

Evaluating the Suitability of Long Document Embeddings for Classification Tasks: A Comparative Analysis

Bardia Rafieian^a and Pere-Pau Vázquez^b

ViRVIG Group Department of Computer Science, UPC-BarcelonaTECH, C/ Jordi Girona 1-3,
Ed Omega 137, 08034, Barcelona, Spain
{bardia.rafieian, pere.pau.vazquez}@upc.edu

Keywords: Long Document Classification, Document Embeddings, Doc2vec, Longformer, LLaMA-3, SciBERT, Deep Learning, Machine Learning, Natural Language Processing (NLP).

Abstract: Long documents pose a significant challenge for natural language processing (NLP), which requires high-quality embeddings. Despite the numerous approaches that encompass both deep learning and machine learning methodologies, tackling this task remains hard. In our study, we tackle the issue of long document classification by leveraging recent advancements in machine learning and deep learning. We conduct a comprehensive evaluation of several state-of-the-art models, including Doc2vec, Longformer, LLaMA-3, and SciBERT, focusing on their effectiveness in handling long to very long documents (in number of tokens). Furthermore, we trained a Doc2vec model using a massive dataset, achieving state-of-the-art quality, and surpassing other methods such as Longformer and SciBERT, which are very costly to train. Notably, while LLaMA-3 outperforms our model in certain aspects, Doc2vec remains highly competitive, particularly in speed, as it is the fastest among the evaluated methods. Through experimentation, we thoroughly evaluate the performance of our custom-trained Doc2vec model in classifying documents with an extensive number of tokens, demonstrating its efficacy, especially in handling very long documents. However, our analysis also uncovers inconsistencies in the performance of all models when faced with documents containing larger text volumes.

1 INTRODUCTION


Text embeddings are pivotal in natural language processing (NLP) tasks, such as text classification, where the quality of embeddings significantly affects performance. Traditional methods, such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), have been foundational in generating embeddings at the token and sentence levels. Recent advancements include transformer-based models like BERT (Devlin et al., 2018) and ELMO (Peters et al., 2018), which have improved the quality of embeddings through contextualized representations.


Despite these advancements, handling very long documents presents substantial challenges. Models like BERT are limited by maximum sequence lengths, which restricts their ability to generate embeddings for extensive texts. While recent models such as Longformer (Beltagy et al., 2020) and large language models offer better performance, they come with high computational costs and resource demands (Samsi

et al., 2023). These models are expensive to train and deploy, and there is a lack of standardized evaluations across various benchmarks (Tay et al., 2021).

On the other side, the scarcity of datasets with very long documents further complicates the issue, where existing labeled datasets consist only of short articles (up to 800 tokens per document), yet training a classifier for long texts requires a labeled dataset consisting of long documents. This gap in the literature highlights the need for more effective methods to handle lengthy texts while considering computational efficiency. This paper provides an evaluation of several state-of-the-art models, including Doc2vec, Longformer, LLaMA-3, and SciBERT, focusing on their effectiveness in handling long to very long document tokens. The study specifically aims to assess how well these models generate embeddings from documents that are exceptionally long in terms of token count. The key question is how agnostic these models are to document length, and how the quality of the generated embeddings influences their performance in downstream text classification tasks.

We also trained a Doc2vec model on a large

^a  <https://orcid.org/0000-0003-4591-8934>

^b  <https://orcid.org/0000-0003-4638-4065>

dataset to evaluate its capability against BERT-based models and large language models (LLMs) in generating embeddings for very long documents. The goal was to assess how well the Doc2vec model performs compared to these advanced models in terms of both the quality of the embeddings produced and their effectiveness in a text classification task. This comparison provides insights into the relative strengths and limitations of traditional embedding methods like Doc2vec versus modern transformer-based approaches when handling lengthy documents. Given the scarcity of very long document datasets, our evaluation utilizes both public datasets and newly created datasets with over 4,000 tokens from arXiv and bioRxiv documents. The paper is structured as follows: Section 2 reviews related work. Section 3 details data preparation and preprocessing steps and the preparation of our pretrained Doc2vec model. In section 4 we study our experiments and finally, we conclude and discuss future directions.

