Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

Moscow, June 14–17, 2023

Exploring Evaluation Techniques in Controlled Text Generation: A Comparative Study of Semantics and Sentiment in ruGPT3large-Generated and Human-Written Movie Reviews

Anastasia Margolina
HSE University
Saint-Petersburg
avmargolina@edu.hse.ru

Anastasia Kolmogorova
HSE University
Saint-Petersburg
akolmogorova@hse.ru

Abstract

The paper describes the proposed strategy for evaluation controlled text generation with the sentiment as attribute. Our approach mainly consists of automatic sentiment analysis (ruBERT) and topic modelling (BERTopic), which are applied to a parallel corpus with artificially produced and human-written texts. The model for evaluation is fine-tuned on the parsed reviews from big Russian movie-related website ruGPT3Large with the sentiment as prompt. The results of the analysis demonstrate that the proposed methods can offer a more comprehensive understanding of the advantages and limitations in the context of semantics and sentiment. Additionally, the paper employs metrics such as BERTscore and self-BLEU to further evaluate the generated text. The proposed methodology provides a novel approach for evaluating the quality of generated text and may have implications for future studies in the field.

Keywords: controlled text generation, strategy for quality measurement, topic modelling, sentiment-analysis, movie reviews.

DOI: 10.28995/2075-7182-2023-22-1082-1090

Оценка контролируемой генерации текста: сравнительное исследование семантики и сентимента в отзывах на фильмы, написанных ruGPT3large и человеком

Марголина А.В.
НИУ ВШЭ
Санкт-Петербург
avmargolina@edu.hse.ru

Колмогорова А.В.
НИУ ВШЭ
Санкт-Петербург
akolmogorova@hse.ru

Аннотация

В статье предлагается новая стратегия оценки контролируемой генерации текста с тональностью в качестве атрибута. Наш подход включает автоматический анализ тональности (ruBERT) и тематическое моделирование (BERTopic). Эти инструменты применяются к параллельному корпусу, состоящему из пар "сгенерированный отзыв – реальный отзыв". Модель используемая для оценки – ruGPT3Large, которая была ранее дообучена на собранных с Кинопоиска отзывах на фильмы с тональностью "вшитой" в затравку. Результаты анализа демонстрируют, что использованные методы предлагают более полное понимание преимуществ и ограничений в контексте семантики и эмоциональной окраски языковой модели. Кроме того, в статье применяются такие метрики, как BERTscore и self-BLEU, для дополнительной оценки сгенерированного текста. Наша методология представляет новый подход для оценки качества генерируемого текста и может дать основу для будущих исследований в этой области. Ключевые слова: контролируемая генерация текста, стратегия оценки качества, тематическое моделирование, сентимент-анализ, кинорецензии.

Ключевые слова: контролируемая генерация текста, стратегия измерения качества, тематическое моделирование, анализ тональности, отзывы на кино

1 Introduction

In this paper we tackle the problem of controlled generation of text in Russian. Our experiments concern such a text genre as movie reviews and the attribute we initially control while generating is text sentiment.

To discuss the problem a few challenges, need to be consequently addressed. First of them - how to evaluate the quality of generation. Despite the exponential growth in the number of pre-trained generative language models (LMs), the problem of accurate metrics for measuring generated text quality persists. There are no studies that aim to explore the artificially made texts, although it could potentially reveal unseen differences and similarities between 'made-up' texts and the 'actual' ones and it could be used as a peculiar metric for the evaluation of the semantic quality of generated texts. This is linked to a certain limitation: to make such research happen one needs a generative model, which is fine-tuned on downstream tasks, and the dataset of real texts that can be directly compared to the dataset of produced data.

In our case, we analyse the effectiveness of two unsupervised metrics (BERTscore and Self-BLEU) and display the results of our experiments when applying Topic Modelling (TM) and Sentiment Analysis methods to compare sentences in two parallel corpora of movie reviews in Russian having the same prompts: written by human users and generated by fine-tuned ruGPT3Large model.

