Journal of Intelligent Learning Systems and Applications, 2024, 16, 143-154
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

State Space Models Based Efficient Long Documents Classification

Bo Song*, Yuanhao Xu, Penghao Liang, Yichao Wu

Khoury College of Computer Science, Northeastern University, Boston, MA, USA
Email: *song.bo1@northeastern.edu

How to cite this paper: Song, B., Xu, Y.H., Liang, P.H. and Wu, Y.C. (2024) State Space Models Based Efficient Long Documents Classification. *Journal of Intelligent Learning Systems and Applications*, **16**, 143-154.
https://doi.org/10.4236/jilsa.2024.163009

Received: May 6, 2024
Accepted: June 16, 2024
Published: June 19, 2024

Abstract

Large language models like Generative Pretrained Transformer (GPT) have significantly advanced natural language processing (NLP) in recent times. They have excelled in tasks such as language translation question answering and text generation. However, their effectiveness is limited by the quadratic training complexity of Transformer models $O(L^2)$, which makes it challenging to handle complex tasks like classifying long documents. To overcome this challenge researchers have explored architectures and techniques such as sparse attention mechanisms, hierarchical processing and efficient attention modules. A recent innovation called Mamba based on a state space model approach offers inference speed and scalability in sequence length due to its unique selection mechanism. By incorporating this selection mechanism Mamba allows for context reasoning and targeted focus on particular inputs thereby reducing computational costs and enhancing performance. Despite its advantages, the application of Mamba in long document classification has not been thoroughly investigated. This study aims to fill this gap by developing a Mamba-based model, for long document classification and assessing its efficacy across four datasets; Hyperpartisan, 20 Newsgroups, EURLEX and CMU Book Summary. Our study reveals that the Mamba model surpasses NLP models such as BERT and Longformer showcasing exceptional performance and highlighting Mamba's efficiency in handling lengthy document classification tasks. These results hold implications for NLP applications empowering advanced language models to address challenging tasks with extended sequences and enhanced effectiveness. This study opens doors for the exploration of Mamba's abilities and its potential utilization, across diverse NLP domains.

Keywords

Mamba, Transformer, NLP

1. Introduction

Long document classification task is always a hot topic in the natural language processing area. The recent breakthrough of NLP starts with the transformer model [1]. It relies on a self-attention mechanism, which can compute the representation of a sequence from different words within that sequence instead of using convolution layer or sequence aligning used in RNN [2]. Transformer model removed encoder-decoder configurations in transduction, and improved performance. Transformer models have been widely used in many NLP tasks, especially text classification. Even though Transformer has better performance and efficiency compared to previous models, it still has a expensive time training complexity of $O(L^2)$, that's mainly because of its self-attention mechanism, where each token in the input sequence interacts with every other token, this is done by a dot product computation of queries and keys in Transformers. This expensive quadratic time complexity reflects the all-to-all comparison of tokens which provides the transformer its powerful contextual understanding but also contributes to higher computational demands, especially for long sequences.

As the advancements in language models progress, effectively managing longer sequences has become increasingly important. Researchers have proposed many methods to address this challenge, including truncating documents to the first 512 tokens [3], segmenting documents into smaller parts [4], and introducing sparse self-attention mechanisms [5]. These methods have expanded the capacity of transformer-based models to handle sequences up to 4096 tokens. However, this is still insufficient for sophisticated large language models. Recent breakthroughs like the Mamba model have presented solutions. Mamba [6], a state space sequence model architecture enhances the simulation of sequence models introduced by Gu *et al.* (2021) [7] by utilizing fixed learned matrices for computing state models thus improving both training and inference speeds.

Mamba takes a significant step forward by introducing a selection mechanism that enables context-dependent reasoning. This allows the model to selectively focus on specific inputs, simplifying its architecture and reducing parameters and computational costs while maintaining strong performance. Notably, Mamba achieves linear scaling in sequence length, making it an attractive solution for long document classification tasks. With its innovative architecture and efficient performance, Mamba offers a promising direction for future research in NLP, enabling large language models to tackle complex tasks like long document summarization with greater ease and efficiency.

