BERT-based Models for Arabic Long Document Classification*

Muhammad AL-Qurishi*, Riad Souissi

Elm Company, Research Department, Riyadh 12382, Saudi Arabia

Abstract
Given the number of Arabic speakers worldwide and the notably large amount of content in the web today in some fields such as law, medicine, or even news, documents of considerable length are produced regularly. Classifying those documents using traditional learning models is often impractical since extended length of the documents increases computational requirements to an unsustainable level. Thus, it is necessary to customize these models specifically for long textual documents. In this paper we propose two simple but effective models to classify long length Arabic documents. We also fine-tune two different models-namely, Longformer and RoBERT, for the same task and compare their results to our models. Both of our models outperform the Longformer and RoBERT in this task over two different datasets.

Keywords
Arabic Text Processing, Long Document Classification, BERT-based Models, Sentence Segmentation

1. Introduction

A large portion of textual content that requires automated processing is in the form of long documents. In some domains such as legal or medical, long documents are the standard. This severely restricts the possibilities for practical use of the most advanced Transformer models for text classification and other linguistic tasks [1]. For example, models such as BERT [2] have significantly improved the accuracy of automated NLP tasks, but their usefulness is limited to relatively short text sequences [3] due to the fact that their complexity increases geometrically. Modifying BERT in such a way to disassociate sequence length from computing complexity would remove this obstacle and bring immediate benefits to numerous fields such as education, science, and business [4]. Innovative approaches that leverage the greatest advantages of Transformers while offsetting their major shortcomings are needed at this stage of development, as they could lead to full maturation of a concept that has been demonstrated to be impressively successful with semantic tasks.

There have been numerous attempts to improve the performance and efficiency of BERT with long documents, using a wide variety of approaches. Some of

the proposed solutions are based on the sliding window paradigm [5, 6]. The downside of this class of solutions is their inability to track long-range dependencies in the text which weakens their analytic insights. Another group of works aim to simplify the architecture of Transformers and decrease complexity as result [7, 8, 9]. So far, none of these attempts could match the same level of performance that BERT achieves with short text. Reusing previously completed steps is another strategy for adapting Transformers for longer text [10] as a prominent example. Longformer model proposed by [11] may be the most promising solution for the problem of using Transformers with long text, and it combines local and global attention to improve efficiency. The issue remains open, and new suggestions for the best method of long document processing are still being made on a regular basis.

In this paper we present two BERT-based language models and fine-tune two others for Arabic long document classification. The first language model consists of four main layers: sentence segmentation layer, BERT layer, a linear classification layer, then the sentence grouping layer with respect to each document, and finally the softmax layer. In this model, we segmented the document into meaningful sentences and then fed these sentences into BERT model along with their document ID. The second model has the same idea of dividing the document into sentences, but instead we hypothesize that a majority of semantically important information is concentrated within specific sentences inside of a longer text, making it unnecessary to check for connections between all words in a document. Instead, we used BERT-based similarity match algorithm that can recognize high-relevance sentences and pass them as input to the BERT-base model that can complete the desired

classification task. Both of those models are based on BERT architecture, and require supervised training for best performance. Input text is divided into sentences that don't exceed the maximum length that BERT can accurately process (512 tokens).

In addition, we have fine-tuned two well-known language models for long documents classification task, which are the Longformer [11] and Recurrent over BERT [6][1]. Before the fintuning process, these two models have been modified to be suitable for the Arabic language. We compared the proposed models against the Longfroemer and RoBERT using two different Arabic datasets. The proposed language models were evaluated against two models, Longfarmer and RoBERT, using two datasets. The first dataset was collected from the Mawdoo3 website[2] and the second dataset was from previous related work [12]. The results showed that the first language model based on sentences aggregating after classifying them is the best among all models on news data with a macro F1-score equal to 98%. While this model achieved a comparative result with the Longformer in the second Mawdoo3 dataset that contains 22 classes.

2. Related Works

Most of the recent works addressing the problem of long document classification start from similar principles common to all deep learning methods. They also diverge in many aspects, as the authors explore different avenues for leveraging the power of the learning algorithms and overcoming the most significant obstacles [13]. Since the authors are essentially attempting to solve the same problem, namely how to maintain high accuracy of semantic predictions while keeping the computing demands reasonable, it would be fair to describe the papers as belonging to the same family despite the considerable differences in approach.