We focus on differences between human generated and AI generated texts of a specific genre. In this context, TM is not only the tool for linguistic research of the overall structure of movie reviews but it is also a strategy for evaluation how well, comparatively to humans, does the model construct the narrative. The Sentiment Analysis use is sanctioned by our desire to compare not the correctness of the label assigned to text by machine, but to verify its adequateness to human subjective expression in analogous text.

Our hypothesis is formulated as following:

1. More discrepancies we observe between topics having the most important weights in human-written texts and AI-generated texts, less qualitative is the controlled generation. To assess the degree of deference in topics we use values of Cosine similarity distance between vectorized representations of topics.

2. The higher is the difference between accuracy values returned by classification when estimating the sentiment in two parallel corpora, the less qualitative is the controlled generation.

2 Related Papers

One of the main challenges in evaluating the quality of generated text is the lack of accurate metrics. It is caused by several factors. First, the evaluation is commonly conducted in a reference-free setting because it is challenging to collect sufficient high-quality references for each input of control variables in this open-ended text generation task (Dathathri et al., 2020). This led to the situation when the majority of existing metrics measures the similarity of generated text against human-written references. Such metrics can be classified into unsupervised, supervised, and human evaluation-based methods, each with its own limitations and advantages. To overcome the shortcomings of classical single-score BLEU (Papineni et al., 2002), researchers propose a family of interpretable metrics for the key aspects of diverse tasks (summarization, style transfer, and dialogue) which either don't require human references (Deng et al., 2022) or can model human assessment with rother high accuracy (Sellam et al., 2020).

An interesting approach tested on Russian language data was suggested in (P. A. et al., 2022): within the RuATD Shared Task 2022 the authors propose to use binary classification methodology designed to detect AI-generated texts to filter well-generated texts (with the high number of false positives in generated texts classified as human written) from bad-generated.

The benchmarking platform to support research on open-domain text generation models Texygen (Zhu et al., 2018)also provides several groups of metrics: Document Similarity based Metrics, Likelihood-based Metrics and Divergence based Metrics (in our experiments we use one of them too).

In contrast with the mentioned above metrics, we suggest two metrics, mostly qualitative, but having a quantitative support, to evaluate the semantic and emotional consistency of human-generated and AI-generated movie reviews in Russian using TM and Sentiment Analysis methods.

3 Data and Methodology

The parallel dataset consists of 1200 actual reviews and 1200 generated reviews. The reviews made by the model were generated according to prompt, which is 5-6 words in the beginning of the actual review and the corresponding sentiment, which is a controlling attribute for the text generation. The dataset is normalised in the context of the sentiment: 400 reviews for each (positive, neutral and negative).

Topic modelling is widely applied in exploratory analysis as a tool for extracting hidden semantic relationships, topics in the set of textual data. Many researchers use this method to analyse not only social data but also literature in order to find covert patterns (Schöch, 2016; Ordun et al., 2020; Sherstinova et al., 2022). Needless to say, all this data is human-written: either the author of some book or the dataset of short-texts from twitter.

In this research such an implementation of topic modelling towards generative language models is presented. We explore the semantic distance between two sets of movie reviews: one generated by ruGPT3Large and one with reviews written by the users. This paper employs BERTopic, an unsupervised topic modelling technique (Grootendorst, 2022), to conduct exploratory data analysis on two distinct datasets of movie reviews. The analysis is conducted in three steps. In the first step, we analyse a dataset of real reviews, identifying overarching trends and topics in web-reviews on films. In the second one, we apply BERTopic to a dataset of generated reviews, revealing the typical semantic net of artificially produced reviews. Finally, we compute a cosine similarity distance between the vectorised representations of topics.

Aside from semantic validation of generated text, the important goal is to investigate whether the model creates appropriate texts in terms of controlling attribute, the sentiment. There are no sustainable metrics for the evaluation of controlled text generation tasks except human assessment. Nevertheless, the psycholinguistic experiment is time and resource consuming. To address this issue, we propose the use of automatic sentiment analysis on the parallel corpora, comparing the given sentiment of texts that have the same prompt and claimed sentiment. For this task the ruBERT[1] was fine-tuned with the movie reviews dataset (60k reviews, 20k for each sentiment) for the multilabel classification task. The final model achieves the accuracy of 95 percent on the test data.

