AUTOMATIC ROMANIAN TEXT GENERATION USING GPT-2

Marius Cristian BUZEA[1], Ștefan TRĂUȘAN-MATU[2], Traian REBEDEA[3]

One of the most significant tasks in natural language processing (NLG) is text generation, which beneficiate from the recent architectures that use large pre-trained transformer models, such as the Generative Pre-trained Transformer-2 (GPT-2) or GPT-3 developed by OpenAI, and Google's Bidirectional Encoder Representations from Transformers (BERT). The paper presents a NLG model based on the GPT-2 architecture that generates Romanian instances, using manually annotated texts. A small Romanian GPT-2 model, using 24 thousand news items, named MCBGPT-2 was developed, tested and evaluated. Additionally, an existing Romanian GPT-2 model, called RoGPT-2, was added to experiments. For evaluation, is presented a comparison of several automatic metrics such as BLEU, ROUGE, BLEURT and BERTScore applied to generated instances from the test and validation datasets. Experimental results revealed that the MCBGPT-2 and RoGPT-2 models provided similar performances in text generation task for Romanian language, using less data for MCBGPT-2 model's training process.

Keywords: Text Generation, MCBGPT-2, RoGPT-2, BERTScore, BERT.

1. Introduction

Text generation or natural language generation (NLG) is a process that generates natural language outputs, being a subfield of natural language processing (NLP). One objective of NLG is to provide contextual sentences and information to minimize or replace the human intervention. The recent text generation systems use various techniques such as generative pre-trained transformers (GPT2/3) [1, 2], recurrent neural network (RNN) models [3, 4], or bidirectional and auto-regressive transformers (BART) [5].

GPT-2 was released in 2019, being a large transformer-based language model developed by OpenAI. This model was trained on a massive 40 gigabyte dataset (corpus), the network having 48 layers, 1600 dimensional vectors for word embeddings and a large batch size of 512 units. The system was designed to

[1] PhD Student, Department of Computer Science and Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: bumarius@gmail.com

[2] Professor, Department of Computer Science and Engineering, University POLITEHNICA of Bucharest, and Researcher, Research Institute for Artificial Intelligence "Mihai Draganescu" of the Romanian Academy, Bucharest, Romania, e-mail: stefan.trausan@upb.ro

[3] Associate Professor, Department of Computer Science and Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: traian.rebedea@upb.ro

generate the next words of a text, based on large datasets crawled from internet (e.g., Reddit). GPT-2 has multiple versions, the smallest takes up only 500MBs of disk space to store its 117M parameters. The largest GPT-2 model is 13 times larger than the smallest model; it takes more than 6.5 GBs of disk space and contains 1.5 billion parameters. Several applications such as video games [6] or generation and answering in a collaboration framework [7] use generative pre-trained transformer.

The GPT-2 architecture is based on the transformer architecture, which is specially used to increase the training speed of these models and outperforms the Google neural machine translation model in specific tasks. One other major advantage of transformers comes from parallelization.

The transformer architecture was initially introduced by Vaswani et al. [8]. There are several libraries for transformer applications, such as TensorFlow (https://www.tensorflow.org/text/tutorials/transformer), from the Tensor2Tensor package or PyTorch (https://pytorch.org/hub/huggingface_pytorch-transformers/). It is based on encoder and decoder blocks The Encoder block has 2 layers such as Multi-Head Attention and Feed Forward Neural Network. The decoder block has one more layer named Masked Multi-Head Attention. The encoder stack receives data such as word embeddings of the input sequence, is transferred to the first encoder, and then is modified and distributed to the next encoders that generate an output for each word/token. Finally, the output of the last encoder becomes the input of the decoder stack. Using this information, the decoders predict the next word based on self-attention mechanisms that allow the model to find other tokens in the input layer for a better recognition of a certain word in the sequence.

Due to the progresses of transformers over the recent years, there have been significant efforts to further develop these models, by giving more importance to self-attention layers and other adapted models. Thus, the use of transformers led to the achievement of important results in neural networks for Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT) [9] and GPT-2/3. Several experiments of using transformers were analyzed, being considered relevant in other additional domains, such as visualization of AI processes [10], biology [11] and medicine – BioBert ("Bidirectional Encoder Representations from Transformers for Biomedical Text Mining") [12-14], science – SciBERT (a pre-trained BERT model for scientific text) [15] and text generation using GPT-2 [16] and GPT-3 [17].

Currently, NLG systems can be evaluated with two methods: automatic evaluation [19] and human evaluation [20]. Due to progress of transformer architectures, the automatic evaluation become more popular in measuring the quality and reliability of the text generation systems, yielding a score between two sentences used for text prompts (the candidate and reference). Instead, the human

evaluation is more expensive and time-consuming due to the large corpora used in the training processes.

Nowadays, NLG systems can be used in a variety of contexts, developing frameworks that automatically generate answers or populate different outline forms, increasing the efficiency and create more personalized offers. This paper presents a new Romanian GPT-2 model, trained on a smaller corpus, representing an alternative for the existing RoGPT-2 model. The proposed model achieved better performances for long sentences, using less data for the training process.

Furthermore, the next sections describe an experiment for designing and training a Romanian GPT-2 model, called MCBGPT-2[4] and RoGPT-2 [18] model, which can be used in Romanian text generation tasks. Using only the first 100 words of news from the test and validation datasets, these models predict the next tokens generating unique instances for the above datasets. In order, to compare and evaluate the proposed model recent automatic metrics are used, such as BLEURT [21] and BERTScore [22]. Additionally, two standard automatic metrics were added in this research, such as BLEU [23], ROUGE [24].

