Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024

Aleksei Vatolin
FRC CSC RAS
Moscow, Russia
vatolinalex@gmail.com

Abstract

This paper presents the winning solution for the RuOpinionNE-2024 competition on structured sentiment analysis in Russian news texts. We propose a novel pipeline with large language models (LLMs) and adapter-based fine-tuning, demonstrating how modern LLMs can be effectively adapted to complex opinion tuple extraction tasks. Our method addresses three key challenges: (1) alignment of model predictions with original text spans through a fuzzy substring matching algorithm, (2) robustness to generation variability via multi-prediction aggregation strategies, and (3) efficient domain adaptation using QLoRA fine-tuning. The proposed approach achieved first place with a test F1 score of 0.405. Experimental results reveal that adapter-based fine-tuning of open-source 70B parameter models (Llama-3.3) surpasses prompt-engineered proprietary models like GPT-4o. Our analysis provides practical insights into adapting LLMs for structured information extraction in morphologically rich languages, showing that targeted fine-tuning with 4-bit quantization enables state-of-the-art performance without task-specific architectures.

Keywords: Structured sentiment analysis, Large language models, QLoRA fine-tuning, Prompt engineering

DOI: 10.28995/2075-7182-2025-23-XX-XX

Структурный анализ настроений с большими языковыми моделями: решение, выигравшее в соревновании RuOpinionNE-2024

Аннотация

Эта статья представляет победное решение для конкурса RuOpinionNE-2024 по структурированному сентимент-анализу текстов российских новостей. Мы предлагаем новый пайплайн с использованием больших языковых моделей (LLM) и адаптерного дообучения, демонстрируя, как современные LLM могут быть эффективно адаптированы для сложных задач извлечения мнений. Наш метод решает три ключевые задачи: (1) согласование предсказаний модели с исходными текстовыми фрагментами с помощью алгоритма нечеткого сопоставления подстрок, (2) устойчивость к вариативности генерации через стратегии агрегации нескольких предсказаний и (3) эффективная адаптация к домену с использованием дообучения QLoRA. Предложенный подход занял первое место с тестовым F1-результатом 0,405. Экспериментальные результаты показывают, что адаптерное дообучение моделей с открытым исходным кодом на 70 миллиардов параметров (Llama-3.3) превосходит результаты GPT-4o с подобранным промптом. Наш анализ предоставля-ет практические инсайты по адаптации LLM для извлечения структурированной информации в морфологически богатых языках, показывая, что целевое дообучение с 4-битной квантизацией позволяет достичь передовых результатов без использования архитектур, специфичных для задачи.

Ключевые слова: Структурный анализ настроений, Большие языковые модели, QLoRA дообучение, Промпт-инжениринг

1 Introduction

Sentiment analysis is a field within Natural Language Processing (NLP) that focuses on determining the emotional tone expressed in text. In recent years, this field has undergone significant advancements, evolving from simple tasks of overall sentiment classification (positive, negative, neutral) to more complex and fine-grained approaches.

One important direction is *Aspect-Based Sentiment Analysis* (ABSA). ABSA aims to identify opinions not at the text level, but concerning specific aspects (features, components) of the entity under consideration. Within ABSA, several subtasks are defined:

- **Aspect Term Extraction** (ATE): Identifying specific words or phrases in the text that indicate aspects.
- **Aspect Sentiment Classification** (ASC): Determining the polarity (positive, negative, neutral) of the expressed opinion towards each extracted aspect.
- **Opinion Term Extraction** (OTE): Identifying words or phrases that express the opinion about the aspect.
- **Aspect-Sentiment-Opinion Triplet Extraction:** A complex task that combines the extraction of aspects, opinion terms, and the determination of their mutual sentiment polarity.
- **Aspect-Category-Opinion-Sentiment Quadruple Extraction**: Extraction of the aspect, its corresponding category, the opinion expressed, and its sentiment polarity.

The development of ABSA has led to the emergence of an even more challenging task – *structured sentiment analysis*. This approach aims to extract complete *opinion tuples*, including not only the aspect and its sentiment but also the *opinion holder* and the *opinion expression*.