4 ruGPT3Large Fine-Tune

The Russian version of GPT – the ruGPT3Large model[2] was chosen for the experiment. The architecture of ruGPT-3 is similar to that of GPT-2: it is a decoder-only transformer-based model, which makes it perfect for text generation (Radford et al., 2019). The data for fine-tuning was collected on Russian-language movie-related website. Aside from the text, the sentiment of the review was also parsed in order to then make a sample less biased. The original dataset consists of 199k reviews (148k positive, 28k neutral and 21k negative) but it stratified for training: 60k total number of reviews, 20k for each sentiment.

Data was transformed from a csv table to a textual file with prompts for model input. This format uses line breaks to separate reviews and special characters to mark start (<s>) and end (</s>) of each string. The structured data looks as follows:

> <s>Тональность: [позитивная, нейтральная или негативная]\nТекст: [текст отзыва]</s>
>
> Translation: <s>Sentiment: [positive, neutral or negative]\nText: [the text of the review]</s>

The objective of incorporating reviews with prompt into the model is to facilitate the memorization of patterns by the ruGPT3. This is achieved by utilising the second segment of the prompt, which serves as a continuation that the model must generate, namely, the review itself. The data then was split with the ratio of 0.3/0.7 for test/train.

[1] https://huggingface.co/Tatyana/rubert-base-cased-sentiment-new

[2] https://github.com/ai-forever/ru-gpts

Table 1 showcases the parameters selected for the fine-tuning of ruGPTLarge to accommodate GPU memory constraints. Opting for the minimal batch size, as the table indicates, enhances the stability of training at the expense of per-step computation efficiency (Li et al., 2022). The learning rate adheres to the default setting. These configurations allowed the large model to complete fine-tuning within a time frame of six and a half hours using the GPU.

Parameter	**Value**
num train epochs	1
per device train batch size	1
per device eval batch size	1
block size	1024
learning rate	2.5e-4

Table 1: ruGPT3Large finetuning parameters.

The perplexity of the fine-tuned model, trained on movie reviews, has been measured to be 19.

Table 2 presents the parameter of generation. We choose them in order to make the model generate more 'creative' and less repetitive (temperature, top k, repetition penalty).

Parameter	**Value**
repetition penalty	5.0
top p	0.95
top k	5
temperature	1
no repeat ngram size	2.0

Table 2: ruGPT3Large generation parameters.

The result is evaluated by two metrics: BERTscore (Zhang et al., 2020) and Self-BLEU (Zhu et al., 2018).

The first one computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. In the context of my research this score shows how different the LM generates the output compared to the original text with the same prompt. The median BERTscore for the parallel dataset (original review – generated review) of 1200 equals 69.7 percent. It means that the model generates similar to the original, however it tends to create its own narratives. At Table 3 we show the example of our parallel corpora: the text column contains human written texts, and the result column contains AI-generated texts. As can be seen they both have the same prompts.

The BLEU score, a classical metric for text evaluation, is not suitable for our task. It is designed for text-to-text tasks like machine translation, measuring translation accuracy. However, our research focuses on creative text generation, where the output can be diverse. To assess text diversity, we utilize a t metric called self-BLEU score.

This metric shows how diverse the output of the generated model is: 'a higher Self-BLEU score implies less diversity of the document, and more serious mode collapse of the GAN model' (Zhu et al., 2018, p. 4). This score is also calculated on the parallel dataset. To evaluate the quality of both types of texts, we computed the self-BLEU-3 metric for each review, which reflects more structural diversity, and calculated the mean values across the entire dataset. Our results, presented at Table 4, reveal that while the metric is slightly better for the neural-generated texts, both the human-written and generated texts exhibit similar levels of diversity.