2 RELATED WORKS

With the introduction of Word2Vec and GloVe, various methods have emerged to encode sentences, paragraphs, and longer texts into embeddings. Among these methods, Doc2vec (Le and Mikolov, 2014) stands out as a Paragraph Vector, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of text. Empirical results have shown that Doc2vec outperforms bag-of-words models and other text representation techniques.

The advent of transformer models brought significant improvements in text encoders. BERT-based models, in particular, demonstrated substantial performance gains. The first application of BERT to document classification, as presented in "DocBERT: BERT for Document Classification" (Adhikari et al., 2019), improved baseline results by fine-tuning BERT, achieving higher classification accuracy across various datasets. However, BERT-based models were limited by a fixed input sequence length of 512 tokens. To address this, models like SciBERT extended the number of tokens to 768 through fine-tuning. In SciBERT that follows the BERT architecture, which uses the Transformer model for encoding text, the process of generating embeddings can be described as follows:

The input is a tokenized text sequence:

$$\mathbf{x} = [x_1, x_2, \dots, x_n]$$

The tokens are then converted to embeddings:

$$\mathbf{E} = [E(x_1), E(x_2), \dots, E(x_n)]$$

Next, these embeddings pass through multiple transformer layers. Each transformer layer applies self-attention:

$$\mathbf{H}^{(l)} = \text{TransformerLayer}(\mathbf{H}^{(l-1)})$$

where $\mathbf{H}^{(0)} = \mathbf{E}$, and l is the layer number.

The final hidden states from the last transformer layer are used for downstream tasks:

$$\mathbf{H}^{(L)} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n]$$

Despite these advancements, transformer-based models struggle with processing long sequences due to the computational complexity of their self-attention mechanism, which can lead to information loss in documents with more than 1,000 tokens. To overcome this limitation, the Longformer was introduced. It features an attention mechanism that scales linearly with sequence length, allowing it to handle documents with thousands of tokens. The Longformer achieves this by sparsifying the full self-attention matrix according to an "attention pattern" that specifies which input locations attend to each other. This makes the model efficient for longer sequences. At the time of its introduction, the Longformer consistently outperformed RoBERTa on long document tasks, setting new state-of-the-art results on datasets like WikiHop and TriviaQA. Here is the process of generating embeddings using longformer:

The input is a tokenized sequence:

$$\mathbf{x} = [x_1, x_2, \dots, x_n]$$

The attention mechanism is restricted to a fixed-size window:

$$\mathbf{A}_i = \text{Attention} \left(\mathbf{H}_i^{(l-1)}, \mathbf{H}_{i-w:i+w}^{(l-1)} \right)$$

where w is the window size. Global attention is applied to selected important tokens across the entire sequence. The embeddings are passed through multiple layers of this modified attention mechanism. The final hidden states are used for downstream tasks:

$$\mathbf{H}^{(L)} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n]$$

More recently, significant efforts have been made to improve the performance of text encoders on long texts. Notable examples include the LLaMA-2 (Touvron and Lavril, 2023) and LLaMa-3 models. Although detailed technical information about these proprietary models is limited, they propose novel methods for generating embeddings from long texts, further advancing the field of NLP. Since we used LLaMA-3 and GEMMA-2B, we describe the process of generating embeddings as below:

LLaMA follows a standard transformer architecture with self-attention and feedforward networks.

