LoveHate: Stance Detection and Generation for Multiple Topics in User-generated Comments in Russian and English

Natalia Evgrafova, Véronique Hoste and Els Lefever
LT3, Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent
{natalia.evgrafova, veronique.hoste, els.lefever}@ugent.be

Abstract

This paper introduces LoveHate, a new multi-topic corpus of user-generated arguments in Russian, collected from the historical data of the debate platform lovehate.ru. The dataset contains nearly 19,000 posts spanning 16 socially and politically relevant topics, each mapped to binary pro and con stances. We test multiple approaches to stance detection and stance generation across Russian and English data, including translated variants, using both classifier-based (Roberta, RuRoberta) and instruction-tuned generative (Llama, Qwen) models. Results demonstrate that language-specific pretraining yields the strongest performance for stance classification (F1 = 0.892 with RuRoberta), while multilingual generative models – when fine-tuned on sufficient data – can effectively generate stance in Russian without explicit Russian pretraining. Cross-domain experiments show that English datasets generalise better across corpora, whereas Russian data capture language- and culture-specific argumentation but are less effective for generalizable models. Generating topics remains a more challenging task for both Russian and English data. The dataset and accompanying results contribute to multilingual stance research and provide a valuable new resource for argument mining in Russian.

Keywords: argument mining, stance detection, stance generation, Russian

1. Introduction

The field of argument mining relies on high-quality annotated data. To date, various annotation schemes have been proposed (Habernal et al., 2014; Ajjour et al., 2019) with more or less fine-grained categories. Traditionally, an argument is represented as a claim supported by one or more premises or reasons. Claim-premise schemes are quite suitable for well-organised and explicit texts, such as argumentative student essays: there is always a clear topic, a claim that a student agrees or disagrees with, and explicitly stated supporting reasons. By contrast, social media texts tend to be less explicit: a claim or a premise might be implicit and need to be inferred from the context. Moreover, such texts can be ambiguous and interpreted differently by different people.

A more general subtask within argument mining is stance detection: the task of inferring pro or con stance toward a given topic. This task does not require very elaborate theoretical frameworks, but it typically relies on pre-annotated or existing topic(s) to annotate the comments reliably. The main limitation of stance detection is that classifiers are usually fine-tuned per topic, which restricts the models' applicability to real-world data. This can be partly overcome if there is enough high-quality data for various topics to fine-tune multi-topic models.

Since annotating user-generated texts is a substantial and costly effort, researchers have looked into using data from various debate platforms, as they provide such stance annotations from *PRO* or *CON* sections of a given debate. Such argument datasets are available for English, with the largest corpus amounting to 387,606 arguments on 72,121 topics (Ajjour et al., 2019).

Although Russian is considered a high-resource language, we do not find sufficient argument data in Russian. The largest existing corpus of arguments in Russian is the Russian version of Microtext (Fishcheva and Kotelnikov, 2019), which is based on the Argumentative Microtext Corpus, a collection of 112 short texts in English and German on various topics (Peldszus, 2015) as well as 171 texts in English only (Skeppstedt et al., 2018), both translated into Russian by a professional translator. Another resource in Russian consists of 9,550 sentences from social media posts on topics related to the COVID-19 pandemic (vaccination, quarantine, and wearing masks) (Kotelnikov et al., 2022). In this paper, we present our corpus of Russian arguments from a debate platform and focus on the following research questions: (1) Do we need annotated data in Russian, or can we use machine-translated data from English to perform stance detection? (2) Do we need Russian pretrained models, or can we effectively fine-tune English or Chinese models? (3) Do fine-tuned stance-detecting classifiers work better than instruction-tuned generative models prompted to generate stance? (4) To what extent is it possible to infer topics automatically in multi-topic stance detection based on our corpus?

The remainder of the paper is organized in the

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8014–8025
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

following sections: Related work (section 2), LoveHate dataset (section 3), where we present the dataset and describe how the data were collected, Methods (section 4), where we outline our experiments, Results (section 5), reporting the scores and presenting error analysis, Conclusions (section 6), Ethical statement (section 7), and Limitations (section 8). We make our data available in a Hugging Face collection[1].

2. Related work

2.1. Data from debate platforms

As mentioned before, one of the largest argumentative corpora today is Webis args.me (Ajjour et al., 2019), which was collected from idebate.org, debatepedia.org, debatewise.org, and debate.org, of which only debatewise.org is currently active, and contains topics and stance labels (pro and con). Despite an extensive number of topics (72,121), some of them are overlapping, and while they include a range of highly debatable issues, there are also topics like "Rap Battle" or "Yo mama jokes", which are a two-team competition rather than a debate about an issue. Based on this corpus, the authors introduced an argument search engine that retrieves relevant pro and con arguments based on a user's query.

Another collection of arguments in a threaded dialogue structure sourced from debate platforms (4forums.com, CreateDebate.com, Convinceme.Net, all of them are no longer active) is presented in the Internet Argument Corpus 2.0 (Abbott et al., 2016) that covers 482K posts on various topics, many of which are annotated for both topic and author stance.

The Winning Arguments Corpus (ChangeMyView) (Tan et al., 2016) contains threads from the ChangeMyView subreddit, but it does not include annotations for stance.