This study addresses a significant research gap by conducting a thorough investigation into the performance of the newly proposed Mamba model on long document classification tasks. Despite its promising architecture, Mamba's capabilities in handling lengthy texts remain largely unexplored. To bridge this knowledge gap, we design a comprehensive experimental framework that pits Mamba against a range of transformer-based models, including the widely used BERT [3] as a baseline. Our evaluation protocol encompasses multiple long document da-

tasets, allowing us to assess the efficiency and effectiveness of Mamba across diverse problem domains. By examining various metrics, including classification accuracy, inference time, and memory usage, we provide a nuanced understanding of Mamba's strengths and weaknesses in tackling long document classification tasks, shedding light on its potential applications and areas for future improvement.

This paper follows the experimental setup of Park *et al.* (2022) [8] and selects four datasets for evaluation, each chosen for its unique characteristics and challenges. The Hyperpartisan dataset (Kiesel *et al.* [9]) presents a binary classification task, where news articles are labeled as either hyperpartisan or not. In contrast, the 20 Newsgroups dataset (Lang, 1995) [10] offers a multi-class classification challenge, with 20 distinct categories of news articles. The EURLEX dataset (Chalkidis *et al.*, 2019) [11] introduces a multi-label classification task, where legal documents are annotated with multiple labels. Finally, the CMU Book Summary dataset (Bamman and Smith) [12] provides a summarization task, where book summaries are categorized into different genres. By using these four datasets, our study ensures a comprehensive evaluation of the Mamba model's performance across various classification tasks, including binary, multi-class, and multi-label classification. This diverse range of datasets allows us to assess the model's versatility, robustness, and ability to generalize across different problem domains.

2. Background

The quest for efficient and effective long document classification has been a longstanding challenge in Natural Language Processing (NLP). Initially, traditional NLP methods, such as Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) [13], were commonly used for document classification. These methods represent texts as vectors of word counts or weighted word frequencies, but they often struggle with long documents due to their inability to capture semantic meanings and contextual relationships between words.

The introduction of machine learning models like Support Vector Machines (SVM), decision trees, and random forests offered more sophisticated ways to classify documents. These methods could handle longer documents better by learning discriminative features. However, they still relied heavily on manual feature engineering and were limited in their ability to understand the nuances of language in extensive texts.

The emergence of deep learning brought significant changes with models that could learn representations from data directly. Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) [2] were among the first deep learning models adapted for NLP. They could process texts of varying lengths by maintaining a memory of previous inputs, making them suitable for long document classification. Convolutional Neural Networks (CNNs), traditionally used for image processing, were also adapted for NLP tasks [14]. They

could capture local and positionally invariant features in text, although they were less used for very long documents compared to RNNs and LSTMs.

The introduction of attention mechanisms, especially the Transformer architecture, marked a turning point. Transformers handle long-range dependencies well and do not require sequential data processing, unlike RNNs and LSTMs. This makes them exceptionally good at handling long documents. Models such as BERT (Bidirectional Encoder Representations from Transformers) and its variants like RoBERTa [15], XLNet [16], and GPT (Generative Pre-trained Transformer) [17] leverage Transformer architectures to understand and generate human-like text.

To specifically address the challenges of long document classification, some methods employ hierarchical processing strategies, Hierarchical Attention Networks (HANs) [18]. These networks use attention mechanisms at different levels (word and sentence level) to capture document structure more effectively, making them particularly useful for longer texts. Segmentation Strategies, Some approaches segment long documents into smaller chunks or paragraphs, process each segment with models like BERT, and then aggregate these representations to make a final classification [4].

Despite advancements, challenges persist. Processing very long documents in a single pass remains computationally expensive due to memory and computational constraints. Moreover, maintaining nuanced understanding of long texts is complex, especially when dealing with abstract concepts or detailed narratives. Ongoing research focuses on improving efficiency, understanding, and generalization across diverse long document types and domains.

3. Dataset and Method

3.1. Dataset

In this paper, we selected four datasets to conduct a comprehensive analysis of Mamba's performance and efficiency.

The Hyperpartisan dataset, introduced by Kiesel *et al.* [9], consists of 748 news articles labeled as either hyperpartisan or non-hyperpartisan. The articles are drawn from a range of online sources and cover a variety of topics, including politics, economics, and social issues. The dataset is notable for its balanced class distribution, with 374 articles in each class, and its inclusion of articles from both liberal and conservative sources.

The 20 Newsgroups dataset, created by Lang in 1995 [10], is a widely used benchmark for text classification tasks. It consists of approximately 20,000 newsgroup documents, evenly divided across 20 different categories, including politics, religion, and science. The documents are relatively short, with an average length of around 500 words, and are drawn from online newsgroups.