In terms of methodological choices, practically all works from this group acknowledge the unmatched power of the attention mechanism for analyzing semantic relationships, and incorporate it in some way into the proposed architecture. There is a division between works that mostly (or completely) embrace an existing architecture and perform only minor operations such as fine-tuning or knowledge transfer in order to reduce the computational demands [14, 15]. On a different end of the spectrum, there are works that propose innovative hybrid solutions in which the attention mechanism and/or Transformer architecture are combined with elements of different deep learning paradigms, such as RNNs and CNNs. In particular, a common strategy is to adopt a hierarchical structure for the overall solution and use the

attention mechanism only in a limited role, thus avoiding the exponential growth of complexity [16, 17].

The aforementioned methodological differences stem largely from the expectations for each paper, which range from proving a theoretical point to attempting to develop specialized model for long document classification. Works with a narrower scope tend to stay closer to the original BERT model design [11], while more ambitious efforts that aim to create new tools are more inclined to experiment with previously untested combinations of elements. In some papers, the scope of intended applications is limited to long documents from a certain domain (i.e. medical) [17], while others are approaching the problem in more general terms. Finally, there is an important distinction between works that aim for greater accuracy, and those that primarily attempt to improve computational efficiency and shorten the inference time [18].

It's a fair assessment that practically all works from this group are grappling with the same problem – the tendency of attention-based models to become prohibitively complex as the length of the analyzed text is increased. In response, the authors tried a variety of ideas that rely on vastly different mechanisms to decrease complexity. From fine-tuning and knowledge distillation to introduction of hierarchical architectures and restrictive elements such as fixed-length sliding window [11, 19], the proposed techniques are quite innovative and typically leverage some known properties of deep learning models to affect how the attention mechanism performs in a particular deployment. The diversity of ideas found in those papers illustrates that researchers are currently casting a wide net and searching for unconventional answers to a difficult problem, without a single dominant strategy. On the other hand, hybrid approaches hold a lot of promise and they combine some proven elements from different methodologies into new, potentially more optimal configurations [16, 20].

Evaluation of the proposed changes to established algorithms is crucially important, and all of the reviewed works include some form of empirical confirmation of their premises. While the numbers seemingly validate that the proposed solutions achieve state-of-the-art results under the best possible conditions, those findings are self-reported and may often be too optimistic. All of the papers are interested in document classification tasks and use it to evaluate their solutions, but datasets used for testing may not be the same in terms of size, diversity, and content. When directly comparing different solutions, it's extremely important to keep in mind the particulars of the evaluation protocols. Studies aiming to provide evaluations with independently administered comparative testing of several different BERT-like algorithms for document classification are slowly emerging and reporting some interesting findings that often di-
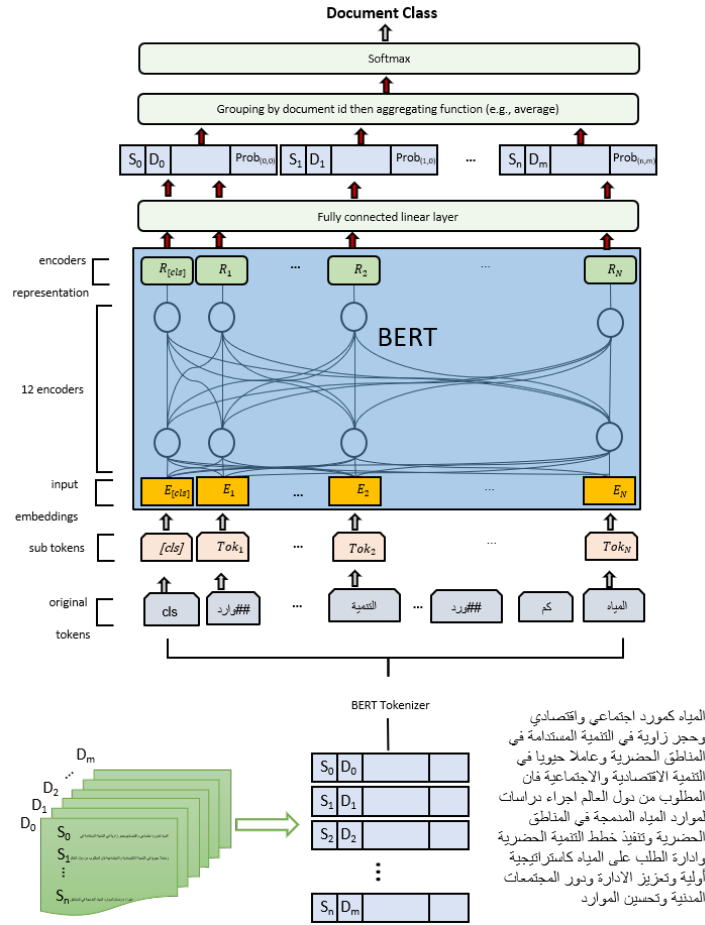
Figure 1: Proposed Model Architecture for Long Document Classification

