INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 18, Issue: 1, Month: February, Year: 2023
Article Number: 5111, https://doi.org/10.15837/ijccc.2023.1.5111

Evaluation of Language Models on Romanian XQuAD and RoITD datasets

D.C. Nicolae, R.K. Yadav, D. Tufis

Dragoş Constantin Nicolae*
Research Institute for Artificial Intelligence
Romanian Academy
Bucharest 050711
*Corresponding author: dragosnicolae555@gmail.com

Rohan Kumar Yadav
0378, Oslo, Norway
errohanydv@gmail.com

Dan Tufiş
Research Institute for Artificial Intelligence
Romanian Academy
Bucharest 050711
tufis@racai.ro

Abstract

Natural language processing (NLP) has become a vital requirement in a wide range of applications, including machine translation, information retrieval, and text classification. The development and evaluation of NLP models for various languages have received significant attention in recent years, but there has been relatively little work done on comparing the performance of different language models on Romanian data. In particular, the introduction and evaluation of various Romanian language models with multilingual models have barely been comparatively studied. In this paper, we address this gap by evaluating eight NLP models on two Romanian datasets, XQuAD and RoITD. Our experiments and results show that bert-base-multilingual-cased and bert-base-multilingual-uncased, perform best on both XQuAD and RoITD tasks, while RoBERT-small model and DistilBERT models perform the worst. We also discuss the implications of our findings and outline directions for future work in this area.

Keywords: NLP, Question Answering, RoBert, RoGPT, DistilBert, Transformer

1 Introduction

The digital revolution transformed our world to what has been called information and knowledge societies [40]. The transition from ruled-based and statistical/fuzzy approaches on language processing [45] to the current neural networks-driven language modelling came with the need for larger and cleaner data, which is required for robust training and evaluation of language-centric AI applications. The field of natural language processing (NLP) has made significant progress in the area of question answering (QA), with a range of datasets available for various types of QA, including extractive, cloze-completion, and open or specialized domain QA [8, 15, 21, 33, 34]. In recent years, the performance of QA systems has even surpassed that of humans in some settings [12]. Despite the popularity of QA, there are few datasets for languages other than English, especially for higher-resource languages [4]. This lack of data limits the internationalization of QA systems, as there is no benchmark data for multilingual QA and it is difficult to train end-to-end QA models without sufficient data. The issue is discussed in detail in [39] and specific steps are mentioned for the Romanian language that are necessary to narrow the current gap. However, multilingual evaluation data is necessary to assess the performance of QA systems in different languages. There have been advances in cross-lingual tasks such as document classification [24, 41], semantic role labelling [1], and natural language inference (NLI) [10], which suggest that multilingual QA evaluation data is crucial for the field.

There is a recognized need for a cross-lingual evaluation benchmark dataset for question answering that is both comprehensive and fair in its coverage of diverse languages [4], and efficient in its use of resources such as annotations. However, many existing cross-lingual datasets have limitations that make them less suitable for this purpose. In order to address these issues and create a dataset that is more suitable for evaluation purposes, it would be valuable to have a dataset that is translated using machine translation or manually annotated for a specific language. One of the popular frameworks for cross-lingual evaluation is XQuAD [3]. XQuAD is a benchmark dataset specifically designed for evaluating cross-lingual question-answering performance. It consists of a subset of 240 paragraphs and 1190 question-answer pairs from the development set of SQuAD v1.1, which have been professionally translated into 10 different languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. As a result, the dataset is entirely parallel across all 11 languages, which makes it a useful resource for comparing question-answering performance across languages and for evaluating the performance of cross-lingual QA models. In addition to this, Nicolae et al,. manually designed a Romanian QA dataset using crowdsourcing in the IT domain. [28].

In the development of neural reading models, attention mechanisms have been widely utilized to construct interdependent representations of both passages and questions. These representations are then utilized in predicting the boundaries of the answer within the passage. Several attention models have been explored, including the Attention Sum Reader [18], Gated Attention Reade [13], Self-Matching Network [30], Attention over Attention Reader [11], and Bi-Attention Network [19]. Recently, pre-trained language models (PrLMs) have been highly successful in the design of encoders for machine reading comprehension tasks. Examples of such PrLMs include ELMo [36], GPT [7] , BERT [12], XLNet [44], Roberta [25], ALBERT [22], and ELECTRA [9]. These models have achieved impressive results on a variety of natural language processing tasks for two main reasons: first, they are trained on large text corpora, allowing them to learn general language patterns and serve as a knowledge base, and second, their Transformer architecture enables them to effectively capture long-range dependencies and higher-order relations in text.

One limitation of transformer models is that they are typically trained on large amounts of data in a single language, making them difficult to adapt to tasks in other languages. To address this limitation, researchers have developed multilingual transformer models, which are trained on large amounts of text data from multiple languages and are able to learn language-agnostic features that can be used for a variety of natural language processing tasks. These models have the potential to significantly improve the performance of NLP tasks in a wide range of languages but may not perform as well as language-specific models on tasks that require a deep understanding of the specific nuances and characteristics of a particular language. For example, a language-specific model may be better equipped to handle idiomatic expressions, cultural references, and other language-specific phenomena that may not be as common in other languages. Thus, we can see several language-specific transformer models such as romBERT [14], KR-BERT [23], Dutch-BERT [43], and GottBERT [37]. There are other BERT variants as well trained for specific languages. The objective of this study is to compare the performance of language-specific transformer models and multilingual models on Romanian QA datasets, XQuAD-ro and RoITD, in order to determine the effectiveness of each model in handling language-specific phenomena and characteristics. Hence, in this paper, the aim is to evaluate the performance of Romanian language-specific transformer models, trained only on Romanian data, against other widely used multilingual models. The results will provide insights into the effectiveness of these models in handling language-specific features for NLP tasks.

2 Background

A transformer is a neural network architecture used for tasks that involve converting one sequence into another, as in the case of natural language processing [42]. It relies on the attention mechanism to take in a sequence and output a related sequence. The architecture consists of multiple encoder and decoder blocks that are identical and stacked on top of each other. The encoder blocks extract linguistic information from the input sequence to form a contextual representation, while the decoder blocks produce the output sequence based on the input. A key aspect of the transformer is the use of multi-head self-attention layers in the encoder blocks, which enables the model to simultaneously attend to different parts of the input sequence, thus improving training compared to recurrent neural networks.