This paper presents a solution to the task of structured sentiment analysis of the Russian language within the framework of the RuOpinionNE-2024 competition (Loukachevitch et al., 2025).

2 Literature review

The field of sentiment analysis has seen significant advancements, evolving from general sentiment classification to more fine-grained approaches like Aspect-Based Sentiment Analysis (ABSA) and its subtasks. Early work focused on identifying overall sentiment polarity (Turney, 2002), while subsequent research delved into aspect-level sentiment (Tang et al., 2015). More recently, the focus has shifted towards extracting complete opinion tuples, encompassing aspect terms, opinion expressions, and sentiment polarity. This progression led to the introduction of tasks like Aspect-Based Sentiment Triplet Extraction (ASTE) (Peng et al., 2019) and the even more comprehensive Structured Sentiment Analysis, which aims to capture all sentiment graphs within a text (Barnes et al., 2022).

Current state-of-the-art approaches for these complex tasks often involve sophisticated neural network architectures, including sequence labeling models, graph-based methods, and encoder-decoder frameworks (Fei et al., 2021). Researchers have explored techniques like position-aware tagging (Xu et al., 2020), multi-task learning, and the incorporation of external knowledge to improve performance (Zhao et al., 2020). The introduction of large language models (LLMs) has further pushed the boundaries of sentiment analysis, with studies investigating their zero-shot capabilities and fine-tuning strategies for tasks like ASTE (Scaria et al., 2024) and structured sentiment analysis (Zhang et al., 2021). Instruction learning paradigms, such as those used in InstructABSA, have shown promise in enhancing the reasoning abilities of LLMs and improving performance across various ABSA subtasks.

3 Methodology

3.1 Data Format

The task of structured sentiment analysis involves extracting opinion tuples from text, specifically in the format (Holder, Target, Polar_expression, Polarity). This task is part of the RuOpinionNE-2024 competition (Loukachevitch et al., 2025), which focuses on processing Russian language news articles. Each sentence in the dataset is annotated with potential opinion tuples, where:

- **Holder**: The entity expressing the opinion, which can be a person, organization, or marked as `NULL` for general opinions or `AUTHOR` for the author's opinion.
- **Target**: The entity or object towards which the opinion is directed, always a non-empty named entity.
- **Polar_expression**: The phrase indicating the sentiment, which must be a contiguous substring of the original text.
- **Polarity**: The sentiment label, either `NEG` for negative or `POS` for positive.

Each tuple includes text spans for the holder, target, and polar expression, defined by character offsets. The task requires precise extraction of these components, ensuring that non-contiguous entities are correctly identified and aligned with the original text.

The dataset is divided into three parts: training (2,556 sentences), validation (1,316 sentences), and test (803 sentences). Notably, 41% of the examples in both the training and validation sets contain no opinions. On average, there are 1.22 opinions per example, with a maximum of 23 opinions in the training set and 15 in the validation set.

3.2 Prompt-Based Extraction

Our approach to structured sentiment analysis leverages prompt engineering, following current trends in large language model (LLM) applications. The designed prompt contains a detailed description of the problem, including the task's objectives, constraints, and specific rules for extracting opinion tuples. To enhance the model's understanding, we employ few-shot learning by providing carefully selected examples that demonstrate correct solutions to the task (Brown et al., 2020).

To ensure reliable JSON output and minimize generation errors, we implement structured generation techniques. This approach guides the model's output to follow a specific format, reducing the likelihood of malformed responses (Willard and Louf, 2023). The prompt includes explicit instructions for the model to first reason about the task in a dedicated reasoning field before producing the final output. This two-step process allows the model to analyze the text and consider the extraction rules before committing to a specific answer.

