Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2026”

June 24–26, 2026

Automatic Sarcasm Detection in Russian Texts

Obridko Maria
Higher School of Economics
obridkomaria@gmail.com

Abstract

This article is devoted to sarcasm detection in short written texts. While the majority of previous works focused on English data mainly using Twitter and Reddit posts as a resource, this research explores a new language (Russian) and a new speech genre of negative reviews that lacks any explicit sarcasm marking. 3596 reviews were scraped from otzovik.ru and pravogolosa.net. The collected data was manually assessed by three linguists, showing ambiguity of sarcasm comprehension even by native speakers. All assembled data is available online. The collected texts were used for fine-tuning a RuBERT model, previously finetuned for a sentiment detection task. The final sarcasm-RuBERT model surpassed results shown in similar researches on Russian data with a F1-score of 85.96 and performed better than GPT-4. In addition to annotated corpora and fine-tuned model sarcasm was theoretically examined. Both the fine-tuned model, GPT-4 model and assessors pointed out similar relatively definite sarcasm markers such as quotation marks, parentheses, words with a high sentiment and others. More ambiguous sarcasm types such as words with a “local” positive sentiment, rhetorical questions, wishes and other were also theoretically researched, all contradicting an explicit sentiment of the utterance and heavily relying on the context.

Keywords sentiment analysis, pragmatics, LLMs’ fine-tuning
DOI: 10.29003/2075-7182-2026-24-461-470

Автоматическая детекция сарказма в текстах на русском языке

Обридко М. С.
Национальный исследовательский университет «Высшая школа экономики»
obridkomaria@gmail.com

Аннотация

Эта статья посвящена выявлению сарказма в коротких письменных текстах. В то время как большинство предыдущих работ были посвящены данным на английском языке, в основном с использованием постов из Twitter и Reddit, в этом исследовании рассматриваются новые язык (русский) и речевой жанр (негативные отзывы), в котором отсутствует явно выраженный сарказм. Было отобрано 3596 отзывов из otzovik.ru и pravogolosa.net, собранные данные были вручную проанализированы тремя лингвистами, что показало неоднозначность понимания сарказма даже носителями языка. Все собранные данные доступны онлайн. Размеченные тексты были использованы для файн-тьюнинга модели RuBERT, ранее дообученной для задачи определения тональности текста. Итоговая модель sarcasm-RuBERT превзошла результаты, показанные в аналогичных исследованиях на русских данных, с показателем F1 85,96, а также показала лучшие результаты, чем GPT-4. В дополнение к аннотированным корпусам и детально настроенной модели, феномен сарказма был теоретически проанализирован. Как усовершенствованная модель, так и модель GPT-4, а также эксперты-оценщики указали на схожие относительно определенные маркеры сарказма, такие как кавычки, круглые скобки, слова с высокими значениями тональности, а также другие. Более неоднозначные типы сарказма, противоречащие явному смыслу высказывания и в значительной степени зависящие от контекста, такие как слова с “местной” позитивной тональностью, риторические вопросы, пожелания и другие, также были теоретически исследованы.

Ключевые слова: анализ тональности, прагматика, файн-тюнинг больших языковых моделей

1 Introduction

Despite computer semantics and pragmatics being developing parts of modern linguistics, the phenomenon of sarcasm has not been thoroughly researched. The vast majority of works on this topic only

use specific English data – posts in X (Twitter) and Reddit that have explicit sarcasm markers, hashtags that seem to be English-specific [9], example 1).

1. *It's not like I wanted to eat breakfast anyway. #sarcasm*

These methods are hard to expand on other languages and especially new language genres.

In this research the problem of automatic sarcasm detection will be explored on a previously scarcely researched Russian data. A new annotated dataset of a distinct language genre of reviews will be used both for fine-tuning RuBERT model that has showed the best performance in the previous works and for theoretical analysis of examples of sarcastic utterances.