A substantial corpus collected from the Kialo debate platform was presented in Agarwal et al. (2022) with 1,560 discussion threads and 324,373 arguments as of 28 January 2020 in English.

Recent resources focus on multilingual arguments, including the NLAS-multi corpus (Ruiz-Dolz et al., 2024) that consists of 1,893 automatically generated and curated arguments in English and 1,917 in Spanish, representing 20 argumentation schemes.

2.2. Stance detection and generation

The task of stance detection is typically performed with respect to a specific topic. Early research in this area focused on fine-tuning topic-specific classifiers for single targets (Mohammad et al., 2016; Wei et al., 2018) or target pairs (Sobhani et al., 2017). Later, multi-target approaches were explored, where models were trained on all target pairs simultaneously (Li and Caragea, 2021). Besides these approaches, cross-target methods have also been explored to see if models can generalise well to similar but unseen topics (Augenstein et al., 2016; Zhang et al., 2020). Finally, zero-shot stance detection refers to detecting stance for new, unseen targets, which has recently gained more attention (Allaway and McKeown, 2020; Zhao and Caragea, 2024; Fan et al., 2025). One way to generalise stance detection to new and unseen topics is by classifying or generating targets first and then using these targets for stance prediction. The task where a target-stance pair is extracted given a sequence is known as Target-Stance Extraction (Li et al., 2023a).

Commonly used methods include training classifiers, e.g., BiLSTM and/or BERT (Augenstein et al., 2016; Zhang et al., 2020; Allaway and McKeown, 2020; Li et al., 2023a), fine-tuning BART for topic generation (Li et al., 2023a), fine-tuning Llama, Mistral, Claude, GPT models using instruction prompts, or in few-shot and zero-shot scenarios to generate targets and stance (Fan et al., 2025; Zarharan et al., 2025).

Beyond these modeling approaches, recent work has sought to further enhance stance detection by augmenting the data by diversifying targets via keywords generation and utilizing a teacher-student framework to use the augmented data (Li et al., 2023b), by incorporating external knowledge in the form of semantic and emotion lexicons for cross-target stance detection (Zhang et al., 2020), and by providing contextual knowledge from large language models and by leveraging reasoning approaches (Taranukhin et al., 2024; Fan et al., 2025).

Overall, at the current stage, the task of stance detection aims to overcome the limitations arising from the data scarcity, lack of contextual information and world knowledge, as well as the need to effectively detect or generate new topics and their corresponding stance.

In this paper, we address the existing challenges by providing labeled data in Russian on a variety of topics, and by comparing the performance of stance detection and stance generation models, as well as the robustness of classifiers given unseen topics and data from a different source. Furthermore, using the topic data, we demonstrate our instruction-tuned models' ability to generate relevant topics and stances at inference time.

[1] LoveHate.RU

3. LoveHate dataset

In this section, we present a new dataset in the Russian language for the tasks of stance detection, stance generation, and stance-target generation. We describe the data collection, labels, and the final dataset description and topic analysis. Our goal is not only to provide information about the dataset, but also to test our hypothesis that language-specific data for argument mining is needed due to language-specific differences and topics that are underrepresented in settings of automatic translation. Since most of the resources are in English, we compare our corpus to existing English data. Specifically, we chose a subset of the Webis args.me dataset that covers politics-related topics, six of which overlap with the LoveHate corpus. We further refer to the chosen subset of the Webis args.me corpus as the Webis dataset.

3.1. Data collection

We present a corpus of Russian arguments collected from the now-closed platform *lovehate.ru*, where users discussed topics ranging from political issues and policies to popular music bands. Since the website is no longer accessible, as is the case with many other online debating platforms, we accessed it using WayBack Machine[2]. The collected data spans the years 2000 to 2019, and for this study, we selected 16 highly-debated topics related to politics and policy-making. The platform was structured such that each user was required to choose either the *love* or *hate* side; these sides can be effectively mapped to *pro* and *con* stances, respectively. While in this setting there is no *neutral* stance, we consider the author's initial pro or con side selection to be indicative of their overall predisposition: even if a comment expresses a balanced opinion, the chosen side still reflects the author's general stance. User comments typically express support or opposition toward the topic with varying degrees of intensity, and if a comment is not irrelevant to a topic, in the majority of cases it can be labelled as *pro* or *con*.

The resulting dataset comprises 18,982 posts. We removed usernames, and since the website is no longer accessible, the original comments and usernames can only be accessed via WayBack machine access to the whole website, which does not provide user personal data apart from their username.

3.2. Topic comparison

In this study, we selected topics that (1) were popular among users during the period 2000–2019 and (2) were politically or socially relevant. This selection enables us to examine whether the resulting topics reflect issues specific to the Russian language and sociocultural context, as well as to explore those that overlap with topics commonly discussed in English-language data.

Figure 1 presents the most commented topics in the LoveHate Corpus as well as those from the Webis dataset. While Webis features *Abortion*, *Gay marriage*, *Atheism*, *God*, and *Death penalty* as the most discussed topics – with *Abortion* receiving the highest number of comments – the LoveHate dataset has overall a more balanced number of comments for its most popular topics that include *Americans*, *God*, *Russians*, *Atheism*, and *Stalin*. Although not all topics were included, we observe that both datasets contain language-specific and relevant topics.