The paper is structured as follows: Section 2 presents the related work, followed by section 3 that presents the methodology. Section 4 describes the experiments, and section 5 presents the results and discussion. Finally, section 6 describes the conclusions.

2. Related work

Using and developing neural networks approaches for natural language processing tasks need a huge number of labeled datasets. In order to perform relevant results in different domains, these datasets were used with architectures based on long-short term memory (LSTM), convolutional neural networks (CNN) or gated recurrent units (GRU) model. Several systems used the LSTM and GRU networks for automatic music composition [25] or text classification [26] with attention mechanisms [27], and some researchers used neural networks for text generation tasks providing encouraging results [28].

The transformer is a replacement for the RNNs architectures, being a neural learning model that uses the attention mechanism. These models were designed to puzzle out the problem of sequence transduction [29], or neural machine translation [30]. This phenomenon can be translated as any other task of natural language processing that transforms an input sequence into an output sequence, including essential fields such as text-to-speech transformation [31], text classification [32] and speech recognition [33]. In recent years, the transformers (e.g., BERT and GPT) have been the most used deep learning

[4] https://github.com/MCBGPT-2/Automatic-Romanian-Text-Generation-using-GPT-2/, last accessed on 29th September 2022.

models for NLP tasks and different BERT models were deployed, such as ALBERT [34] (from Google Research and Toyota Technological Institute at Chicago), RoBERT [35] (from University Politehnica of Bucharest), RoBERTa [36] (from Facebook) and DistilBERT [37] (from Hugging Face). The last one, compared to BERT, maintains 97% performance with 40% fewer parameters, achieving impressive results. Moreover, several GPT-2 models were developed such as GePpeTto [38] an Italian GPT-2 model consisting of 13GB of text, a Dutch GPT-2 model [39] or RoGPT-2, a Romanian GPT-2 model. The transformer architectures, many of whom are being provided as an open-source solution by the Hugging Face transformers library [40], have a major advantage, being highly parallelizable; therefore, larger datasets can be trained at a faster rate. Unlike BERT, which is bidirectional, the GPT models are unidirectional [41]. The most important improvement of GPT models is the number of datasets: thus GPT-3 model, the third-generation, was trained on 175 billion parameters, about 10 times the size of GPT-2 model. Having these large pre-trained models, the researchers can develop complex NLP tasks with small datasets.

Current NLG systems based on machine learning use parallel datasets, from different sources such as news, medical or IT [42] and several systems are presented by Novikova [43], such as LOLS - an imitation learning framework, by Lampouras and Vlachos [44], TGEN - a statistical natural language generator for spoken dialogue systems, by Lampouras and Vlachos [45] and RNNLG - an open source benchmark toolkit for NLG in spoken dialogue system application domains, by Wen et al. [46]. Two of the largest natural language processing artificial intelligence models is Switch Transformer from Google [47] with 1.6 trillion parameters and WuDao 2.0 from Beijing Academy of Artificial Intelligence (BAAI) with 1.75 trillion parameters[5].

Nowadays, recent text generation architectures have been developed using GPT and BERT models, machine translation and text summarization achieving important results [48]. In addition, there were papers that used both models and compared their results with various neural network techniques [49, 50] and the researchers had the opportunities to search and design even hybrid model of transformers [51] for text generation tasks such as grammatical errors correction, enabling the creation of error patterns that could be used to thoroughly clean other datasets. It is important to note that for evaluation process of the latest designed systems based on GPT models need new automatic evaluation metrics such as BLEURT and BERTScore. Several papers proposed fine-tuning processes for BLEURT [52] or BERTScore [53] to better correlate these architectures with human judgments. Moreover, in this paper two standard evaluation metrics were added, such as BLEU and ROUGE, which represent the most popular metrics used to compare NLG models [54-56].

[5] https://gpt3demo.com/apps/wu-dao-20, last accessed on 29th September 2022.

Furthermore, the next sections present a Romanian text generation system based on a GPT-2 model, trained on a small news dataset, called the MCBGPT-2 model, an existing Romanian model trained on large text datasets, named RoGPT-2 and a comparison of several automatic metrics such as BLEU, ROUGE, BLEURT and BERTScore which were applied to the generated instances. These unique instances were generated by using the MCBGPT-2 and RoGPT-2 models and news items from the test and validation datasets.

3. Methodology

Text generation is considered an important component of search engines, chatbots, text summarize, and other applications such as home assistants or smart speakers which can include text generation in some forms. In this research, the quality and quantity of datasets are responsible for the differences between original and generated instances.

The dataset was collected between March and October 2021 and it contains 24,600 news items automatically crawled from Romanian online news platforms such as agerpres.ro, defenseromania.ro or ziuadeconstanta.ro. Each article's content and title was verified for grammatical errors and classification tasks were performed, being manually labeled as fake or true news with different polarities (e.g. negative vs. positive). In this process were involved 10 employees, consisting of males and females, aged between 36 and 53 years, in a public institution in Romania