In the works in this field, it is customary to omit the difference between sarcasm and irony, both being considered as a contradiction between an explicit sentiment of the utterance and its "real" one. The same definition will be used in this work.

2 Previous works

The task of automatic sarcasm detection is most frequently solved by binary classification of words' embeddings. All works emphasize the importance of context and extralinguistic knowledge for better sarcasm detection, which leads to the better performance of models with pretrained embeddings.

Some researches use lexical text elements such as interjections, capitalization, punctuation marks, intensifications and elongated words occasionally adding them into classical embeddings. The best accuracy achieved by this solution on English data is 78.74% (Random Forest architecture by [3]).

Another approach uses incongruity of sentiment of distinct words and their context. W. Chen et. al. (2022) used concatenated vectors of word's sentiment and its accordance to its context and achieved F1-score 73.85%. In similar works specialized sentiment dictionaries were used, thanks to them rule-based models were able to show good performance [6], though noticeably more expensive than LLM and other neural networks.

On Russian data only few sarcasm researches took place. BERT's embeddings showed the best performance comparing to ones of Word2Vec, GloVe. Interestingly [7] achieved the best F1-score of 76% with RuBERT fine-tuned on the smaller better annotated dataset of 1928 texts (74% when fine-tuned on a 'noisier' 3344 texts).

The best current result achieved on Russian data reaches F1-score 84.78% and was likewise achieved via BERT [11]. These numbers, however, should be perceived with caution due to 'sarcastic' texts consisting of translated English news' headlines and Russian jokes, neither of which could be considered truly sarcastic.

Recent studies of automatic sarcasm detection increasingly focus on large language models, contextual prompting and model interpretability. The classical transformer architectures tend to be outperformed by the prompted LLMs in tasks of detecting implicit semantic and pragmatic meaning as shown in [5] due to the ability of letter to process broader contextual and extralinguistic information. Similar to the given research, [8] explored explainable sarcasm detection of English data via the analysis of the transformer attention. That work showed the models to rely on such linguistic markers as punctuation, quotation marks and sentiment inconsistency. Contemporary research also addresses figurative language interpretation in large language models. [13] demonstrated that during the sarcasm and irony processing LLMs detect complex semantic and pragmatic patterns from the features of classified texts. These works confirm the importance of contextual and pragmatic information for the task of automatic sarcasm detection and showcase the advantages of the modern explainability methods for analyzing transformer-based models.

3 Goals and methods of research

The main goals of this work are to achieve a good result in automatically detecting sarcasm in Russian texts and to analyze this linguistic phenomenon theoretically to explain different levels of ambiguity of sarcastic texts.

The data source for this work are negative reviews of multiple products, firms and TV programs. This language genre was chosen to simplify primary automatic texts selection, which will be described in the

next section in more detail. Moreover, this genre was expected to contain plentiful occurrences of sarcasm due to its emotionality and informality.

Collected data was annotated by three linguists, separated into four groups according to the level of annotators' agreement, each group was theoretically analyzed. A special Ru-BERT model was chosen as a base for fine-tuning due to Transformers showing the best results in previous work on the matter (for example [11]). The specific chosen model (https://huggingface.co/sismetanin/rubert-ru-sentiment-rusentiment) was pretrained on a Rusentiment dataset for a goal of automatic sentiment detection [10], which would be very helpful due to a close connection between tasks of sarcasm detection and sentiment evaluation.

4 Data collection and annotation

As mentioned above, negative reviews were chosen as dataset of the research to exclude the primary task of analyzing the sentiment of each text. Knowing it to be negative, the task of automatically detecting texts with a contradiction of an explicit and implied sentiment consisted of finding texts with non-negative sentiment. This incongruity is one of the most frequent criteria in sarcasm detection works (for example [1]). The ru-BERT-ru-sentiment model that was later used for the final fine-tuning was used for an automatic sentiment evaluation. The length of collected texts was restricted to 5 sentences, which was considered enough for consisting sufficient context for sarcasm and not leaving away enough of sarcastic sentences (Figure 1).

