

A Dataset of Russian-Language Debates For Argument Mining

Rodion Sulzhenko¹ and Boris Dobrov²

¹ Lomonosov Moscow State University, Moscow, Russia rodion13par@gmail.com

² Lomonosov Moscow State University, Moscow, Russia dobrov_bv@rcc.msu.ru

Abstract. We present DebateRu, annotated dataset of Russian-language student debates designed for argument mining in culturally specific contexts. Comprising 10 hours of spontaneous televised debates (429 arguments across 10 topics), the corpus captures authentic rhetorical strategies and socio-political discourse patterns unique to Russian youth culture. Unlike scripted debate datasets, DebateRu preserves the emotional intensity and contextual nuances of real-world argumentation, addressing a critical gap in non-English resources. We evaluate the dataset through two tasks: stance detection and argument generation, testing several state-of-the-art Russian-adapted large language models. DebateRu provides a benchmark for developing context-aware argumentation systems and studying cross-cultural discourse patterns. We release the dataset to support research in multilingual NLP, rhetorical education, and computational social science. Collected dataset is publicly available on github.³

Keywords: Argument Mining · Stance Detection · Argument Generation · Large Language Models

1 Introduction

Argument mining remains critical for developing language models (LLMs) capable of nuanced reasoning in culturally specific contexts. Although existing datasets like DebateSum[2] focus on structured debate evidence, they often overlook the dynamic, emotionally charged nature of real-world discourse, particularly in non-English settings. This limits their utility for training models to analyze organic and culturally situated arguments.

To address this gap, we introduce DebateRu, a meticulously annotated dataset of Russian-language student debates from a popular television show. Although it is relatively small in scale, covering 10 debate topics and spanning 10 hours of dialogue, DebateRu provides distinctive advantages.

First, it captures cultural specificity by reflecting Russian socio-political contexts, thereby addressing a critical lack of non-English argument mining resources. Second, it features organic discourse, with unscripted and emotionally charged exchanges that serve as a realistic benchmark for modeling human-like

³ <https://github.com/rodion-s/DebateRu>

argumentation. Third, it includes high-quality markup, with annotated argument structures allowing for fine-grained evaluation.

We evaluate DebateRu’s potential through two tasks. First, stance detection with Russian-adapted LLMs: RuadaptQwen2.5-32B-Pro-Beta⁴, tlite⁵, classic state-of-the-art models qwen2.5-72b⁶, llama3-70b⁷ and fine-tuned BERT⁸ as a baseline. Second, argument generation with the same models and GPT-4 as a judge. To complement automated judging, we additionally conducted a small-scale human evaluation, where a single assessor compared model outputs against reference debate arguments across the same qualitative dimensions, providing an independent check of the automatic scores.

2 Related Works

Several existing datasets address argument mining and debate analysis. Current argument mining datasets primarily focus on structured, evidence-based debates, often neglecting spontaneous and culturally specific discourse. Key resources include:

1. **IBM Debater**[3]: A well-known argument mining dataset featuring structured claims and evidence, but limited to formalized, English-centric debates.
2. **ArgAnalysis35K**[4]: A dataset of 35,000 single argument pairs from parliamentary debates, useful for claim detection but lacking multi-turn interactions.
3. **IAM (Integrated Argument Mining)**[5]: A richly annotated dataset for holistic argument analysis, but small in scale (1,000 articles).
4. **DebateSum**[2]: A large corpus of pre-season debate evidence (240,566 examples), capturing structured arguments but omitting live, unscripted discourse.
5. **OpenDebateEvidence**[6]: One of the largest English debate datasets (3.5M documents), yet still focused on U.S. high school policy debates, with no Russian-language representation.

While these resources have advanced argument mining substantially, they are predominantly English-language and drawn from highly formalized settings, leaving limited coverage of spontaneous, culturally specific debate practices.