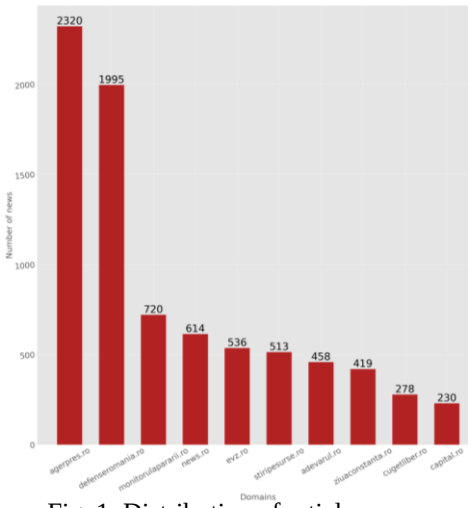

Fig. 1. Distribution of articles across news sources

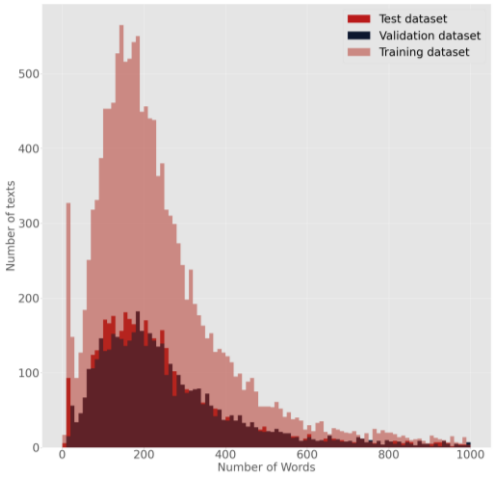

Fig. 2. Distribution of words across test,validation and training datasets

Fig. 1 presents the top news sources and the distribution of number of articles from each source, such as agerpres.ro – 2320, monitorulapararii.ro – 720, news.ro – 614 or evz.ro – 536 articles.

The aforementioned dataset was split into train (60%), validation (20%), and test (20%). The training process of the MCBGPT-2 model used 14,756 news items, while 4922 news items in each of the test and validation datasets were presented.

Fig. 2 and Table 1 presents the distributions of the test, validation, and train datasets, revealing that our datasets contain the same average words per news item (e.g., ~200 words) and the vocabulary size (unique words from The Explanatory Dictionary of the Romanian Language - DEX) of the test and validation datasets are well balanced (e.g., ~1M words), while the training dataset is about three times larger (e.g., ~3M words).

Table 1

Distribution of words across test, validation and train datasets

Words	Dataset	No.
Romanian unique words	Test	957,787
	Validation	1,026,680
	Train	2,837,236
Average words per news item	Test	194
	Validation	208
	Train	192

4. Experiments

For the experiments, 4 NVIDIA Tesla V100 GPU Accelerator with 32GB RAM and 5120 CUDA cores were used. The experiments were performed on a single server using tf.distribute.Mirrored Strategy, a TensorFlow API to distribute training across multiple GPUs and several software or packages such as CUDA (vers. 11.2), Python (vers. 3.9.7) and TensorFlow (vers. 2.4.1).

The proposed MCBGPT-2 model was trained and validated using the aforementioned news corpus, that contain 18 fake news and 24,582 true news, 368 news with negative polarity and 2135 news with positive polarity, being split in equal parts between train, test and validation datasets.

In this section, the steps followed to train the proposed model are described. The first step is to load the train dataset from specified folders (“*.txt”). The second step is to tokenize the data with Byte Level BPE (Byte Pair Encoding) [57] and save it into a folder. Two files are created, merges.txt (e.g., “just ifica”, “corp ului”, “cost uri”) and vocab.json (e.g., “perechile: 53,963”, “individuali: 53,975”, “muzicianului: 53,984”) in a specified directory. The BPE tokenizer does not evaluate several forms of word as different (e.g., “costuri” is considered as two tokens, “cost” and “uri”), keeping the similarity and meaning for “cost” and “costuri” and every word is identified by an unique id (e.g., “cuibului: 42,453”). The third step is to initialize the TensorFlow library, to slice datasets into equal parts and set the batch size (e.g., 12). The fourth step is to train the model for 15

epochs. After the model is trained, we can use it to generate a piece of news by encoding the input text and passing it to the proposed model for text generation.

For this paper, the Adam optimizer [58] is used with a small learning rate (e.g., 3 x 10^5), a loss function such as sparse categorical cross-entropy and a vocabulary size of 60,000 words. Several training statistics of the proposed model, including the hyperparameters and training time, are presented in Table 2.

Table 2

Training statistics of MCBGPT-2 model

Parameters name	Value of parameter
Number of parameters	131M
Number of epoch	15
Duration of an epoch	5h
Context size	512
Batch size	12

The RoGPT-2 model [18] is a Romanian language version of the GPT2 model, with 3 public versions available: base, medium, and large. The RoGPT-2 model was trained on a large Romanian corpus, consisting of 92.02M of sentences, 2.5 billion numbers of words with a total of 17.03GB of collected from several sources such as Wikipedia (1.79M number of sentences with 68M number of words), Oscar [59], books (37.39M number of sentences with 667M number of words) or news from online platforms such as Digi24 (e.g., https://www.digi24.ro/) or Ziarul Financiar (e.g., https://www.zf.ro/), consisting in 0.77M number of sentences with 23M number of words.

For the experiments, the MCBGPT-2 and RoGPT-2 models received as input small text prompts from the test and validation datasets. Starting from these prompts, a new instance with each model was generated and compared with the original text. Moreover, a minimum (e.g., 600 words) and maximum (e. g., 4000 words) length for the generated news was applied. Also, Table 3 presents several training statistics of the RoGPT-2 model and for this research the large version was used.