The LoveHate dataset also has a wider variety in the pro and con comments proportions, with distinctly most loved or supported topics, as well as those that are most hated or opposed, as shown in Figure 2, with *United Russia* (political party) and *School uniform* showing the largest proportion of con comments, and *Russian language* having the largest proportion of pro comments, while in the Webis dataset we observe a more even distribution of pro and con comments per topic, except for *Withdrawal from Iraq* that has more pro than con comments.

These topics' popularity coincides with the interest in these topics as supported by Google Trends data[3]. Figure 3 demonstrates spikes of interest in the search for *Americans*, and a weaker interest in *Abortion*, especially given the period of our observed data, 2000-2019. There are also significant differences in the interest in *Stalin* in Russia and the USA.

The analysis of topics across both datasets reveals that language-specific data can differ substantially in the range and popularity of topics they cover. Even with a high-quality translation of an English dataset, such an approach may not sufficiently capture language-dependent contexts or locally important issues.

3.3. Arguments comparison

We retained a few overlapping topics between the two datasets, which are listed in Table 1 together with the corresponding number of arguments in each corpus.

To gain initial insights into argumentation patterns, we analyzed the most prominent keywords across topics using the YAKE! keyword extraction algorithm (Campos et al., 2020, 2018a,b).

[2] https://web.archive.org/

[3] https://trends.google.com/trends/

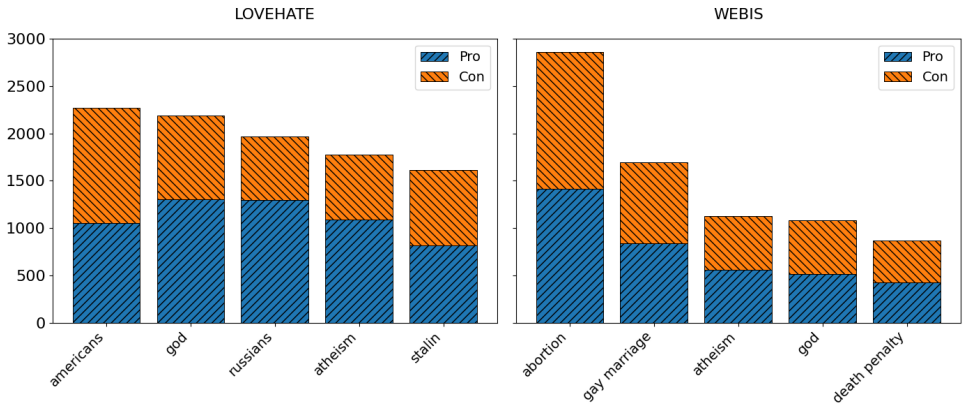

Figure 1: The bar charts show the most commented topics in the Russian-language LoveHate corpus and in the English-language Webis corpus. The number of pro and con arguments per topic is indicated by colors and hatches.

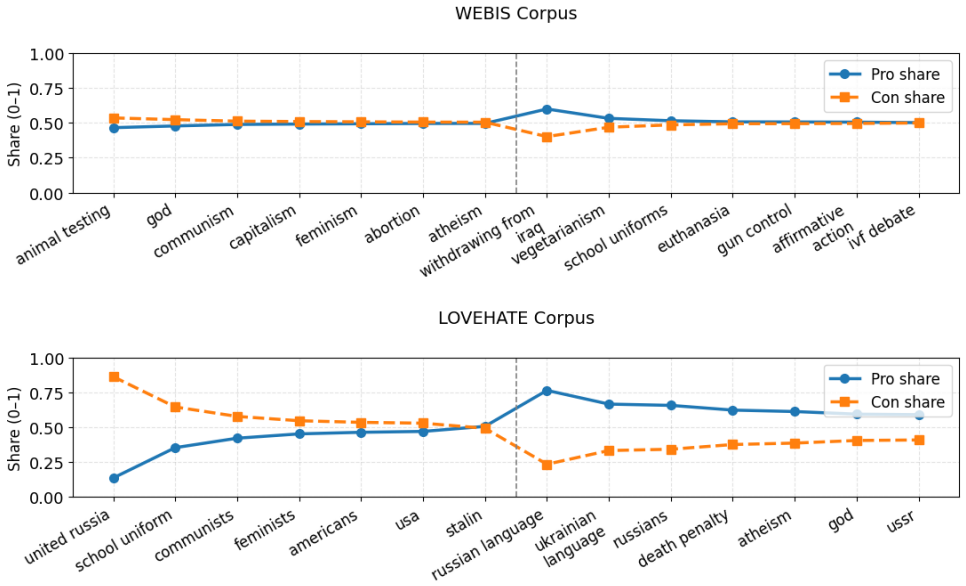

Figure 2: The graphs demonstrate the relative proportion of pro and con comments per topic in the LoveHate and Webis datasets.

Topic	LoveHate	Webis
School uniform	350	305
Atheism	1,778	1,124
God	2,189	1,085
Death penalty	685	869
Feminists/Feminism	956	144
Communists/Communism	1,602	170

Table 1: Counts of arguments in overlapping topics in the LoveHate and Webis datasets.