⁴ <https://huggingface.co/RefalMachine/RuadaptQwen2.5-32B-Pro-Beta>

⁵ <https://huggingface.co/AnatoliiPotapov/T-lite-instruct-0.1>

⁶ <https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>

⁷ <https://huggingface.co/meta-llama/Meta-Llama-3-70B>

⁸ <https://huggingface.co/google-bert/bert-base-multilingual-cased>

Russian-language resource gap. As highlighted by Kotelnikov et al.[7, 8], the scarcity of Russian argument-annotated corpora has long constrained progress on Slavic-language argument mining. Although repositories such as AIFdb[9] host over 150 annotated corpora, until recently none targeted Russian-language materials at scale. Attempts to address this gap typically rely on either manual annotation of native texts—an intensive process requiring approximately 270 hours per corpus[10]—or professional translation of existing English datasets at substantial costs (around \$3,000 per 7,000 sentences). Machine translation (e.g., Google Translate⁹, Yandex.Translate¹⁰, Prompt¹¹) is a potential shortcut but still requires rigorous human validation to ensure label and discourse fidelity. This underscores the need for native Russian-language resources that directly capture authentic argumentative discourse.

RuArg. An important step in this direction is **RuArg** [11], introduced in the first Russian shared task on argument mining at the Dialogue conference. RuArg contains 9,550 sentences from social media posts on COVID-19 (vaccination, quarantine, masks), annotated for stance detection and argument (premise) classification; top systems achieved macro F1 of 0.6968 (stance) and 0.7404 (argument classification). RuArg represents a valuable starting point for Russian-language argument mining, though its scope remains limited to short, sentence-level social media comments within a single topical domain. Broader, multi-turn, and speech-based corpora remain largely unexplored, leaving open opportunities for expanding coverage across genres, modalities, and domains.

3 Russian Student Debates Dataset

3.1 Data Collection

The DebateRu dataset originates from Debattle¹², a Russian-language student debate program designed to cultivate rhetorical and critical thinking skills through structured discourse. The episodes feature two teams engaging with socially significant topics over approximately one hour of real-time debate. The format consists of two primary phases: an initial round where each team presents one-minute opening arguments sequentially, followed by a dynamic cross-examination round where teams exchange three questions each while delivering spontaneous rebuttals.

We selected Debattle as a data source because it offers full-length and publicly accessible recordings, which ensures reproducibility of research results. Its debate format promotes both structured stance-bearing arguments and spontaneous exchanges, providing a natural setting for argumentation analysis. Furthermore, the topics addressed in the program reflect socially relevant issues for

⁹ <https://translate.google.com/>

¹⁰ <https://translate.yandex.ru/>

¹¹ <https://www.translate.ru/>

¹² <https://mosmolodezh.ru/projects/debattl/>

the target population, and the audio quality of the recordings is adequate for reliable automatic speech recognition.

The video content was systematically sourced from the show’s official YouTube channel¹³, selected to capture culturally grounded argument styles specific to the socio-political contexts of Russia. The final corpus comprises ten full-length debates totaling ten hours of dialogue, with topics reflecting contemporary youth concerns in Russia. This selection criteria prioritizes authentic, unscripted interactions over rehearsed performances, distinguishing DebateRu from scripted debate datasets.

3.2 Data Preprocessing

Audio processing leveraged OpenAI’s Whisper-large model[12] for its demonstrated proficiency in multilingual speech recognition and structural preservation, including accurate punctuation and speaker diarization. Comparative evaluation against GigaAM (SberAI)¹⁴ revealed Whisper’s superior performance. Transcripts underwent subsequent refinement using DeepSeek[13] for lexical and grammatical correction, particularly addressing colloquialisms and domain-specific terminology prevalent in youth discourse.

3.3 Annotation Protocol

The dataset was annotated by a single annotator with prior experience in text analysis. The annotation process was relatively straightforward, as the debate format is highly structured: one team consistently argues in favor of the motion while the other argues against it. This setting makes stance identification unambiguous, since the team affiliation directly determines the stance. The main challenge consisted in segmenting the transcripts into individual arguments, which required determining coherent units of reasoning. In practice, this was achieved by splitting the text into argumentative segments (typically corresponding to short passages or paragraphs) that each expressed a distinct claim supported by justification. Given the single-annotator setup, inter-annotator agreement is not reported in the current version; however, future extensions of the dataset will involve multiple annotators to enable reliability assessment.

3.4 Dataset Description

The DebateRu dataset comprises 10 debate topics with 429 annotated arguments, featuring an average argument length of 386 characters. Following debate convention, we designate the team supporting the motion as Team A (affirmative) and the opposing team as Team B (negative). This labeling persists regardless of the specific topic being debated. An example of the layout of the arguments for the first and second rounds is given in the Appendix Fig.1-2.