The prompt design incorporates several key elements:

- Clear definitions of each tuple component (Holder, Target, Polar_expression, Polarity)
- Specific rules for handling edge cases (e.g., NULL holders, AUTHOR markers)
- Examples demonstrating cases of opinion extraction
- Instructions for handling sentences without opinions

This structured approach to prompt design (see Appendix A for the full prompt) significantly improves the model's ability to extract opinion tuples accurately while maintaining the required output format. The inclusion of reasoning steps also provides valuable insights into the model's decision-making process, which can be useful for error analysis and system improvement.

We evaluated multiple state-of-the-art models for the prompt-based approach, selecting architectures based on their performance in the MERA Russian-language benchmark (Fenogenova et al., 2024) and model scale diversity. We tested one proprietary model and several open-source models including Llama and Qwen, ranging in size from 7B to 70B parameters (Table 3.2).

Model Name	**Params**	**Hugging Face Hub**	**Citation**
Llama-3.2	1B	meta-llama/Llama-3.2-1B-Instruct	(Grattafiori et al., 2024)
Llama-3.1	8B	meta-llama/Llama-3.1-8B-Instruct	
Llama-3.3	70B	meta-llama/Llama-3.3-70B-Instruct	
GPT-4o	-	-	(OpenAI et al., 2024)
Qwen2.5	7B	Qwen/Qwen2.5-7B-Instruct	(Qwen, 2024)
Qwen2.5	32B	Qwen/Qwen2.5-32B-Instruct	
Qwen2.5	72B	Qwen/Qwen2.5-72B-Instruct	
Vikhr-Nemo	12B	Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24	(Nikolich et al., 2024)
RuadaptQwen2.5	32B	msu-rcc-lair/RuadaptQwen2.5-32B-Instruct	(Tikhomirov and Chernyshev, 2023)
T-pro-it-1.0	32B	t-tech/T-pro-it-1.0	(T-bank, 2024)

Table 1: Models Used in Experiments

3.3 Fuzzy Substring Matching Algorithm

The inherent limitations of language models in precisely reproducing exact text substrings necessitated the development of a specialized fuzzy matching algorithm. When extracting opinion components through prompt-based approaches, models frequently introduce subtle alterations such as case variations, omitted punctuation, or minor lexical substitutions. These discrepancies render exact string matching ineffective for aligning model predictions with original text spans.

Our algorithm employs a two-stage approximate matching strategy that balances computational efficiency with matching accuracy. The first phase utilizes regular expressions to identify potential start and end points of text spans, treating sequences of letters and digits as words and punctuation marks as natural separators. This preprocessing step ensures robust handling of common model errors such as misplaced or omitted punctuation by defining candidate spans based on these word/punctuation boundaries.

The core matching mechanism iterates through all possible text spans derived from these boundaries. For each candidate span extracted from the source text, the algorithm computes a similarity score against the model-generated substring using the `fuzz.ratio` function from `thefuzz` library[1], which calculates a normalized score based on Levenshtein distance. It systematically compares the target substring against all potential spans within the source text, identifying the span with the highest similarity score.

Finally, this highest-scoring alignment is selected only if its similarity score meets a predefined minimum confidence threshold, empirically set to 70 based on validation experiments. This threshold filters out spurious matches where the model might have generated text fragments significantly divergent from or absent in the original source. If a suitable match is found, the algorithm returns the similarity score, the character offsets (start:end) of the best-matching span in the original text, and the matched text itself. Otherwise, it indicates that no satisfactory match was found.

3.4 Aggregation Strategies

To address the inherent stochasticity in large language model outputs and improve prediction robustness, we developed three aggregation strategies that process multiple model predictions. Based on validation set performance, we determined that using 5 predictions per input was optimal. This approach mitigates random fluctuations in individual predictions and substantially improves the quality of the prediction.

Our implementation provides three distinct aggregation approaches:

Most Common strategy selects the single most frequent prediction across all generations. This conservative approach works best when models consistently produce high-quality outputs, prioritizing precision over recall.