This finding is surprising, given the longstanding perception that human-written texts are of higher quality than their machine-generated counterparts (Hardcastle and Scott, 2008). However, we attribute the small difference in favour of the generated texts to the parameters used in the generation process, namely the repetition penalty and the no-repeat n-gram size. These parameters are designed to penalise

sentiment	**Generated reviews**	**Human reviews**
нейтральная [neutral]	Лично я ожидал большего. Фильм явно не дотягигивает до того уровня, на который предендует. [Personally, I expected more. The film is clearly not up to the level it's supposed to be.]	Лично я ожидал большего. Фильм явно не доработан, не хватает продуманного смысла. [I was expecting more of the same. The film is clearly incomplete, lacking thoughtful meaning.]
положительная [positive]	Перед нами довольно бесцветная и тревожная, но в то же время захватвывающая история о том, как группа людей...[We have before us a rather colorless and disturbing, but at the same time gripping, story of how a group of people...]	Перед нами довольно бесцветная и тревожная, но тем временем, жизнеутверждающая история...[We are faced with a rather colorless and disturbing, but in the meantime, life-affirming story...]
отрицательная [negative]	Признаюсь фильм решила посмотреть из-за трейлера. Он меня очень впечатлил и я ожидала от него чего-то невероятного. [I admit that I decided to watch the movie because of the trailer. It really impressed me and I was expecting something incredible from it.]	Признаюсь фильм решила посмотреть из-за трейлера. Вторая ошибка моей жизни.[I admit that I decided to watch the movie because of the trailer. The second mistake of my life.]

Table 3: Example of our parallel corpora of texts with the same prompts.

the model for repeating words and sequences, thereby encouraging the model to produce more diverse texts.

To explore semantic features of generated texts, we suggest the strategy based on Topic Modelling and Sentiment Analysis methods.

	Human-written reviews	**Artificial reviews**
mean	0.074323	0.032231
max	0.020033	0.011562
min	0.209597	0.127813

Table 4: Self-BLEU metric applied to human-written and neural-generated texts.

5 Controlled Text Generation Evaluation

5.1 Topic Modeling

This study aims to compare the topics extracted from generated by artificial intelligence (AI) texts and human-written texts using the BERTopic algorithm with multilingual embedding model (Reimers and Gurevych, 2019). The results, presented in Figures 1 and 2, reveal notable differences in the topics discussed in these two sets of texts.

The topics showed in Figure 1 reveal the most frequent topics in our dataset of 1200 human-written movie reviews. Although most topics are not very interpretable, there are several clusters that can be analyzed. Reviewers tend to focus their comments on various aspects of the movie's production, such as the soundtrack, acting performances, and visual effects.

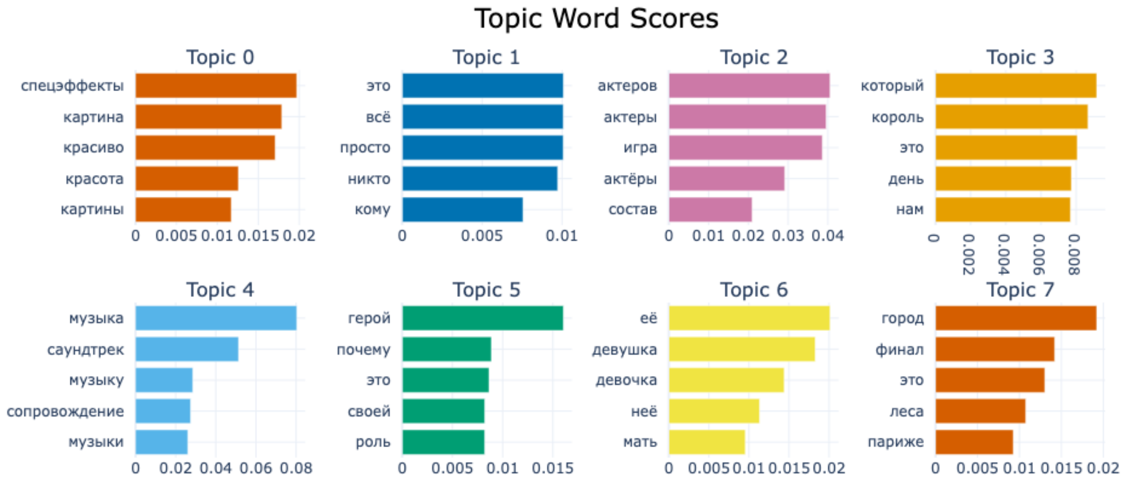

Figure 1: The most frequent topics in the dataset of human-written movie reviews.