This helped us identify the lexical focus of stance-specific arguments for each dataset. A detailed list of extracted keywords per topic and stance is provided in the Appendix (Table 7). Overall, we observe that while *LoveHate* arguments tend to contain more personal and emotionally charged expressions, *Webis* arguments rely more heavily on formal reasoning and general statements. For instance, in the *Death penalty* topic, *LoveHate* arguments frequently mention words such as *children*, *sin*, or *maniacs*, whereas *Webis* arguments emphasize *prevent*, *guilty*, and logical linking words as in *because they kill*. This lexical difference aligns with the datasets' origins – *LoveHate* containing more spontaneous and emotional user-generated discourse, and *Webis* comprising more neutral debate material. We have also observed that some emotional slang words were not always correctly automatically translated, but transliterated (see "Tupo" (stupid) in Table 7, LoveHate (Con) for *school uniform*), the general meaning of sentences, however, was preserved.

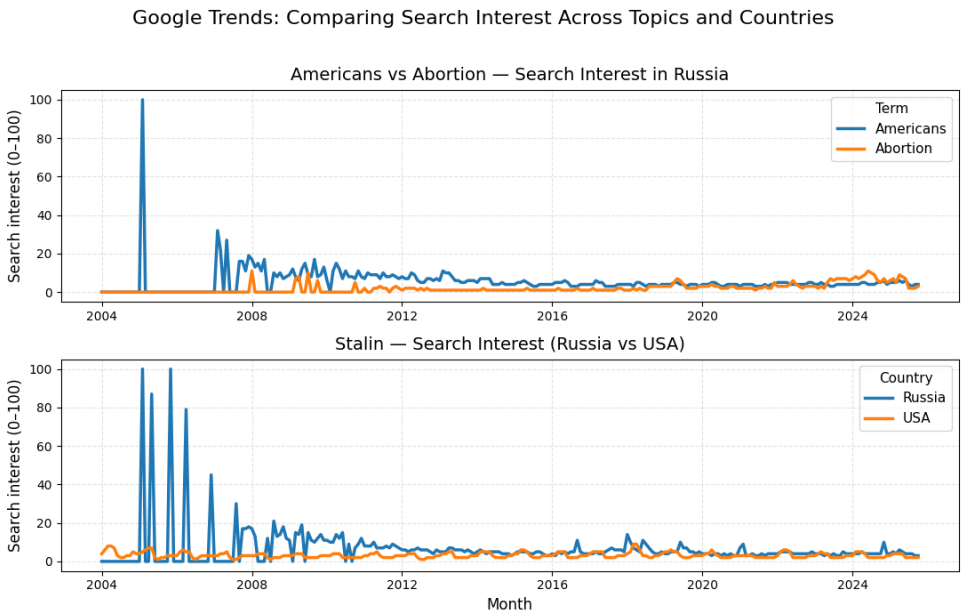

Figure 3: The search interest in the topic terms on Google as supported by data from Google Trends.

4. Methods

In this section, we outline the approaches, models, and parameters used across three experimental settings: (1) stance detection: we fine-tune classifiers on multiple topics that are prepended to the input text in both the original language and translation (Ru<->En); (2) stance generation given a topic: we instruction-tune decoder-only models to generate a stance, given a topic; (3) stance and topic generation: we instruction-tune decoder-only models to generate both the stance and topic at inference time.

4.1. Stance detection

For the task of stance detection, we used two models – the English-based Roberta-large (Liu et al., 2019) `FacebookAI/roberta-large` and the Russian-pretrained ruRoberta (Zmitrovich et al., 2024) `ai-forever/ruRoberta-large`. RuRoberta is based on the configuration of Roberta-large and pretrained on the large Russian-language corpus collected from Wikipedia (Attardi, 2015), news[4], books (Panchenko et al., 2017), and other, including web texts and subtitles in Russian.

We fine-tuned these models first on the corresponding English and Russian data, as well as translated data (Ru->En and En->Ru). Comments had been translated using the `google-trans` client[5]. The train, test, and val splits are shown in Table 2.

[4] https://github.com/natasha/corus/tree/master

[5] Copyright (c) 2015 SuHun Han

Dataset	Train	Val	Test
Webis	6,921	1,863	1,864
LoveHate	12,290	3,309	3,310

Table 2: Train, validation and test splits used for Roberta and RuRoberta models fine-tuning on stance-detection task.

We fine-tuned the models for binary stance classification (*pro*/*con*). Each input sequence was prepended with its corresponding topic ("<*topic*> || <*comment*>") to provide explicit contextual information. The models were trained for eight epochs with a learning rate of 1×10^{-5}, weight decay of 0.01, and an effective batch size of 64 (per-device batch size of 8 with gradient accumulation of 8).

4.2. Stance generation

For stance generation with a prepended topic, we used two decoder-only large language models: `meta-llama/Llama-3.1-8B` and `Qwen/Qwen2.5-7B-Instruct`. In addition to English, the LLaMA-3.1-8B model (Touvron et al., 2023) supports seven other languages (German, French, Italian, Portuguese, Hindi, Spanish, and Thai), but not Russian. According to the model card on Hugging Face[6], it was pretrained on a multilingual corpus and can be further fine-tuned on additional languages. The Qwen model (Yang et al., 2025) is a multilingual instruction-tuned model that, in addition to Chinese, supports over

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

28 languages, including Russian[7]. Both models are actively used in the research community, achieve strong performance on inference tasks, and are comparable in terms of the number of parameters.