Table 3

Training statistics of RoGPT-2 model

RoGPT-2	Base	Medium	Large
Number of parameters	124M	354M	774M
Number of epoch	15	10	5
Duration of an epoch	7h	22h	45h
Context size	1024	1024	512
Batch size	72	24	16

Furthermore, for the experiments the automatic metrics have the following input parameters:

- Bilingual Evaluation Understudy Score (BLEU) – only one reference was used, the maximum n-gram length was set to four and no pre-processing techniques were applied.
- Recall-Oriented Understudy for Gisting Evaluation (ROUGE) – the best values was achieved by measuring the match-rate of unigrams (ROUGE-1).
- Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) – a BLEURT checkpoint was used – a self-contained folder that contained a regression model which was tested on several languages, but should work for the 100+ languages of multilingual C4 (a cleaned version of Common Crawl's web crawl corpus), including the Romanian language with 45M of training and 45K of validation examples. Specifically, BLEURT-20[6] was used as checkpoint, being a 32 layers pre-trained transformer model, named RemBERT, which contained 579M parameters [60] fine-tuned on human ratings and synthetic data (~590K sentence pairs) collected during years 2015 and 2019 from WMT Metrics Shared Task.
- BERTScore – "bert-base-multilingual-cased" was used as model type [61], a 12-layer transformer with token embeddings of size 768, trained by Google on the Wikipedia dumps from 104 languages, including Romanian which was explicitly added (e.g., lang="ro") and the selected number of layers was 9 (e.g., num_layers=9).

5. Results and discussion

In this paper, a new Romanian GPT-2 model is proposed, named MCBGPT-2. Based on the MCBGPT-2 and RoGPT-2 models, unique news instances were generated for the test and validation datasets. The results are presented in this section, comparing and performing a statistical analysis between original and generated news items.

Fig. 3 and Table 4 present the distribution of words of original and generated news items for the test and validation datasets. These show that the RoGPT-2 architecture generates better news instance than the proposed model in terms of average number of words (e.g., ~180 words). The texts generated by the MCBGPT-2 model are twice larger than the RoGPT-2 model and original news, which have an average number of words equal with 200 words. Fig. 4 shows several BLEU metric values (30 values) of generated instances from the validation dataset whose best score was 66.78. Additionally, the experimental analysis shows the differences in correlation between the RoGPT-2 and MCBGPT-2 models for BLEU metric.

[6] https://storage.googleapis.com/bleurt-oss-21/BLEURT-20.zip, last accessed on 29th September 2022.

For the BLEURT and BERTScore metrics, the differences in correlation are very small (e.g., gray and green in Fig. 5), however, to conclude the superiority in performances than common metrics, the minimum and maximum values of each metric are presented in Table 5. In this table, one can be observed that the BERTScore metric (max=0.9593) performs better results than ROUGE (max=0.84) and BLEURT (max=0.81).

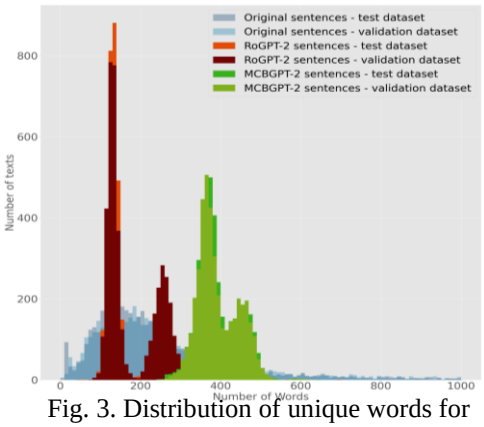

Fig. 3. Distribution of unique words for RoGPT-2 and MCBGPT-2 models

Table 4

Distribution of words across generated news items for test and validation datasets

Words	Dataset	MCB GPT-2	Ro GPT-2
Unique Words	Test	2,147,889	839,778
	Validation	2,297,947	890,837
Average words	Test	436	168
	Validation	466	180

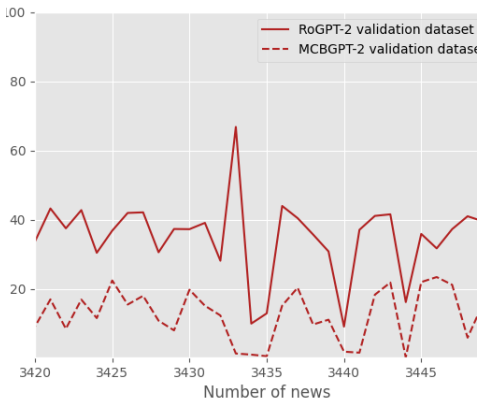

Fig. 4. BLEU metric values of generated news items (30 values) using RoGPT-2 and MCBGPT-2 models

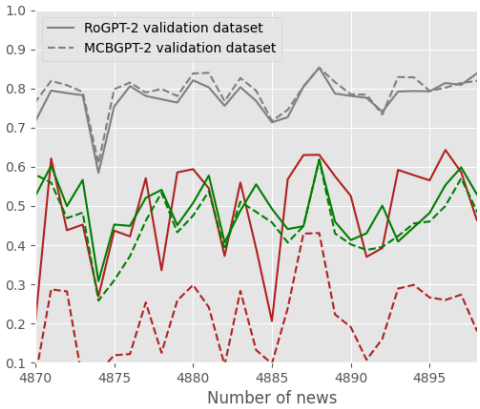

Fig. 5. ROUGE (red), BLEURT (green) and BERTScore (gray) metrics values (30 values) of generated news item

Additionally, Fig. 5 presents several values (30 values) achieved by ROUGE, BLEURT and BERTScore metrics of generated news items using RoGPT-2 (continuous line) and MCBGPT-2 (interrupted line) models.

To provide further insights, Fig. 6 presents the distribution of ROUGE, BLEURT, and BERTScore metrics values of generated news items from the validation dataset using the MCBGPT-2 model. The plot shows that the BERTScore metric is more consistent than other metrics, achieving a narrow and

higher interval of values, being recommended to be used for evaluating Romanian text generation systems.