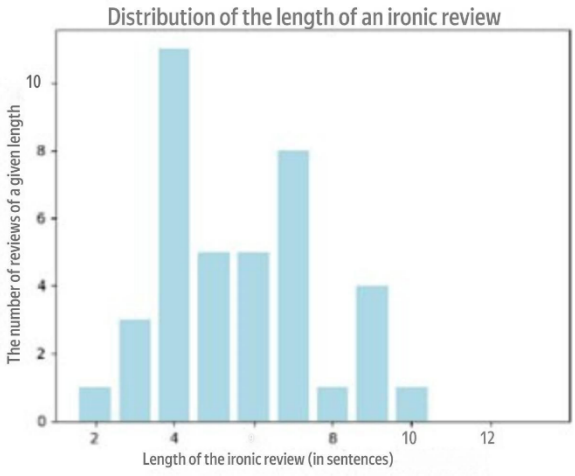

Figure 1: Length of a sarcastic text (sentences)

After the end of annotation of first part of data (1744 texts), the second one (1822 texts) was collected using both sentiment contradiction and specific sarcasm markers, found in the initial texts. A general amount of 3 596 utterances were scraped out of following open internet review resources: otzovik.ru and pravogolosa.net. According to [7] the quality of data manually annotated by multiple linguists should substitute for its relatively small quantity.

The data was independently annotated by three annotators into two classes of containing sarcasm and neutral. Experts were also asked to write down the causes of their choices, which led to finding common conscious sarcasm markers and patterns. As expected, classes of sarcastic (having at least one sarcastic word) and neutral texts occurred to be imbalanced: 547 reviews contained the researched phenomena, in 2273 it has not appeared.

All data was split into four groups according to its level of annotators' agreement (ambiguity of sarcasm) that was calculated by the ratio of the annotators that have considered the occurrence to be sarcastic. Groups were uneven with the biggest having annotators' agreement of 0.0 (all annotators marked text as non-sarcastic) – 2273 reviews, followed by agreement of 1.0 (all annotators marked text as sarcastic) – 547 reviews. These two groups with unambiguous sarcasm were used in model's

finetuning. Groups with intermediate level of agreement had almost equal number – 378 text with level 0.3 and 398 with level 0.6.

Random samples of 100 texts from the most definite groups (annotators' agreement 1.0 and 0.(6)) were analyzed for the scope of sarcasm and the amount of sarcastic occurrences per text.

The final dataset reviews_final.db contains a table with all data (general), tables split by value of annotators' agreement (agreement_all, agreement_two, agreement_one, agreement_none), train, test, validation tables and row data. It is stored at https://github.com/MaryObr/coursework.

5 Sarcasm typology

During the theoretical data analysis, a number of types of sarcasm were distinguished based on the previous research of the phenomena, which distinguishes types of irony (sarcasm) according to the relation between speaker and target, contextual dependence and the scope of ironic meaning [1]. The following classification adapts these principles for the task of automatic sarcasm detection in Russian online reviews. Main approaches determining the present sarcasm types consist of the relation between the speaker and the object of sarcasm, the taxis between the time of the sarcastic occurrence and action described by it and finally the linguistic scope of the sarcastic meaning within text.

- **By coincidence of a speaker and a subject of sarcasm**
  - **Autosarcasm** (the speaker is the subject (example 2)

2. *Мтс все еще воруют , 7 лет назад такая же ситуация была и сейчас уже переводы делают. А служба помощи в чате не способные амёбы которые просят обратиться в полицию.* ***Отлично просто.***

Mts are still stealing, 7 years ago the situation was the same and now [scammers] are already making transactions. And he help service in a chat are incapable ameba that ask to call the police. ***Just great***

  - **Ironic quotation** (the speaker and the subject do not coincide (example 3)

3. Не возможно дозвонится до оператора ! Не днём не ночью ! Уже аж смешно ! Говорят **якобы все специалисты заняты**))