2.1 BERT

Indeed, BERT achieved state-of-the-art results on multiple tasks at that time, but now it is no longer valid as a lot of models employ BERT-based architectures. It distinguishes itself from other language models by using a unique training approach that involves predicting a masked word rather than the next word based on previous words. This technique allows BERT to consider both the left and right context around the target word, giving it the ability to generate a comprehensive, bidirectional representation of the input sequence. This is in contrast to other models that only consider the context leading up to the target word. The BERT is trained in two model sizes: BERT-BASE (L=12, H=768, A=12, Total Parameters=110M) and BERT-LARGE (L=24, H=1024, A=16, Total Parameters=340M), where the number of layers is denoted as L, the hidden size is denoted as H, and the number of self-attention heads is denoted as A.

BERT is designed to accept two segments of the tokenized text: $x_1, x_2, \cdots, x_n$ and $y_1, y_2, \cdots, y_m$ as the input. These segments usually consist of multiple natural sentences and are combined into a single input sequence for BERT with the help of special delimiting tokens: [**CLS**], $x_1, x_2, \cdots, x_n$, [**SEP**], $y_1, y_2, \cdots, y_m$, [**EOS**]. To avoid exceeding the maximum allowed sequence length during training, the lengths of the segments m and n are limited to $m + nT$, where T is the parameter that controls the maximum allowed sequence length. Before being fine-tuned for a specific task using labelled data, BERT is trained on a large dataset of unlabeled text. BERT model uses the transformer architecture to process and understand natural language. It has L layers, each of which uses self-attention with A heads, and a hidden dimension of size H. These components work together to analyze and understand the input text.

2.1.1 Masked Language Model (MLM)

BERT uses a technique called Masked Language Modeling (MLM). In MLM, a percentage of the input tokens are randomly masked and the model is trained to predict the masked tokens based on the context provided by the remaining tokens. To mitigate the mismatch between pre-training and fine-tuning, the training data generator randomly chooses a percentage of the token positions to predict, replacing the chosen tokens with the [MASK] token with a probability of 80%, a random token with a probability of 10%, or the original token with a probability of 10%, respectively. The model is then trained to predict the original token using cross-entropy loss.

2.1.2 Next Sentence Prediction (NSP)

The NSP is a way to predict whether two segments of text should appear consecutively in the original text, based on their content. It is used in a binary classification system, where positive examples are pairs of sentences that appear consecutively in the same document, and negative examples are pairs of sentences that come from different documents or paragraph. The goal of the NSP is to improve the accuracy of natural language processing tasks that involve understanding the relationship between pairs of sentences, such as natural language inference.

2.1.3 Multilingual BERT

Multilingual BERT is a variant of BERT that is specifically designed to handle multiple languages. It is trained on a dataset that contains text in a wide range of languages, including English, Spanish, German, and French. Multilingual BERT is able to handle multiple languages by using a shared vocabulary and a single model architecture that can process text in any of the languages it has been trained on. One of the main advantages of multilingual BERT is that it can be fine-tuned for a specific NLP task on a per-language basis, using only a small amount of labelled data for each language. This makes it particularly useful for tasks that involve multiple languages, such as machine translation and cross-lingual information retrieval. Similar to BERT, it has two variants: bert-base-multilingual-cased and bert-base-multilingual-uncased.

The bert-base-multilingual-cased model is trained on a dataset containing a wide range of languages, including English, Spanish, German, and French. The model is trained on text that has been lowercased and cased (meaning that it takes into account the case of the words, such as "upper" and "lower" case). This version of BERT is well-suited for tasks that require the use of case information, such as named entity recognition. On the other hand, the bert-base-multilingual-uncased model is also trained on a diverse dataset of languages, but the text has been lowercased (meaning that all words are in lowercase). This version of BERT is well-suited for tasks that do not require the use of case information, such as language translation. In addition to this, there are several language-specific BERT model trained using a large corpora. In particular, in the Romanian language, there are two BERT models romBERT [14] and RoBERT [26].

2.2 DistilBERT

In recent years, transfer learning approaches using large-scale pre-trained language models have become a fundamental model in Natural Language Processing (NLP). These models have significantly improved performance in a variety of NLP tasks, but they often have hundreds of millions of parameters, and research shows that even larger models can lead to better downstream task performance. However, the trend toward larger models has raised concerns about the environmental impact of the increased computational requirements and the potential difficulty of running these models on-device in real-time [38]. Hence a smaller language model of BERT pre-trained with knowledge distillation is designed with similar performance on downstream tasks while being lighter, faster at inference time, and requiring a smaller computational training budget. These compressed models are also small enough to run on the edge, such as on mobile devices.

2.2.1 Knowledge Distillation

Knowledge distillation [17] is a compression technique in which a smaller model, known as the student, is trained to replicate the behaviour of a larger model, or an ensemble of models, referred to as the teacher. In supervised learning, a classification model is typically trained to predict the class of an instance by maximizing the estimated probability of the correct labels. The standard training objective involves minimizing the cross-entropy between the model's predicted distribution and the one-hot empirical distribution of the training labels. A model that performs well on the training set will predict a high probability for the correct class and near-zero probabilities for other classes. However, some of these "near-zero" probabilities are larger than others and reflect the model's generalization capabilities and how it will perform on the test set.

2.2.2 Student Architecture

DistilBERT, has the same general architecture as BERT, but with the token-type embeddings and the pooler removed and the number of layers reduced by a factor of 2. While most operations used in the transformer architecture are highly optimized in modern linear algebra frameworks, the authors found that variations in the hidden size dimension have a smaller impact on computation efficiency (for a fixed parameter budget) than variations in other factors, such as the number of layers. Therefore, the focus is on reducing the number of layers.

2.2.3 Distillation

The training of DistilBERT followed best practices for training the BERT model as recently proposed by Liu et al. [25]. Distillation was performed using very large batches with gradient accumulation (up to 4K examples per batch) and dynamic masking but without the next sentence prediction objective. It is 40% smaller and 60% faster but retains 97% of the language understanding capabilities of BERT. It was shown that a general-purpose language model can be trained using the technique of distillation, and an ablation study was conducted to analyze the various components. Additionally, it was demonstrated that DistilBERT is a viable option for edge applications.