The dataset covers ten socially significant topics in Russian context:

¹³ <https://www.youtube.com/@debattleleague>

¹⁴ <https://github.com/salute-developers/GigaAM>

Table 1. Dataset Schema

Field	Description
Topic	Debate motion (e.g., "Weapon legalization")
Team_name_support	Team A (affirming the motion)
Team_name_opposite	Team B (opposing the motion)
Markup	Full annotated transcript with speaker tags
First_round	Verbatim transcript of opening arguments
Second_round	Complete Q&A round transcript
First_round_support_args	Extracted affirmative arguments from Team A
First_round_opposite_args	Extracted negative arguments from Team B
Second_round_AtoB_[1-3]	Team A’s questions to Team B with full dialogue chains
Second_round_BtoA_[1-3]	Team B’s questions to Team A with full dialogue chains

- Weapon legalization
- Capital punishment introduction
- Same-sex marriage legalization
- Childlessness tax
- Mandatory higher education
- Credit systems
- AI: Beneficial or harmful
- Nuclear weapons abolition
- Casino legalization
- Universal basic income

4 Experiments

4.1 Experimental setup

We evaluated our dataset on two key tasks: stance detection (first round) and argument generation (second round). All experiments were conducted using zero-shot learning methodology. The computational experiments were performed on a DGX-2 cluster where all models were deployed.

We evaluated a mix of Russian-adapted and multilingual instruction-tuned LLMs of different sizes to probe size effects and accessibility. Models were chosen based on public availability of inference endpoints or weights and their reported performance on Russian-language tasks, enabling replication.

For the stance detection task, models were instructed to determine whether each argument supported or opposed the debate topic ("for" or "against"). The dataset was divided into 80% training and 20% test data. The fine-tuned BERT baseline (bert-base-multilingual-cased) was trained for binary classification using the following hyperparameters: a learning rate of 3e-4, 3 training epochs, a batch size of 16, 500 warmup steps, and a weight decay of 0.01.

In the argument generation task, models were prompted to generate counterarguments based on given cross-examination arguments from the second debate round. The generated arguments were evaluated against human-produced

responses from real debates using GPT-4 as a judge model (LLM-as-Judge approach). The evaluation employed five criteria:

1. Position Consistency (1-5): Adherence to the assigned stance
2. Human Persuasiveness (1-5): Emotional impact rather than logical rigor
3. Argument Uniqueness (1-5): Avoidance of template responses
4. Contextual Relevance (1-5): Use of specific references
5. Realism (1-5): Resemblance to authentic debate content

Detailed prompting strategies are provided in the Appendix A-C.

4.2 Experimental results

In the stance detection task, models demonstrated reasonable but not exceptional performance – particularly notable given the binary nature of the classification task.

Table 2. Stance detection performance across evaluated models

Model	Precision	Recall	F1-score
Fine-tuned BERT	0.51	0.89	0.64
RuadaptQwen2.5-32B-Pro-Beta	0.86	0.80	0.80
Qwen2.5-72b	0.92	0.90	0.90
Tlite	0.87	0.84	0.84
Llama3-70b	0.89	0.88	0.89

For the argument generation task, we employed a comparative evaluation metric measuring the win rate – the proportion of cases where the model-generated response received a higher average score than the gold-standard reference response. Our analysis revealed Qwen2.5-72b as the top-performing model, achieving a 42% win rate against human-generated arguments. While all models demonstrated strong performance across evaluation criteria (scoring consistently above 3.7 on our 5-point scale), qualitative analysis identified a key limitation: the tendency of LLMs to produce generic, contextually-disengaged arguments rather than position-specific responses tailored to the debate’s nuanced context.

As shown in Table 3, the marginal superiority of generated arguments (win rates ranging 29.6-42.6%) suggests that while modern LLMs can produce rhetorically competent outputs, they still struggle to match the contextual precision and tactical relevance of human debaters. This performance gap is particularly evident in the "Contextual Appropriateness" metric, where all models underperformed relative to other criteria. More detailed examples of the algorithm’s results are available on GitHub at the same repository where the dataset is uploaded¹⁵.