We instruction-tuned the models to generate a one-token stance label (*pro*/*con*) given a prepended topic. The prompts included the instruction:

`Determine the author's stance (pro or con) on the given topic. Answer with exactly one word: pro or con.`, and a user input that concatenated the topic and comment:

`Topic: <topic>\Comment: <comment>.`

We used LoRA (Hu et al., 2022) (`r`=8 `lora_alpha`=16), and training was performed with a learning rate 2×10^{-4}, batch size 8, and gradient accumulation of 4 (effective batch size 32), for 2 epochs; we evaluated and checkpointed every 200 steps and loaded the best model at the end. Data were split into train, val, and test as shown in Table 2, ensuring balanced stance labels.

4.3. Stance and topic generation

For stance and topic generation, we used the same models as in subsection 4.2 and instruction-tuned them in a similar manner, but this time we instructed the models to infer both stance and topic. We used the following prompt:

`Infer the short topic name and the author's stance. Reply in EXACTLY two lines and exactly this format: stance: pro|con topic: <topic>`

The training parameters and data splits were the same as in subsection 4.2.

Given the difficulty of jointly inferring stance and topic from open-ended user arguments, we conducted an additional series of controlled prompting experiments. The aim was to test whether different forms of task conditioning could affect model consistency and reduce topic drift. We considered three experimental settings:

(1) **Topic list provided**: In this setting, the model is presented with the full list of candidate topics and instructed to select the most appropriate one for the given comment. This constrained format turns topic generation into a classification-like decision. Although this approach requires a predefined topic list, it is particularly useful in scenarios where the set of possible topics is known in advance and the goal is to explore how well the model can discriminate among them. For this setting, we constructed the following prompt:

`Your task is to infer the attitude of this comment towards a topic. Think briefly and choose exactly one topic from the list. Then output: stance: pro|con topic: <selected topic>`

(2) **Keyword-augmented prompting**:

Here, we enriched each comment with a short list of topic-related keywords automatically extracted using YAKE!, restricted to bigrams and a total of 6 words. The goal was to guide the model with minimal semantic context – representative cues – to support topic inference. Here, the prompt was:

`Your task is to infer the attitude of this comment towards a topic. Consider the provided keywords. Then output: stance: pro|con topic: <topic>`

(3) **Context-augmented prompting**:

In the final setup, we instructed the model to briefly describe the context of the comment before providing the final stance and topic labels. Topic labels also served as gold labels for the context. In this way, we checked if generating the comment's topic first could yield improvements. We used the prompt below:

`Your task is to infer the attitude of this comment towards a topic. Think briefly about what this comment is about (the context), then output exactly in this format: context: <> stance: pro|con topic: <topic>`

5. Results

5.1. Stance detection

Table 3 presents the macro-F_1 scores for stance detection across the LoveHate and Webis datasets. The best performance ($F_1 = 0.892$) was achieved by `RuRoberta` when trained and evaluated on Russian data, confirming the benefit of language-specific training. `Roberta` reached comparable performance on the English Webis dataset ($F_1 = 0.855$), but its effectiveness dropped notably when applied to the Russian data ($F_1 = 0.658$–0.734). Training on the data translated into English (`trEN`) showed promising results, with `Roberta` reaching $F_1 = 0.863$ and, surprisingly, `RuRoberta` achieving $F_1 = 0.792$. However, translation alone did not fully bridge the gap with native-language performance.

Overall, these results demonstrate that the best performance is achieved when the language of the data and the model match, i.e., Russian data with `RuRoberta` and English data with `Roberta`. The translated LoveHate data used with `Roberta` also shows good results, which may be attributed to the fact that the Russian data from the LoveHate corpus may be more expressive and often rely on sentiment, contributing to higher model performance.

After evaluating the models on their respective test sets, we also evaluated them on the test sets

[7] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Dataset	Train	Model	Macro F_1
LoveHate	RU	RuRoberta	**0.892**
LoveHate	trEN	Roberta	0.863
Webis	EN	Roberta	0.855
Webis	trRU	Roberta	0.734
LoveHate	RU	Roberta	0.658
LoveHate	trEN	RuRoberta	0.792

Table 3: Macro F_1 scores of RuRoberta and Roberta on LoveHate and Webis datasets. TrEn and TrRu mean the original data was automatically translated into English and Russian, respectively.

of the other dataset to investigate cross-dataset generalization. This experiment aimed to evaluate whether a model trained on one collection could accurately detect stance in a different dataset written in the same language.

Table 4 presents the results of the cross-dataset generalisation experiments. Interestingly, while all models exhibit a performance drop compared to their in-domain scores, the models trained on Webis data generalise notably better to the LoveHate test sets (Macro F_1 of 0.775 for English and 0.724 for Russian) than in the reverse direction, where LoveHate-trained models achieve lower scores on Webis (0.546 and 0.503, respectively). This pattern suggests that the more stylistically homogeneous Webis data, originally written in English, enables models that generalise more effectively to both the Russian original and its translated variants, although none surpass their in-domain benchmarks.