The EURLEX dataset, introduced by Chalkidis *et al.* [11] in 2019, is a collection of 66,015 legal documents from the European Union's legal database, EUR-Lex. The documents are annotated with multiple labels from a hierarchy of

126 categories, making it a challenging multi-label classification task. The dataset covers a range of legal topics, including trade, environment, and human rights, and includes documents in multiple languages. In this experiment, we also used inverted-EURLEX which shuffles the section of legal documents, making the dataset better determine model's performance on global feature extraction.

The CMU Book Summary dataset, created by Bamman and Smith [12], consists of 16,559 book summaries from the online book review platform, Goodreads. The summaries are categorized into 14 different genres, including fiction, non-fiction, mystery, and romance. The dataset is notable for its large size and diverse range of genres, making it a valuable resource for evaluating text classification models. This experiment also uses a new variant of CMU book summary proposed by Park *et al.* [8]—paired book summary, which combines two book summaries together to make the sample even longer.

3.2. Models

In this paper, we chose 6 models and methods to test their performance as a comparison of Mamba. These models are all state-of-the-art language models or variants of them, and they have been widely used in text classification tasks. We selected these models to evaluate their performance on long document classification tasks and to compare their results with Mamba. We aim to investigate whether Mamba can outperform these models and provide better results in terms of accuracy and efficiency.

BERT: BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that has been fine-tuned on a wide range of Natural Language Processing (NLP) tasks, achieving state-of-the-art results [3]. It utilizes a multi-layer bidirectional transformer encoder to generate contextualized representations of words in text. However, BERT has a limitation of 512 tokens in its input layer, which can be a challenge when dealing with long documents.

BERT + TextRank: BERT + TextRank is a variant of BERT that incorporates TextRank, a graph-based algorithm, to identify key sentences in a document [19]. It uses these sentences to fill the 512 tokens of a BERT model, rather than using the full document as input. This approach allows for more efficient processing of long documents and can improve performance on tasks such as text classification.

BERT + Random: BERT + Random is a simpler baseline that uses random sentences to fill the 512 tokens of a BERT model. While less efficient than BERT + TextRank, it can still provide a decent performance on some tasks.

Longformer: Longformer is a transformer-based architecture designed to handle long-range dependencies in input sequences [5]. It utilizes a combination of local and global attention mechanisms to process input sequences of up to 4096 tokens, making it particularly useful for tasks such as long document classification.

ToBERT: ToBERT is a variant of BERT that uses chunking to process documents of any length [4]. It breaks up long documents into smaller chunks and processes each chunk separately, allowing for more efficient processing and im-

proved performance on tasks such as text classification.

CogLTX: CogLTX is a method that uses two jointly trained BERT models to select key sentences from long documents for various tasks, including text classification [20]. It is designed to be efficient and effective for long document classification tasks and can handle documents of any length. CogLTX utilizes a combination of local and global attention mechanisms to select the most relevant sentences and can outperform traditional BERT models on some tasks.

Mamba: Mamba is a novel approach to sequence modeling that combines the strengths of transformer-based architectures and graph-based algorithms to efficiently process long documents and extract relevant information [6]. This approach is designed to address the limitations of traditional State Space Models (SSMs) [7] and Transformer models, which struggle with content-aware reasoning and selective attention. Mamba's architecture relies on making its parameters dependent on the input, allowing it to filter relevant and irrelevant information. This is achieved through the use of dynamic matrices B and C, which are adjusted based on the input sequence length and batch size. As a result, the model can selectively compress information and focus on specific parts of the input. To optimize performance on modern GPUs, Mamba's architecture incorporates a hardware-aware algorithm. This algorithm reduces the number of times the model needs to access the slower DRAM memory through kernel fusion, and avoids saving intermediate states through recomputation. This results in significant improvements in performance. The core of Mamba's architecture is the Selective SSM (S6) model, a variant of the traditional SSM. The S6 model utilizes the aforementioned techniques to selectively compress information and optimize performance. This allows the model to efficiently process long sequences and extract relevant information. The S6 model can be implemented as a block, similar to self-attention in a decoder block. Multiple Mamba blocks can be stacked to process long sequences, and the output of each block is used as input for the next block. This allows the model to capture long-range dependencies and extract relevant information from the input sequence. **Figure 1** shows a detailed view of a mamba block.

