ISBN: 978-981-94-4198-3
2025 the 15th International Workshop on Computer Science and Engineering (WCSE 2025)
doi: 10.18178/wcse.2025.06.011

Integrating BERT Model for Document Classification in Document Management System

Janelle Kyra Sagum [+] and Jallane Roncales

Polytechnic University of the Philippines, Philippines

Abstract. The Document Management Systems (DMS) has become important in various organizations today. This study analyzes the application of Bidirectional Encoder Representations from Transformers (BERT) technology into Document Management Systems (DMS) for more effective and precise classification of memorandum circulars. By leveraging the power of BERT's understanding of natural language to accomplish this task. This system comprises a pre-training model, fine-tuning, and seamless integration into a DMS which simplifies document classification. The used model enables effective document management by classifying huge numbers of documents with the use of data augmentation and stratified kfold sampling. The evaluation results of the study showed marked improvements in accuracy and scalability as opposed to customary techniques. This study presents a methodology for implementing BERT models to automate and improve classification of memorandum circulars in the document management system. The BERT-based document classification model does well in categorizing the memorandum circulars varied in nature and does really have promising results based on their F1-scores of 0.91 indicating high result for some classification. Overall, this study focuses on the model's effectiveness in automating document classification and contributing to an improved accessibility and efficiency in the document management.

Keywords: Document Management System, BERT Model, Machine Learning, Document Classification, Natural Language Processing (NLP)

1. Introduction

The need to convert data into information is growing in the era of information technology. DMS nowadays are having a hard time classifying kind of documents. Additionally, by embracing new and innovative technologies that will enable them to fulfil their purpose as effectively as possible, a document management system may maintain a strong competitive edge and develop its core skills. [1] Therefore, new and creative technologies are essential to the success and an accurate document management system. Furthermore, it is a means of enhancing productivity, quality, cost, time, and flexibility, all of which can support the objectives and goals. Applying the Bidirectional Encoder Representations from Transforms innovates to improve daily performance in terms of effectiveness and efficiency as well as to make information easier, faster, and more accurate to disseminate [2]. In addition, it provides a means of preserving the records in electronic form to be retrieved and accessed in the future. This paper presents Integrating Bidirectional Encoder Representations from Transformers (BERT) for document classification of memorandum circulars in Document Management Systems.

The goal of this study is to create a DMS that optimizes the performance of the BERT model. Its specific goal is to determine the dependability of each automated procedure in document classification of different types of memorandum circulars and to determine the efficacy of the model used. This study provides accurate and trustworthy information about the system. It provides vital information that other researchers can benefit greatly from this study since it might be used as a guide to help them choose and develop future proposals, especially for those who plan to create a study that is similar to this one.

[+] Corresponding author.
E-mail address: jkasagum@pup.edu.ph

2. Literature Review

2.1. Document Management System

Nowadays, the majority of DMS in use simplify the processing of information that is document-oriented. Researchers from all around the world have contrasted traditional and automated DMS methods [3] [4]. The manual classification and metadata indexing used by early DMSs were ineffective when dealing with massive amounts of unstructured data. These days, automated document classification, retrieval and analysis are improved by modern systems with the help of machine learning algorithms and artificial intelligence [5].

2.2. Application of BERT and NLP

Many machine learning methods have been developed over time. Preprocessing, data extraction, data selection, and classification are the four stages of a typical text classification system [2]. Machine learning, data mining, and natural language processing (NLP) methods combine to automatically classify and spot patterns in electronic documents. Enabling people to extract information from documents through operations and summarization is the primary objective of text mining [6][7]. In 2018, Google introduced Bidirectional Encoder Representations from Transformers (BERT). Since this model has been able to achieve advanced outcomes in text classification problems, sparking widespread research interests [8].

The traditional DMS struggles to identify accurately and retrieve government-issued memorandum circulars, relying on inconsistent and inefficient human tagging. By capturing dependencies, BERT outperforms traditional methods. [9] Recent studies emphasize BERT's superiority in understanding semantics and contextual links, leading to faster and more reliable DMS performance [10].

2.3. Document Classification

Document classification is the way of organizing documents into categories or classification that based upon its context or content. [11] The document classification is a growing learning problem that is present at the many information management tasks. The document classification is an important component in building contextual and performing essential roles in different applications that deal in organizing, classifying, and concisely representing significant information.

