An LCF-IDF Document Representation Model Applied to Long Document Classification

Renzo Alva Principe[1,2], **Nicola Chiarini**[2], **and Marco Viviani**[1]
[1] Università degli Studi di Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)
IKR3 LAB (https://ikr3.disco.unimib.it/)
Edificio U14 (ABACUS), Viale Sarca, 336 – 20126, Milan, Italy
[2] DATASINC (https://www.datasinc.it/)
Via Soperga, 20 – 20127 Milan, Italy
renzo.alvaprincipe@unimib.it, nicola.chiarini@datasinc.it, marco.viviani@unimib.it

Abstract

A document representation model that has been used for years in NLP and Text Mining tasks is TF-IDF (Term Frequency-Inverse Document Frequency). This model is indeed effective for various tasks like Information Retrieval and Document Classification. However, it may fall short when it comes to capturing the deeper semantic and contextual meaning of a text, which is where Transformer-based Pre-trained Language Models (PLMs) such as BERT have been gaining significant traction in recent years. Despite this, these models also face specific challenges related to Transformers and their attention mechanism limits, especially when dealing with long documents. Therefore, this paper proposes a novel approach to exploit the advantages of the TF-IDF representation while incorporating semantic context, by introducing a Latent Concept Frequency-Inverse Document Frequency (LCF-IDF) document representation model. Its effectiveness is tested with respect to the Long Document Classification task. The results obtained show promising performance of the proposed solution compared to TF-IDF and BERT-like representation models, including those specifically for long documents such as Longformer as well as those designed for particular domains, especially when it comes to Single Label Multi-Class (SLMC) classification.

Keywords: Text Representation, TF-IDF, BERT, Pre-Trained Language Models, Clustering, Classification.

1. Introduction

The fields of *Natural Language Processing* (NLP) and Text Mining have long relied on frequency-based methods for document representation, with the *Term Frequency-Inverse Document Frequency* (TF-IDF) model being a prominent and well-established choice (Sidorov, 2019). TF-IDF has demonstrated its effectiveness across various tasks, including Information Retrieval and Document Classification. Its simplicity and interpretability have made it a favored choice for many practical applications. However, it is not without limitations, particularly in capturing the intricate semantics and context of textual data, where more advanced models like word embeddings and Deep Learning models have displayed distinct advantages.

In recent years we have in fact witnessed the emergence and widespread adoption of newer approaches, notably exemplified by models like BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019), which have revolutionized NLP. These models leverage Transformers (Vaswani et al., 2017) and pre-training on massive corpora to capture rich contextual information within the text (Niu et al., 2021). However, as disruptive as these models are, they are not without challenges, especially when dealing with long documents. The computational demands, potential information loss, and difficulties in maintaining context over extended text can pose significant obstacles (Lin et al., 2022).

In light of these considerations, this paper introduces a novel approach to blend the traditional TF-IDF model with advanced Transformer-based models like BERT, aiming to combine TF-IDF's established strengths with BERT's contextual semantic understanding seamlessly. To do this, *Latent Concepts* (LCs) are extracted from the corpus of documents by relying on *Pre-trained Language Models* (PLMs) and clustering, substituted for the word tokens in the document representation, and a *Latent Concept Frequency - Inverse Document Frequency* (LCF-IDF) representation is obtained on top of them.

This LCF-IDF model thus obtained is instantiated and evaluated primarily in the context of Long Document Classification. The results of our study indicate that the proposed solution effectively combines the strengths of TF-IDF and advanced Deep Learning models, achieving results comparable to or even surpassing traditional TF-IDF and BERT-like representation models, including those specifically for long documents such as Longformer as well as those designed for particular domains, in particular in the case of *Single Label Multi-Class* (SLMC) classification. This research paves the way for more nuanced and context-aware document representations, which are crucial for achieving high

LREC-COLING 2024, pages 1129–1135
20-25 May, 2024.

accuracy in complex NLP tasks, particularly when handling extensive textual data.

In the sections that follow, we delve into the related work (Section 2), methodology (Section 3), experiments and findings (Section 4), and conclusions and further research directions (Section 5).