verge from self-assessed results [4, 13, 18]. Still, there are no widely accepted evaluation standards and every comparison suffers from 'apples-to-oranges' problem up to an extent.

When it comes to practical use of the proposed solutions, there is a general lack of field data and even discussions of use cases are rare. This is understandable considering the main focus is on discovering more efficient methods, but without real world testing it's difficult to predict whether any of the solutions can deliver similar results to their reported findings. Some works may be directed as specific niches such as legal or medical, but even in this case little attention is paid to practicalities associated with real world application. This weakness may reflect the current state of the field, which is highly experimental and mostly built on data collected in a controlled environment.

3. Data

Experimental parts of our study are conducted using two different datasets, with the choice of the datasets based on the domain of research which is long length Arabic documents. The datasets are vastly different in terms of size and diversity of classes.

3.1. Mawdoo3 Dataset

The first dataset was scraped from Mawdoo3 which is the largest Arabic content website[3]. The number of classes from mawdoo3 is 22 class and each category contains between 700 to 12K articles. We have selected almost one thousand of long articles from each category as presented in Figure 2.

---

[3]https://mawdoo3.com/

Article category	Number of articles	Category in English
إسلام	996	Islamic content
الآداب	1000	Literature
الأسرة	1000	Family
الحمل والولادة	1000	Pregnancy and Childbirth
الحياة والمجتمع	999	Life and Society
الزواج والحب	865	Marriage and Love
العناية بالذات	1000	Self Care
تسلية وألعاب	1000	Fun and Games
تعليم	1000	Education
تغذية	1000	Food and Nutrition
تقنية	1000	Technology
حكم وأقوال	992	Proverbs and Sayings
حول العالم	1000	Worldwide
حيوانات ونباتات	1000	Animals and Plants
رياضة	1000	Sports
سؤال و جواب	1000	Question and Answer
صحة	1000	Health
علوم الأرض	1000	Geology
فن الطهي	1000	Gastronomy
فنون	780	Arts
مال وأعمال	1000	Business and Finance
وزن ورشاقة	1000	Fitness

Figure 2: Mawdo3 Dataset that contains 22 class.We selected almost 1000 article under each class.

Article category	Number of articles	Category in English
تقنية	2728	Technology
ثقافة	3846	Culture
دينية	3860	Religion
أخبار محلية	4000	Local
الشرق الأوسط	4000	MiddleEast
إقتصار	4000	Economy
رياضة	4000	Sport
أخبار دولية	4000	International

Figure 3: Arabic News Dataset we choose almost 4000 articles from each category.

3.2. Arabic News Dataset

The second dataset was about news articles and we downloaded them from different sources [21, 12, 22]. These data have almost the same 8 categories so we merged them together and the resulted dataset is described in Figure 3. We have selected almost four thousands of long articles from each class.

4. Models

In this section we introduce two BERT-based language models. Both of those models are based on BERT architecture, and require supervised training for best perfor-

mance. Input text is divided into sentences that don't exceed the maximum length that BERT can accurately process (512 tokens). We also fine-tune two others for Arabic long document classification. the following sections explain that in details.