2.2.4 Multilingual DistilBERT

While many efforts have been made to increase the efficiency of pre-trained models, the majority of these efforts have focused on English and only a few for other languages, such as BERTino [27] for Italian, MBERTA for Arabic [2], or GermDistilBERT [1] for German. Hence Avram et al., proposed three lightweight and fast versions of distilled BERT models for the Romanian language are introduced: Distil-BERT-base-ro, Distil-RoBERT-base, and DistilMulti-BERT-base-ro.

- **distilbert-base-romanian-cased [2]:** It is obtained by distilling the knowledge of romBERT [14] using its original training corpus and tokenizer.
- **distilbert-base-romanian-uncased [3]:** It is created from RoBERT-base [26] in similar conditions (i.e., using both original training corpus and tokenizer);
- **distilbert-multi-base-romanian-cased [4]:** It was obtained by distilling the knowledge of an ensemble consisting of romBERT and RoBERT-base, using the combined corpus and tokenizer of romBERT. [6]A

2.3 GPT

The task of generating text is acknowledged to be a challenging aspect of natural language processing (NLP) as it requires the comprehension of context and the ability to continue generating text while maintaining format, coherence and adhering to any specific restrictions, if applicable. In recent years, advancements in deep learning have resulted in the development of architectures that are capable of generating clear, concise and long-form text. Initial significant results were obtained through the implementation of seq2seq architectures, which were based on recurrent neural networks (RNNs). The introduction of an attention layer resulted in improved performance, however, limitations associated with the RNN architecture, such as the vanishing gradient problem, inefficiency in sequential processing, and inconsistency for long sequences, persisted. The transformer architecture subsequently enabled the creation of models that were able to surpass the limitations imposed by the RNN architecture by allowing for the processing of long sequences, optimizing the utilization of available hardware and creating pre-trained models that can be easily fine-tuned for other tasks. The transformer architecture resulted in the development of models with an impressive number of

[1] https://huggingface.co/distilbert-base-german-cased
[2] https://huggingface.co/racai/distilbert-base-romanian-cased
[3] https://huggingface.co/racai/distilbert-base-romanian-uncased
[4] https://huggingface.co/racai/distilbert-multi-base-romanian-cased

parameters, the most recent ones being in the order of billions. Based on the transformer's decoder architecture, GPT autoregressive model was introduced by OpenAI [31]. The next generation of architectures, GPT2 [32], came with a number of parameters 100 times larger than the initial version and the ability to process 1024 tokens at once. This model was trained on a 40 GB corpus (WebText) and four versions were made available: base (117M parameters), medium (345M parameters), large (762M parameters), and xllarge (1542M parameters). GPT2 demonstrated remarkable performance on the synthetic GLUE benchmark and in human evaluations for generated text. It was trained to predict the next token given the previous sequence of tokens and subsequently fine-tuned for other tasks such as summarizing, question answering, translation, or generating text with a specific format. The initial model developed by OpenAI was trained solely for English. ChatGPT is a cutting-edge language model developed by OpenAI that leverages the GPT-3 architecture to generate text in a manner that resembles human language. The model was trained on massive amounts of text data from various sources, which allowed it to learn patterns and relationships between words and phrases in multiple languages. ChatGPT also employs reinforcement learning, where it is rewarded for generating text that meets specific goals, such as answering questions accurately or generating coherent text. The combination of supervised and reinforcement learning makes ChatGPT one of the most advanced and versatile conversational language models available today. Due to its ability to generate human-like text, it can also be used for content creation and personalization in various fields such as marketing, customer service, and social media.

2.3.1 Romanian GPT

The introduction of a Romanian version of the GPT2 model, referred to as RoGPT2 [29], is trained in three versions of the model, named base, medium, and large, have been trained using the largest corpus available for the Romanian language. The performance of RoGPT2 was evaluated using six tasks from the LiRo benchmark and was compared to that of other models including BERT-Base, RoBERT, BERT-robase, and RoDiBERT. Results indicate that RoGPT2 achieved similar or superior performance on these tasks, except for zero-shot learning cross-lingual question answering. Additionally, RoGPT2 was found to obtain state-of-the-art results on the task of grammar error correction using the RONACC corpus, indicating its ability to generate grammatically correct text. The use of RoGPT2 in the context of continuing Romanian news articles is also explored. Fine-tuning of the model was found to result in the generation of long text which effectively accounted for the context of the news.

The corpus utilized for the training of RoGPT2 was sourced from the corpus used for the development of RoBERT [18]. A variety of datasets were included in the corpus, with OSCAR [21] making up the largest percentage. This multi-language corpus was extracted via language classification, filtered, and cleaned from Common Crawl. Additionally, the latest version of the Wikipedia Dumps for Romanian was obtained, parliamentary debates from 1996-2017 were included from ReadME RoText, over 10,000 books in Romanian from various fields were also included and articles from a financial and economic website were obtained up until May 18, 2021, for diversification of topics.

3 Evaluations and Results

3.1 Method with Corpora

We evaluated the above-mentioned multilingual and language-specific Romanian transformer models on two publicly available QA datasets: XQuAD and RoITD. The overall structure of the fine-tuning process is demonstrated in Fig 1. The input representation for machine reading comprehension (MRC) tasks typically consists of three stages. In the first stage, the input text, typically comprised of a question and its corresponding context, is separated using a [SEP] token. In the second stage, the input text is transformed into input_ids, attention_mask, and token_type_ids, which are the input format used by selected transformer models. Finally, in the third stage, these inputs are used to train a QA model, which fine-tunes the weights of the transformer to better perform MRC tasks.