3.3. Experiment Configuration

During the training phase, we ran 30 epochs on every model for each dataset. For learning rate selection, we ran experiments with three learning rates: 0.005, 3e−5, and 5e−5, and chose the best performing rate for each model. A dropout rate of 0.1 was used consistently across all models, based on recommendations from previous research. Results were averaged over five different random seeds to ensure reliability.

3.4. Evaluation Metric

In our experiments, we evaluate the performance of our model on various classification tasks, including binary, multi-class, and multi-label datasets. For binary

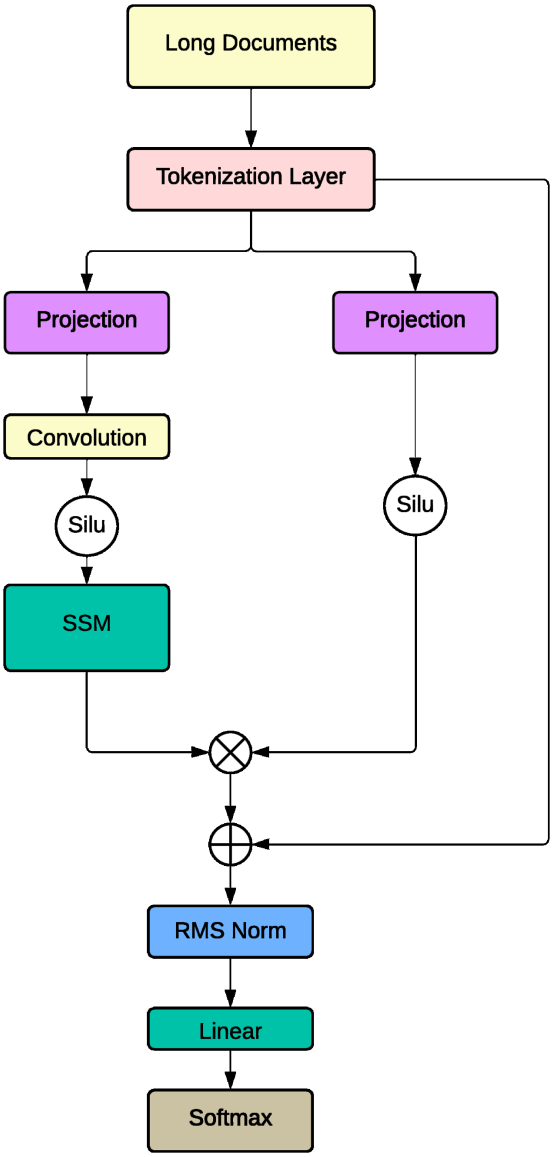

Figure 1. Overview of Mamba architecture.

and multi-class classification tasks, such as Hyperpartisan and 20 NewsGroups, we report the accuracy (%) on the test set, which is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where *TP*, *TN*, *FP*, and *FN* represent the number of true positives, true negatives, false positives, and false negatives, respectively.

For multi-label classification datasets, we use the micro-F1 score (%), which is a widely used metric for evaluating the performance of multi-label classification models. The micro-F1 score is calculated based on the true positives, false positives, and false negatives for each class, and is defined as:

$$\text{F1} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2}$$

The micro-F1 score is a harmonic mean of precision and recall, and provides a balanced measure of both. We report the micro-F1 score (%) on the test set for each multi-label classification dataset.

4. Result

The experimental results presented in Table 1 and Table 2 offer a detailed comparative analysis of the Mamba model's efficacy in long document classification tasks relative to established models like BERT, BERT + TextRank, BERT + Random, Longformer, ToBERT, and CogLTX. Mamba stands out not only for its superior performance across several datasets but also for its efficient resource management, despite higher GPU memory usage.

From Table 1, Mamba consistently achieves the highest scores across all datasets. For instance, in the Hyperpartisan dataset, Mamba records an accuracy of 96.71%, compared to its nearest competitor, Longformer, which scores 95.69%. Similarly, in the 20 News dataset, Mamba attains an impressive 93.16%, significantly outperforming ToBERT, the next best at 85.52%. This trend of outperformance is consistent across other datasets such as EURLEX and Inverted EURLEX, where Mamba scores 74.82% and 74.56%, respectively, surpassing other models by a considerable margin. The Paired Book Summary dataset particularly highlights

Table 1. Performance of different models on various datasets.