Document classification is also a longstanding problem in information retrieval which has been thoroughly investigated. [12] The research communities rely heavily on extraction, integration, and classification of electronic documents from various sources, as well as discovery from these papers [6].

Currently, the internet serves as the primary repository for textual documents, with the volume of available textual data continually expanding, and roughly 80% of an organization's information is maintained in an unstructured textual format [13]. Since unstructured formats hold almost 90% of the world's data, information-intensive business operations demand to move beyond the simple document retrieval in DMS.

Classification of documents is essential in DMS, as it organizes and automates large volumes of documents based on their content, kind, or purpose allowing for faster access and increased searchability and simpler workflows. [14] [15].

In this paper, the researchers focus on the need for machine learning for automatically extraction of useful information from the large volume of textual data to support human interpretation is a necessity [16]. Stating that many DMS rely on manual categorization, which often leads to inefficiencies, inconsistencies, and misclassification. As the number of official papers, including the memorandum circular in the Philippines from 2001 to 2024, keeps increasing and because of their limited capacity to comprehend context and semantics, the traditional classification techniques find it difficult to classify effectively.

3. Methodologies

3.1. Dataset

The dataset is composed of memorandum circulars sourced from the Official Gazette of the Philippines, published from 2001 until 2024. Each document has a type of designation. For effective model training, the

dataset was randomly split into training (70%), validation (10%), and testing (20%). Due to unequal representation across types, stratified-kfold sampling was used to maintain proportionality, but it does not remove imbalance class.

Table 1: Document classification and total samples

Classification	Total
Administrative	282
Emergency_Special	303
Training_Development	363
Policy_Clarification	318
Compliance	259
Total	1525

Table 2: Distribution of training, validation, and testing data for each class

Classification	Training (70%)	Validation (10%)	Testing (20%)
Administrative	197	28	57
Emergency_Special	212	30	61
Training_Development	254	36	73
Policy_Clarification	223	32	63
Compliance	181	26	52

3.2. Pre-Processing and Data Augmentation Techniques

The BERT model's performance in classifying memorandum circulars depends on data quality, pre-processing, fine-tuning, and computational resources. High quality, well-tagged datasets are needed to avoid misclassification and overfitting. Additionally, the pre-processing techniques like effective tokenization, text cleaning, and handling of long documents also plays a significant role. Techniques like text segmentation or key sentence extraction are utilized to mitigate and limit prospective information loss. Data augmentation methods such as paraphrasing, synonym replacement, and back translation also enhance the training set and improve model stability without consuming extra data.

3.3. System Architecture

The BERT-based classification model operates as a transformer-based natural NLP architecture that is designed for the task of document classification. The processing pipeline ensure tokenization of documents and converts them into meaning-preserved embeddings. Built on a pre-trained BERT model, it captures deep contextual relationships between words. For memorandum circulars with complex language, the attention mechanism enhances the classification accuracy by simply the model process the whole word sequence in a single step.

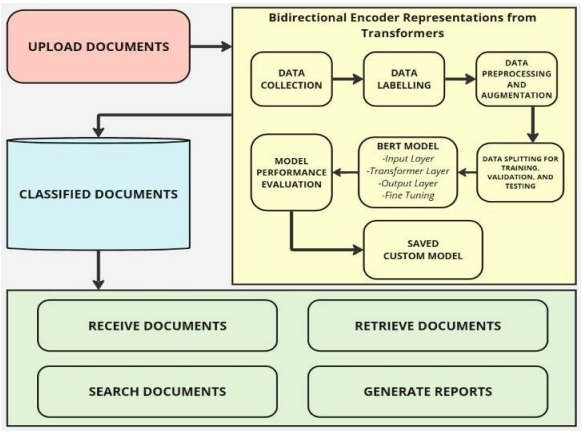

Fig. 1: System Architecture

BERT utilizes a transformer-based architecture that includes the utilization of feed-forward networks and multiple layers of self-attention mechanisms. After tokenization, it produces embeddings that have token, position, and segment information. The model employs stacked transformer encoder layers that each incorporate feed-forward networks, residual connections, and multi-head self-attention. The result is a collection of contextualized word representations that are fine-tuned to perform different tasks.