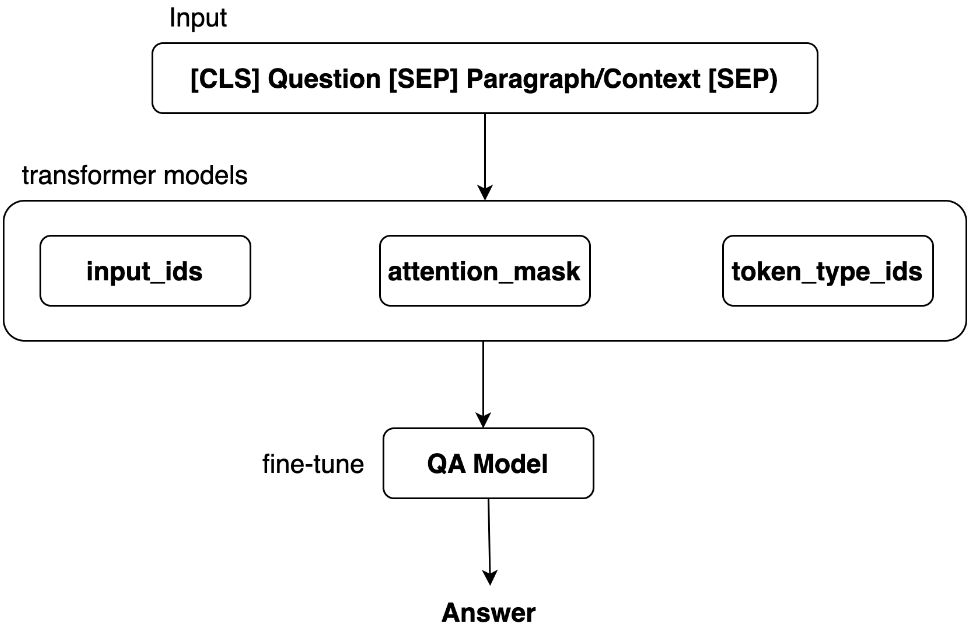

Figure 1: Overall structure of QA model fined-tuned using selected transformer models.

3.1.1 XQuAD

XQuAD (Cross-lingual Question Answering Dataset) is a benchmark dataset for evaluating cross-lingual question-answering performance. The dataset includes a subset of 240 sentences and 1190 question-answer pairs from the SQuAD v1.1 development set [33] as well as certified translations into ten different languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. The dataset is therefore fully parallel across 11 languages. This method is based on learning a new embedding matrix through masked language modelling (MLM) in the new language while freezing the parameters of all other layers. The approach is competitive with state-of-the-art multilingual models such as multilingual BERT (mBERT) on standard cross-lingual classification benchmarks and a new cross-lingual question-answering dataset (XQuAD) despite not relying on a shared vocabulary or joint training. These results contradict the common belief that the generalization ability of multilingual models is based on shared vocabulary items, joint training, and deep cross-lingual representations. The new cross-lingual benchmark XQuAD, which includes 240 paragraphs and 1190 question-answer pairs from SQuAD v1.1. Recently, XQuAD later added the professionally translated version of SQuAD 1.1 in the Romanian language, represented as XQuAD-ro.

3.1.2 RoITD

A Romanian-language QA dataset, known as the Romanian IT Dataset (RoITD) [28], has been created that focuses on IT and electrical themes. The set of 9575 pairs were classified as possible (5013) and not possible (4472). The possible questions are those which have a possible answer in the associated paragraph. The questions considered not possible, 4472, are those for which the paired paragraphs do not contain information pertinent to the question. The dataset includes not just the question and answer, but also the passage providing full context for the QA pair, including the corresponding response order, so that it meets the requirements of the BERT-special architecture. The machine that uses this data set must be able to comprehend the natural language text and extract relevant information from the main paragraph in a similar manner to how people do when answering questions.

Crowd workers were asked to create up to four questions for each provided paragraph, with at least one question being unanswerable. These impossible questions were constructed based on the provided paragraph in such a way that crowd workers could not identify a plausible response within the text. Additionally, crowd workers were instructed that the questions should not always use the exact phrasing found in the context, but they should be grammatically correct. After creating the questions, crowd workers were instructed to combine the question with the relevant category. If the question is unanswerable and the answer is likely, they should use the letter "U," and if the question is answerable and the answer is correct, they should use the letter "A". The user interface used by the crowd workers was hosted using an Amazon Web Service t2.micro machine with an Ubuntu 20.04.1 operating system.

Table 1 gives a general overview of the size of the RoITD dataset. The table illustrates that the dataset is divided into training and testing sets. It also includes other pertinent information, such as the total amount of articles and questions, the average length of questions and answers, and the size of the vocabulary used.

Table 1: Dataset Statistics for split train and test sets.

	Train	Test	All
Number of articles	4170	813	5043
Number of questions	7175	2400	9575
Average passage length	52.72	61.97	55.04
Average question length	8.08	8.12	8.11
Average answer length	13.44	7.85	9.25
Vocabulary size	38265	18396	48821

3.2 Neural Models

We utilized various transformer models, including BERT, GPT, and DistilBERT, which were all pre trained on the Romanian language. These models were fine-tuned on the XQuAD and RoITD datasets for our study. In addition to this, we also used BERT-multilingual model to fine tune on these selected dataset so as to compare the performance of other models trained using Romanian language. We assume BERT-multilingual as the baseline for comparison. The results of our experimentation are summarized in the table above. The models were distilbert-base-romanian-cased, distilbert-base-romanian-uncased, distilbert-multi-base-romanian-cased, bert-base-multilingual-cased, bert-base-multilingual-uncased, RoBERT-small, RoGPT2 Base, and RoGPT2 Medium.

3.3 Experimental Setup

We trained all of the models for 25 epochs with a learning rate of 3e-5 and a batch size of 12. However, both datasets have different numbers of training steps i.e., 120K for RoITD and 20K for XQuAD. This is because we have used the same number of epochs for both datasets and the datasets have different sizes. We use two main QA evaluation metrics for the evaluation of the models:

- **F1-score:** It is particularly useful when the class distribution is imbalanced, as it gives equal importance to both precision and recall. For example, in a binary classification problem, if the positive class is rare, a model that always predicts the negative class will have high accuracy but low recall, thus a low F1-score. In contrast, the F1-score can provide a more comprehensive evaluation of a model's performance by considering both precision and recall. In a QA task, precision refers to the proportion of correctly answered questions among all the questions that the model answered, while recall refers to the proportion of correctly answered questions among all the questions in the dataset. A high precision indicates that the model is able to correctly identify and extract relevant information from the input text, while a high recall indicates that the model is able to find all the relevant information in the input text. Therefore, a high F1-score for a QA model implies that the model is able to accurately identify and extract all the relevant information from the input text.