Most Common N strategy addresses the limitations of basic frequency selection by dynamically adapting to the number of opinions in the first prediction. Rather than using a fixed N, this method selects the top N most frequent predictions, where N is equal to the number of opinions detected in the initial model response. This approach solves the key problem of the basic "Most Common" method, which often collapses multiple legitimate opinions into a single highest-frequency candidate due to its strict maximal frequency requirement. By preserving multiple high-probability candidates, each of which received substantial model support, this strategy better handles cases with multiple valid opinions while maintaining frequency-based filtering of spurious predictions.

[1] `https://github.com/seatgeek/thefuzz`

Combine strategy acts as a union operator, aggregating all unique predictions while eliminating duplicates. This approach maximizes recall at the cost of potentially lower precision, which is particularly useful when dealing with texts containing multiple valid opinions.

3.5 Model adapter finetuning

While prompt-based approaches demonstrated initial feasibility, their limitations in handling the task's specific linguistic patterns and structural constraints motivated our transition to adapter-based fine-tuning. The QLoRA approach (Dettmers et al., 2023) enables efficient parameter updates through 4-bit quantized training, allowing adaptation of large language models to the structured sentiment analysis task without prohibitive computational costs.

All models underwent identical training conditions: 10 epochs with a batch size of 8 distributed across 8×NVIDIA A100 80GB GPUs. The final 'Llama-3.3-70B' model, for example, completed training in approximately 1.5 hours using this setup. While this configuration accelerated the process, such extensive computational resources are not strictly mandatory; a minimum of 48GB of VRAM is sufficient to fine-tune 70B parameter models using QLoRA, though requiring longer training times. The training format maintained similarity to the prompt-based approach in output structure, but with crucial modifications: models now received only the raw sentence text as input and were required to directly output a JSON-formatted list of opinion tuples. This format explicitly removed the dedicated reasoning field used in initial prompt-based approaches, forcing models to learn the extraction patterns directly rather than through explicit reasoning steps.

Model	**Size**	**Single**	**Most Common**	**Most Common N**	**Combine**
Llama-3.1	8B	0.123	0.133	0.133	0.167
RuadaptQwen2.5	32B	0.170	0.171	0.172	0.173
Vikhr-Nemo	12B	0.183	0.184	0.189	0.174
Qwen2.5	32B	0.192	0.197	0.195	0.186
Qwen2.5	72B	0.200	0.208	0.209	0.191
T-pro-it-1.0	32B	0.199	0.199	0.203	0.201
gpt-4o	-	0.199	0.204	0.203	0.212
Llama-3.3	70B	0.234	0.241	0.240	**0.252**

Table 2: Model performance with prompt-based approach

4 Results

Our experimental results demonstrate significant performance improvements through model fine-tuning and multiple prediction aggregation. The prompt-based approach (Table 3.5) achieved the best performance with Llama-3.3-70B using Combine aggregation (F1=0.252), while QLoRA fine-tuning (Tables 3-4) substantially boosted results, with the same model reaching F1=0.379 on validation and 0.405 on test data using the Combine strategy. Inference using this final model configuration ('Llama-3.3-70B' with 'Combine' aggregation) on the entire validation or test dataset took approximately 5 minutes on the 8×A100 GPU setup.

Three key findings emerge from the data:

Model	**Size**	**Single**	**Most Common**	**Most Common N**	**Combine**
Llama-3.2	1B	0.245	0.247	0.243	0.253
Qwen2.5	7B	0.304	0.305	0.309	0.311
Qwen2.5	32B	0.331	0.338	0.336	0.326
T-pro-it	32B	0.322	0.331	0.328	0.326
RuadaptQwen2.5	32B	0.323	0.324	0.327	0.328
Llama-3.1	8B	0.335	0.333	0.334	0.335
Vikhr-Nemo	12B	0.345	0.346	0.350	0.352
Qwen2.5	72B	0.364	0.359	0.362	0.356
Llama-3.3	70B	0.372	0.368	0.372	**0.379**