¹⁵ <https://github.com/rodion-s/DebateRu>

Table 3. Argument generation performance across evaluated models

Model	Win Rate	Final Score	Consistency	Persuasiveness	Uniqueness	Relevance	Realism
RuadaptQwen2.5-32B-Pro-Beta	0.35	3.93	4.85	3.76	3.43	3.30	3.76
Qwen2.5-72b	0.42	4.09	4.93	3.91	3.50	3.35	3.93
Tlite	0.30	3.85	4.83	3.63	3.31	3.19	3.76
Llama3-70b	0.37	4.00	4.83	3.83	3.50	3.41	3.89

4.3 Human Evaluation

In addition to LLM-as-a-Judge evaluation, we conducted a small-scale human assessment of generated arguments. We selected 20 reference arguments from DebateRu and compared them against 80 model-generated counterparts (20 per model). A single assessor with background in NLP rated all arguments on the same five criteria as in the LLM-based evaluation, using a 1–5 scale.

Table 4 reports the average scores and pairwise win rates, computed analogously to the LLM-based evaluation. The results confirm the general trends observed with automated judging: although model outputs occasionally approached or even exceeded gold references in terms of persuasiveness and surface-level variety, they systematically underperformed in stance consistency and contextual grounding. These deficits suggest that current LLMs, despite their fluency, struggle to sustain debate-specific constraints such as strict adherence to side and topical anchoring. Among the evaluated systems, Qwen2.5-72B and Llama3-70B achieved the strongest overall results, whereas Tlite and RuAdaptQwen2.5-32B-Pro-Beta lagged behind. This superiority of Qwen2.5-72B is likely attributable not only to its scale (substantially larger parameter count) but also to the broader and more domain-adapted training data available to the Qwen family, which together contribute to its strong performance in argumentation tasks. Nevertheless, human evaluation still showed a clear and statistically consistent preference for authentic debate arguments, highlighting the gap between generated text and the realism of naturally occurring discourse.

5 Conclusion and Future Work

In this paper, we presented DebateRu, a novel Russian-language debate dataset that captures culturally situated argumentation patterns through authentic student discourse. Unlike existing resources focused on English-language contexts or scripted debates, DebateRu provides: annotated scripts for 10 hours of spontaneous, emotionally charged exchanges reflecting Russian socio-cultural dynamics; meticulous annotation of 429 arguments with structural and pragmatic fea-

Table 4. Human evaluation of generated arguments

Model	Win Rate	Final Score	Consistency	Persuasiveness	Uniqueness	Relevance	Realism
Reference (debates)	–	4.96	5.00	4.60	4.70	4.50	5.00
RuAdaptQwen2.5-32B-Pro-Beta	0.33	3.62	3.90	3.20	4.10	1.70	3.20
Qwen2.5-72B	0.45	3.90	4.80	4.10	4.40	2.30	3.90
Tlite	0.28	3.34	3.50	2.90	4.00	1.50	2.80
Llama3-70B	0.36	3.84	4.60	3.80	4.20	1.90	3.70

tures; and a benchmark for evaluating context-aware argument generation in non-English settings.

The study’s findings should be considered in light of certain methodological constraints. While the dataset offers substantial qualitative depth through its carefully curated debates, its relatively limited scale of ten debates imposes restrictions on the types of statistical analyses that can be meaningfully performed. Future expansions incorporating regional debate formats from across Russia’s diverse cultural landscape would significantly strengthen the dataset’s representational breadth and analytical utility. Additionally, our evaluation focused exclusively on zero-shot learning scenarios, leaving open important questions about how specialized fine-tuning techniques tailored specifically for argumentative contexts might enhance model performance beyond the current baseline results. Such adaptation methods could potentially bridge the observed gap in contextual relevance that persisted across all evaluated models.

We release DebateRu to support: development of culturally-aware argumentation systems, comparative studies of cross-linguistic debate strategies, and educational applications for rhetoric training. In future work, we plan to expand the dataset both in the number of topics and in the volume of annotated arguments, as well as to include multiple annotators to enable inter-annotator agreement measurement. We also aim to integrate human evaluation more systematically and to benchmark fine-tuned models alongside zero-shot settings, providing a clearer picture of how DebateRu can advance argument mining in Russian.

Acknowledgements

The study was conducted under the state assignment of Lomonosov Moscow State University. We acknowledge the use of generative AI tools for English proofreading and minor formatting adjustments. All dataset design, annotation, and experimental analysis were performed entirely by the authors.