It is impossible to reach an operator. Neither in the daytime nor at night! It's already funny! They say **as if all specialists are busy))**

- **By timing of occurrence of sarcasm and the utterance in its scope:**
  - **Synchronic sarcasm** (both sarcasm and its goal utterance happen simultaneously, example 2)
  - **Asynchronic sarcasm** (a moment of speech with sarcasm is preceded by the utterance in its scope, example 3)

- **By type of scope of sarcasm:**
  - **A narrow scope** (a word or a word phrase, example 4)

4. *Есть матч ТВ эчди,там нет этого коммента* ***"хваленного"*** *достал уже! Несёт всякую чужь?!отдайте его в Комеди клуб! Там всё* ***прекрасно, великолепно****!!!!*

There is match TV hd, where there is no this **"praised"** commentator, [who] annoys me already! He's talking every nonsense?! Give him to Comedy club! There everything is **perfect, great!!!**

  - **A wide scope** (a clause, example 3)

- **A full scope** (the rarest, each clause in text is sarcastic, example 5)

5. **Вауууууууу!!!!!!! 21 век!!!!))))),а в такой иновационной и передовой стране,как Россия,президент должен вмешиваться в ситуацию с бытовыми отходами,или просто с мусором!!!!! - КРУТО!!!!!,или где до сих пор воду летом отключают!!!!))))))**

Wawwwwwwww!!!!!!! 21 century!!!!))))),and in such an innovative and advanced county,as Russia,the president has to intervene into a situation with household waste,or just with garbage!!!!! - GREAT!!!!!,or somewhere still water is shut down in summer!!!!))))))

By the level of annotators' agreement and mistakes of the model sarcastic utterances can be split into a core and periphery. The most prototypical and least ambiguous case of phenomena of sarcasm is a case of **synchronic autosarcasm** (example 4) with narrow or wide scope (which are sometimes hard to distinguish due to the sentiment of the sentence fully relying on the sentiment of one word, see the last sentence of example 4). It is opposed to frequently cooccurring (but not coinciding – example 4 is synchronic ironic quotation) types of **asynchronic ironic quotation** that are mostly present in groups with lower annotators' agreement.

The periphery of sarcasm phenomena also consists of rhetorical questions and wishes (for example, wishes to gain new fooled clients) the interpretations of which heavily relies on the context rather than explicit lexical meaning [1; 2] and metaphorical utterances (example 6). These cases show the contradiction between explicit and implied sentiment of the sentence from the sarcasm definition, however, generally lack negative emotion and are worse detected as sarcasm both by the experts and a model.

6. *я как дурак с этой **электронной проституткой** разговариваю...*

I as a fool am talking with this **electronic prostitute…**

6 Model fine-tuning

After collecting and theoretically analyzing data it was used to finetune a model from a BERT family due to it showing the best results in previous similar works [11]. To achieve the best stability and quality [12], not only a Ru-BERT was chosen as a base for fine-tuning, but a pretrained on a large dataset with a similar task of sentiment determination ru-BERT-ru-sentiment. As expected, this model showed more consistent performance than original Ru-BERT (Figure 2), having both a staring and ending F1-score higher than its counterpart. Therefore it was chosen to be a source model for the fine-tuning process, the resulting model of which is to be examined later.