2. Related Work

As previously introduced, the challenge with TF-IDF is related to both the high numerosity of terms and the intricate nuances of language and context over extended content, which TF-IDF struggles to manage. First advancements in the literature like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), enabled a more profound understanding of language by capturing semantic associations between words. Even more effective in this sense nowadays, BERT employs a bidirectional contextual model, making it a state-of-the-art choice for a wide range of NLP tasks.

However, the primary issue with Transformer-based models is the substantial computational demand associated with processing extended texts, given their inherent attention mechanisms (Tay et al., 2022). Researchers have explored various solutions, notably *efficient Transformers* (Beltagy et al., 2020; Zhang et al., 2021), which propose an alternative attention mechanism capable of managing longer text by reducing the quadratic computational complexity of self-attention. Additionally, several strategies involving *hierarchical processing* (Grail et al., 2021; Pappagari et al., 2019; Zhang et al., 2019) and *key-content extraction* (Ding et al., 2020) have been investigated. Even these solutions are not without their drawbacks. These include potential performance loss due to self-attention approximation, long dependencies loss due to chunking, increased model complexity (Dong et al., 2023; Park et al., 2022).

When considering Long Document Classification – i.e., the task that we employ for both instantiating and evaluating the proposed model – some of the above-mentioned techniques have been employed in the literature. Among the main solutions, (Beltagy et al., 2020; Qin et al., 2022; Zaheer et al., 2020) have applied to classification a mixture of alternative pattern attentions, such as random, dilated, global, and clustering-based. Others, such as (Hu et al., 2022; Pappagari et al., 2019; Wu et al., 2021) applied Transformers in a hierarchical manner, proposing different methods for enhancing the flow of information across document segments. Finally, other approaches use extractive summarization to fit the Transformer model's input length limits (Ding et al., 2020; Park et al., 2022).

3. The LCF-IDF Representation Model

In this section, we provide an in-depth description of the LCF-IDF (*Latent Concept Frequency - Inverse Document Frequency*) document representation model proposed in this paper, which aims to combine the advantages of TF-IDF and BERT-like models when employed to perform NLP and Text Mining tasks. In general, the purpose of this model is to apply dimensionality reduction to documents at the word token level by identifying *Latent Concepts* (LCs) prior to any document embedding representation (as opposed to, for example, techniques such as BERTopic (Grootendorst, 2022)). In particular, the architecture of LCF-IDF is made up of three modules: (i) the module for constructing a *Latent Concept Space* (LCS); (ii) the module for translating word token-based documents into LC-based documents; and (iii) the module for generating the LCF-IDF representation of documents.

3.1. Building the Latent Concept Space

To build the LCS, it is necessary to learn a function that maps the word tokens in the corpus into the Latent Concepts. Specifically, for each document d_i in the corpus D, we consider each word token $w_{ij} \in V$ and obtain a *contextualized word embedding* $\mathbf{w_{ij}} \in \mathbf{V}$, by means of a *Pre-trained Language Model* (PLM). This way, when w_{ij} and w_{kl} are the same term, either in the same document or in different documents, their corresponding $\mathbf{w_{ij}}$ and $\mathbf{w_{kl}}$ representations differ when their contextual usage varies. For efficiency purposes, each high-dimensional vector $\mathbf{w_{ij}}$ is transformed into a lower-dimensional embedding $\mathbf{w'_{ij}} \in \mathbf{V'}$. Finally, to extract the LCs, a mapping function $\Psi : \mathbf{V'} \rightarrow \mathbf{C}$ is learned, where $\mathbf{C}$ is the set of LC embeddings and its cardinality is a hyperparameter of the model employed to learn Ψ. Note that $|\mathbf{C}| \ll |\mathbf{V'}|$.