Train Dataset	Test Dataset	Macro F_1
LoveHate (RU)	Webis (RU)	0.503
LoveHate (EN)	Webis (EN)	0.546
Webis (RU)	LoveHate (RU)	**0.724**
Webis (EN)	LoveHate (EN)	**0.775**

Table 4: Cross-dataset generalisation results (macro F_1) for Roberta (for EN) and RuRoberta (for RU) models trained on LoveHate and Webis datasets. Each model is evaluated on the test set of the other dataset in the corresponding language.

To summarize, and in line with our initial hypothesis, the model pretrained and fine-tuned on the Russian data achieves the best overall performance in stance detection. However, contrary to our expectations, the models trained on the original English data and its translated variant exhibit stronger cross-dataset generalisation, suggesting that the English data provides features that transfer more effectively across corpora.

5.2. Stance generation

The results in Table 5 demonstrate that both Llama and Qwen models achieve strong performance in stance generation, with Llama consistently outperforming Qwen across most datasets. The highest score is achieved by Llama on the original Russian LoveHate dataset ($F_1 = 0.899$), indicating that the model effectively adapts to Russian stance data despite its predominantly English-centric pretraining. Performance slightly decreases on automatically translated data (LoveHate(trEN)), though both models maintain comparable levels of accuracy. On the Webis dataset, results show the opposite trend: Qwen performs better on the translated Russian version (Webis(trRU)), while Llama achieves higher scores on the original English dataset. We attribute that to a more balanced average length of comments in the LoveHate dataset as well as its expressiveness.

Model	LoveHate (RU)	LoveHate (trEN)
Qwen F_1	0.807	0.781
Llama F_1	**0.899**	0.873
Model	**Webis (trRU)**	**Webis (EN)**
Qwen F_1	0.824	0.771
Llama F_1	0.772	0.796

Table 5: Macro F_1 scores of `Qwen` and `Llama` models on LoveHate and Webis datasets. *trEN* and *trRU* indicate automatically translated data.

At an earlier stage, we experimented with an instruction that prompted the models to generate strings like:

`"I support <topic>"`, or `"I oppose <topic>"`.

This proved more difficult to post-process and to map back to stance and topics. Overall, the prompt that explicitly instructed the model to generate concrete variants with a given format resulted in the cleanest and most consistent output.

5.3. Stance and topic generation

Generating both stance and topic proved as expected to be a more challenging task for both the Llama and Qwen models. Table 6 demonstrates F_1 and exact match[8] for stance and topic generation with no additional context given, with topic list provided, with keywords from YAKE!, and with context provided by the models at inference time.

Providing keywords, topics, and context consistently improves model performance. When no topic or keywords are given, both models perform moderately on stance detection (F_1 0.55–0.73) and even worse on topic recovery (F_1 0.37-0.68). Supplying explicit topic information, as a topic list (see *topics given*) or as a gold standard for context (see *context given*), yields the highest scores, with stance F_1 approaching 0.88 on the Russian LoveHate dataset

[8] https://huggingface.co/spaces/evaluate-metric/exact_match

Setting	Model	Dataset	Stance F_1 (macro)	Topic EM	Topic F_1 (macro)
No additional context given	Llama	LoveHate (RU)	0.734	0.529	0.668
	Qwen	LoveHate (RU)	0.709	0.536	0.675
	Llama	Webis (EN)	0.550	0.225	0.368
	Qwen	Webis (EN)	0.563	0.201	0.365
Topics given	Llama	LoveHate (RU)	**0.880**	**0.872**	**0.889**
	Qwen	LoveHate (RU)	0.862	**0.872**	**0.889**
	Llama	Webis (EN)	0.695	0.800	0.843
	Qwen	Webis (EN)	0.781	0.767	0.827
Keywords given	Llama	LoveHate (RU)	**0.882**	0.808	0.840
	Qwen	LoveHate (RU)	0.850	0.809	0.867
	Llama	Webis (EN)	0.683	0.189	0.313
	Qwen	Webis (EN)	0.779	0.222	0.524
Context given	Llama	LoveHate (RU)	**0.882**	**0.874**	**0.890**
	Qwen	LoveHate (RU)	0.846	0.852	0.873
	Llama	Webis (EN)	0.818	0.780	0.844
	Qwen	Webis (EN)	0.823	0.778	0.838

Table 6: Macro F_1 for stance, exact match (EM), and macro F_1 for topics across settings, models, and datasets.

and 0.82 on the English Webis dataset, accompanied by topic F_1 values near 0.89 and 0.84, respectively. *Keywords-given* condition helps stance recognition but does not always succeed at topic inference, suggesting that isolated keywords offer limited semantic grounding. Comparing *topics given* and *context given* settings, we observe more stable stance generation across models and datasets under the latter. The qualitative analysis of the output shows that providing the topic list during inference helps reduce off-topic generations and inconsistent topic formulations.

Across languages, performance is consistently higher on the Russian LoveHate corpus than on the English Webis dataset. We attribute this mostly to a more emotional language in the LoveHate dataset, which makes it easier for models to learn stance cues. Comparing models, Llama demonstrates slightly stronger and more stable results than Qwen. These findings indicate that (i) a proper topic reconstruction remains a challenge, (ii) prompt specificity has a major influence on both tasks, and (iii) lightweight contextual scaffolding – such as explicit topics – substantially enhances multi-aspect stance modeling. This setting can benefit from adding a neutral stance for contexts where the topics of interest are already known.