4.1. BERT-based Sentence Aggregation

We propose a simple but effective model to do a long document classification task. Our proposed model consist of multiple layers as shown in figure 1; namely, sentence segmentation layer, BERT layer, a linear classification layer, then the sentence grouping layer with respect to each document, and finally the softmax layer. The first layer

is to make a segmentation of sentences from the long text, taking into account the structure of the sentence in Arabic language. So that the sentence does not lose its meaning or break. The second layer is the BERT tokenizer followed by the embedding representation layer. Since we are using BERT base model named Arabert-V2 [23], this layer consists of 12-layer stacked encoders that receive the embedding inputs and process it and send to the an MLP layer.

We train the model on all the sentences and each sentence is considered as document. The training outputs are the classification probability for each class as well as the sentence ID and orginal document ID. We make a grouping of text sentences with the probabilities of each category in each sentence, and in the end we aggregate all sentence in the category with the highest probability with respect to the document ID.

4.2. BERT-based Key Sentences Model

This model has the same idea of dividing the document into sentences, but instead we hypothesize that a majority of semantically important information is concentrated within specific sentences inside of a longer text, making it unnecessary to check for connections between all words in a document. Instead, we used BERT-based similarity match algorithm that can recognize high-relevance sentences and pass them as input to the BERT model that can complete the desired classification task. The high-relevance sentences were selected by applying a maximal marginal relevance (MMR) [24] similarity algorithm as shown in equation 1. The length of the sentences is between 30 to 150 tokens.

$$MMR = argmax_{D_i \in X}[\lambda Sim_1(D_i, S) - (1 - \lambda) \max_{D_j \in C} Sim_2(D_i, D_j)] \quad (1)$$

Where S is the sentence vector and D_i is the document vector related to S. X is a subset of documents in our dataset we already selected and λ is a constant in range of [0–1], for diversification of results. The Sim_1 and Sim_2 are the similarity function which can be replaced by cosine, euclidean, Jacard and any other distance similarity measures. In our model we have used the proper cosine similarity that explained by equation 2 [4].

$$\frac{cos^{-1}\left(\frac{\sum_{i=1}^n u_i \times v_i}{||u||_2 \times ||v||_2}\right)}{\pi} \quad (2)$$

4.3. Fine-tuned Models

In this part, we reproduced and fine-tuned two of the important research works in the literature for processing

---

[4]XCS224 Mod2 lecture by Prof. Pott

long texts, which are explained in [6, 11]. We have trained and fine-tuned them to classify Arabic long length documents using the datasets mentioned in Tables 2 and 3.

4.4. Longformer

The Longformer [11] was proposed to reduce the complexity of the self-attention matrix. This can be done by making the matrix sparser through the introduction of attention pattern with specified locations that need to be prioritized. By using a sliding window with a fixed length, the model doesn't enter exponential progression and instead scales linearly with input sequence length. Additional gains can be achieved by dilating the sliding window, which frees up some attention heads to process the overall semantic context while non-dilated heads remain focused on local tokens.

However, the implemented restrictions interfere with the model's ability to be trained for specific tasks, which required the addition of global attention to the model. Linear projections are used to calculate the attention scores, and in this work an extra set of projections related to global attention are used to make training more reliable. The resulting linguistic model has an impressive capacity for contextual analysis, but expends far less computational resources when used with long-form documents than traditional BERT and other Transformer architectures. Nonetheless, Longformer was trained for autoregressive modeling with left-to-right word sequence and train it with Arabic needed some preprocessing. We converted the base model of Arabert-V2 into a Longformer then we fine-tuned the output model for our Arabic long document classification task.

4.5. RoBERT

In this model the authors [6] are looking into possible ways to extend the usefulness of the BERT linguistic model to text samples longer than a few hundred words. To do this, they introduce an extension to the fine-tuning procedure and separate the input into smaller chunks. After those chunks are processed by the base BERT model, they are passed through another Transformer or a single recurrent layer before a classification decision is made in the softmax layer. Those variations were named RoBERT (Recurrence over BERT) and ToBERT (Transformer over BERT), collectively described as Hierarchical Transformers because they maintain the hierarchical structure of representations both on the level of extracted segments and the whole document. Those models were found to converge very quickly when trained on a narrowly focused dataset and to perform better than the original BERT with long text sequences. Suitability of those derivative models was examined for different tasks, including topic identification and satisfaction prediction

during a customer call, which are possible real world applications. Unlike the Longformer, with RoBERT the fine-tuning process was straightforward because it was a BERT-Based model.