From the results of this paper (see Table 6 and Table 7), it can be observed that BERTScore metric provides better performances for MCBGPT-2 than RoGPT-2 model. Additionally, the results show that the MCBGPT-2 and RoGPT-2 models provide similar performances for BLEURT, using less data for the MCBGPT-2 model's training process.

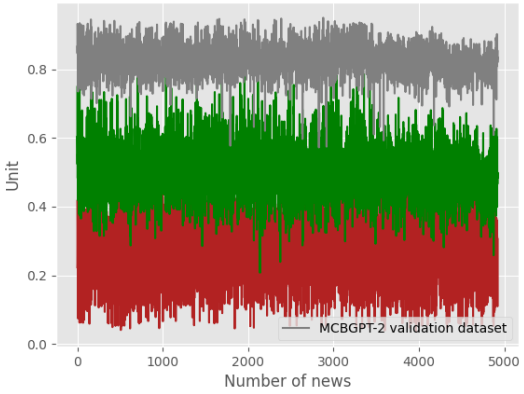

Fig. 6. Distribution of ROUGE (red), BLEURT (green) and BERTScore (gray) metrics values of generated news items

Table 5

Automatic metrics values of generated news items for the validation dataset using RoGPT-2 and MCBGPT-2 models

Metric	Model	Min.	Max.
BLEU	RoGPT-2	0.0055	66.78
	MCBGPT-2	0.0001	26.75
ROUGE	RoGPT-2	0.1009	0.8445
	MCBGPT-2	0.0373	0.6176
BLEURT	RoGPT-2	0.0706	0.8119
	MCBGPT-2	0.2075	0.8035
BERT Score	RoGPT-2	0.5827	0.9593
	MCBGPT-2	0.5793	0.9506

Table 6

Scores of generated news items using RoGPT-2 and MCBGPT-2 models for the test and validation datasets

Model name	Dataset	BLEU	ROUGE	BLEURT	BERTScore
RoGPT-2	Test	32.29	0.53	0.68	0.8106
	Validation	28.70	0.50	0.52	0.8136
MCBGPT-2	Test	8.79	0.25	0.63	0.8124
	Validation	9.11	0.14	0.50	0.8277

For the experiments, in the evaluation process there are two other hyperparameters that can be mentioned such as temperature which is used to scale the probabilities of a given word or top p filtering used to sort the word probabilities in descending order. Further, it should also add the human subjective judgments in analyzing and evaluation process for our corpus. Additionally, the previously mentioned employees were included in a short qualitative experiment. Each evaluator received 10 original sentences, 10 sentences generated by the MCBGPT-2 model and 10 sentences generated by the RoGPT-2 model from test and evaluation datasets. The generated sentences were randomly selected based on the length of sentences which contains 1000 and 2000 words and BERTScore metric for MCBGPT-2 is higher than RoGPT-2 model. The evaluators annotated the most logical sentences based on subjective principles and the result achieved by the proposed MCBGPT-2 model was 84%, compared with RoGPT-2 model which achieved 81%. The results reflect that our model is efficient for longer

sentences, sustaining the quantitative scores achieved by BERTScore metric (as shown in Table 6). The future studies will include only qualitative evaluation based on the same grammar and coherence principles for all evaluators and sentences from test and evaluation datasets.

6. Conclusions

In this research, two GPT-2 architectures have been tested and evaluated, generating news items based on short Romanian text prompts. Using automatic metrics such as BLEU, ROUGE, BLEURT and BERTScore, the MCBGPT-2 and RoGPT-2 models are compared, thus, providing a better solution for Romanian text generation systems and higher performances in long sentences for the MCBGPT-2 model.

The RoGPT-2 outperforms MCBGPT-2 model, as shown by length and metrics score (e.g., BLEU, ROUGE and BLEURT) of the generated sentences, however, the MCBGPT-2 model achieves slightly better scores for the BERTScore metric. The MCBGPT-2 and RoGPT-2 models can be very useful for the Romanian companies and businesses in choosing the AI processes to automate multiple tasks that employ online or social media articles. Additionally, we encourage public institutions to use AI technologies [62, 63] for effective defense in front of the most dangerous online phenomenon such as fake news.

Table 7

Examples of Romanian generated news (Appendix - English version)

State	Text			
Original	Coordonatorul campaniei naționale de vaccinare, medicul militar Valeriu Gheorghiță a vorbit despre finalul pandemiei în România.Întrebat, într-o emisiune televizată, când crede că se va reveni la viața normală, doctorul Gheorghiță a spus Eu sper cât mai repede. Dacă mă întrebați de o lună anume, cred că undeva în iulie august ușor ușor vom intra într-o normalitate care va fi din ce în ce mai concretizată pe măsură ce nivelul de acoperire vaccinală va crește undeva peste *a spus medicul Valeriu Gheorghiță.*			
Link	https://www.ziarulprofit.ro/index.php/cand-se-va-termina-pandemia-in-romania-doctorul-valeriu-gheorghita-da-raspunsul-video/			
	Generated text			
Ro GPT-2 model	*….a spus medicul Valeriu Gheorghiță*, care a mai spus că s-a întâlnit cu ministrul sănătății, Sorina Pintea, pentru a discuta despre starea de sănătate a populației și despre măsurile care se vor lua pentru ca să nu se mai ajungă la astfel de situații, iar ministrul a promis că va avea o discuție cu el și va vedea care sunt măsurile pe care ministerul le va lua, dar și ce măsuri vor fi luate de către minister, atât pentru prevenirea, cât și pentru tratarea celor care au fost depistați cu COVID-19 și care nu au putut fi vaccinați, precum și măsuri de combatere a epidemiei de coronavirus și a altor boli care pot fi transmise prin contact direct cu o persoană infectată cu virusul imunodeficienței umane (HIV).			
Metrics	BLEU	ROUGE	BLEURT	BERTScore
Scores	22.43	0.39	0.83	0.8487
MCB	*….a spus medicul Valeriu Gheorghiță*. În ceea ce privește vaccinarea cu cea de-a			