Model	Hyper-Partisan	20 News	EURLEX	Inverted EURLEX	Book Summary	Paired Summary
BERT	92.00	84.79	73.09	70.53	58.18	52.24
BERT + TextRank	91.15	84.99	72.87	71.30	58.94	55.99
BERT + Random	89.23	84.65	73.22	71.47	59.36	56.58
Longformer	95.69	83.39	54.53	56.47	56.53	57.76
ToBERT	89.54	85.52	67.57	67.31	58.16	57.08
CogLTX	94.77	84.63	70.13	70.80	58.27	55.91
Mamba	96.71	93.16	74.82	74.56	64.70	69.23

Table 2. Resource usage of different models.

Model	Train Time	Inference Time	GPU Memory
BERT	1.00	1.00	<16
BERT + TextRank	1.96	1.96	16
BERT + Random	1.98	2.00	16
Longformer	12.05	11.92	32
ToBERT	1.19	1.70	32
CogLTX	104.52	12.53	<16
Mamba	6.49	8.35	64

Mamba's strength, achieving 69.23%, which is markedly higher than the 57.76% of Longformer, the next closest in performance.

Table 2 reveals the computational efficiency and resource usage of the various models. While Mamba demands higher GPU memory at 64 units, compared to less than 16 for BERT and CogLTX, its training and inference times are relatively modest at 6.49 and 8.35 respectively. In contrast, CogLTX, despite having comparable memory usage to BERT, shows excessively high training times at 104.52. Longformer, despite its extended capacity to handle up to 4096 tokens, shows significantly higher training and inference times at 12.05 and 11.92 respectively, along with a substantial GPU memory usage of 32 units.

The efficient resource management of Mamba, despite its high memory requirements, is noteworthy. The model utilizes its resources to ensure faster training and inference times compared to Longformer and CogLTX, which, while effective in certain aspects, do not balance performance and resource usage as effectively as Mamba. This balance makes Mamba particularly attractive for real-world applications where both performance and computational efficiency are critical.

The architecture of Mamba allows for linear scaling with sequence length, a stark contrast to the quadratic scaling of traditional transformer models like BERT and Longformer. This efficiency is particularly advantageous for processing documents that exceed typical length limitations of other models, thus providing practical benefits for real-world applications where document length can be unpredictable. While BERT has been a benchmark model in NLP, its limitations with token length (typically capped at 512 tokens) restrict its use in long document classification without modifications like segmentation or summarization (e.g., BERT + TextRank). Mamba, with no such constraints, offers an inherent advantage in seamless processing of entire documents.

Although Longformer is designed to handle longer texts (up to 4096 tokens), it still lags behind Mamba in terms of both efficiency and performance. Mamba's selective attention mechanism provides a more refined approach to managing long-range dependencies than Longformer's broader attention span. However, Mamba's resource usage is higher in terms of GPU memory compared to BERT and even Longformer, as seen in Table 2. This aspect is crucial for scalability and deployment in resource-constrained environments.

The ability of Mamba to efficiently process long documents without the need for pre-segmentation opens new avenues for complex NLP tasks such as detailed document summarization, comprehensive content analysis, and extensive legal or scientific document processing. The effectiveness of Mamba in classifying varied types of documents, from news articles to legal texts and literary summaries, demonstrates its versatility and potential for integration into various professional and academic workflows.

Further research could explore ways to optimize Mamba's architecture to reduce its high resource demands without compromising performance. This could involve more efficient implementation of its selection mechanism or exploring

new methods of model compression. Extending the evaluation of Mamba to other forms of text-based analysis, such as sentiment analysis, event detection, or thematic categorization, could further validate its adaptability and utility across different domains.

5. Conclusions

This study presents a comprehensive analysis of the Mamba model's performance in long document classification, highlighting its superior capabilities and advancements over traditional and contemporary NLP models. Mamba, a state space model-based architecture, demonstrates a notable improvement in handling long sequences due to its innovative selection mechanism and efficient sequence processing. The experimental results across diverse datasets such as Hyperpartisan, 20 Newsgroups, EURLEX, and CMU Book Summary show that Mamba not only surpasses traditional transformer models like BERT and Longformer in terms of classification accuracy and F1 scores but also exhibits exceptional efficiency in training and inference times.

The remarkable performance of Mamba can be attributed to its unique approach to sequence modeling, which integrates context-dependent reasoning with a focus on relevant inputs, thereby reducing computational overhead and enhancing processing speed. This allows Mamba to achieve linear scaling in sequence length, making it particularly effective for complex NLP tasks that involve extensive texts.