5. Experimental approach

In our work we aim to find a balance between model accuracy on classification task performed over long text sequences and computational simplicity. Therefor we tried to utilize the base version of BERT which have less memory size of 500MB and faster prediction process where the length of the embedding is 768. We used Google Colab pro to train and fine-tune our models. In terms of accuracy, we use standard metrics to track all of those qualities for the tested models. Macro F1 score is used as a general measure of accurate prediction on all comparisons , as it provides a basis for comparison of results between studies.

Several hyperparameters have been setup to fine-tune the experimented models. Our proposed classification solutions were tested using two collection of documents mentioned in Sec. 3 where 80% of the dataset was used to train the model and 10% as a validation set and 10% utilized for conducting the tests. Table 1 shows the general parameters used in the training and fine-tuning processes.

Table 1
Hyperparameters used in the training and fine-tuning processes

Parameter Name	Value
number of epochs	5
maximum sequence length aggregation and similarity models	128
maximum sequence length longFormer model	1024
maximum sequence length truncation , RoBERT models	1024
mawdoo3 data number of training steps:	107466
adam epsilon	1e-8
train batch size	64
valid batch size	128
epochs	20
learning rate	5e-5
warmup ratio	0.1
max grad norm	1.0
accumulation steps	1

6. Results Discussion and Analysis

The evaluation was conducted using standardized hyper parameters such as batch size and sequence length and others as shown in Table 1, with two different datasets suitable for Arabic long document classification task as described in Sec 3. We will try to report the results and analyze them according to each data set separately.

6.1. Mawdoo3 Dataset

All models were empirically evaluated on long document classification task. We compared our proposed models with Longformer as well as with the RoBERT on Mawdoo3 dataset. The results were very close between the two proposed solutions and the Longformer, with a very slight superiority to the language model based on extracting key sentences using MMR method with macro F1 score equal to 83%. While Robert performed very poorly on Mawdoo3 dataset with macro F1 score of 21%. The overall results of all models in the long document classification task are explained in Table 2. We can say that this results support our hypothesis of identify the most relevant parts of the text. The resulting solution retains the ability to capture relationships between distant tokens, but doesn't have to actively back-propagate all of them and instead focuses only on key sentences. Because of this, the model avoids geometric progression of complexity and continues to be efficient with much longer texts than the original BERT is able to. It is worth noting that we have pre-processed and removed the information at the beginning of each article in this dataset because that the parts of the document containing easily identifiable indicators of the class.

6.2. Arabic News Dataset

The results of the experiment were completely different with the Arabic news dataset. All models performed very well, and in this experiment, the first model outperformed the rest with macro F1 score of 98.4% which revealed that additional modification can have a positive impact on model performance, but it's important which dataset is used. It was discovered that classifying each sentence is better than classifying the whole sequence, which could even increase performance when working with short sentences. However, both Longformer and our second model with MMR are still performing very well with macro F1 score of 96% and 96.2%, respectfully. Whereas RoBERT model has macro F1 score of 74.4%. The overall results of all models in the long document classification task are described in Table 3.

7. Conclusion

Unmatched flexibility of BERT is one of the main reasons for its rapid acceptance as state-of-the-art language model. With additional algorithm and some modifications and fine-tuning, the model can be adjusted for certain topics or tasks and its accuracy pushed to even higher level. This work explores this possibility in detail, taking long text classification as the target task and searching for the best parameters for this type of usage. In particular, different possibilities for supervised pre-training

Table 2

Overall results of all models in the long text classification task on Mawdoo3 Dataset

Model	Macro F1	Macro Precision	Macro Recall	Accuracy
Our Model Aggregating	0.82187	0.82338	0.83049	0.83083
Our Model-MMR	**0.82732**	**0.83162**	**0.83555**	**0.83522**
Longformer	0.82347	0.82497	0.83263	0.83291
RoBERT	0.21157	0.19309	0.29507	0.36461

Table 3

Overall results of all models in the long text classification task on Arabic News Dataset