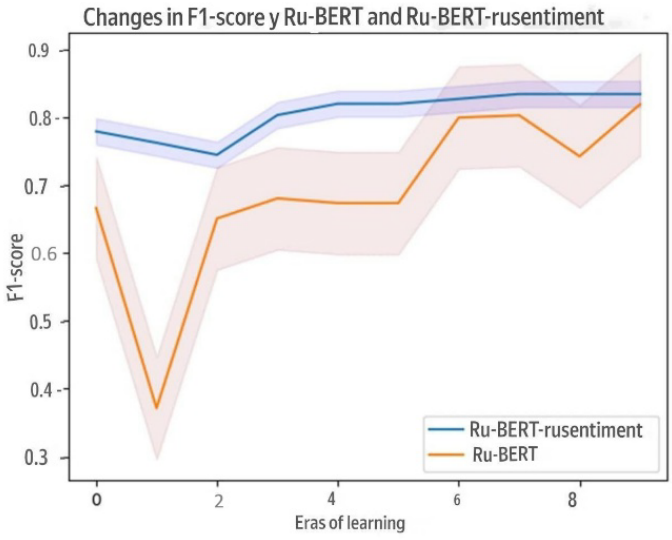

Figure 2: F1-score changes of Ru-BERT and Ru-BERT-rusentiment

The final model was fine-tuned via Trainer with following parameters:

- Loss: CrossEntropyLoss().
- Optimizer: AdamW() (learning rate=5e − 5)
- Sheduler: get_linear_schedule_with_warmup()
- Number of trainig epochs: 8
- Batch size: 8
- Maximum length: 512
- Seed: 29
- Warmup_ration: 0.1

After fine-tuning sarcasm-RuBERT model it was compared to GPT-4 on a data of texts with unambiguous presence of sarcasm (with annotators' agreement 1.0 and 0.0). GPT-4 was asked to classify each given text as containing sarcastic occurrences or not and provide a short (one sentence) explanation of its decision that was later used for analyzing markers of sarcasm for models and native speakers. The used prompt can be seen in the example 7:

7. "Ты профессиональный лингвист, специализирующийся на вопросах прагматики и семантики. Я буду давать тебе текст, тебе нужно будет определить, является этот текст саркастическим или нет. Ответ должен состоять из двух частей. Первая часть: одна цифра, если текст саркастический -- ответ 1, если несаркастический -- ответ 0. Вторая часть -- твоё объяснение твоего выбора в одном предложении. Между частями должен стоять разделитель //"

Table 1 presents metrics of sarcasm-RuBERT in comparison to previous researches of automatic sarcasm detection in Russian texts. Sadly, none of the models or their weights from the previous works are available online, whether on GitHub repositories or uploaded to Hugging Face. Therefore, the performance of the fine-tuned model could only be compared to ones of GPT-4, the ru-BERT-ru-sentiment being the source BERT model and recreated baseline architectures, mentioned in [7].

	F1-score	accuracy	recall
Logistic Regression, reproduced from [7] on all data	0.69	0.68	0.72
SVM, reproduced from [7] on all data	0.68	0.68	0.69
Random Forest, reproduced from [7] on all data	0.63	0.64	0.61
ru-BERT-ru-sentiment without fine-tuning on all data	0.0%	50%	0.0%
Sarcasm-RuBERT on all data	85.96%	85.45%	89.09%
Sarcasm-RuBERT on data with 1.0 annotators' agreement	79.6%	79.4%	80.7%
GPT-4 on data with 1.0 annotators' agreement	78%	77.6%	79.8%

Table 1: The performance of different models on the sarcastic dataset

As expected, the fine-tuned Sarcasm-RuBERT model exceeds the results of all baseline methods, the best being Logistic Regression, it also slightly outperforms prompted GPT-4 model, interestingly achieving the better metrics on the whole data and not its unambiguous part. The performance of the non-fine-tuned ru-BERT-ru-sentiment model does not surpass one of the Random Baseline with it only classifying each text from test dataset as non-sarcastic (class 0).

7 Texts with different level of annotators' agreement

The metrics of sarcasm-RuBERT (F1-score, recall, accuracy, precision, loss) and its confidence were compared on groups of texts with different level of annotators' agreement.

Each metric except one worsened almost linearly with the growth of ambiguity of texts (Figure 3). Precision being the exception can be explained by the decline of the share of non-sarcastical texts.