The input is a tokenized text sequence:

$$\mathbf{x} = [x_1, x_2, \dots, x_n]$$

Each transformer layer consists of multi-head self-attention and feedforward networks:

$$\mathbf{H}^{(l)} = \text{MultiHeadAttention}(\mathbf{H}^{(l-1)}) + \text{FFN}(\mathbf{H}^{(l-1)})$$

The final hidden states are used for downstream tasks:

$$\mathbf{H}^{(L)} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n]$$

A recent study on transformer-based models (Fields et al., 2024) addresses key questions such as "How Wide, How Large, How Long, How Accurate, How Expensive, and How Safe are they?" The study emphasizes the latest advancements in large language models (LLMs) by evaluating their accuracy across 358 datasets spanning 20 different applications. The findings challenge the assumption that LLMs are universally superior, revealing unexpected results related to accuracy, cost, and safety. LLMs now encompass both unimodal and multimodal tasks, where unimodal models use only textual information, and multimodal models incorporate text, video, signals, images, audio, and columnar data for classification. The paper highlights that while recent models like GPT-4 and Longformer can handle input text lengths of up to 8,192 tokens with high accuracy in classification tasks, the cost of training these LLMs, along with the associated economic and environmental concerns, has become a significant issue in recent years. Another notable study by (Wagh et al., 2021) examines the classification of long documents. The authors reaffirm that while BERT-based models can perform well across various datasets and are suitable for document classification tasks, they come with a high computational cost. They also point out that long document classification is a relatively simpler task, and even basic algorithms can achieve competitive performance compared to BERT-based approaches on most datasets.

3 METHODOLOGY

In our paper, we compare and discuss the capabilities of state-of-the-art models in generating high-quality embeddings for very long texts. Subsequently, we evaluate the generated embeddings using various methods, including Doc2vec, in the context of document classification tasks.

3.1 Datasets

Given the ongoing challenge of benchmarking very long texts due to the lack of agreement on datasets

and baselines (Tay et al., 2021), we have prepared and introduced datasets with more than 1,000 tokens per text to evaluate embedding quality. Table 1 shows the detailed information about each dataset.

Table 1: Dataset information including token size, sample size, and number of labels. Note*: 20 news and arxiv_100 information on section appendix 6.

Dataset	# Avg. Tokens	Size	Labels
Dataset#1	7630	554	11
Dataset#2	11305	1101	11
s2orc	3450	58905	4
20 news*	149	11297	20
arxiv_100*	121	100004	10

S2ORC. Semantic Scholar Open Research Corpus (Lo et al., 2020) is a comprehensive corpus designed for natural language processing and text mining on scientific papers. It includes over 136 million paper nodes, with more than 12.7 million full-text papers connected by approximately 467 million citation edges, derived from various sources and academic disciplines. The number of tokens in our selected dataset ranges from 1 to 287,400. We chose documents with at least 200 tokens in two classes of computer science and physics to ensure they are not smaller than the shortest document in our test set.

arxiv + Biorxiv. This dataset includes documents from the years 2022 and 2023 in both combination of arxiv and biorxiv, containing 550 and 1,000 documents respectively. Each document includes the full text of the papers, with an average of more than 7,000 tokens after preprocessing (tokenization, lemmatization, stop words removal and extra phrases removal). These datasets encompass multiple classes where for arxiv+biorxiv 2022 (labeled as **Dataset#1**) includes Evolutionary Biology, Paleontology, Mathematics, Computer Science, Zoology, Statistics, Pharmacology and Toxicology, Biochemistry, Economics, Physics and Electrical Engineering. On the other side, the arxiv+biorxiv 2023 (labeled as **Dataset#2**) dataset contain Biochemistry, Paleontology, Genomics, Quantitative Biology, Quantitative Finance, Statistics, Computer Science, Electrical Engineering and Systems Science, Mathematics, Physics and Zoology labels.

To prepare them, we first converted PDF documents to text format and then removed author names, images, tables, captions, references, acknowledgments, and formulas. Furthermore, we eliminated sentences with fewer than three tokens. All preprocessing steps, as well as subsequent operations, were executed using Python 3.

3.2 Embeddings

Doc2vec. Given Doc2vec’s scalability with large datasets, we explored its functionalities by training it on extensive technical corpora. For training purposes, we focused on technical documents of S2ORC and collected 341,891 documents, totaling approximately 10GB, from fields including Engineering, Computer Science, Physics, and Math. It is important to note that we excluded the test set from the training set to train the Doc2vec model effectively. Finally, we generated the embeddings using Doc2vec.