3.4. BERT Fine-Tuning

To enhance the model performance, we fine-tune BERT with a train batch size of 8, validation batch size of 8, testing batch size of 8, and its default learning rate over five (5) epochs optimal for small datasets to prevent overfitting while capturing complex patterns. The model is trained on the pre-processed memorandum circular dataset, leveraging the universal language representations for more improved classification accuracy. The key hyperparameters are learning rate, batch size, and epochs, are carefully adjusted. Additionally, the learning rate scheduling, dropout regularization, and early stopping are applied to ensure efficient training and prevent overfitting. Stratified Kfold Parameters is set to splits five (5) that controls on how many folds the data is split into, Shuffle set to true to avoid biased splits and random state to forty-two (42) to ensure reproducibility.

3.5. Evaluation Metrics

To assess the model's performance, the following metrics are applied:

- True Positive: The actual result was positive, as well predicted by the model.
- True Negative: The actual result was negative, as well predicted by the model.
- False Positive: The actual result was negative, but the model predicted a positive result.
- False Negative: The actual result was positive, but the model predicted a negative result.

4. Results and Discussion

Table 3: Results of Training

Classification	Precision	Recall	F1-Score
Administrative	0.82	0.56	0.67
Emergency_Special	0.96	0.96	0.96
Training_Development	0.89	0.95	0.92
Policy_Clarification	0.82	0.72	0.77
Compliance	0.52	0.68	0.59

The table 3 represents the model's performance on the training dataset. The model performed best on the Emergency_Special and Training_Development classes, achieving F1-score of 0.96 and 0.92 respectively, these results indicate that the model has effectively learned the patterns from the data for training. However, such high scores suggest a potential risk of overfitting, meaning he model might not adapt well to unknown information. The Administrative (F1-score = 0.67) and Compliance (F1-score = 0.59) had the lowest result, indicating more difficulty in correctly identifying samples from this class.

Table 4: Results of Validation Data

Classification	Precision	Recall	F1-Score
Administrative	0.90	0.58	0.70
Emergency_Special	0.88	0.97	0.92
Training_Development	0.85	0.95	0.89
Policy_Clarification	0.80	0.60	0.68
Compliance	0.50	0.65	0.56

Table 4 demonstrates the model's performance on the validation dataset, which is used to fine-tune the model and prevent overfitting. Compared to the training results, the scores are slightly lower, particularly in Policy_Clarification (F1-score = 0.68) and Compliance (F1-score = 0.56), While Emergency_Special and Training_Development achieving F1-scores of 0.92 and 0.89 respectively. suggesting that the model's ability to generalize varies across classes, other classes experienced a decline, which indicates potential difficulty in correctly classifying certain memorandum types when exposed to new data.

Table 5: Results of Testing Data

Classification	Precision	Recall	F1-Score
Administrative	0.97	0.60	0.74
Emergency_Special	0.84	0.98	0.91
Training_Development	0.87	0.96	0.91
Policy_Clarification	0.75	0.54	0.63
Compliance	0.49	0.68	0.56

Lastly, Table 5 represents the model's performance on the testing dataset, which assesses its effectiveness in real-world scenarios. The results show relatively strong performance across most classifications, with Training_Development and Emergency_Special achieving an F1-score of 0.91, indicating perfect classification. However, some classes, such as Policy_Clarification (F1-score = 0.63) show a notable drop in recall (0.54), suggesting that the model may struggle to correctly identify all instances of this class. Despite these variations, the overall performance suggests that the model can effectively classify most memorandum types, although some improvements may be necessary to enhance recall and precision for specific classes.

Several factors influenced the model's performance, especially when classification accuracy differed between categories. Class imbalance played a key role, as classifications with more training tests, such as Training_Development and Emergency_Special, allowed the model to learn distinguishing characteristics more effectively, whereas less frequent classes, such as Policy_Clarification and Administrative, had less instances, increasing the likelihood of error. Furthermore, feature similarity across classifications, due to similar language and structure, also complicated classification. The limited training data, may have further restricted the model's ability to generalize effectively. Moreover, the model may have experienced overfitting to some classes potentially leading to a bias that resulted in misclassifications of minority classes. Addressing these factors through data augmentation, balancing techniques and applying the Stratified Kfold Technique or model adjustments could enhance performance and generalizability.