Furthermore, the findings from this research suggest significant implications for the future of NLP applications, particularly in domains where the rapid and accurate classification of large volumes of text is critical. Mamba's ability to effectively manage and classify long documents paves the way for its adoption in various high-stake environments, including legal document analysis, comprehensive literature reviews, and extensive data categorization tasks.

By advancing our understanding of Mamba's capabilities, this study not only confirms its efficacy but also sets a new benchmark for future research in the NLP field. It encourages further exploration into the optimization of state space models and their potential applications across different areas of artificial intelligence and machine learning. The ongoing development of Mamba and similar models holds the promise of significantly expanding the horizons of what can be achieved with NLP technology, making it an exciting area for continued academic and practical innovation.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9

December 2017, 6000-6010.

[2] Sherstinsky, A. (2020) Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Physica D: Nonlinear Phenomena*, **404**, Article ID: 132306. https://doi.org/10.1016/j.physd.2019.132306

[3] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[4] Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y. and Dehak, N. (2019) Hierarchical Transformers for Long Document Classification. 2019 *IEEE Automatic Speech Recognition and Understanding Workshop* (*ASRU*), Singapore, 14-18 December 2019, 838-844. https://doi.org/10.1109/ASRU46091.2019.9003958

[5] Beltagy, I., Peters, M.E. and Cohan, A. (2020) Longformer: The Long-Document Transformer.

[6] Gu, A. and Dao, T. (2023) Mamba: Linear-Time Sequence Modeling with Selective State Spaces.

[7] Gu, A., Goel, K., *et al.* (2021) Efficiently Modeling Long Sequences with Structured State Spaces.

[8] Park, H.H., Vyas, Y. and Shah, K. (2022) Efficient Classification of Long Documents Using Transformers. *Proceedings of the* 60*th Annual Meeting of the Association for Computational Linguistics*, Volume 2, 702-709. https://doi.org/10.18653/v1/2022.acl-short.79

[9] Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B. and Potthast, M. (2019) SemEval-2019 Task 4: Hyperpartisan News Detection. *Proceedings of the* 13*th International Workshop on Semantic Evaluation*, Minneapolis, June 2019, 829-839. https://doi.org/10.18653/v1/S19-2145

[10] Lang, K. (1995) NewsWeeder: Learning to Filter Netnews. *Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, 9-12 July 1995, 331-339. https://doi.org/10.1016/B978-1-55860-377-6.50048-7

[11] Chalkidis, I., Fergadiotis, E., Malakasiotis, P. and Androutsopoulos, I. (2019) Large-Scale Multi-Label Text Classification on EU Legislation. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 6314-6322. https://doi.org/10.18653/v1/P19-1636

[12] Bamman, D. and Smith, N.A. (2013) New Alignment Methods for Discriminative Book Summarization.

[13] Liu, C.-Z., Sheng, Y.-X., Wei, Z.-Q. and Yang, Y.-Q. (2018) Research of Text Classification Based on Improved TF-IDF Algorithm. 2018 *IEEE International Conference of Intelligent Robotic and Control Engineering* (*IRCE*), Lanzhou, 24-27 August 2018, 218-222. https://doi.org/10.1109/IRCE.2018.8492945

[14] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[15] Liu, Y.H., Ott, M., Goyal, N., Du, J.F., Joshi, M., Chen, D.Q., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[16] Yang, Z.L., Dai, Z.H., Yang, Y.M., Carbonell, J., Salakhutdinov, R. and Le, Q.V. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding.

[17] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., *et al.* (2020)

Language Models Are Few-Shot Learners.

[18] Yang, Z.C., Yang, D.Y., Dyer, C., He, X.D., Smola, A. and Hovy, E. (2016) Hierarchical Attention Networks for Document Classification. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 1480-1489. https://doi.org/10.18653/v1/N16-1174

[19] Mihalcea, R. and Tarau, P. (2004) TextRank: Bringing Order into Text. *Proceedings of the* 2004 *Conference on Empirical Methods in Natural Language Processing*, Barcelona, 25-26 July 2004, 404-411.

[20] Ding, M., Zhou, C., Yang, H.X. and Tang, J. (2020) CogLTX: Applying BERT to Long Texts. *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*, Vancouver, 6-12 December 2020, 12792-12804.