SciBERT. (Beltagy et al., 2019), is a BERT-based model pre-trained on a large corpus of scientific text, which includes papers from the corpus of Semantic Scholar. The model aims to address the unique challenges posed by scientific text, such as specialized terminology and longer sentence structures. By leveraging this specialized pre-training, SciBERT achieves better performance on downstream scientific NLP tasks compared to the vanilla BERT model, particularly in domains like biomedical and computer science literature. In this model, full-text documents are encoded into chunks of 512 tokens. We generated text embeddings using the SciBERT model *SciBERT_scivocab_uncased* with a maximum sequence length of 512 tokens. The final layer’s hidden states were used as embeddings, with mean pooling applied to obtain sentence-level embeddings. We utilized the Hugging Face ‘transformers’ library (version 4.x.x) for model loading and inference.

Longformer. Introduced by (Beltagy et al., 2020), addresses the challenge of processing long documents by extending the input sequence token size up to 4096 tokens, significantly more than BERT’s 512-token limit. Longformer employs a combination of local and global attention mechanisms that scale linearly with the sequence length, allowing it to handle much longer documents efficiently. This model is specifically designed to mitigate the computational inefficiencies of the quadratic complexity of the standard self-attention mechanism in BERT. In our experiments, we utilized the Longformer-large model *allenai/longformer.large_4096* to generate document embeddings. This model comprises 24 layers, each with a hidden size of 1024, and uses 16 attention heads. It is capable of processing sequences up to 4096 tokens in length, leveraging a sliding window attention mechanism with a window size of 512 tokens and supporting global attention for key tokens. Embeddings were generated by extracting the CLS token’s output from the last hidden layer, optionally followed by mean pooling for a fixed-size representation.

LLaMA-3 and GEMMA-2B. Large Language Model for AI Assistance (Touvron and Lavril, 2023) represents a significant advancement in the realm of large-scale language models. Unlike earlier models like BERT or even Longformer, which are constrained by their maximum input sequence lengths, LLaMA-3 is designed to handle extremely large contexts, accommodating up to 16,000 tokens per sequence. This makes it particularly suitable for tasks involving extensive documents, such as entire books, comprehensive reports, and complex dialogues. Moreover, GEMMA-2B (Team, 2024) (Generative Embedding Model with Multi-headed Attention) distinguishes itself with a focus on generating high-quality embeddings for downstream NLP tasks. This model operates with a maximum input sequence length of 2048 tokens, striking a balance between the extensive context capabilities of models like LLaMA-3 and the more focused scope of traditional models. We generated text embeddings using the LLaMA 3 8B model provided by Ollama (Ollama, 2024). This model, which has 8.03 billion parameters, is optimized for instruction-following tasks and operates efficiently through quantization techniques, such as Q4.0. The embedding generation process utilizes the output from the model’s last hidden layer, ensuring rich contextual representations of the input text. Ollama’s quantization reduces the model’s size to 5.5GB, allowing for effective deployment on local hardware while maintaining high-quality performance. Table 2 illustrates detailed information of each model.

Table 2: Characteristics of different models including vocabulary size, corpus size, maximum length, and embedding size.

Model	Vocab	Corpus	Max Len	Embedding
SciBERT	30K	1.14M	512	768
Doc2vec	33K	1.2M	10k	400
LLaMA-3	128K	15 Tn	8k	4096
GEMMA-2B	256K	6 Tn	8k	2048
Longformer	30K	33 Tn	4k	768

4 EXPERIMENTS AND RESULTS

In this section, we present the results of our experiments on several datasets using state-of-the-art models to generate high-quality embeddings for text classification. The models evaluated include Doc2vec, LLaMA-3, Longformer, SciBERT, and GEMMA-2B. We utilized both SVM and MLP classifiers to assess the performance of these embeddings. The evaluation metrics include accuracy, precision, recall, and F1 score. The reason we selected these classifiers,

rather than model-based ones like LongformerClassifier, is to remain agnostic regarding classifier selection. This approach allows us to reuse the embeddings for other NLP tasks, providing greater flexibility and utility. Below we give more information on each:

SVM. We utilized a Support Vector Machine (SVM) classifier with a linear kernel to perform the classification tasks. The model was configured with a regularization parameter, C , set to 1.0 to balance the trade-off between minimizing training error and achieving low testing error. The SVM classifier was trained on the given feature set and corresponding labels, facilitating effective class separation within the feature space. **MLP.** We utilized the MLP Classifier from scikit-learn to build a neural network classifier for our dataset. The model features two hidden layers with 100 and 50 neurons, respectively, and was trained for a maximum of 60 iterations. We set the random seed to 42 for reproducibility.

4.1 Results

We observed Doc2vec consistently demonstrated robust performance on the Dataset#2 dataset, achieving an MLP accuracy of 0.67, and an F1 score of 0.65. Longformer also delivered competitive results, with SVM accuracy of 0.64, and an F1 score of 0.65. In contrast, SciBERT and LLaMA-3 showed slightly lower performance, with SVM accuracies of 0.61 and 0.56, and MLP accuracies of 0.64 and 0.60. The GEMMA-2B model, however, had the least favorable outcomes. We can express the lower results of GEMMA-2B model comparing with LLaMA-3 due to its lower embedding dimension and model parameters size. We were surprised by the strong performance of TF-IDF embeddings, which outperformed all other models, likely due to its effectiveness in handling massive documents. On Dataset#1 Doc2vec emerged as the top performer, achieving an SVM accuracy of 0.7590, an MLP accuracy of 0.71, and an F1 score of 0.78. SciBERT followed closely, with an SVM accuracy of 0.72, an MLP accuracy of 0.71, and an F1 score of 0.72. Longformer, however, showed a decline in performance, reflected by an SVM accuracy of 0.5500, an MLP accuracy of 0.5833, and an F1 score of 0.5550. LLaMA-3 provided moderate results with an SVM accuracy of 0.4940, an MLP accuracy of 0.3976, and an F1 score of 0.7804. Meanwhile, GEMMA-2B continued to struggle, recording the lowest performance metrics with an SVM accuracy of 0.4700, an MLP accuracy of 0.4600, and an F1 score of 0.4500.

Finally LLaMA-3 demonstrated superior performance on the S2ORC dataset with two classes,

achieving nearly perfect scores with an SVM accuracy, MLP accuracy, and F1 score all at 0.99. Doc2vec also showed strong results, with an SVM accuracy of 0.97, an MLP accuracy of 0.99, and an F1 score of 0.9776. Both Longformer and SciBERT maintained high levels of accuracy, with SVM scores of 0.96 and 0.97, and MLP accuracies of 0.99 and 0.97, respectively, complemented by high F1 scores. These findings highlight the remarkable efficiency of LLaMA-3 and Doc2vec in managing large-scale scientific documents.

The results in table 3 indicates that LLaMA-3 consistently outperforms other models across various datasets, particularly on the 20 news (6) and S2ORC datasets, demonstrating its robustness and effectiveness in handling long and shorter documents by generating high-quality embeddings. Doc2vec also shows competitive performance, especially on the S2ORC dataset. Longformer and SciBERT exhibit moderate performance, with SciBERT performing better on the arxiv_100 dataset (APPENDIX). GEMMA-2B, while a powerful model for embedding generation, did not perform as well in this classification task, suggesting that its embeddings might need further fine-tuning for specific tasks or datasets. To further analyze the effectiveness of the embeddings generated by the different models, we projected the high-dimensional embeddings into a 2D space using the PACMAP dimensionality reduction technique (Wang et al., 2021). This visualization allows for a deep understanding of how well the models differentiate between classes in various datasets (see Appendix 6).