- **Exact Match:** Exact Match is a binary evaluation metric that is often used to evaluate the performance of question answering (QA) models. It measures the proportion of examples for which the model's prediction exactly matched the ground truth. In other words, it calculates the percentage of questions that the model answered correctly without any errors. A high exact match score implies that the model is able to provide the correct answer to a large number of questions, and it is able to understand the context of the questions and provide the correct answer. Exact match is a simple and straightforward metric, it only checks whether the model's prediction exactly matches the ground truth or not, it does not take into account the quality or similarity of the model's prediction with the ground truth, it can also be seen as a strict metric since it only considers the exact match as a correct answer.

We trained all of the models for 25 epochs with a learning rate of 3e-5 and a batch size of 12. First, we present the evaluation of the selected models on RoITD dataset. As we can see from Table 2, the BERT-base-multilingual models, specifically the cased and uncased versions, performed the best in terms of both F1-score and exact match, with scores of 0.5966 and 0.6011 respectively. These models achieved an F1-score and exact match that is significantly higher than the other models in the comparison set. Additionally, distilbert-multi-base-romanian-cased and RoBERT-small also performed well, with F1-scores of 0.4747 and 0.5107 respectively. However, the RoGPT2 Base and RoGPT2 Medium models had relatively lower performance, with F1-scores of 0.303 and 0.408 respectively.

Further, Figure 2 shows a graph for the training history of the RoITD dataset with the y-axis representing EM and the x-axis representing steps as a visual representation of how the EM of a model changes over time. As we can see at the beginning of the training, the EM score is low, but as the model is trained on more data, the score increases. This increase in performance is depicted by the upward trend of the graph. In later cases, the performance plateaus, indicating that the model has reached its optimal performance and additional training is not improving the model's accuracy. However, we can see some drastic fluctuation of EM by Robert-small and distilbert-base-romanian-cased. This might be the indication of overfitting due to the model memorizing the training data or the model is not yet exposed to enough data and it is not able to generalize from the limited data it has seen. As the model is exposed to more data and continues to train, it begins to generalize better and the performance on the test set improves. The same trend is observed in Fig 3 in the case of the F1-score.

Table 2: Performance of selected models on RoITD QA and XQuAD datasets.

Model Name	F1 (%) RoITD	EM (%) RoITD	F1 (%) XQuAD	EM (%) XQuAD
distilbert-base-romanian-cased	44.68	24.16	12.91	7.58
distilbert-base-romanian-uncased	46.19	24.22	25.41	48.00
distilbert-multi-base-romanian-cased	47.47	25.18	25.39	4.84
bert-base-multilingual-cased	59.66	35.38	57.31	**41.96**
bert-base-multilingual-uncased	**60.11**	**36.40**	**57.65**	41.52
RoBERT-small	51.07	23.29	16.65	9.82
RoGPT2 Base	30.30	19.60	35.97	23.69
RoGPT2 Meduim	40.80	26.70	44.74	29.66

Similarly, for XQuAD dataset, we can see from Table 2, the BERT-base-multilingual models, specifically the cased and uncased versions, performed the best in terms of both F1-score and exact match, with scores of 0.5731 and 0.5765 respectively. These models achieved an F1-score and exact

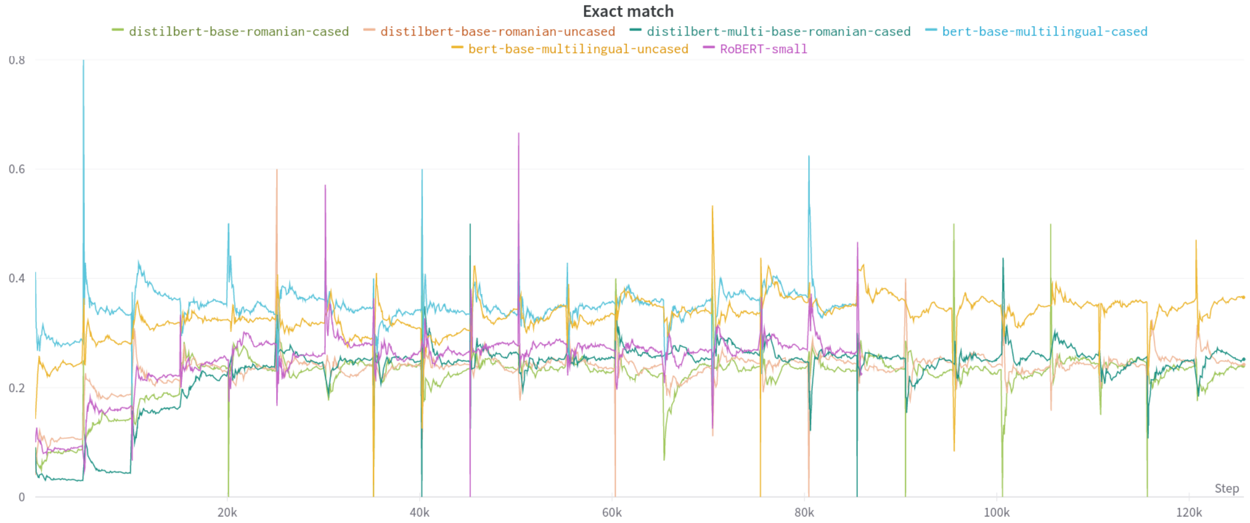

Figure 2: Step-by-step visualization of the evolution of EM in RoITD dataset.