GPT-2 model	treia doză de vaccin, medicul a explicat că, într-adevăr, nu se pune problema de a face o formă gravă de boală, dar nu este un lucru foarte important pentru noi toți cei care se vaccinează cu prima schemă cu vaccinul de la compania Johnson Johnson, pentru că nu avem niciun fel de imunitate de grup peste 40%, iar pentru persoanele care au fost vaccinate cu schemă completă au un risc scăzut de cel puțin 20 de ori să se infecteze și cu o creștere a numărului de persoane care s-au vaccinat cu schema completă. De asemenea, el a mai spus că este important să avem o rată de pozitivare a testelor de maxim 20 la sută, a adăugat medicul.			
Metrics	BLEU	ROUGE	BLEURT	BERTScore
Scores	20.74	0.38	0.75	0.8588

Text generation systems have also been explored in prior studies and supplementary data can be found in papers focused on generating short sentences that contain hundreds of tokens (e.g., 350 tokens) [64, 65] or longer sentences (e.g., 1000 tokens) [66]. Also, to achieve better performances for text generation systems it is necessary to develop more GPT-2 models and, considering the quantitative evaluation, the presented model obtained slightly better scores for long sentences, using a smaller dataset that contains specific news with particularly subjects from the online environment.

We expect to see more papers in this field due to the progress of technologies and frameworks that will integrate Romanian GPT-2 architectures for different purposes and scope. Focusing on the main elements, such as corpora and fine-tuning processes, we will continue to develop the proposed model to create a better system for NLG tasks.

R E F E R E N C E S

[1]. *A. Radford, J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever*, "Language Models are Unsupervised Multitask Learners," OpenAI blog, **vol. 1**, no. 8, p. 9, 2019.

[2]. *T. B., Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan,P. Shyam,G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss,G. Krueger, T.J. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei,* "Language Models are Few-Shot Learners," 2020.

[3]. *R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner and Y. Choi*, "Defending against neural fake news," in NeurIPS, 2019, arXiv:1905.12616.

[4]. *T. Fagni, F. Falchi, M. Gambini, A. Martella and M. Tesconi,* "TweepFake: About detecting deepfake tweets," PLoS ONE, **vol. 16**, no. 5, 2021, DOI: 10.1371/journal.pone.0251415.

[5]. *M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy and L. Zettlemoyer*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019 in ACL, pp. 7871–7880, DOI: 10.18653/v1/2020.acl-main.703.

[6]. *S. Judith and M. Jakub*, "Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation," in 16th International Conference on the Foundations of Digital Games (FDG'21), Association for Computing Machinery, New York, NY, USA, **vol. 2**, 2021, pp. 1–8, DOI: 10.1145/3472538.3472595.

[7]. *T. Klein and M. Nabi,* "Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds", 2019, arXiv: 1911.02365.

[8]. *A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin*, "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 5998–6008, arXiv: 1706.03762.

[9]. *J. Devlin, M. W. Chang, K. Lee and K. Toutanova*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, **vol. 1**, 2018, DOI: 10.18653/v1/N19-1423.

[10]. *A. M. P. Braşoveanu and R. Andonie*, "Visualizing Transformers for NLP: A Brief Survey," in the 24th International Conference Information Visualisation (IV), 2020, pp. 270-279, DOI: 10.1109/IV51561.2020.00051.

[11]. *A. Rives, S. Goyal, J. Meier, D.Guo, M. Ott, C L. Zitnick, J. Ma and R. Fergus*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," in Proceedings of the National Academy of Sciences, 2021, DOI: 10.1073/pnas.2016239118.

[12]. *A. Madani, B. McCann, N. Naik, N.S. Keskar, N. Anand, R.R. Eguchi and R. Socher*, "Progen: Language modeling for protein generation," 2020, arXiv: 2004.03497.

[13]. *J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, Chan H. S. and J. Kang*, "Biobert: pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, **vol. 36**, no. 4, 2020, pp. 1234-1240, DOI: 10.1093/bioinformatics/btz682.

[14]. *K. Wang, B. Yang, G. Xu and X. He*, "Medical question retrieval based on siamese neural network and transfer learning method," in International Conference on Database Systems for Advanced Applications, Springer, 2019, pp. 49–64, DOI:10.1007/978-3-030-18590-9_4.

[15]. *I. Beltagy, A. Cohan and K. Lo*, "Scibert: Pretrained contextualized embeddings for scientific text," 2019, arXiv: 1903.10676.

[16]. *Y. Qu, P. Liu, W. Song, L. Liu and M. Cheng*, "A Text Generation and Prediction System: Pre-training on New Corpora Using BERT and GPT-2," in IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIE/; C), 2020, pp. 323-326, DOI: 10.1109/ICEIEC49280.2020.9152352.

[17]. *A. Uchendu, Z. Ma, T. Le, R. Zhang and D. Lee*, "TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation," 2021, EMNLP, arXiv: 2109.13296.

[18]. *M. A. Niculescu, S. Ruseti and M. Dascalu*, "RoGPT2: Romanian GPT2 for Text Generation," in 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), 2021, pp. 1154-1161, DOI:10.1109/ICTAI52525.2021.00183.

