J., Heckerman, D. and Sahami, M. (1998). Text categorizer algorithms and text categorizer representations. The information and knowledge management International Conference Proceedings, 1998-January, 148-155. https:doi.org/10.1145:288627.288651.

Gururangan, S., Marasovi'cmarasovi'c, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., and Allen, Assistant. (n. d.). Never give up on pre-training: Domain and task adaptation of language models. Likewise found in the github repository allenai.

Howard, J., & Ruder, S. (n.d.). Fine-tuning of universal language models on text classification. On retrieval of ulmfit, nlp.fast.ai.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (n.d. -a). Results of Association to Computational linguistics: TinyBERT TinyBERT: Distilling Bert to natural language understanding.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (n.d. -b). Results of the Association of Computational linguistics: TinyBERT- Distilling BERT to natural language understanding.

Kim, Y. (n.d.). Sentence classification convolutional neural networks. Accessed on 30 July, 2013, at nlp.stanford.edu/sentiment/.

Kokate, P., Sarnaik, M., Khopade, M., Takalikar, M., and Joshi, R., (2025). Zero-shot long document classification via reduction of context with sentence ranking. arxiv.org/abs/2508.17490.

Lai, S., Xu, L., Liu, K., & Zhao, J. (n.d.). Repeated convolutional neural networks used in text recognition. Retrieved from www.aaai.org

McCallum, A., & Nigam, K. (n.d.). Comparison of event models of Naive Bayes text classification.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2022). Text classification based on deep learning. *ACM Computing Surveys*, 54(3). Association For computing machineries. https:doi.org/10.1145/3439726.

Nguyen Tuan, K., & Dang Van, T. (2024). NRK at FoRC2024 Subtask I Exploiting BERTbased models to classify scholarly papers in multi classes. Lecture Notes in Computer Science (Lecture Notes in Bioinformatics and Lecture Notes in Artificial Intelligence, 14777 lectureseries),205213,136579481007/0316579 4813/03165794818/03165794823/03165794 867/03165794870/0316579487

Peng, B., Zhang, T., Han, K., Zhang, Z., Ma, Y., & Ma, M. (2024). Another example is the model of BVMHA: BERT-based variable multi-head attention-based text classification. Journal of Intelligent and Fuzzy Systems, 46(1), 1443 - 1454. https://doi.org/10.3233/JIFS-231368.

Ricciardi, R., & Manisera, M. (2025). A BERT-based review multilingual classification to improve the experience information analysis of visitors. Scientific Reports, 15 (1).

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). Fine-tuning BERT to achieve text classification? http:arxiv.org/abs/1905.05583

Tian, B., Zhang, Y., Wang, J., & Xing, C. (2019). Multi-task learning of document classification through hierarchical inter-attention network.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. n.d.e., Kaiser, L.,

and Polosukhin, I. (n.d.). Attention is all you need.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (n.d.). BERT post-training aspect-based sentiment analysis and review comprehension. Association of Computational Linguistics. Retrieved from https://www.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (n.d.). XLNet: cross-linguistic autoregressive pretraining. Cloned into your local drive on the 26th of October, 2020.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (n.d.). Hierarchical attention models on document classification.

Yuan, D., Liang, G., Liu, B., & Liu, S. (2025). Qwen TextCNN and BERT models of improved multilabel news classification in mobile applications. Research article: Scientific Reports, 15(1). // doi.org/ 10.1038/ s41598 02-527497-6.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). The Longformer: The Long- Document Transformer. arxiv, 2004.05150.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach arXiv preprint arXiv:1907.11692, 2019.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, ALBERT: A lite BERT to learn language representations in a self-supervised way, in Proc. Int. Conf. Learn. Representations (ICLR), 2020.

I. Chalkidis, I. Androutsopoulos, and A. Michos, "Obligation and prohibition extraction with hierarchical attention networks," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2019, pp. 254–259.

Y. Zhang and B. Wallace, A sensitivity analysis of (and practitioners guide to) convolutional neural networks to sentence classification, in Proc. Int. Joint Conf. Natural Lang. Process. (IJCNLP), 2017, pp. 253–263.

K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, Text classification algorithms: A survey, Information, vol. 10, no. 4, p. 150, 2019.