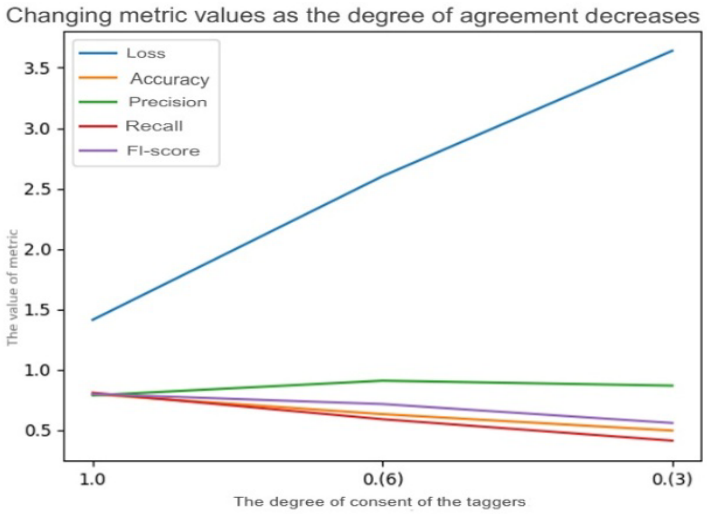

Figure 3: Metrics' dependency on ambiguity of sarcasm

Despite that the confidence of the model is close to 100% on all groups, its maximum value was achieved not on 1.0 agreed sarcastic data. The highest value on when choosing non-sarcastic class is on the group of 0.(3) agreement, when choosing sarcastic class – surprisingly on the group of 0.(6) (Figure 4).

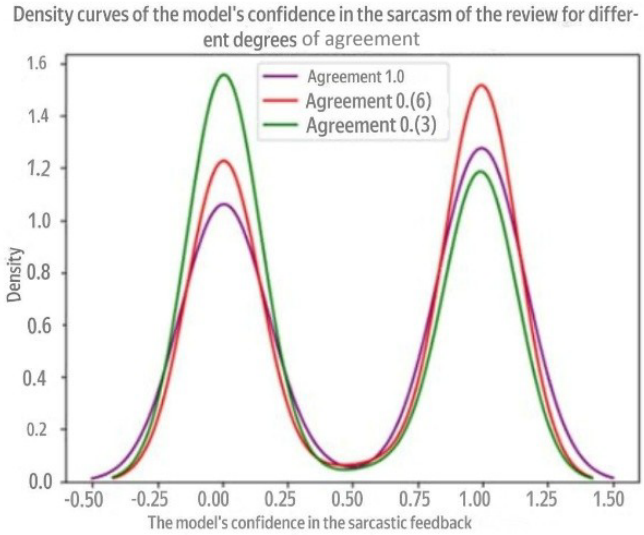

Figure 4: Density of sarcasm-RuBERT on data with different annotators' agreement

8 Interpretation of model's decisions

When comparing metrics of GPT-4 to sarcasm-RuBERT on a dataset with annotators' agreement of 1.0, part of a prompt of a first model was to briefly explain its choice of a class in one sentence. The decisions of fine-tuned model were also interpreted with the help of SequenceClassificationExplainer[1] -- a model explainability tool designed specifically for the highlighting of the tokens influencing the decision of the model from the Hugging Face transformers library. It shows the token's importance by visually coloring the most influential tokens the most intense color with green standing for positive label classification and red denoting the opposite.

Following markers of sarcasm were identified for the models:

- **Quotation marks** (both models, the most common marker). Being the most cohesive sarcasm marker, they were repeatedly mentioned as such in previous pragmatics and computational research as graphical indicators of irony and evaluative distancing of the speaker and the discourse [9; 10; 11]

8. "Текст саркастический, поскольку автор использует кавычки, чтобы подчеркнуть иронию и недовольство по поводу обслуживания."

- **Sentiment contradiction** (both models). This marker indicates the classical theoretical base of the sarcasm phenomena, stating the opposition between internal implication and its external expression [1]. As an example of it, both tokens with the high sentiment values "доволен" and "вагон старинный" have high importance for the sarcasm-BERT's classification.