References

1. Kotelnikov, E., Loukachevitch, N., Nikishina, I., Panchenko, A.: RuArg-2022: Argument Mining Evaluation. In: Proc. of Dialogue (2022)
2. Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In Proceedings of the 7th Workshop on Argument Mining, pages 1–7, Online. Association for Computational Linguistics.
3. Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From Arguments to Key Points: Towards Automatic Argument Summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4029–4039, Online. Association for Computational Linguistics.
4. Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. ArgAnalysis35K : A large-scale dataset for Argument Quality Analysis. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13916–13931, Toronto, Canada. Association for Computational Linguistics.
5. Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A Comprehensive and Large-Scale Dataset for Integrated Argument Mining Tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.
6. Allen Roush, Yusuf Shabazz, Arvind Balaji, Peter Zhang, Stefano Mezza, Markus Zhang, Sanjay Basu, Sriram Vishwanath, Mehdi Fatemi, and Ravid Shwartz Ziv. 2024. OpenDebateEvidence: A Massive-Scale Argument Mining and Summarization Dataset. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 34270–34293. Curran Associates, Inc.
7. Kotelnikov, Evgeny. (2018). Extraction of argumentation from texts and the problem of the lack of Russian-language text corpora. Mathematical bulletin of Vyatka State University. 10.25730/VSU.0536.18.26.
8. Fishcheva, Irina & Kotelnikov, Evgeny. (2019). Cross-Lingual Argumentation Mining for Russian Texts. 10.1007/978-3-030-37334-4_12.
9. Lawrence, John & Reed, Chris. (2014). AIFdb Corpora. Frontiers in Artificial Intelligence and Applications. 266. 465-466. 10.3233/978-1-61499-436-7-465.
10. Stab C., Gurevych I. Parsing Argumentation Structure in Persuasive Essays // Computational Linguistics. 2017. Vol. 43(3), pp. 619–659.
11. E. Kotelnikov, A. Goncharov, S. Nartov, I. Pugachev, V. Ivanov, D. Kuznetsov, I. Smirnov, A. Khazov, and A. Frolov. RuArg: Argumentation Analysis for Russian. *arXiv preprint arXiv:2206.09249*, 2022.
12. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.
13. DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, et al. 2025. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*

Appendices

A: Stance Detection Prompt

Determine whether the following argument represents a position "for"

or "against" the topic "{topic}".
Respond only with "for" or "against".

Topic: {topic}
Argument: {argument}
Answer:

B: Argument Generation Prompt

Debate Topic: {topic}
Your Position: {stance}
Opponent's Question: {question}
Formulate a persuasive response while maintaining the {stance} position:

C: LLM-as-Judge Evaluation Prompt

You are a debate analysis expert.
Evaluate this argument against strict criteria:

Topic: {topic}
Question: {question}

Argument:
{generated}

Rate on a 5-point scale:

1. Position consistency (for/against) - strict adherence
2. Human persuasiveness - emotional impact (not logic)
3. Argument uniqueness - non-templated
4. Contextual relevance - specific references
5. Realism - resembles real debates

Output ONLY JSON with:

- scores by criteria
- brief explanation
- final_score (weighted average with weight 3 for #2 and #5)

<first_round_team_A_arguments>[Команда «Если не мы, то кот»]: Добрый день, уважаемые жюри, зрители, участники. Так сложилось, что каждое свое выступление здесь я начинал с того, что пафосно. Я взял что-то из кармана. В прошлый раз это была записка с цитатой из Библии. На этот раз рука уже была потянулась, чтобы достать нож, но я решил, что не буду вас пугать. Но я хочу попросить вас, чтобы вы представили. Представьте, что я оказался перед вами не здесь, в этой светлой студии, под объективами камер, а где-то в темном переулке. Я уверен, что каждый из вас в этот момент захотел бы, чтобы у него в кармане оказался пистолет. Среднее время приезда полиции составляет 15-20 минут. И я не хочу, чтобы эти 15-20 минут стали моими последними минутами.

[Команда «Если не мы, то кот»]: По статистике, в странах, где самое большое количество оружия в руках у населения, самый низкий уровень преступлений, за исключением нескольких южноамериканских стран, где нет просто нелегального оружия. Поэтому явно можно сделать вывод о том, что много оружия у населения не обязательно коррелирует с количеством преступлений.