5.4. Error analysis

The qualitative analysis of the output revealed the following trends. Llama tended to generate topics that it had seen and that were somehow close in meaning, e.g. *Communism* instead of *Liberalism*, *Russian women* instead of *Feminism*, *Russian atheists*/*Communists* instead of *Atheism*. Qwen would sometimes generate more specific topics, e.g. *Stalingrad battle* instead of *Stalin* or *USSR*, or related topics that are not a topic of the original comment: *Russian men* instead of *Feminism* in a comment that indeed attacks *Russian men*. Accounting for such topics that are not per say wrong but differ from the gold standard presents another challenge for automatic topic generation.

Overgeneration or repetition of tokens was quite a common source of error. Less frequent issues included mixed or truncated topics and stance-topic mismatches. These suggest that generation control and topic disentanglement remain open challenges for instruction-tuned models in stance reasoning.

6. Conclusions

We presented LoveHate, a new Russian-language corpus for stance and topic detection and generation, together with a comparative evaluation against English debate data. Our experiments demonstrate four key findings.

First, annotated data in languages other than English, in our case Russian, is essential for language-specific fine-tuning to achieve optimal performance in stance detection because (1) such data offers language-specific topics and arguments, (2) models trained and evaluated in the same language consistently outperform cross-lingual or translated counterparts.

Second, instruction-tuned generative models such as Llama achieve competitive results in stance generation even without explicit Russian pretraining, suggesting that multilingual adaptation through fine-tuning can effectively bridge linguistic gaps.

Third, while generative models achieve perfor-

mance comparable to classifier-based approaches, they require substantially more time and computational resources for fine-tuning and inference.

Finally, topic generation remains a challenging task, with optimal performance observed only on previously seen topics provided to the model at inference.

Future work will focus on extending the LoveHate corpus with additional topics and neutral stances, improving topic inference quality, exploring zero-shot stance and target generation on unseen data, as well as incorporating pre-annotated resources in other languages, including Dutch and French.

7. Ethical statement

The LoveHate corpus was created from previously publicly available, user-generated content originally posted on the debate platform lovehate.ru. The data do not contain personally identifiable information beyond publicly shared usernames, which were deleted in the released version. The dataset includes topics that may involve sensitive or polarizing opinions (e.g., religion, gender equality, or the death penalty); these are retained to preserve the diversity of argumentative content, but researchers are advised to exercise caution when analyzing or disseminating example texts. All models were fine-tuned and evaluated under responsible-use guidelines, and the dataset is released solely for non-commercial research purposes to advance understanding of multilingual argumentation and stance detection. Potentially, the dataset can be leveraged for hate speech, toxicity, and stereotype detection tasks with additional layers of annotations.

8. Limitations

While the LoveHate dataset provides valuable Russian-language resources for stance detection and generation, it has several limitations. First, the data originate from a single debate platform (lovehate.ru) that is no longer active, which may introduce historical and platform-specific biases in language use, topic distribution, and stance expression. The selected 16 topics, though diverse, do not represent the full spectrum of contemporary sociopolitical debates and exclude neutral or ambiguous stances. Second, automatic translation between Russian and English may have introduced noise and semantic drift, potentially affecting cross-lingual comparisons. Third, while we fine-tuned large models such as Robertaa, RuRoberta, Llama, and Qwen, our experiments were limited by computational resources and covered only a subset of hyperparameter configurations. Finally, our models for topic generation were evaluated only on in-domain data, which does not demonstrate their applicability to unseen or out-of-domain topics.

Acknowledgements

This work was supported by the Research Foundation — Flanders (FWO) under grant FWO.OPR.2023.0004.01 (G019823N).

9. Bibliographical References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Vibhor Agarwal, Sagar Joglekar, Anthony P. Young, and Nishanth Sastry. 2022. GraphNLI: A graph-based natural language inference model for polarity prediction in online debates. In *Proceedings of the ACM Web Conference 2022*, page 2729–2737. ACM.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling Frames in Argumentation. In *24th Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP 2019)*, pages 2922–2932. ACL.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018a. A text feature based automatic keyword extraction method for single

documents. In *Advances in Information Retrieval*, pages 684–691, Cham. Springer International Publishing.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018b. YAKE! Collection-independent automatic keyword extractor. In *Advances in Information Retrieval*, pages 806–810, Cham. Springer International Publishing.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Qinlong Fan, Jicang Lu, Yepeng Sun, Pi Qiankun, and Shang Shouxin. 2025. Enhancing zero-shot stance detection via multi-task fine-tuning with debate data and knowledge augmentation. *Complex Intelligent Systems*, 11:151. Received: 22 July 2024; Accepted: 23 December 2024; Published: 15 January 2025.

Irina Fishcheva and Evgeny Kotelnikov. 2019. Cross-lingual argumentation mining for Russian texts. In *Analysis of Images, Social Networks and Texts*, pages 134–144, Cham. Springer International Publishing.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing (ArgNLP 2014)*, pages 26–39, Forlì-Cesena, Italy. INRIA.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Evgeny Kotelnikov, Natalia Loukachevitch, Irina Nikishina, and Alexander Panchenko. 2022. RuArg-2022: Argument mining evaluation. In *Computational Linguistics and Intellectual Technologies*, page 333–348. RSUH.