5. Conclusion

The BERT-based document classification model effectively categorizing diverse memorandum circulars, showing strong results based on their F1-scores. However, lower recall for the Policy_Clarification class suggests challenges likely due to class imbalance and overlapping features. In contrast, class like Training_Development and Emergency_Training achieved the highest classification, likely due to distinct structural pattern. While the model generalizes well, minor misclassifications indicate areas for improvement. Techniques like data augmentation, stratified kfold sampling, and fine-tuning could improve accuracy, particularly for underrepresented classes. Overall, this study highlights the model's effectiveness in automating document classification, contributing to improved accessibility and efficiency in document management.

6. Acknowledgements

The successful completion of this study would not have been possible without the support and guidance of many individuals. The authors sincerely appreciate everyone especially the Polytechnic University of the Philippines for providing invaluable support and encouragement throughout this journey. Additionally, we recognize the contributions of volunteers for their hard work in processing and annotating the dataset. Lastly, we express our deepest gratitude to God Almighty for granting us strength, wisdom, guidance throughout this journey.

7. References

[1] I. Triyadi, B. Prasetiyo, and T. Nikmak. 2023. News text classification using long-term short memory (LSTM) algorithm. *J. Soft Comput. Explor.* 4, 2 (June 2023), 79–86.

[2] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the ACL 2021*, 7088–7105. DOI:https://doi.org/10.18653/v1/2021.acl- long.551

[3] A. Ismael and Okumus. 2017. Design and implementation of an electronic document management system. *In Proceedings of Maku University*, 9–17. DOI:https://doi.org/10.31200/makuubd.321093

[4] J. M. Noyes and K. J. Garland. 2008. Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics* 51, 9 (Sept. 2008), 1352–1375. DOI:https://doi.org/10.1080/00140130802170387

[5] M. Sambetbayeva, I. Kuspanova, A. Yerimbetova, S. Serikbayeva, and S. Bauyrzhanova. 2022. Development of intelligent electronic document management system model based on machine learning methods. *East.-Eur. J. Enterp. Technol.* 1, 2(155) (April 2022), 68–76. DOI:https://doi.org/10.15587/1729-4061.2022.251689

[6] A. Dasgupta, P. Drineas, B. Harb, V. Josifovski, and M. Mohoney. 2007. Feature selection methods for text classification. *In Proceedings of ACM KDD 2007*. DOI:https://doi.org/10.1145/1281192.1281220

[7] A. Adhikari, A. Ram, R. Tang, and J. Lin. 2019. DocBERT: BERT for document classification. *arXiv preprint*. DOI:https://doi.org/10.48550/arXiv.1904.08398

[8] W. Antoun, F. Baly, and H. Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of ACL 2020*, 9–15.

[9] N. Limsopatham. 2021. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, 210–216. DOI:https://doi.org/10.18653/v1/2021.nllp-1.22

[10] M. Ostendorff, P. Bourgonje, M. Berger, J. Moreno-Schneider, G. Rehm, and B. Gipp. 2019. Enriching BERT with knowledge graph embeddings for document classification. *Speech and Language Technology, DFKI GmbH, Germany*. DOI:https://doi.org/10.48550/arXiv.1909.08402

[11] A. Basarkar. 2017. Document classification using machine learning. Retrieved from https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1531&context=etd_projec ts

[12] R. Power, J. Chen, T. Karthik, and L. Subramanian. 2009. Document classification for focused topics. Retrieved from https://chenjay.org/publications/aaai4d-power.pdf

[13] L. H. Cheeks, T. L. Stepien, and D. M. Wald. 2016. Discovering news frames: Exploring text, content, and concepts in online news sources to address water insecurity in the Southwest region. In *Proceedings of IRI 2016*. DOI:https://doi.org/10.1109/IRI.2016.67

[14] A. Khan, B. Baharudin, L. H. Lee, and K. Khan. 2010. A review of machine learning algorithms for text- documents classification. *J. Adv. Inf. Technol.* 1, 1 (Feb. 2010).

[15] R. S. Wilkho, S. Chang, and N. G. Gharaibeh. 2024. FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events. *Adv. Eng. Inform.* 59 (2024). DOI:https://doi.org/10.1016/j.aei.2023.102293

[16] T. Verma and N. S. Gill. 2020. Machine learning techniques for better data-driven decisions revisited. *Int. J. Eng. Adv. Technol. (IJEAT)* 9, 4 (April 2020). DOI:https://doi.org/10.35940/ijeat.D6766.049420