[Команда «Если не мы, то кот»]: Также стоит отметить, что по данным наших спецслужб, которые приводят статистику в России, сейчас около 20 миллионов единиц оружия находится на черном рынке. И если мы расширим, уберем законодательные рамки и легализуем рынок оружия, то получится вывести огромное количество вооружений из тени. И закончу выступление цитатой: «Убивает не оружие, а человек в первую очередь». Спасибо.</first_round_team_A_arguments>

[Ведущий]: Ну хорошо. Проектор перестройки. Поехали. Ваша презентация.

<first_round_team_B_arguments>[Первый раунд, аргумент команды В][Команда «Проектор Перестройки»]: Мы считаем, что такому древнему, примитивному по рождению инстинкту убивать, как оружие, в принципе, не место в современном мире. В современном цивилизованном гражданском обществе. Поэтому мы против продажи оружия.

[Команда «Проектор Перестройки»]: Просто начнем с обычных фактов. Статистически разрешение оружия сразу же приводит к увеличению количества смертей. Любой конфликт оборачивается просто пулей в лоб. Любая пьяная драка, любое столкновение на шоссе, в любой ситуации решение конфликта — просто пуля в лоб, в состоянии аффекта. Любая перестрелка на улице — это в том числе риск для гражданских, что они погибнут от шальных пуль. Что им придется прятаться по кустам, вместо того, чтобы просто спокойно идти по улице и чувствовать себя в безопасности, как и должно быть в современном цивилизованном обществе.

[Команда «Проектор Перестройки»]: К тому же, возможно, тем, кто поддерживает разрешение оружия, кажется, что оружие — это что-то очень доступное, как хлеб или вода. Но оружие — это очень дорогой инструмент. Мы же не можем продавать всем, например, все сложные медицинские инструменты. Для этого есть больницы. Точно также для ношения оружия есть армия, есть полиция. В каких-то локальных случаях его могут использовать охотники или спортсмены. Но гражданским оружие не нужно, потому что сам факт ношения оружия каждым человеком уже увеличит общественную тревожность. У нас и так депрессивное общество, уставшее от кризисов и конфликтов. А здесь в нем будет постоянный невроз, что только еще сильнее увеличит вероятность каких-либо столкновений. И у них для этого будут все способы для того, чтобы убить своих оппонентов.

[Команда «Проектор Перестройки»]: Среди людей так высокое недоверие друг к другу. Если у всех будет оружие, это приведет просто к общественной паранойи. К тому же, у многих людей нет денег на оружие. Многие люди из-за каких-то религиозных взглядов не могут, не захотят нажать на курок, направив оружие на другого человека. Поэтому данному радикальному концепту просто не место в нашей жизни. Пусть оно останется прерогативой людей, которые готовы с ним обращаться и умеют это делать.</first_round_team_B_arguments>

Fig. 1. First round markup example

<second_round_question_BtoA_2>

[Команда «Полуфабрикаты»] Если ценность высшего образования для карьеры так велика, то почему в таком случае процент трудоустроенных гораздо выше среди тех, кто обладает дипломом среднего профессионального образования? Время пошло.

[Команда «Раскол»]: Твой вопрос в том, почему малое количество трудоустроенных, потому что человек, когда начинает обучаться в процессе, он может понять, что это не его, однако у него останутся эти знания, и эти знания он может все равно в своей жизнедеятельности применять. Нет, он эти знания может применять, и также они могут быть смежными. Например, если я учусь на журналиста, то я также могу работать и SMM-менеджером, о котором мы говорили, писать посты. Я могу быть телеведущей и разговаривать. Это все одни и те же навыки и строятся вокруг одного и того же. Например, гуманитарий и технарь. Я бы хотел добавить, что по поводу трудоустроенных людей с образованием средним, в этих сферах, то есть если ты, допустим, электрик, сварщик, у этих профессий, да, проще устроиться, но у этих профессий абсолютно четкий потолок. То есть вы дальше, если вы электрик, дальше там 200 тысяч, 250. Вы не прыгнете никогда. Вам нужно будет переквалифицироваться, что-то делать, что-то менять. У всех этих профессий есть четкий потолок. Который как бы какую-то великую успешную карьеру не позволит сделать. Но да, он позволит вам устроиться на работу, если вы об этом.

[Команда «Полуфабрикаты»]: А почему вы забываете про самообразование, про то, что есть великая штука интернет или, допустим, курсы, великий Skillbox?