We can also see that people tend to focus on the setting and main characters (Topics 1, 3, 6, and 7). However, the descriptions provided by reviewers are often vague and generalised, lacking details or named entities that could enhance nuance of their analysis.

On the other hand, the topics identified in AI-generated movie reviews focus on plot elements, action scenes, and character descriptions (Topic 1, 2 and 5). This suggests that the LM has a better grasp of narrative elements and character development. Interestingly, one named entity, Hans Zimmer, is present in the "musical" topic, which could indicate that the LM has prior knowledge of famous people in the domain.

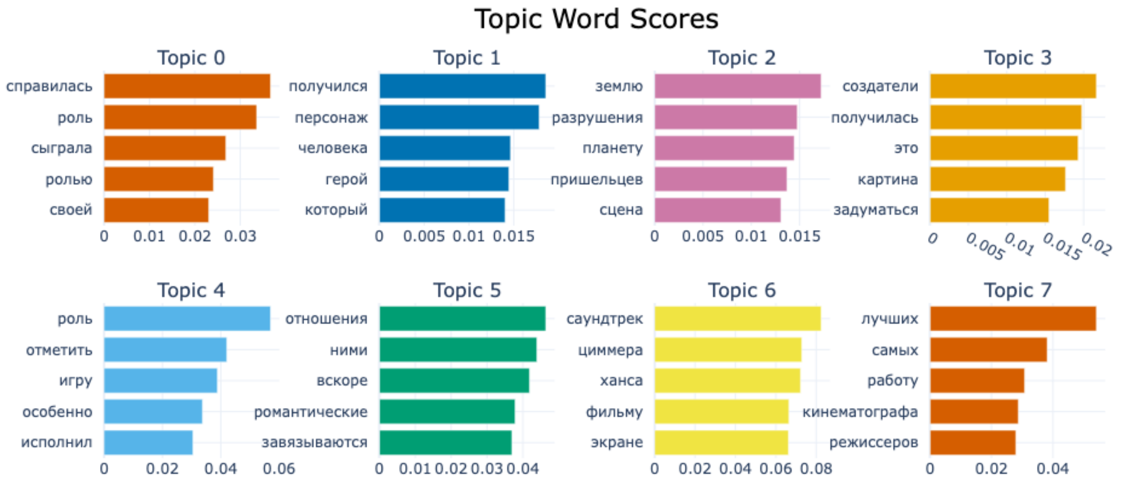

Figure 2: The most frequent topics in the dataset of AI-written movie reviews.

These findings highlight the strengths and limitations of both human and AI-generated movie reviews.

While human-written reviews are better at evaluating technical aspects of the movie's production, AI-generated reviews excel in capturing plot and character details. Future research could explore ways to combine the strengths of both types to improve the quality of generated movie reviews.

We are interested in topic modelling not only as an exploratory data analysis tool but also as a metric for evaluating the similarity between two datasets' topic distributions. To address it, we compute the Cosine similarity between vectorized representations of topics. This metric yields a value ranging from 0 to 1, where 0 denotes minimal similarity and 1 signifies complete identity. For our dataset, the cosine value is 0.56, indicating that while the topics are largely congruent, they also exhibit certain distinctions. Thus, we conclude that an ideal cosine similarity range lies between 0.50 and 0.70. Scores within this range indicate a balanced similarity level. A score of 1 would suggest overfitting, implying entirely identical token distributions, while a substantially lower score around 0 would suggest underfitting, indicating a lack of topic congruence and potentially suggesting that the topics are not related to movies.

5.2 Sentiment Analysis

The fine-tuned BERT is used to evaluate controlled attribute quality on a parallel corpus.

The classification achieves 74 percent accuracy on human data and 66 percent on neural data. This could indicate either inaccurate classification or issues with claimed sentiment. To investigate further, a confusion matrix is examined.

Figure 3 presents the confusion matrix for human-written review classification, with labels 0 (neutral), 1 (positive), and 2 (negative) sentiment.