Table 3: Model performance with QLoRA adapter on validation dataset

Model	**Size**	**Single**	**Most Common**	**Most Common N**	**Combine**
Llama-3.2	1B	0.263	0.263	0.265	0.278
Qwen2.5	7B	0.315	0.320	0.329	0.319
Qwen2.5	32B	0.329	0.342	0.340	0.349
Llama-3.1	8B	0.335	0.337	0.342	0.349
T-pro-it	32B	0.350	0.360	0.348	0.357
Vikhr-Nemo	12B	0.385	0.405	0.380	0.381
Qwen2.5	72B	0.375	0.376	0.382	0.383
RuadaptQwen2.5	32B	0.375	0.373	0.378	0.384
Llama-3.3	70B	0.396	0.394	0.390	**0.405**

Table 4: Model performance with QLoRA adapter on test dataset

Team Number	Validation F1	Test F1
Team 1 (ours)	**0.38**	**0.41**
Team 2	0.17	0.35
Team 3	0.28	0.33
Team 4	0.21	0.28
Team 5	0.34	0.24
Team 6	-	0.24
Team 7	-	0.20
Team 8	-	0.11

Table 5: Public leaderboard from the RuOpinionNE-2024 competition showing team rankings based on Validation and Test F1 scores.

- **Aggregation Impact**: The Combine strategy consistently outperformed other methods in final test evaluations (Table 4), particularly for larger models. However, the Most Common N strategy showed better validation performance (Table 3), suggesting different generalization patterns between datasets.
- **Scale vs Specialization**: While model scale generally correlated with performance, specialized adaptations proved crucial - the 32B RuadaptQwen2.5 outperformed the base Qwen2.5-32B by 3.8% in prompt-based settings despite identical parameter counts.
- **Training Effectiveness**: QLoRA fine-tuning produced dramatic improvements, with Vikhr-Nemo 12B showing the largest gain of 88% increase from prompt-based (0.183) to fine-tuned (0.345) validation performance.

The public leaderboard results (Table 3.5) confirm our approach's effectiveness, with our solution achieving 0.41 test F1 versus 0.35 for the nearest competitor. Notably, our validation score (0.38) strongly predicted final test performance, demonstrating the robustness of our evaluation methodology.

The results reveal several patterns:

- Smaller fine-tuned models (Qwen2.5-7B) outperformed much larger prompt-based models (Qwen2.5-72B), highlighting the value of adapter-based fine-tuning.
- Proprietary GPT-4o underperformed open-source models after fine-tuning, indicating the importance of task-specific adaptation over general capability.

4.1 Impact of Fuzzy Substring Matching

To quantify the contribution of the fuzzy substring matching algorithm described in Section 3.3, we conducted an ablation study comparing its performance against a baseline exact substring search. The exact matching approach performs a simple, case-insensitive search for the model-generated substring within the original sentence text. If an exact match is found, its character offsets are used; otherwise, the predicted component is discarded. We re-evaluated the best-performing aggregation strategy (Combine) for both prompt-based and fine-tuned models using this exact matching method. The 'Impact of Exact Match' column in the following tables shows the change in F1 score when the fuzzy matching algorithm is replaced with the exact matching baseline.

The results (Tables 6 and 7) quantify the substantial benefit derived from the fuzzy matching algorithm, particularly for the prompt-based approach. The ’Impact of Fuzzy Matching’ column in Table 6 shows F1 score improvements ranging from +0.020 to +0.042 when fuzzy matching is employed instead of exact matching. This highlights that prompt-based models frequently generate text spans with minor deviations (e.g., punctuation, changes in word form or case) from the source text, which only the fuzzy matching algorithm can successfully align.

For the QLoRA fine-tuned models (Table 7), the F1 score gains from using fuzzy matching over exact matching were less pronounced but still consistently positive. The improvements ranged mostly between +0.001 and +0.007 F1. This suggests that fine-tuning trains models to generate more precise text spans that adhere closer to the original text, thus reducing the reliance on fuzzy matching. However, even our best-performing fine-tuned model, Llama-3.3-70B, benefited from a +0.007 F1 increase, confirming that fuzzy matching remains a valuable final alignment step to capture residual generation variations, even for highly adapted models.