4.2 Training and Inference Time

Transformer-based models, such as SciBERT, LLMs, and Longformer, possess a complex architecture involving multi-head self-attention mechanisms and multiple layers, which enable them to capture entangled dependencies and contextual information. These models typically require massive datasets for pre-training where depending on the model size and hardware, the training time can range from several days to months, although fine-tuning usually takes a few hours to a few days on powerful GPUs (Devlin et al., 2018). In contrast, simpler models like Word2Vec and Doc2vec use much less complex architectures. Word2Vec, for example, leverages shallow neural networks with a single hidden layer, while Doc2vec extends Word2Vec by considering document context but remains relatively straightforward. These models also utilize large datasets but not to the extent required for transformer models, typically training on corpora

Table 3: Evaluation metrics Macro average of (Precision, Recall, F1 Score) and SVM Classification/MLP classification accuracy for different models across bioarxiv 2022, bioarxiv 2023, and s2orc datasets.

Dataset	Model	Macro avg. P	Macro avg. R	Macro avg. F1	SVM acc	MLP acc
Dataset#2	Doc2vec	0.6702	0.6545	0.6593	0.6545	0.6780
	GEMMA-2B	0.5100	0.5100	0.5000	0.5000	0.5000
	LLaMA-3	0.5964	0.5697	0.5744	0.5697	0.6000
	Longformer	0.6919	0.6424	0.6519	0.6424	0.6420
	SciBERT	0.6295	0.6182	0.6213	0.6180	0.6400
	TF-IDF	0.8900	0.8800	0.8800	0.8800	0.8900
Dataset#1	Doc2vec	0.8181	0.7711	0.7825	0.7590	0.7100
	GEMMA-2B	0.4800	0.4700	0.4500	0.4700	0.4600
	LLaMA-3	0.7819	0.7819	0.7804	0.4940	0.3976
	Longformer	0.6789	0.5500	0.5550	0.5500	0.5833
	SciBERT	0.7597	0.7229	0.7279	0.7229	0.7100
	TF-IDF	0.7400	0.7200	0.7200	0.7600	0.7800
s2orc	Doc2vec	0.9775	0.9777	0.9776	0.9778	0.9998
	LLaMA-3	0.9976	0.9976	0.9976	0.9976	0.9976
	Longformer	0.9674	0.9677	0.9675	0.9678	0.9993
	SciBERT	0.9749	0.9749	0.9749	0.9749	0.9797
	TF-IDF	0.9700	0.9700	0.9700	0.9800	0.9800

containing millions to billions of words. Training these models is much faster, with Doc2vec being trainable on a large corpus in a matter of hours using a few CPUs or a single GPU, still considerably quicker than transformer models. Figures 1a and 1b show the comparison of fine-tuning—training/ time and memory/time of full self-attention and different implementations of Longformer’s methods vs Doc2vec.

As shown in 2, Doc2vec can perform competitively while offering significant advantages in terms of inference time, resource requirements, and energy consumption. Specifically, Doc2vec demonstrates much faster average embedding inference times per second on a CPU, needing significantly less computational resources and consuming less energy compared to other models.

5 LIMITATIONS

One of the significant challenges encountered in this study was finding datasets with tokens exceeding 1,000 to effectively compare the models’ ability to extract embeddings from very long texts. Such datasets are crucial for evaluating model performance on extended sequences.

Additionally, inferring heavy models like LLaMA-3 and GEMMA-2B required substantial time, effort, and computational resources. These models have considerable demands, and their inference process was constrained by the limitations of available libraries and computing environments.

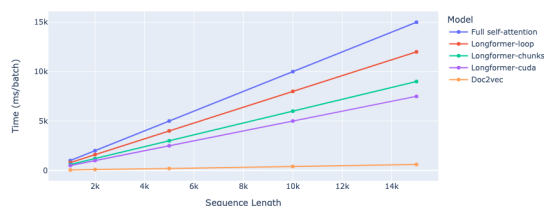
6 CONCLUSIONS

In this study, we have evaluated the performance of various state-of-the-art models, including Doc2vec, SciBERT, Longformer, LLaMA-3, and GEMMA-2B, on the task of generating high-quality embeddings for text classification. Our experiments spanned multiple datasets such as 20 news, arxiv_100, Dataset#1, Dataset#2, and S2ORC, providing a comprehensive analysis of each model’s strengths and limitations.