[Команда «Раскол»]: Интернет сейчас бесплатный, однако не все пользуются этими возможностями. У нас образование не на таком уровне, если бы все пользовались этими доступными источниками.

</second_round_question_BtoA_2>

<second_round_question_AtoB_3>

[Команда «Раскол»]: Смотрите, среда вуза, получается, располагает к знакомствам и связям. Когда человек работает, у него попросту на это времени нет. Где брать связи?

[Команда «Полуфабрикаты»]: Ну, ответ, опять же, состоит из двух частей и довольно простой. Если ты интроверт, социопат и, в принципе, не особо-то общительный, то тебе не спасет ни вуз, ни работа, ничего. Ты будешь точно так же закрытым, у тебя никого не будет из знакомых, друзей, твой круг не будет расширяться просто потому, что ты такой человек. Как вы сами сказали, все зависит от личности, от человека.

И вторая часть моего ответа заключается, что профессиональный круг общения гораздо более ценный, потому что те люди, которые попадают в университет, довольно много тех, кто не заинтересован в учебе, довольно много тех, кто никогда не пойдет в той же стезе, что и вы, и будет слишком далек от вас по мировоззрению, по роду деятельности и так далее. А профессиональный круг в этом плане гораздо конкретнее. То есть ты обретаешь контактами, которые тебе в будущем однозначно помогут, поскольку ты уже существуешь в этой сфере.

[Команда «Раскол»]: Однако ваши контакты очень ограничены этой профессией, а в вузе человек может выбрать другие. И в вузе можно выбрать человека по другим направлениям. И познакомиться, обрести огромным просто нетворкингом не в одной сфере, а во многих сразу.

[Команда «Полуфабрикаты»]: Но это возвращается к вопросу личности. Если личность такова, что человеку хочется больше, хочется узнать других людей и так далее, это не проблема. Независимо от вуза он найдет точно так же себе широкий круг общения и вне его.

Тут дело в том, что... образование еще очень сильно отстает от самой сферы. От социальной среды вне. Да, от социальной среды. Пока тебя 4 года обучают одному, мир очень быстро меняется сейчас. Ты выходишь и у тебя происходит несовпадение образов того, что ты получил в вузе, с тем, как это было на практике. И твои социальные контакты уже скорее всего не очень важны.

</second_round_question_AtoB_3>

Fig. 2. Second round markup example

Review answers

———— REVIEW 1 ————

SUBMISSION: 38

TITLE: A Dataset of Russian-Language Debates For Argument Mining

AUTHORS: Rodion Sulzhenko and Boris Dobrov

———— Overall evaluation ————

SCORE: 0 (borderline paper)

—— TEXT:

The paper addresses the lack of Russian-language datasets for argument mining. The authors collected and annotated student debates, resulting in 429 arguments, and evaluated the dataset on stance detection and argument generation tasks.

Issues:

- The abstract contains line breaks and an inline GitHub link.
- The claim that "Argument mining remains critical for developing language models (LLMs) capable of nuanced reasoning in culturally specific contexts" is questionable. To my knowledge, argument mining datasets are not directly used to enhance LLMs' reasoning capabilities.
- The second paragraph of the introduction is duplicated.
- Footnotes begin with number 3.
- The terms "West" and "Western" are not scientific descriptors.
- It is evident that Generative AI was used in preparing the paper. While I am not opposed to this, such use should be explicitly acknowledged.
- The generated arguments were evaluated using an LLM-as-Judge approach. Since LLMs are not fully reliable, this method is not a robust evaluation metric.
- There is no evaluation of the annotation process itself, for example, in terms of inter-annotator agreement.
- The paper contains only 10 pages, of which 3 are appendices, making it a short paper.
- The claim that "the absence of Russian argument-annotated corpora severely limits Argumentation Mining development for Slavic languages... none currently exist for Russian-language materials" is inaccurate. Paper [7] already presents a Russian-language dataset. Moreover, the RuArg competition (<https://github.com/dialogue-evaluation/RuArg>), in which one of the authors of this paper is listed as an organizer, also features a Russian-language argument mining dataset.

Overall, I acknowledge that any Russian-language resources for argument mining are valuable. However, this paper should be considered for acceptance as a short paper only after the identified issues are addressed.

Answers:

- **Abstract formatting.** We revised the abstract to remove line breaks and moved the GitHub link into a footnote for clarity.