[19]. *N. Mathur, T. Baldwin and T. Cohn*, "Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics," 2020, arXiv: 2006.06264,.

[20]. *C. Van Der Lee, A. Gatt, E. Van Miltenburg, S. Wubben, S. and E.Krahmer*, "Best practices for the human evaluation of automatically generated text," in Proceedings of the 12th International Conference on Natural Language Generation, 2019, pp. 355-368, DOI: 10.18653/v1/W19-8643.

[21]. *T. Sellam, D. Das and A.P. Parikh*, "BLEURT: Learning Robust Metrics for Text Generation," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics - ACL, 2020, pp. 7881-7892, DOI: 10.18653/v1/2020.acl-main.704.

[22]. *T. Zhang, V. Kishore, F. Wu, K.Q.Weinberger, Y. Artzi*, "BERTScore: Evaluating Text Generation with BERT," 2019, arXiv: 1904.09675.

[23]. *K. Papineni, S. Roukos, T. Ward and W.J. Zhu*, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311-318, DOI: 10.3115/1073083.1073135.

[24]. *C. Y. Lin*, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, Barcelona, Spain, Association for Computational Linguistics, 2004, pp. 74-81.

[25]. *K. Choi, G. Fazekas and M. Sandler*, "Text-based LSTM networks for Automatic Music Composition," in 1st Conference on Computer Simulation of Musical Creativity, 2016, arXiv: 1604.05358.

[26]. *G. Liu and J. Guo*, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," Neurocomputing, **vol. 337**, 2019, pp. 325-338, DOI: 10.1016/j.neucom.2019.01.078.

[27]. *X. Bai*, "Text classification based on LSTM and attention," in Thirteenth International Conference on Digital Information Management (ICDIM), 2018, pp. 29-32, DOI: 10.1109/ICDIM.2018.8847061.

[28]. *S. Santhanam*, "Context based Text-generation using LSTM networks," 2020, arXiv: 2005.00048.

[29]. *A. Graves*, "Sequence Transduction with Recurrent Neural Networks," in International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning, 2012, arXiv: 1211.3711.

[30]. *P. Koehn and R. Knowles*, "Six Challenges for Neural Machine Translation," 2017, pp. 28-39, arXiv: 1706.03872.

[31] *T. Saeki, S. Takamichi and H. Saruwatari*, "Incremental Text-to-Speech Synthesis Using Pseudo Lookahead with Large Pretrained Language Model," in IEEE Signal Processing Letters, **vol. 28**, 2021, pp. 857-861, DOI: 10.1109/LSP.2021.3073869.

[32] *A. A. Irissappane, H. Yu, Y. Shen, A. Agrawal and G. Stanton*, "Leveraging GPT-2 for Classifying Spam Reviews with Limited Labeled Data via Adversarial Training," CS - Artificial Intelligence, 2020, arXiv: 2012.13400.

[33] *D. Fohr and I. Illina*, "BERT-based Semantic Model for Rescoring N-best Speech Recognition List," Computer Science/Human-Computer Interaction, Available: https://hal.sorbonne-universite.fr/INRIA/hal-03248881, 2021.

[34] *Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut*, "ALBERT: A lite BERT for self-supervised learning of language representations," in 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, OpenReview.net, Available: https://openreview.net/forum?id=H1eA7AEtvS, April 26-30, 2020, arXiv: 1909.11942.

[35] *M. Masala, S. Ruseti and M. Dascalu*, "RoBERT - A Romanian BERT Model," in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6626-6637, DOI: 10.18653/v1/2020.coling-main.581.

[36] *Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, arXiv: 1907.11692.

[37] *V. Sanh, L. Debut, J. Chaumond and T. Wolf*, "Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter," 2019, arXiv: 1910.01108.

[38]. *L. De Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim and M. Guerini*, "Geppetto carves italian into a language model," 2020, arXiv: 2004.14253.

[39]. *W. De Vries and M. Nissim*, "As good as new. How to successfully recycle English GPT-2 to make models for other languages," 2020, arXiv: 2012.05628.

[40]. *T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz and J. Brew*, "Huggingface's transformers: State-of-the-art natural language processing," 2019, CoRR, arXiv: 1910.03771.

[41]. *M. Topal, A. Bas and Imke van Heerden*, "Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet," 2021, arXiv: 2102.08036.

[42]. *J. Novikova, O. Lemon and V. Rieser*, "Crowd-sourcing NLG data: Pictures elicit better data," In Proceedings of the 9th International Natural Language Generation Conference, Edinburgh, UK, 2016, pp. 265–273, arXiv: 1608.00339.

[43]. *J. Novikova, O. Dušek, A. C. Curry and V. Rieser*, "Why we need new evaluation metrics for NLG," 2017, arXiv: 1707.06875.

[44]. *G. Lampouras and A. Vlachos*, "Imitation learning for language generation from unaligned data," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, the COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 1101–1112, 2016.

[45]. *O. Dušek and F. Jurcicek*, "Training a natural language generator from unaligned data," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, **vol. 1**: Long Papers, Beijing, China, 2015, pp. 451–461, DOI: 10.3115/v1/P15-1044.

[46]. *T. H. Wen, M. Gasic, N. Mrksic, P. H., Su, D. Vandyke and S. Young*, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," 2015, arXiv: 1508.01745.

[47]. *W. Fedus, B. Zoph and N. Shazeer,* "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," 2021, arXiv: 2101.03961.

[48]. *Y. C. Chen, Z. Gan, Y. Cheng and J. Liu*, "Distilling the knowledge of BERT for text generation," 2020, Available: https://openreview.net/pdf?id=Bkgz_krKPB.