Model	**Size**	**Exact Match F1**	**Impact of Fuzzy Matching ($\vec{\Delta}$F1)**
Llama-3.1	8B	0.132	+0.035
Vikhr-Nemo	12B	0.152	+0.022
RuadaptQwen2.5	32B	0.152	+0.021
Qwen2.5	72B	0.158	+0.028
T-pro-it	32B	0.159	+0.042
Qwen2.5	32B	0.166	+0.020
gpt-4o	-	0.187	+0.025
Llama-3.3	70B	0.187	+0.031

Table 6: Performance comparison using exact matching for the prompt-based approach (validation set).

Model	**Size**	**Exact Match F1**	**Impact of Fuzzy Matching ($\vec{\Delta}$F1)**
Llama-3.2	1B	0.248	+0.004
Qwen2.5	7B	0.316	+0.005
RuadaptQwen2.5	32B	0.322	+0.006
Qwen2.5	32B	0.325	+0.001
T-pro-it	32B	0.327	0.000
Llama-3.1	8B	0.333	+0.002
Vikhr-Nemo	12B	0.350	+0.002
Qwen2.5	72B	0.355	+0.001
Llama-3.3	70B	0.372	+0.007

Table 7: Performance comparison using exact matching for the QLoRA fine-tuned approach (validation set).

5 Challenges in Cross-Dataset Adaptation

An experiment involved training a QLoRA adapter on combined data from SemEval 2022 (Barnes et al., 2022) and RuOpinionNE-2024. Despite the structural similarity in the annotation format, this approach yielded substantially worse performance (F1=0.27-0.31) compared to RuOpinionNE-only training (F1=0.38), revealing critical challenges in cross-dataset adaptation.

The failed methodology consisted of:

- Merging 7 SemEval datasets (Basque, Catalan, English, Norwegian, Spanish) with Russian data
- Removing neutral sentiment examples (2.1% of total) to match RuOpinionNE's binary labels
- Maintaining identical training parameters (10 epochs, batch size 8) as successful single-dataset runs

Three primary failure factors emerged from our analysis:

- **Label Space Mismatch**: While neutral examples were removed, models struggled with residual three-way classification patterns in the multilingual data, particularly for near-neutral expressions.
- **Structural Divergence**: SemEval's allowance for multi-segment Polar_expressions (6.6% vs 0.6% in RuOpinionNE) introduced conflicting annotation patterns that degraded model performance on Russian texts.
- **Annotation Guideline Discrepancy**: Despite apparent structural similarity in the output format, the underlying annotation guidelines and the interpretation of what constitutes an "opinion" might differ between SemEval and RuOpinionNE. This semantic divergence in annotation intent could lead to the model learning conflicting patterns.
- **Domain Discrepancy**: The news-focused Russian dataset clashed with SemEval's mixed domains (product reviews, social media), preventing effective knowledge transfer.

The results underscore the importance of dataset homogeneity, with the combined training approach performing 27% worse than monolingual adaptation. Leaderboard simulations suggest this method would have ranked 6th (F1=0.24) versus our actual 1st place finish (F1=0.41), highlighting the critical impact of focused dataset selection.

6 Generalizability and Cross-Domain Application

Given that the model was fine-tuned on a dataset derived from Russian news texts, it is naturally suited for extracting opinions from similar news sources beyond the specific competition dataset. This suggests its potential utility in broader media monitoring and analysis tasks within the news domain.

To explore the model's applicability in a significantly different domain, we applied it to scientific texts. We utilized a publicly available dataset RuSciBench (Vatolin et al., 2024) comprising approximately 182 thousand abstracts of scientific articles from various disciplines, sourced from the Russian scientific library elibrary.ru. We processed these abstracts with our best-performing fine-tuned model ('Llama-3.3-70B' with 'Combine' aggregation) to identify expressed opinions.

Table 8 shows the percentage of abstracts within different GRNTI (State Rubricator of Scientific and Technical Information) categories where at least one opinion was detected.