- **Words with high sentiment** (both models, though their interpretations differ from the one of experts). Even a single occurrence of an intensifier corresponds to the hyperbolic strategies previously described as characteristic of sarcastic speech [5], which is only magnified by the simultaneous usage of words with the polar sentiment mentioned in the previous paragraph. As an example, both consider 'чудеса' to be highly important for text to have a sarcastic label.

- **Logical contradiction** (GPT-4). This sarcasm marker close to the first one corresponds to the global contextually and semantically incongruous nature of sarcasm [1; 3].

9. "Текст явно противоречит новогодним пожеланиям, что указывает на сарказм в его пожеланиях компании Ростелеком"

[1] https://github.com/cdpierse/transformers-interpret

• **Capitalization** (GPT-4). This sarcastic marker was also previously noted to be an expressive sarcasm marker in the social media discourse [5; 9].

10. В тексте используются большие буквы и выражения с иронией, чтобы показать недовольство и насмешку над политической ситуацией

• **Extralinguistic factors** (GPT-4), which were expected following to the findings of the recent works [7; 15] that show the LLM's to use their pragmatic and extralinguistic knowledge in the cases of sarcasm and irony detection.

11. Текст саркастический, поскольку автор использует иронические выражения, такие как "вроде начинает летать, секунд на 10", чтобы подчеркнуть неудовлетворение связью.

• **"Understatement"** (GPT-4). This sarcasm marker has not been mentioned by the annotators, however does correspond to the pragmatic irony strategy described in the theoretical sarcasm researches [1; 2].

12. Выражение раздражения и недовольства с использованием преуменьшения и иронии указывает на саркастическое отношение к ситуации.

• **Exclamation marks** (GPT-4). As part of other punctuation markers, especially ellipsis they were also considered to be important for sarcasm theoretical expression and classification in the previous works [2; 5; 10] being expressive and sometimes hyperbolic expressions of it.

13. Использование восклицательных знаков и названия «Первого!!! Лучшего!!!!» может указывать на саркастическое отношение к заявленным заслугам компании

• **Punctuation in general** (sarcasm-BERT). Both commas, exclamation marks, quotation marks, ellipsis and dots have had ranging importance for the classification with sometimes identical punctuation influencing the opposite choices, as final dots show in Figure 5.

• **Numbers** (sarcasm-BERT). Unlike other sarcasm markers noted by the models, this one lacks any theoretically researched pragmatical interpretation.

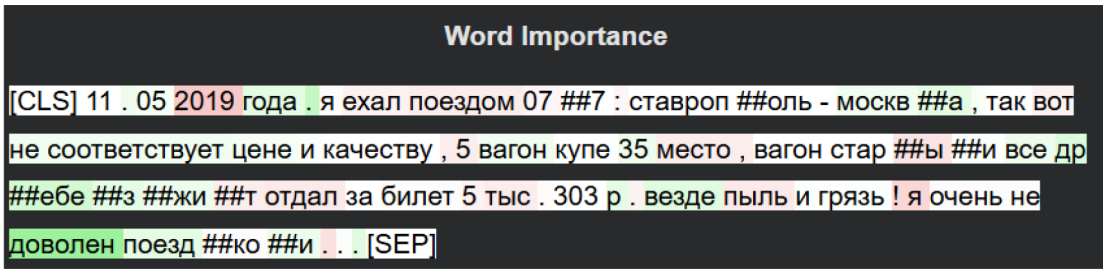

Figure 5: Token "2019" has the most influence on the "non-sarcastic" label choice

The patterns affecting the class selection of models are in many ways similar to the ones pointed out by human annotators. Besides many mentioned above, both models and experts tend to consider longer texts non-sarcastic, due to them often containing serious proposals to fight the scammers. This tendency may correspond to the observations that sarcasm is often realized through compact incongruity contexts rather than extended discourse [2].