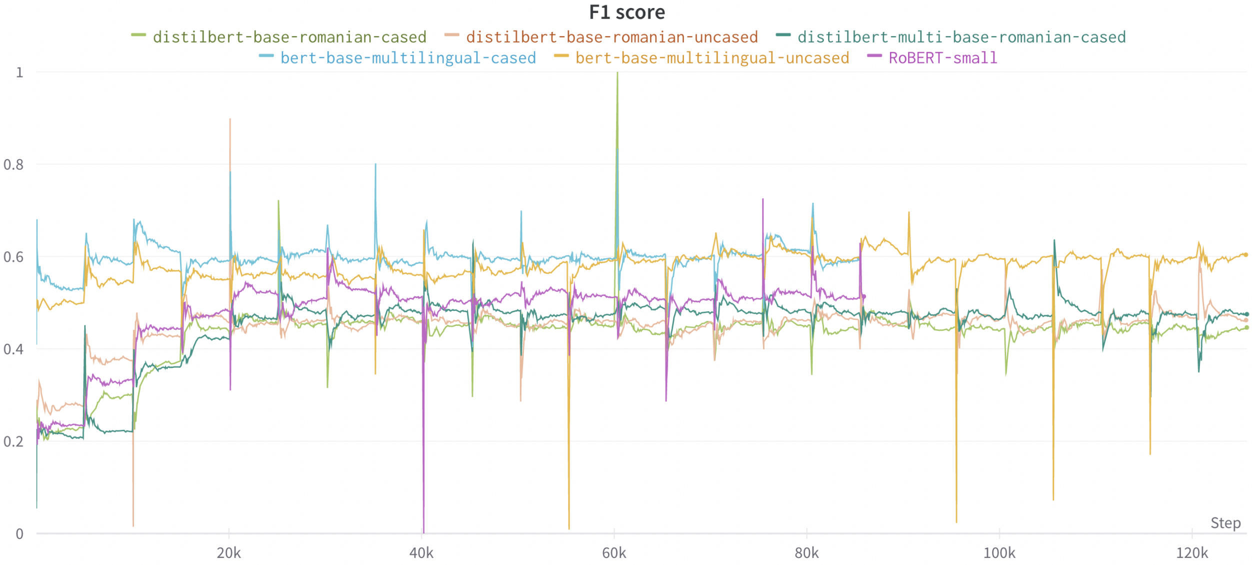

Figure 3: Step by step visualization of the evolution of F1-score in RoITD dataset.

match that is significantly higher than the other models in the comparison set. Additionally, RoGPT2 Medium also performed well, with F1-score of 0.4474. However, the other models had relatively lower performance, distilbert-base-romanian-uncased and RoBERT-small had F1-scores of 0.1137 and 0.1665 respectively, and distilbert-multi-base-romanian-cased and RoGPT2 Base had F1-scores of 0.05102 and 0.3597 respectively.

In addition to this, Figure 4 shows a graph for XQuAD dataset with the y-axis representing EM and the x-axis representing steps as a visual representation of how the EM of a model changes over time. As we can see the similar trend in the early phase of learning, the EM score is low, but as the model is trained on more data, the score increases. This increase in performance is depicted by the upward trend of the graph. However, the overfitting occurs more frequently here by the model bert-base-multilingual-cased and bert-base-multilingual-uncased. This overfitting is also observed in the F1-score of XQuAD as shown in Fig 5. Moreover, the difference between the performance of distilbert-based models and other selected models can be clearly seen from both graphs.

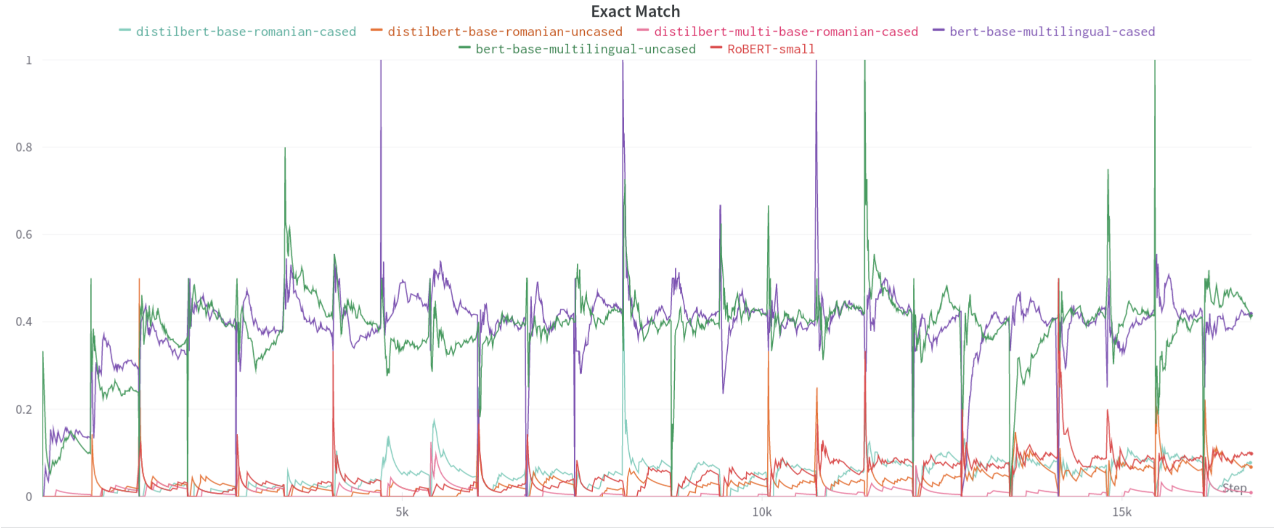

Figure 4: Step by step visualization of the evolution of EM in XQuAD dataset.