[49] *D. M. Montesinos*, "Modern Methods for Text Generation," 2020, arXiv: 2009.04968.

[50] *J. S. Lee and J. Hsiang*, "Patent claim generation by fine-tuning OpenAI GPT-2," 2019, arXiv: 1907.02052.

[51] *L. Wang, W. Zhao, R. Jia, S. Li and J. Liu*, "Denoising based Sequence-to-Sequence Pre-training for Text Generation," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, arXiv: 1908.08206.

[52]. *S. V. Mehta, J. Rao, Y. Tay, M. Kale, A. Parikh, H. Zhong and E. Strubell*, "Improving Compositional Generalization with Self-Training for Data-to-Text Generation," 2021, arXiv: 2110.08467.

[53]. *C. R. Chan, C. Pethe and S. Skiena*, "Natural language processing versus rule-based text analysis: Comparing BERT score and readability indices to predict crowd funding outcomes," Journal of Business Venturing Insights, **vol. 16**, 2021, DOI: 10.1016/j.jbvi.2021.e00276.

[54]. *M. Kilickaya, A. Erdem, N. Ikizler-Cinbis and E. Erdem*, "Re-evaluating automatic metrics for image captioning," 2016, arXiv: 1612.07600.

[55]. *A. Yang, K. Liu, J. Liu, Y. Lyu and S. Li*, "Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task," 2018, arXiv: 1806.03578.

[56]. *J. Wieting, T. Berg-Kirkpatrick, K. Gimpel and G. Neubig*, "Beyond BLEU: training neural machine translation with semantic similarity," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4344-4355, arXiv: 1909.06694.

[57]. *R. Sennrich, B. Haddow and A. Birch*, "Neural machine translation of rare words with subword units," 2015, arXiv: 1508.07909.

[58]. *D. Kingma and J. Ba*, "Adam: A Method for Stochastic Optimization, International Conference on Learning Representations," 2015, arXiv: abs/1412.6980.

[59]. *P. J. O. Suarez, B. Sagot and L. Romary*, "Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures," in 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), 2019, DOI: 10.14618/IDS-PUB-9021.

[60] *H. W. Chung, T. Févry, H. Tsai, M. Johnson and S. Ruder*, "Rethinking embedding coupling in pre-trained language models," in International Conference on Learning Representations, 2020, arXiv: 2010.12821.

[61]. *V. Gopalan and M. Hopkins*, "Reed at SemEval-2020 Task 9: Fine-Tuning and Bag-of-Words Approaches to Code-Mixed Sentiment Analysis," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1304-1309, arXiv: 2007.13061.

[62]. *M. C. Buzea, S. Trausan-Matu and T. Rebedea*, "Automatic Fake News Detection for Romanian Online News," Information, **vol. 13**, no. 3, 2022, p. 151, DOI: 10.3390/info13030151.

[63]. *C. Busioc, V. Dumitru, S. Ruseti, S. Terian-Dan, M. Dascalu and T. Rebedea*, "What Are the Latest Fake News in Romanian Politics? An Automated Analysis Based on BERT Language Models," in Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education: Proceedings of the 6th International Conference on Smart Learning Ecosystems and Regional Development, Bucharest, Romania, 24–25 June 2021, Springer Nature: Singapore, **vol. 249**, 2021, p. 201, DOI:10.1007/978-981-16-3930-2_16.

[64]. *A. Holtzman, J. Buys, L. Du, M. Forbes and Y. Choi*, "The curious case of neural text degeneration," 2019, arXiv: 1904.09751.
[65]. *A. See, A. Pappu, R. Saxena, A. Yerukola and C.D. Manning*, "Do massively pretrained language models make better storytellers?," 2019, arXiv: 1909.10705.
[66]. *B. Tan, Z. Yang, M. AI-Shedivat, E.P. Xing and Z. Hu*, "Progressive Generation of Long Text with Pretrained Language Models," 2020, arXiv: 2006.15720.

APPENDIX

English translation of the Romanian generated news items

Original text: The coordinator of the national vaccination campaign, military doctor Valeriu Gheorghiță, talked about the end of the pandemics in Romania. When asked, in a television broadcast, about the moment life would go back to normal, doctor Gheorghiță said: I hope as soon as possible. If you ask me about a certain month, I believe that somewhere in July or August we'll easily go back to normal which can be materialized as the level of vaccination coverage increases somewhere over *doctor Valeriu Gheorghiță said.*

RoGPT-2 generated text: *....doctor Valeriu Gheorghiță said*, who also mentioned that he met with the Minister of Health, Sorina Pintea, in order to talk about the health status of the population and the measures that will be taken so that these situations can be avoided, while the minister promised he would talk to him and see what are the measures the ministry will take, but also the steps that are to be taken by the ministry in order to both prevent and treat those people infected with COVID-19 and who could not have been vaccinated, as well as the measures taken to combat the coronavirus pandemics and other diseases that can be transmitted by direct contact with an person infected with the Human Immunodeficiency Virus (HIV).

MCBGPT-2 generated text: *....doctor Valeriu Gheorghiță said.* As regards the vaccination with the third dose, the doctor explained that, it is true, one cannot get a serious form of disease, but it is not a very important thing for all of us who get vaccinated with the first dose of Johnson Johnson vaccine, because we have no herd immunity achieved over 40%, while for the people who got vaccinated with two doses have 20 times less chances to get infected and with an increase of the number of people who completed their vaccination schedule. Also, he said that it is important to have a maximum 20% rate of positivation, the doctor added.