The construction of the LCS is performed by the module illustrated in Figure 1 (a). In this paper, we employed Longformer, BERT, RoBERTa, and LegalBERT as PLMs, an *autoencoder* (Rumelhart et al., 1986) to perform dimensionality reduction on word token embeddings length, and the *k-means* clustering algorithm (MacQueen et al., 1967) to learn the mapping function Ψ. At this point, in the LCS, each concept is formally represented by the *centroid* $\mathbf{c_i}$ of the cluster representing a distinct concept.

3.2. Substituting Word Tokens with LCs in the Documents

In this second step, which is carried out by the module shown in Figure 1 (b), each "word token-based document" in the corpus is translated into a new "LC-based document". Specifically, the word tokens

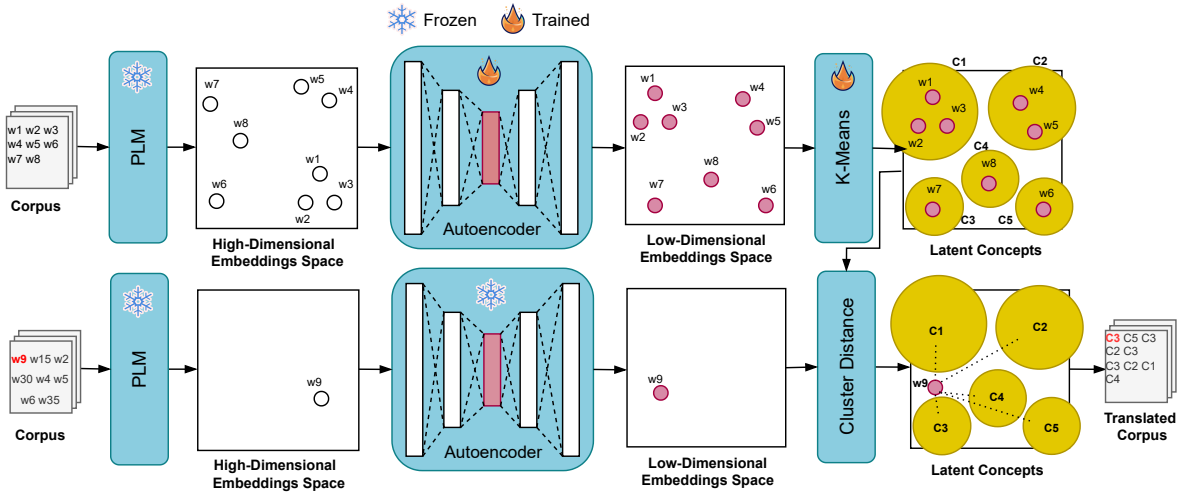

Figure 1: Latent Concept Space generation (a) and word token substitution with LCs (b).

in the document are substituted for the centroid of the cluster of the concept to which the word token is closest.

3.3. LCF-IDF Vectorization

The last step, which concerns the actual construction of the document representation using LCF-IDF, is carried out by the module shown in Figure 2. Each document consisting of LCs is passed to the LCF-IDF Vectorizer and then can be used against any NLP/Text Mining task.

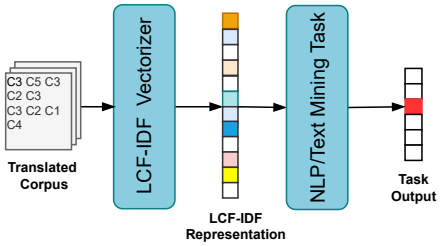

Figure 2: The LCF-IDF document vectorization.

The LCF-IDF Vectorizer is a modification of the familiar TF-IDF document representation model. Rather than determining the significance of a term within a document compared to the document collection, it assesses the importance of a Latent Concept within a document relative to the entire document collection. In practice, it considers the frequency of a Latent Concept in a document compared to its inverse frequency in the collection.

4. Experimental Evaluation

This section discusses experimental evaluations related to the proposed LCF-IDF model. In particular, the task of *Long Document Classification* is considered over a set of publicly available datasets, which were selected either based on their popularity or the fact that they actually consisted of long documents in distinct domains.