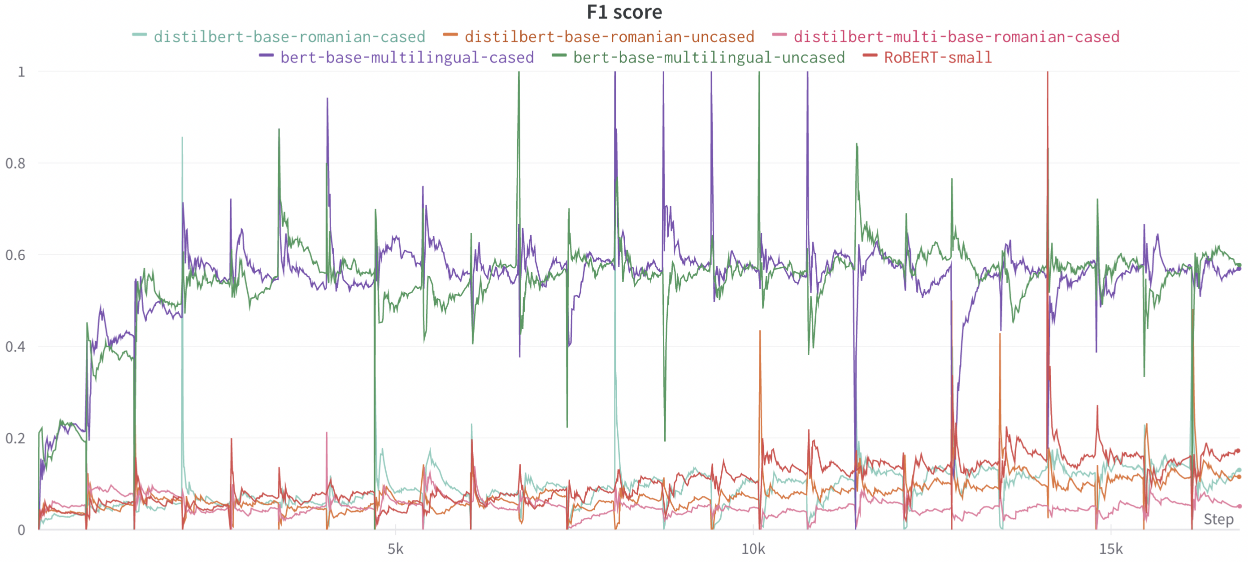

Figure 5: Step by step visualization of the evolution of F1-score in XQuAD dataset.

3.4 Discussion

In general, we can observe from the experiments and results that, the bert-base-multilingual-cased model has an F1-score of 0.5731 and an Exact Match score of 0.4196 on XQuAD and an F1-score of 0.5966 and an Exact Match score of 0.3538 on RoITD. This suggests that BERT models are well-suited for both question-answering and span-identification tasks. On the other hand, the RoBERT-small model performs less well on both tasks, with lower F1-scores and Exact Match scores. For example, on XQuAD task it has F1-score of 0.1665 and EM of 0.09821 and on RoITD it has F1-score of 0.5107 and EM of 0.2329. This suggests that RoBERT-small might not be suitable for these tasks. RoGPT2 models perform well on the RoITD task as compared to XQuAD, with F1-scores of 0.303 and 0.408 and Exact Match scores of 0.196 and 0.267 respectively on RoITD, but on XQUAD, the F1-scores are 0.3597 and 0.4474 and Exact Match scores are 0.2369 and 0.2966 respectively. The DistilBERT models perform poorly on both tasks, with low F1-scores and Exact Match scores on both XQuAD and RoITD, This suggests that DistilBERT models might not be suitable for these tasks.

This might be because DistilBERT models, such as distilbert-base-romanian-cased, distilbert-base-romanian-uncased, and distilbert-multi-base-romanian-cased, are smaller versions of BERT models, which are trained to have similar performance to the original BERT models but with fewer parameters.

The main difference between BERT and DistilBERT is the architecture, DistilBERT uses a distillation method that reduces the number of parameters of the model. Due to this architecture, the model's ability to understand the context and relationships between words in a text passage may not be as good as BERT, which in turn affects its performance on XQuAD and RoITD tasks. Additionally, DistilBERT models are not pre-trained on a large corpus of text data as BERT models, which may also contribute to their poor performance. Additionally, RoBERT-small model also performs less well on both tasks because it is a smaller version of the RoBERT model which is trained on a smaller corpus of Romanian text data, this may not be enough to capture the complexity of the language and relationships between words in the text passages. On the other hand, RoGPT2 is based on the unidirectional and only considers the context before a token in contrast to BERT which is bidirectional, meaning that they consider the context before and after a token during pre-training. It helps BERT models to better understand the relationships between words in a sentence. Hence RoGPT2 performs decent but does not outperform the state-of-the-art.

4 Conclusion

In this paper, we evaluated eight NLP models on two Romanian datasets, XQuAD and RoITD. The results provide valuable insights into the performance of these models on Romanian language data and could be of interest to researchers and practitioners working with this language. Our findings suggest that BERT models, specifically bert-base-multilingual-cased and bert-base-multilingual-uncased, perform well on both XQuAD and RoITD tasks, while RoBERT-small model and DistilBERT models may not be suitable for these tasks. RoGPT2 models performed well on RoITD task but not on XQUAD. These results are consistent with the literature on NLP models and the characteristics of the Romanian language. Future work can be done to improve the performance of these models by fine-tuning them on larger Romanian datasets or by using techniques such as transfer learning.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Akbik, A., Chiticariu, L., Danilevsky, M., Li, Y., Vaithyanathan, S. & Zhu, H. Generating High Quality Proposition Banks for Multilingual Semantic Role Labeling. *Proceedings Of The 53rd Annual Meeting Of The Association For Computational Linguistics And The 7th International Joint Conference On Natural Language Processing (Volume 1: Long Papers).* pp. 397-407 (2015,7), https://aclanthology.org/P15-1039

[2] Alyafeai, Z. & Ahmad, I. Arabic Compact Language Modelling for Resource Limited Devices. *Proceedings Of The Sixth Arabic Natural Language Processing Workshop.* pp. 53-59 (2021,4), https://aclanthology.org/2021.wanlp-1.6

[3] Artetxe, M., Ruder, S. & Yogatama, D. On the Cross-lingual Transferability of Monolingual Representations. *Annual Meeting Of The Association For Computational Linguistics.* (2019)

[4] Asai, A., Eriguchi, A., Hashimoto, K. & Tsuruoka, Y. Multilingual Extractive Reading Comprehension by Runtime Machine Translation. *ArXiv.* **abs/1809.03275** (2018)

[5] Avram, A., Catrina, D., Cercel, D., Dascualu, M., Rebedea, T., Puaics, V. & Tufics, D. Distilling the Knowledge of Romanian BERTs Using Multiple Teachers. *LREC.* (2021)

[6] Avram, A., Catrina, D., Cercel, D., Dascălu, M., Rebedea, T., Păiş, V. & Tufiş, D. Distilling the Knowledge of Romanian BERTs Using Multiple Teachers. (arXiv,2021)

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. Language Models Are Few-Shot Learners. *Proceedings Of The 34th International Conference On Neural Information Processing Systems.* (2020)

[8] Chen, D., Fisch, A., Weston, J. & Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. *Proceedings Of The 55th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers).* pp. 1870-1879 (2017,7), https://aclanthology.org/P17-1171

[9] Clark, K., Luong, M., Le, Q. & Manning, C. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *8th International Conference On Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* (2020), https://openreview.net/forum?id=r1xMH1BtvB