Patterns mentioned only by GPT-4 belonging to the more complex levels of language such as pragmatics can be explained by the better interpretability of its architecture and diverse ways of highlight the important features – they can be used by BERT model as well.

9 Conclusion

The phenomenon of sarcasm can be seen as having core (synchronic autosarcasm) and peripheral occurrences. Some apparent general markers are recognized better both by humans and by models. The

one that stands out the most are quotation marks (Figure 5, brightness symbolizes token's importance), hence their development into a sign.

Generally, annotators and models interpreted different sarcasm markers mostly similar, for example rhetorical utterances and words with neutral sentiment were ambiguous for both.

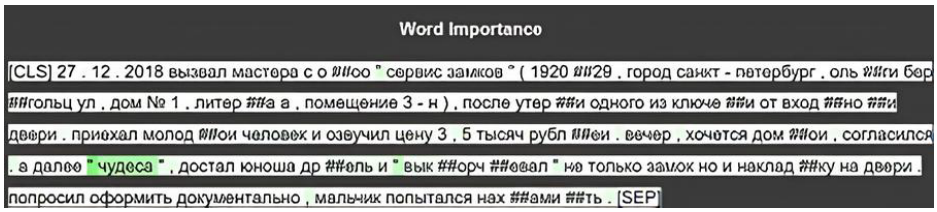

Figure 6: Tokens' importance for BERT classification

During the work on this research a dataset of Russian sarcastic negative reviews was collected and annotated, texts were theoretically analyzed, final sarcasm-RuBERT achieved the best results.

In the future the work can be developed by increasing the amount of data, its genre diversity and the number of annotators. Another branch lies in adding sarcasm detection into modern sentiment analysis for its improvement.

10 Bibliographical references

[1] Attardo S. (2000) Irony as Relevant Inappropriateness // Journal of Pragmatics. 2000. Vol. 32, №6. P. 793–826.

[2] Burgers C., van Mulken M., Schellens P. (2012) Verbal Irony: Differences in Usage Across Written Genres // Journal of Language and Social Psychology. 2012. Vol. 31, №3. P. 290–310.

[3] Govindan V., Balakrishnan V. (2022) A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection // Journal of King Saud University-Computer and Information Sciences. 2022. Vol. 34, №8. P. 5110–5120.

[4] Gurin A.A., Sadykov T.M., Zhukov A. The big data approach towards sarcasm detection in Russian // Computational Linguistics and Applications. 2023.

[5] Khan A., Kumar N., Sharma R. (2024) Sarcasm Detection Using Large Language Models and Context-Aware Prompting // Expert Systems with Applications. 2024.

[6] Kolchinski Y. A., Potts C. (2018) Representing Social Media Users for Sarcasm Detection // Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018.

[7] Kosterin M., Paramonov I., Lagutina N. (2023) Automatic Irony and Sarcasm Detection in Russian Sentences: Baseline Methods // Computational Linguistics and Intellectual Technologies. 2023.

[8] Liu Y., Chen H. (2024) Explainable Sarcasm Detection in Social Media Texts Using Transformer Attention Analysis // Information Processing & Management. 2024.

[9] Maynard D., Greenwood M. (2014) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. P. 4238–4243.

[10] Rogers A., Romanov A., Rumshisky A., Volkova S., Gronas M., Gribov A. (2018) Automatic Sarcasm Detection in Russian Reviews // Proceedings of the 27th International Conference on Computational Linguistics. 2018. P. 755–763.

[11] Trenev I., Chaplinskaya N. (2022) Russian sarcastic comments detection // Proceedings of the International Conference on Natural Language Processing. 2022.

[12] Zhang T., Wu F., Katiyar A., Weinberger K.Q., Artzi Y. (2021) Revisiting Few-Sample Bert Fine-Tuning, 2021.

[13] Zhou X., Li J. (2025) Interpreting Large Language Models in Figurative Language Understanding // Computational Linguistics. 2025.