4.1. Datasets

In our experiments, we chose datasets that were developed for both *Single Label Multi Class* (SLMC) and *Multi Label Multi Class* (MLMC) classification. Due to computational limitations, we employed scaled-down versions (denoted with μ) of the original ones. In particular, we applied *stratified sampling* to SLMC datasets, while *random sampling* to the MLMC dataset. The characteristics of the datasets used are outlined below.

- **μ-20NG.** An SLMC news dataset extracted from 20NewsGroups (Lang, 1995). It contains 2,037 documents in the training set, 1,132 in the validation set, and 1,130 in the test set;
- **μ-A-512.** An SLMC review dataset extracted from Amazon-512 (Li et al., 2023), where documents are long at least 512-tokens. It contains 1,900 documents in the training set, 634 in the validation set, and 633 in the test set;
- **μ-ECtHR.** An MLMC legal dataset derived from ECtHR (Task B) (Chalkidis et al., 2021). It contains 1,800 documents in the training set and 1,000 in both the validation and the test set.

Dataset	# Classes	# Docs	# Tokens Avg.	% Docs >512 tok.	% Original dataset
μ-20NG	20	4299	548	26.12%	22.8%
μ-A-512	5	3167	861	100%	4%
μ-ECtHR	10	3800	2200	84%	34.5%

Table 1: Dataset statistics.

Table 1 summarizes dataset statistics. Two out of three datasets have a conspicuous average number of tokens as well as a percentage of documents exceeding 512 tokens (i.e., the limit of the input sequence of BERT base). 20NewsGropus is included because it is widely used in the literature on Text Classification despite being mostly composed of short documents.

Dataset	AVG #concepts/doc (min)	AVG #concepts/doc (max)	Shared concepts (min)	Shared concepts (max)
μ-20NG	12	22	66	82
μ-A-512	61	85	169	179
μ-ECtHR	57	88	146	159

Table 2: Statistics on concepts across classes on the considered datasets.

Table 2 provides additional statistics regarding the frequency of concepts within each document (min/max), relative to its class membership, as well as the extent of concept sharing among classes (min/max). Examining shared concepts across classes reveals that it is common for a class to share numerous concepts with the rest of the dataset, except for μ-20NG, which tends to possess more unique concepts for each class.

4.2. Technical Details

We implemented the LCF-IDF model in *Pytorch*, on a server with an RTX 3090 GPU (24GB), 125 GB of RAM, and 36 cores. The *autoencoder* consists of six layers with the following number of neurons: 768, 600, 500, 100, 500, 600, 768. *K-means* generates 200 clusters on all the experiments and uses the default distance function (Euclidean). The TF-IDF *vectorizer* uses 100k terms as vocabulary maximum size and tokenizes the input text in unigrams, bi-grams, and tri-grams. Concerning *hyperparameter tuning* during training, we used a batch of size 8 (2 for the Longformer baseline), a number of input tokens equals to 512 (4,096 for models that include Longformer), a learning rate of $3e-05$, and Adam (Kingma and Ba, 2014) as an optimizer. As for the implementation of the baselines and LCF-IDF classifiers, illustrated in the next section, we used a *single-layer Feedforward Neural Network* (FNN) classifier, with *Cross Entropy* (CE) loss for SLMC datasets, and *Binary Cross Entropy* (BCE) loss for MLMC datasets. Concerning the processing of documents exceeding the PLM token input size (e.g, 512 for BERT), documents are split into non-overlapping chunks of 510 tokens each. CLS and SEP tokens are added at the start and end of each chunk. For instance, a 700-token document is divided into two chunks: C1 (510 tokens + 2) and C2 (190 tokens + 2). BERT processes each chunk separately to generate contextual word embeddings. Finally, we used the `sklearn` TF-IDF implementation for the LCF-IDF Vectorizer, hence computing scores as follows: LCF-IDF$(c,d) =$ LCF$(c,d) *$ IDF(c), where LCF-IDF(c,d) is the frequency of a concept c in a document d, IDF$(c) = \log\left[\frac{1+n}{1+\text{DF}(c)}\right] + 1$, n the number of documents in the collection, and DF(c) the number of documents containing c.