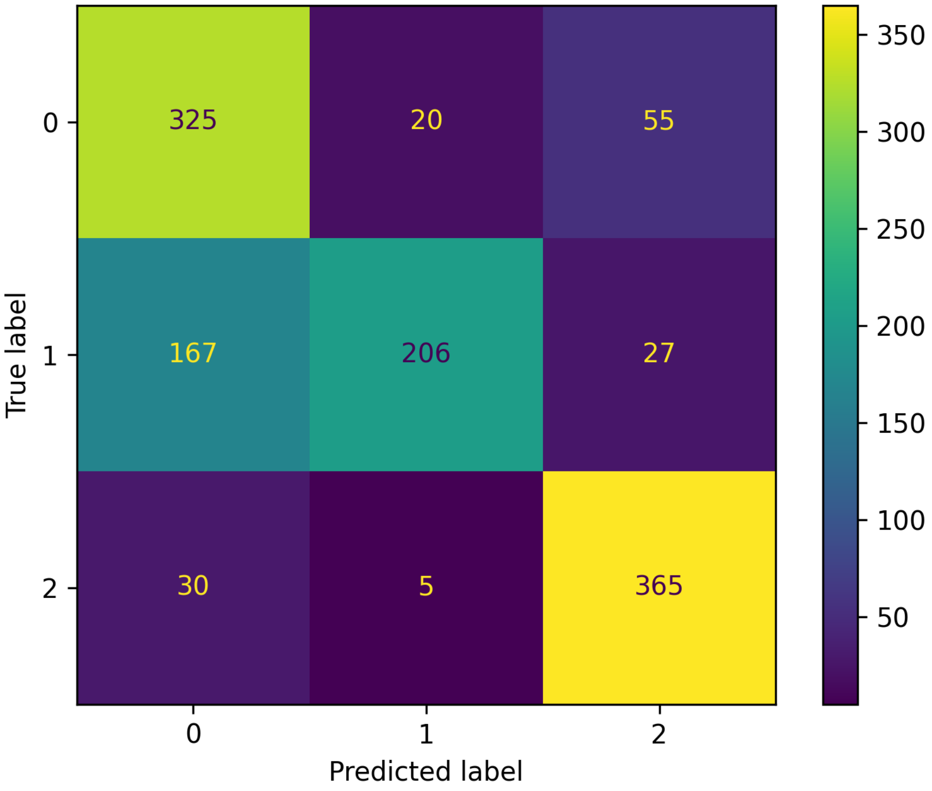

Figure 3: Confusion matrix for the results of human-written reviews classification by sentiment.

Confusion matrix analysis shows high accuracy for negative sentiment (365 out of 400 true negatives) but challenges in distinguishing positive sentiment. People's tendency to use less explicit language in positive reviews creates a subtle and emotionless tone, leading to confusion with neutral sentiment.

Negative sentiment, requiring stronger conviction, is more clearly expressed.