Yingjie Li and Cornelia Caragea. 2021. A multi-task learning framework for multi-target stance detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2320–2326, Online. Association for Computational Linguistics.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023a. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085, Toronto, Canada. Association for Computational Linguistics.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023b. TTS: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1500–1509, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev, Denis Paperno, Natalia Konstantinova, Natalia Loukachevitch, and Chris Biemann. 2017. Human and machine judgements for Russian semantic relatedness. In *Analysis of Images, Social Networks and Texts*, pages 221–235, Cham. Springer International Publishing.

Andreas Peldszus. 2015. An annotated corpus of argumentative microtexts. In *First European Conference on Argumentation: Argumentation and Reasoned Action*, Portugal, Lisbon.

Ramon Ruiz-Dolz, Joaquin Taverner, John Lawrence, and Chris Reed. 2024. NLAS-multi: A multilingual corpus of automatically generated natural language argumentation schemes. *Data in Brief*, 57:111087.

Maria Skeppstedt, Andreas Peldszus, and Manfred Stede. 2018. More or less controlled elicitation of argumentative text: Enlarging a microtext corpus via crowdsourcing. In *Proceedings of the 5th Workshop on Argument Mining*, pages 155–163, Brussels, Belgium. Association for Computational Linguistics.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*,

pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624. International World Wide Web Conferences Steering Committee.

Maksym Taranukhin, Vered Shwartz, and Evangelos Milios. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15257–15272, Torino, Italia. ELRA and ICCL.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMa: Open and efficient foundation language models. ArXiv:2302.13971 [cs.CL].

Penghui Wei, Wenji Mao, and Daniel Zeng. 2018. A target-guided neural memory model for stance detection in Twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Majid Zarharan, Maryam Hashemi, Malika Behroozrazegh, Sauleh Eetemadi, Mohammad Taher Pilehvar, and Jennifer Foster. 2025. FarExStance: Explainable stance detection for Farsi. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10125–10147, Abu Dhabi, UAE. Association for Computational Linguistics.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714, Bangkok, Thailand. Association for Computational Linguistics.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

A. Appendix

Topic	LoveHate (Pro)	Webis (Pro)	LoveHate (Con)	Webis (Con)
school uniform	year and I do n't, people should go to school, stand out even in school, children that they have children, school is still an educational, school is so to study	dress down usually wear, clothes causes nobody to feel, debates I had where people, uniform is to take focus, children would have more time, school to get an education	form that now is Tupo, uniform is very often uncomfortable, put it on a shirt, clothes in which you feel, walk in this form, time when the school	child is in your school, wear the something every day, clothes that their kids, people tell me that public, people than your clothes, discipline and did not wear
atheism	person must only hope, religions of the state, times of the USSR, IMHO they are just fanatic, understand that you love, Testament say that these books	meaning he cannot be made, faith that some good, reason that we should follow, case is a well thought, causation also gives the atheist, matter if there no objective	n't believe my word, atheists for the most part, answer to the first question, live all their lives, live and do not die, scientist who said with complete	statement that it is impossible, definition to another religious, argues that the first premise, find that the only thing, atheists do and they end, Atheism is indeed a set
god	books are just a kind, existence is under a big, Inquisition or to the Crusades, understand that all religions, order to at least understand, gave me to the death	show your the Lord, arguments I will present, note that she only quotes, explanation is given in terms, argue that we are simply, problem with your six day	created us all in order, times has a person, Religion is a kind, question to the right column, understand that the church, feel sorry for these people	matter how well I present, assertion is when a fact, debate with this resolution, fails because its first premise, show that it is false, false that only the truth
death penalty	Lord where we live, committed at the same time, children who became the victims, people in our country, Maniacs must be killed, person can be very cruel	man so that he dies, prevent them from being killed, killer let out of jail, kill them to make, round for them to read, made that someone was killed	sin for which a person, n't take your life, fact that the person, decide to whom to live, completely and from this life, kill him or just put	found that those who killed, person because they killed, country way behind our time, killed and then later found, victim is just as guilty, human and still make
feminists	Feminism is a female, Olga and I can give, person and her place, money for the same work, equality is still not completely, life and in your career	sports because of her gender, support that it is made, true that more men, people like my opponent, rape and while i agree, idea is what the United	woman into a kind, women who want to achieve, life is in a completely, fact that a girl, Feminists can do a damn, understand what they should fight	countries such as the USA, Simply for being a man, feminists is that people, gender and that gender, point out that the definition, round is n't the time
communists	States would not be afraid, party that would not dream, system is not the ideal, fact is that in Yeltsinism, ideology was our power, time were a little worse	resources available to the people, states that it is wrong, bad since so many people, Lenin was a good, Capitalism is just as bad, capitalism is an economic	cares about each particular person, dad Zu in the Duma, party to the Kremlin, communes now are people, development to have my opinion, history and did not remember	fail because of these reasons, communism can be the correct, opponent provides is that socialism, reason to take any job, work under a Communistic, rates when it was Communist

Table 7: Representative YAKE-extracted keywords comparing LoveHate and Webis datasets across overlapping debate topics.