GRNTI Category	**% Abstracts with Opinions**
A comprehensive study of individual countries and regions	58.06
Religion. Atheism	57.97
Literature. Literary criticism. Oral folk art	57.57
Art. Art history	56.01
Politics and political sciences	50.96
History. Historical sciences	48.70
Mass communication. Journalism. Media	42.49
Complex problems of social sciences	41.94
Patent business. Invention. Rationalization	41.67
Biotechnology	40.00
Philosophy	39.93
Social sciences in general	39.50
Demography	39.32
Culture. Culturology	38.52
Scientific	38.10
Rural and forestry	37.78
Nuclear equipment	37.68
Food industry	33.73
Medicine and healthcare	33.36
Fisheries. Aquaculture	32.49

Table 8: Percentage of scientific abstracts (elibrary.ru) containing at least one detected opinion, by GRNTI category (Top 20).

The results indicate a higher prevalence of detectable opinions in humanities and social sciences disciplines (e.g., regional studies, religion, literature, politics, history) compared to natural sciences and technical fields, although opinions are found across the board. This aligns with the expectation that humanities research often involves interpretation, evaluation, and argumentation, which are likely to contain sentiment-bearing expressions. While a model specifically trained on scientific discourse might achieve higher accuracy and capture domain-specific nuances more effectively, these results demonstrate that the news-trained model can still provide interesting insights even in a significantly different domain.

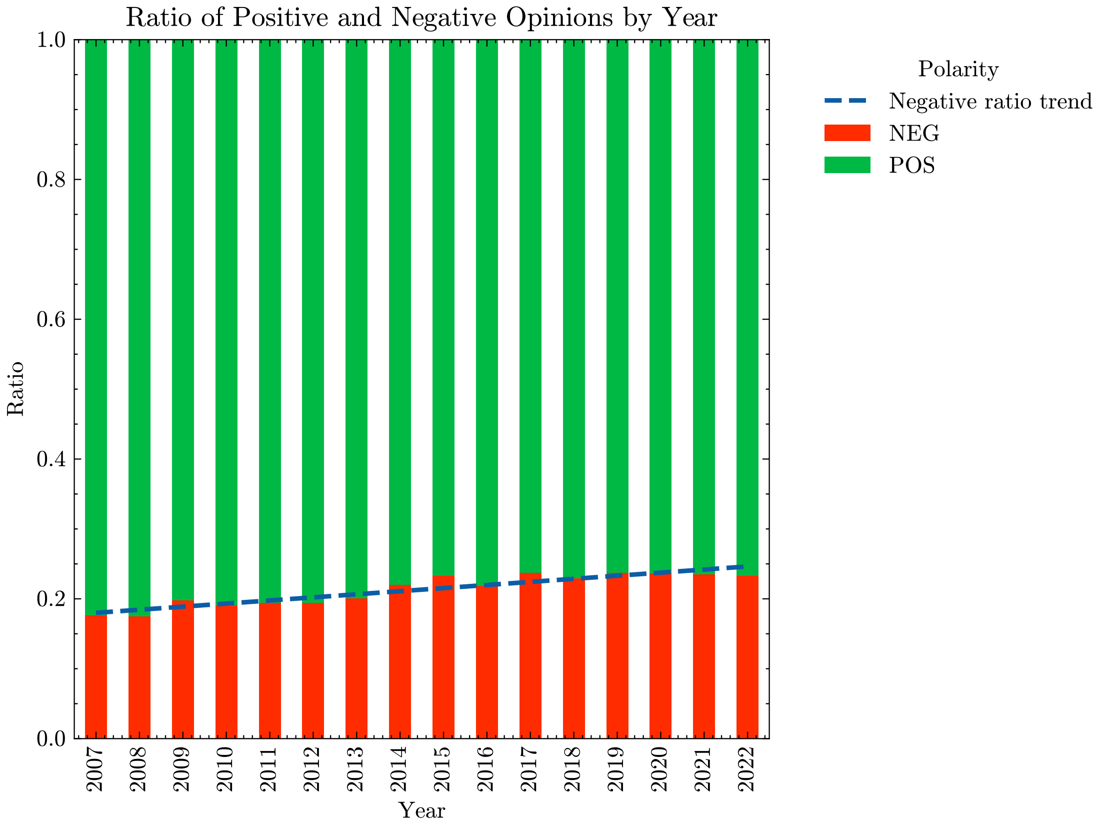

Figure 1: Negative sentiment ratio by year.