Model	Macro F1	Macro Precision	Macro Recall	Accuracy
Our Model Aggregating	**0.98411**	**0.98591**	**0.98264**	**0.98434**
Our Model-MMR	0.96217	0.96240	0.96206	0.96263
Longformer	0.95908	0.95956	0.95880	0.95961
RoBERT	0.73062	0.75382	0.75124	0.75142

and fine-tuning were examined on two different datasets. Through detailed experimentation, we were able to identify the most optimal procedures that enable BERT to be more accurate with our particular downstream task. While the value of the proposed training and tuning actions was confirmed only for text classification, it stands to reason that analogous procedures could prove to be useful for other linguistic tasks as well. Finally, we want to denote that we did not explore all hyperparameters which can be a future work to have along with trying another language models such as Roberta and Electra.

References

[1] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, D. Metzler, Long range arena: A benchmark for efficient transformers, arXiv preprint arXiv:2011.04006 (2020).

[2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[3] M. Ding, C. Zhou, H. Yang, J. Tang, Cogltx: Applying bert to long texts, Advances in Neural Information Processing Systems 33 (2020) 12792–12804.

[4] V. Wagh, S. Khandve, I. Joshi, A. Wani, G. Kale, R. Joshi, Comparative study of long document classification, in: TENCON 2021-2021 IEEE Region 10 Conference (TENCON), IEEE, 2021, pp. 732–737.

[5] Z. Wang, P. Ng, X. Ma, R. Nallapati, B. Xiang, Multi-passage bert: A globally normalized bert model for open-domain question answering, arXiv preprint arXiv:1908.08167 (2019).

[6] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, N. Dehak, Hierarchical transformers for long document classification, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 838–844.

[7] S. Sukhbaatar, E. Grave, P. Bojanowski, A. Joulin, Adaptive attention span in transformers, arXiv preprint arXiv:1905.07799 (2019).

[8] Y. Tay, M. Dehghani, D. Bahri, D. Metzler, Efficient transformers: A survey, ACM Computing Surveys (CSUR) (2020).

[9] J. W. Rae, A. Potapenko, S. M. Jayakumar, T. P. Lillicrap, Compressive transformers for long-range sequence modelling, arXiv preprint arXiv:1911.05507 (2019).

[10] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, R. Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, arXiv preprint arXiv:1901.02860 (2019).

[11] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[12] M. Abbas, K. Smaili, Comparison of topic identification methods for arabic language, in: Proceedings of International Conference on Recent Advances in Natural Language Processing, RANLP, 2005, pp. 14–17.

[13] X. Dai, I. Chalkidis, S. Darkner, D. Elliott, Revisiting transformer-based models for long document classification, arXiv preprint arXiv:2204.06683 (2022).

[14] A. Adhikari, A. Ram, R. Tang, J. Lin, Docbert: Bert for document classification, arXiv preprint arXiv:1904.08398 (2019).

[15] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China national conference on Chinese computational linguistics, Springer, 2019, pp. 194–206.

[16] W. Huang, Z. Tao, X. Huang, L. Xiong, J. Yu, Hierarchical self-attention hybrid sparse networks for document classification, Mathematical Problems in Engineering 2021 (2021).

[17] Y. Si, K. Roberts, Hierarchical transformer networks for longitudinal clinical document classification, arXiv preprint arXiv:2104.08444 (2021).

[18] H. H. Park, Y. Vyas, K. Shah, Efficient classification of long documents using transformers, arXiv preprint arXiv:2203.11258 (2022).

[19] Z. Wang, C. Wang, H. Zhang, Z. Duan, M. Zhou, B. Chen, Learning dynamic hierarchical topic graph with graph convolutional network for document classification, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 3959–3969.

[20] J. He, L. Wang, L. Liu, J. Feng, H. Wu, Long document classification from local word glimpses via recurrent attention learning, IEEE Access 7 (2019) 40707–40718.

[21] A. Chouigui, O. Ben Khiroun, B. Elayeb, An arabic multi-source news corpus: experimenting on single-document extractive summarization, Arabian Journal for Science and Engineering 46 (2021) 3925–3938.

[22] M. Abbas, D. Berkani, Topic identification by statistical methods for arabic language., WSEAS Transactions on Computers 5 (2006) 1908–1913.

[23] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).

[24] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 335–336.