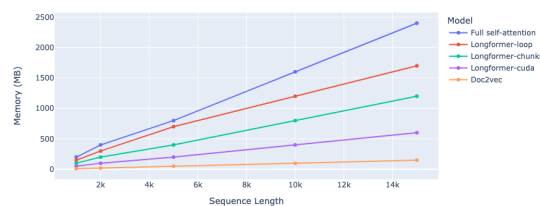
The results indicate that LLaMA-3 consistently outperforms other models across different datasets, particularly excelling in the 20 news and S2ORC datasets with superior accuracy and F1 scores. SciBERT also demonstrated robust performance, especially with the arxiv_100 dataset. Notably, Doc2vec, while slightly behind in absolute performance metrics, offers competitive results with significantly better computational efficiency, making it an excellent choice for applications requiring faster inference times and lower resource consumption. This balance between performance and efficiency is critical for practical deployment in real-world scenarios.

Additionally, our study highlighted the challenges associated with handling very long documents, where models like Longformer and LLaMA-3, designed for extended context processing, showed significant advantages. However, GEMMA-2B, despite its powerful embedding capabilities, requires further fine-tuning.

In future, we aim to investigate the quality of embeddings in additional NLP tasks, such as question



(a) Fine-tuning time of self-attention Longformers and Doc2vec (Beltagy et al., 2020).



(b) Memory usage of self-attention Longformers and Doc2vec (Beltagy et al., 2020).

Figure 1: Performance comparison of Longformers and Doc2vec models.

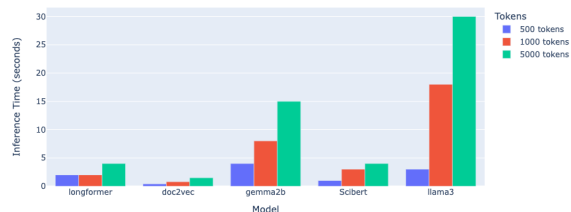


Figure 2: Inference time of different models for embedding extraction.

answering and summarization on very long texts. We will also review the tuned combinations of embeddings for specific tasks and domains.

ACKNOWLEDGEMENTS

This project has been supported by PID2021-122136OB-C21 from the Ministerio de Ciencia e Innovación, by 839 FEDER (EU) funds.

REFERENCES

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Docbert: BERT for document classification. *CoRR*, abs/1904.08398.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pre-trained language model for scientific text.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fields, J., Chovanec, K., and Madiraju, P. (2024). A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531.

Lang, K. (1995). Newsweeder: learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning, ICML'95*, page 331–339, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Ollama (2024). Ollama: Ai models locally. Accessed: July 26, 2024.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. *arXiv e-prints*, page arXiv:2310.03003.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. (2021). Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Team, G. (2024). Gemma: Open models based on gemini research and technology.

Touvron, H. and Lavril, T. (2023). Llama: Open and efficient foundation language models.

Wagh, V., Khandve, S. I., Joshi, I., Wani, A., Kale, G., and Joshi, R. (2021). Comparative study of long document classification. *CoRR*, abs/2111.00702.

Wang, Y., Huang, H., Rudin, C., and Shaposhnik, Y. (2021). Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73.

APPENDIX

Datasets 20 News: (Lang, 1995) is widely used for text classification and natural language processing (NLP) tasks. It contains approximately 20,000 news-group documents, divided into 20 different news-groups.

arxiv_100: dataset comprises 100,000 arXiv paper abstracts and averages 121 tokens per document, covering subjects such as Electrical Engineering and Systems Science, Statistics, Computer Science, Physics, Quantum Physics, Mathematics, High Energy Physics - Theory, High Energy Physics, Condensed Matter Physics, and Astrophysics.