4.3. Results

The results obtained from the application of LCF-IDF to the Long Document Classification task are illustrated in this section. We show them with respect to a series of baseline classifiers and a number of PLM variants for the proposed LCF-IDF model.

- **TF-IDF.** The FNN classifier using the simple TF-IDF representation;
- **BERT$_{\text{FT}}$.** The FNN classifier using the BERT representation (Devlin et al., 2019), *fine-tuned* (FT) w.r.t. the downstream task and dataset;
- **RoBERTa$_{\text{FT}}$.** The FT (as above) FNN classifier using the RoBERTa representation (Liu et al., 2019);
- **LegalBERT$_{\text{FT}}$.** The FT (as above) FNN classifier using the LegalBERT representation (Chalkidis et al., 2020);
- **Longformer$_{\text{FT}}$.** The FT (as above) FNN classifier using the Longformer representation (Beltagy et al., 2020);
- **PLMs + LCF-IDF.** These FNN classifier variants make use of Longformer, BERT, RoBERTa, and LegalBERT in the LCF-IDF model without any training beyond the original pre-training. They are denoted as LCF-IDF$_{\text{Longformer}}$, LCF-IDF$_{\text{BERT}}$, LCF-IDF$_{\text{RoBERTa}}$, and LCF-IDF$_{\text{LegalBERT}}$;
- **FT PLMs + LCF-IDF**. Here, instead of using PLMs "as is", we use their FT versions. In this sense, LCF-IDF$_{\chi-\text{FT}}$ refers to the FNN classifier employing the LCF-IDF model that uses the PLM χ preemptively fine-tuned.

Table 3 illustrates the obtained results in terms of *macro*-F1 (m-F1) and *weighted*-F1 (w-F1).

4.4. Discussion

First and foremost, it can be observed that each istance of the LCF-IDF$_{\chi-\text{FT}}$ classifier achieves comparable results or outperformes the baseline models depending on the considered dataset, at least with respect to an evaluation metric.

Model	μ-20NG m-F1	μ-20NG w-F1	μ-A-512 m-F1	μ-A-512 w-F1	μ-ECtHR m-F1	μ-ECtHR w-F1
TF-IDF	64.8	65.2	18.8	38.5	37.6	56.6
$\text{Longformer}_{\text{FT}}$	50.9	51.7	12.6	29.1	**59.7**	70.5
$\text{LCF-IDF}_{\text{Longformer}}$	44.1	44.4	17.4	36.7	33.4	51.0
$\text{LCF-IDF}_{\text{Longformer}-\text{FT}}$	71.2	52.0	14.4	29.7	57.6	**71.5**
BERT_{FT}	**77.8**	**78.8**	26.1	41.7	51.6	62.7
$\text{LCF-IDF}_{\text{BERT}}$	57.0	57.5	18.7	38.8	37.3	55.5
$\text{LCF-IDF}_{\text{BERT}-\text{FT}}$	77.3	78.4	34.6	49.6	47.3	65.6
$\text{RoBERTa}_{\text{FT}}$	76.2	77.1	29.5	45.0	51.8	62.9
$\text{LCF-IDF}_{\text{RoBERTa}}$	43.0	43.3	17.3	36.5	32.4	49.9
$\text{LCF-IDF}_{\text{RoBERTa}-\text{FT}}$	77.0	78.2	**41.1**	**53.6**	54.5	70.3
$\text{LegalBERT}_{\text{FT}}$	-	-	-	-	50.6	65.3
$\text{LCF-IDF}_{\text{LegalBERT}}$	-	-	-	-	37.9	56.0
$\text{LCF-IDF}_{\text{LegalBERT}-\text{FT}}$	-	-	-	-	57.0	71.4

Table 3: Performance of the classifiers considered. Best performance per dataset in bold. Best performance for each baseline/LCF-IDF group is underlined. LegalBERT is used only on legal-domain data.