- **Claim about LLM reasoning.** We softened this statement to emphasize that argument mining resources can inform research on culturally specific discourse, but do not directly enhance LLM reasoning capabilities.
- **Duplicated paragraph.** The repeated paragraph in the introduction has been removed.
- **Terminology (“West/Western”).** We replaced informal terms with precise references to “English-language” or “Anglophone” resources.
- **Use of Generative AI.** As requested, we explicitly acknowledged in the Acknowledgements section that generative AI tools were used for English proofreading and formatting only.
- **Evaluation with LLM-as-Judge.** We agree that LLM-based judging has limitations. To mitigate this, we added a complementary human evaluation on a 20-argument subset, described in a new section.
- **Annotation evaluation.** The dataset was annotated by a single annotator. We clarified this process in corresponding section and explained why stance labeling was relatively straightforward. Future releases will involve multiple annotators to allow inter-annotator agreement measurement.
- **Length of paper.** While the main text is concise, we expanded Related Work, Annotation, and Evaluation sections, bringing the total length closer to the standard.
- **Russian-language corpora.** We corrected our earlier claim and now explicitly discuss RuArg [11] as a key Russian-language dataset, contrasting it with DebateRu.

————— REVIEW 2 —————

SUBMISSION: 38

TITLE: A Dataset of Russian-Language Debates For Argument Mining

AUTHORS: Rodion Sulzhenko and Boris Dobrov

————— Overall evaluation —————

SCORE: 1 (weak accept)

— TEXT:

The paper introduces DebateRu, a novel dataset of Russian-language student debates, annotated for argument mining tasks. The dataset addresses a research gap in non-Western, culturally specific resources for argumentation analysis. The authors evaluate the dataset on stance detection and argument generation tasks using state-of-the-art language models and highlight the challenges LLMs face in capturing contextual and culturally nuanced argumentation.

The paper is of scientific and practical interest. However, some details need to be specified. In particular, the work provides a few information regarding the annotation process.

1. The related work section does not mention the RuArg dataset. It covers another domain, however, RuArg is a notable source for the argument mining task in Russian.

2. The selection of data source (Debattle) is not explained.
3. The annotation process is described very briefly. How was the annotation performed? How many annotators were involved? How was the inter-annotator agreement calculated? How were the topics selected?
4. The evaluation revealed the superiority of the Qwen model, which was the largest in the evaluation. Is this superiority related to the model size or any other characteristics?
5. The authors indicate the small size of the dataset. Please clarify your plans on dataset expansion.

Answers:

1. **RuArg.** We added discussion of RuArg in the Related Work section and compared it with DebateRu.
2. **Data source (Debattle).** We explained our rationale for choosing Debate: reproducibility, structured stance-bearing format, topical relevance, and audio quality.
3. **Annotation details.** We expanded the description of annotation, including challenges, single-annotator setup, and future plans for multiple annotators.
4. **Qwen model performance.** In the Human Evaluation section, we added that Qwen2.5-72B's superiority is likely due to model scale but may also reflect training data adaptation.
5. **Dataset expansion.** We clarified in the Conclusion that we plan to increase the dataset size, cover more debate formats, and include multiple annotators.

———— REVIEW 3 ————

SUBMISSION: 38

TITLE: A Dataset of Russian-Language Debates For Argument Mining

AUTHORS: Rodion Sulzhenko and Boris Dobrov

———— Overall evaluation ————

SCORE: 2 (accept)

— TEXT:

The paper proposes a new open dataset to train and test argument mining models on Russian texts. The paper is well-written. It describes in detail the dataset, the procedures of obtaining the data, making scripts, and labeling. The paper also provides results of several baseline experiments on the new dataset.

The paper has a couple of minor issues to be taken care of:

1. Introduction. The passage "To address this gap, we introduce DebateRu...." is repeated twice.
2. The scores for the "Argument generation" subtask are obtained with a third-party LLM-based tool. Therefore, the scores might be biased, showing more GPT-4 vision of consistency and persuasiveness of the generated than the real scores. It would be nice to have some human validation, at least for a test subset.

Answers:

1. **Duplicated introduction passage.** Corrected as noted.
2. **Argument generation evaluation.** In addition to LLM-as-Judge, we now report results of a small-scale human evaluation, which confirms the main trends.