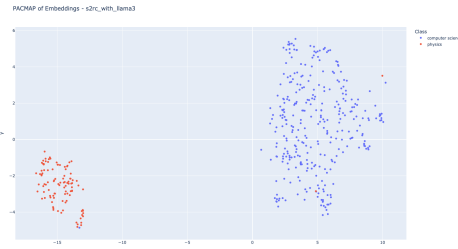
Results: On 20 news dataset, LLaMA-3 significantly outperformed other models, achieving an SVM accuracy of 0.97 and an F1 score of 0.97. Doc2vec showed decent performance with an SVM accuracy of 0.75, while its F1 score was 0.67. Longformer and SciBERT demonstrated moderate results, with SVM accuracies of 0.75 and 0.66, and MLP accuracies of 0.65 and 0.65, respectively. LLaMA-3's results reflect its superior ability to handle the complexity of the news-group data.

On arxiv_100, SciBERT led with an SVM accuracy of 0.81, both with an F1 score of 0.81. Doc2vec followed closely, with SVM and MLP accuracies of 0.81 and 0.76. LLaMA-3 also performed well, showing an SVM accuracy of 0.78, and an F1 score of 0.78. Longformer lagged behind with an SVM accuracy of 0.72 and an MLP accuracy of 0.74, with an F1 score of 0.72. These results underscore SciBERT's effectiveness in handling scientific abstracts and technical documents. Table 4 summarizes the classification and F-score results on these datasets.

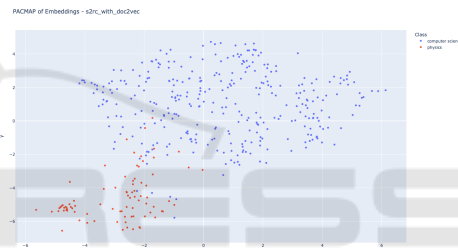
Dimensionality Reduction and Embedding Analysis: We applied the PACMAP dimensionality reduction method (Wang et al., 2021) to embeddings extracted from various models on the S2ORC test set. As illustrated in Figure 3, LLaMA-3 effectively separated the embeddings in the 2D space, demonstrating distinct class separation. While Doc2Vec and SciBERT also achieved some degree of separation between classes, the resulting data points remained in close proximity within the 2D space. Finally, Longformer, despite distinguishing the classes, performed the weakest separation performance among the others.

Table 4: Evaluation metrics (Precision, Recall, F1 Score) and SVM/MLP classification results for different models across arxiv_100 and 20 news datasets.

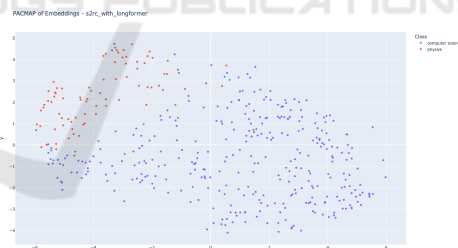
Data	Model	P	R	F1	SVM	MLP
20n	Doc2vec	0.6700	0.6665	0.6665	0.747	0.694
	LLaMA-3	0.9775	0.9741	0.9749	0.974	0.971
	Longf	0.6481	0.6324	0.6303	0.746	0.646
	SciBERT	0.6628	0.6581	0.6589	0.658	0.653
arxiv_100	Doc2vec	0.8009	0.8007	0.8007	0.805	0.756
	LLaMA-3	0.7819	0.7819	0.7804	0.781	0.780
	Longf	0.7183	0.7173	0.7171	0.716	0.739
	SciBERT	0.8094	0.8093	0.8093	0.809	0.785



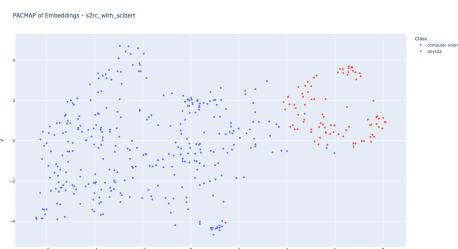
a) PACMAP with LLaMA-3



b) PACMAP with Doc2vec



c) PACMAP with Longformer



d) PACMAP with SciBERT

Figure 3: 2D embedding visualization on S2ORC dataset(tests), extracted from a) LLaMA-3, b) Longformer, c) SciBERT and d) Doc2vec, results showing a great performance of LLaMA-3 on class separations.