More specifically, when considering basic models such as TF-IDF and a BERT-like model (excluding Longformer), we observe that incorporating semantic contextualization into the TF-IDF representation yields superior performance compared to traditional TF-IDF alone and generally outperforms BERT-like models (except on data from the 20NewsGroups dataset, which, however, does not actually consist of long documents). When considering specialized models like Longformer tailored for longer documents, utilizing LCF-IDF can still enhance performance in most cases, suggesting that significant content relevant to the classification task may extend beyond 4,096 tokens.

In general, it is possible to observe how the proposed model works well for both generic domains and specific domains (i.e., the legal domain) using the appropriate linguistic resources. Not surprisingly, we can observe how the fact of not fine-tuning the PLM with respect to the downstream task and dataset leads to results that are not satisfactory.

5. Conclusions and Further Research

In our article, we introduced LCF-IDF, a novel model crafted to harness the combined strengths of TF-IDF and BERT-like models, elevating the efficacy of NLP and Text Mining tasks in textual document representation. Specifically tailored to tackle challenges encountered with long documents by both aforementioned models, LCF-IDF was instantiated and validated through the Long Document Classification task, yielding promising outcomes.

There are significant avenues for further research on the proposed model from various perspectives. First, we aim to enhance the efficiency of the clustering phase in the future by pinpointing the optimal embeddings to incorporate into the vector space. Furthermore, we intend to conduct experiments anew with optimized parameters across all datasets, including those from diverse domains. It will also be necessary to investigate aspects related to the tokenization process and how it affects the generation of Latent Concepts, particularly for specific domains. Moreover, we aspire to evaluate this representation model for other NLP tasks involving long documents, such as Information Retrieval. Finally, we count on investigating Explainable AI techniques to provide more details about the reasons for the effectiveness of the model.

Data and Code Availablity

Code, data, requirements and execution scripts for reproducibility are available at the following link: https://github.com/rAlvaPrincipe/lcf-idf.

Acknowledgments

This work was carried out as part of a Doctoral Program in Higher Apprenticeship (*Dottorato in Alto Apprendistato*) and partly supported by Datasinc (https://www.datasinc.it/) and Regione Lombardia (https://www.regione.lombardia.it/) under the Project: "Natural Language Processing and Computer Vision applied to information extraction, classification, linking, and retrieval in legaltech, fintech, proptech, and insurtech." The project is identified with Project ID 3874245 and CUP: H45E22000940002.

We also acknowledge Marco Braga and Alessandro Raganato for their insights about the proposed model, which greatly contributed to its refinement.

6. Bibliographical References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. In *Advances in Neural Information Processing Systems*, volume 33, pages 12792–12804. Curran Associates, Inc.

Zican Dong, Tianyi Tang, Lunyi Li, and Wayne Xin Zhao. 2023. A survey on long text modeling with transformers.

Quentin Grail, Julien Perez, and Eric Gaussier. 2021. Globalizing BERT-based transformer architectures for long document summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1792–1810, Online. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Yongli Hu, Wen Ding, Tengfei Liu, Junbin Gao, Yanfeng Sun, and Baocai Yin. 2022. Hierarchical multiple granularity attention network for long document classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Hyunji Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 702–709, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ruyu Qin, Min Huang, Jiawei Liu, and Qinghai Miao. 2022. Hybrid attention-based transformer for long-range document classification. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA.

Grigori Sidorov. 2019. Vector space model for texts and the tf-idf measure. *Syntactic n-grams in Computational Linguistics*, pages 11–15.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. pages 848–853.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Poolingformer: Long document modeling with pooling attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12437–12446. PMLR.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

7. Language Resource References

Chalkidis, Ilias and Fergadiotis, Manos and Tsarapatsanis, Dimitrios and Aletras, Nikolaos and Androutsopoulos, Ion and Malakasiotis, Prodromos. 2021. *Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases*. Association for Computational Linguistics.

Lang, Ken. 1995. *NewsWeeder: Learning to Filter Netnews*. Morgan Kaufmann Publishers Inc., ICML'95.

Li, Irene and Feng, Aosong and Radev, Dragomir and Ying, Rex. 2023. *HiPool: Modeling Long Documents Using Graph Neural Networks*. Association for Computational Linguistics.