[10] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H. & Stoyanov, V. XNLI: Evaluating Cross-lingual Sentence Representations. *Proceedings Of The 2018 Conference On Empirical Methods In Natural Language Processing.* pp. 2475-2485 (2018), https://aclanthology.org/D18-1269

[11] Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T. & Hu, G. Attention-over-Attention Neural Networks for Reading Comprehension. *Proceedings Of The 55th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers).* pp. 593-602 (2017,7), https://aclanthology.org/P17-1055

[12] Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers).* pp. 4171-4186 (2019)

[13] Dhingra, B., Liu, H., Yang, Z., Cohen, W. & Salakhutdinov, R. Gated-Attention Readers for Text Comprehension. *Proceedings Of The 55th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers).* pp. 1832-1846 (2017,7), https://aclanthology.org/P17-1168

[14] Dumitrescu, S., Avram, A. & Pyysalo, S. The birth of Romanian BERT. *Findings Of The Association For Computational Linguistics: EMNLP 2020.* pp. 4324-4328 (2020), https://aclanthology.org/2020.findings-emnlp.387

[15] Ion R., Badea V.G., Cioroiu G., Mititelu V., Irimia E., Mitrofan M. & Tufis D. A Dialog Manager for Micro-Worlds *In Studies in Informatics and Control, 29(4)* . **ISSN: 1220-1766 eISSN: 1841-429X** pp. 411-420 (2020)

[16] Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). *ArXiv: Learning.* (2016)

[17] Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. *ArXiv.* **abs/1503.02531** (2015)

[18] Kadlec, R., Schmid, M., Bajgar, O. & Kleindienst, J. Text Understanding with the Attention Sum Reader Network. *Proceedings Of The 54th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers).* pp. 908-918 (2016,8), https://aclanthology.org/P16-1086

[19] Kim, J., Jun, J. & Zhang, B. Bilinear Attention Networks. *Proceedings Of The 32nd International Conference On Neural Information Processing Systems.* pp. 1571-1581 (2018)

[20] Kingma, D. & Ba, J. Adam: A Method for Stochastic Optimization. *CoRR.* **abs/1412.6980** (2014)

[21] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A., Uszkoreit, J., Le, Q. & Petrov, S. Natural Questions: A Benchmark for Question Answering Research. *Transactions Of The Association For Computational Linguistics.* **7** pp. 452-466 (2019), https://aclanthology.org/Q19-1026

[22] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.. *ICLR.* (2020)

[23] Lee, S., Jang, H., Baik, Y., Park, S. & Shin, H. KR-BERT: A Small-Scale Korean-Specific Language Model. *ArXiv: Computation And Language.* (2020)

[24] Lewis, D., Yang, Y., Rose, T. & Li, F. RCV1: A New Benchmark Collection for Text Categorization Research. *J. Mach. Learn. Res..* **5** pp. 361-397 (2004)

[25] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv.* **abs/1907.11692** (2019)

[26] Masala, M., Ruseti, S. & Dascalu, M. RoBERT – A Romanian BERT Model. *International Conference On Computational Linguistics.* (2020)

[27] Muffo, M. & Bertino, E. BERTino: An Italian DistilBERT model. *Italian Conference On Computational Linguistics.* (2020)

[28] Nicolae D. C., Tufis D. RoITD: Romanian IT Question Answering Dataset. *ConsILR-2021.* (2021)

[29] Niculescu, M., Ruseti, S. & Dascalu, M. RoGPT2: Romanian GPT2 for Text Generation. *2021 IEEE 33rd International Conference On Tools With Artificial Intelligence (ICTAI).* pp. 1154-1161 (2021)

[30] Park, C., Song, H. & Lee, C. S 3 -NET: SRU-Based Sentence and Self-Matching Networks for Machine Reading Comprehension. (2020)

[31] Radford, A. & Narasimhan, K. Improving Language Understanding by Generative Pre-Training. (2018)

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. Language Models are Unsupervised Multitask Learners. (2019)

[33] Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings Of The 2016 Conference On Empirical Methods In Natural Language Processing.* pp. 2383-2392 (2016,11), https://aclanthology.org/D16-1264

[34] Richardson, M., Burges, C. & Renshaw, E. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. *Conference On Empirical Methods In Natural Language Processing.* (2013)

[35] Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv.* **abs/1910.01108** (2019)

[36] Sarzyńska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M. & Okruszek, L. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research.* **304** (2021)

[37] Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V. & Boeker, M. GottBERT: a pure German Language Model. *ArXiv.* **abs/2012.02110** (2020)

[38] Schwartz, R., Dodge, J., Smith, N. & Etzioni, O. Green AI. *Communications Of The ACM.* **63** pp. 54 - 63 (2019)

[39] Tufiș, D. Romanian Language Technology – a view from an academic perspective. *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS and CONTROL.* **17** (2022,1), https://doi.org/10.15837/ijccc.2022.1.4641

[40] Tufis D., Filip F. G. (coordinators). *Limba Romana in Societatea Informationala - Societatea Cunoasterii, editura Expert* . **ISBN: 973-8177-83-9** pp. 512 (2002)

[41] Schwenk, H. & Li, X. A Corpus for Multilingual Document Classification in Eight Languages. *ArXiv.* **abs/1805.09821** (2018)

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is All You Need. *Proceedings Of The 31st International Conference On Neural Information Processing Systems.* pp. 6000-6010 (2017)

[43] Vries, W., Cranenburgh, A., Bisazza, A., Caselli, T., Noord, G. & Nissim, M. BERTje: A Dutch BERT Model. *ArXiv.* **abs/1912.09582** (2019)

[44] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. & Le, Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings Of The 33rd International Conference On Neural Information Processing Systems.* (2019)

[45] Zadeh L., Tufis D., Filip F.G., Dzitac I.(editors) *From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence, Editura Academiei* . **ISBN: 978-973-27-1678-6** pp. 268 (2009)

Copyright ©2023 by the authors. Licensee Agora University, Oradea, Romania.
This is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License.
Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

C O P E
Member since 2012
JM08090

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Nicolae, D.C.; Yadav, R.K.; Tufis, D. (2023). Evaluation of Language Models on Romanian XQuAD and RoITD datasets, *International Journal of Computers Communications* & *Control*, 18(1), 5111, 2023.

https://doi.org/10.15837/ijccc.2023.1.5111