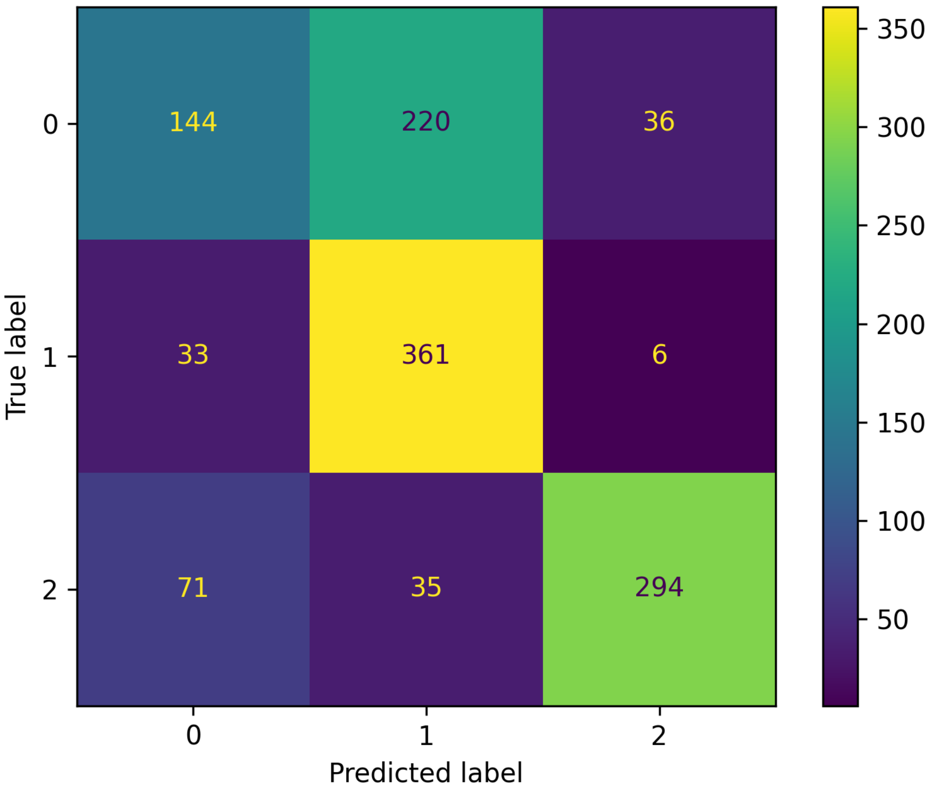

Figure 4: Confusion matrix for the results of AI-written reviews classification by sentiment.

The different distribution of true predicted labels is seen on the Figure 4, which displays confusion matrix of generated text by AI. In this case, the positive sentiment achieves the highest accuracy, while the neutral sentiment is the least accurate. This disparity provides valuable insights into the performance of the classifier. While Figure 3 might have suggested that BERT was biased towards neutral sentiment, the current findings indicate that the problem may be attributed to the nature of the texts themselves.

It is widely accepted that models may struggle to identify neutral sentiment. In this context, it is notable that our fine-tuned model appears to generate neutral sentiment less accurately than positive and negative sentiment.

Upon manually analyzing the neutral texts, we discovered a recurring pattern where many of them ended with the phrase "highly recommend to watch" or included a mention of "10 out of 10" ratings. This observation suggested that the model has a bias towards generating positive reviews even when the sentiment should have been neutral.

6 Conclusion

In this study, we have presented results of an approach to the validation of controlled text generation, which involves the use of popular natural language processing methods as reliable metrics to investigate the success of LM's generation. Our experiments showed the potential of Topic modeling and Sentiment Analysis tools to provide a deeper and more accurate estimation of the semantic consistency of generation validated on a parallel dataset that includes the controlled attribute, the original human-written text, and the generated text with the same beginning as in the original review.

Our approach has been implemented using the decoder's part (ruGPT3) of transformer architecture as a generative model and the encoder part (ruBERT) as a validation tool. Our findings offer important insights into the structure of movie reviews in general.

Funding: This article is an output of a research project "Text as Big Data": modeling of convergent processes in language and speech by digital methods" implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University) in 2023.

References

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation, March. arXiv:1912.02164 [cs].

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2022. Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation, January. arXiv:2109.06379 [cs].

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure, March. arXiv:2203.05794 [cs].

David Hardcastle and Donia Scott. 2008. Can we Evaluate the Quality of Generated Text? January.

Conglong Li, Minjia Zhang, and Yuxiong He. 2022. The stability-efficiency dilemma: Investigating sequence length warmup for training gpt models.

Catherine Ordun, Sanjay Purushotham, and Edward Raff. 2020. Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs, May. arXiv:2005.03082 [cs].

Posokhov P. A., Skrylnikov S. S., and Makhnytkina O. V. 2022. Artificial text detection in Russian language: a BERT-based Approach. // *Computational Linguistics and Intellectual Technologies*, P 470–476. RSUH, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. October.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11.

Christof Schöch. 2016. Topic Modeling Genre: An Exploration Of French Classical And Enlightenment Drama. January.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation, May. arXiv:2004.04696 [cs].

Tatiana Sherstinova, Anna Moskvina, Margarita Kirina, Irina Zavyalova, Asya Karysheva, Evgenia Kolpashchikova, Polina Maksimenko, and Alena Moskalenko. 2022. Topic Modeling of Literary Texts Using LDA: on the Influence of Linguistic Preprocessing on Model Interpretability. // *2022 31st Conference of Open Innovations Association (FRUCT)*, P 305–312.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT, February. arXiv:1904.09675 [cs].

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A Benchmarking Platform for Text Generation Models, February. arXiv:1802.01886 [cs].