Further analysis of the extracted opinions revealed a slightly increasing trend in the proportion of negative opinions expressed in abstracts over the years covered by the dataset (see Figure 1). Additionally, examining the sentiment associated with specific targets (Figure 2) shows distinct patterns.

Beyond these expected extremes (like digital technology or terrorism), the analysis reveals more nuanced sentiment distributions for other entities within the context of scientific abstracts. For instance, terms central to the academic environment like "university" (83.9% positive) and "student" (73.2% positive) show predominantly positive sentiment, possibly related to research achievements or educational successes. However, the significant minority of negative mentions

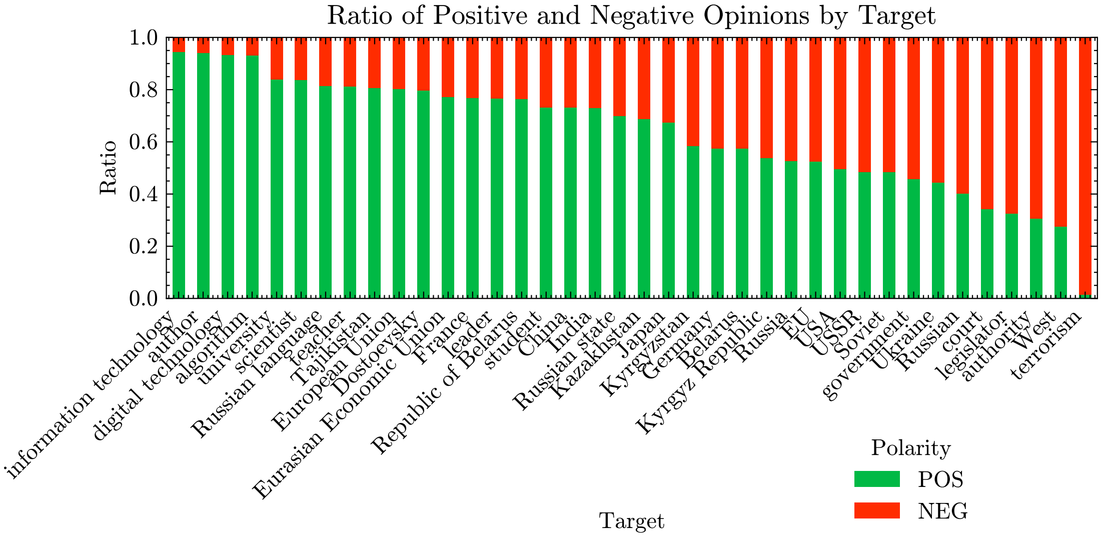

Figure 2: Most common opinion targets.

might stem from abstracts discussing challenges, critiques of educational systems, or studies focusing on student problems. Similarly, "scientist" (83.7% positive) follows this pattern.

Nations and political bodies often receive more mixed evaluations. For example, "Russia" (52.7% positive), "USA" (49.7% positive), and historical entities like "USSR" (48.5% positive) show near-even splits. This balance could reflect the objective and critical nature of academic discourse in fields like international relations, political science, or history, where abstracts might analyze policies, conflicts, or societal issues from multiple perspectives, leading to both positive and negative framing depending on the research focus. The comparatively lower positive sentiment towards "government" (45.8% positive) and higher negativity towards "West" (27.5% positive) might also indicate a tendency towards critical analysis of power structures and policies within certain academic disciplines prevalent in the dataset.

A closer look at the specific expressions associated with "Russia" provides further insight into this mixed picture.

Focusing specifically on "Russia" as an opinion target (Figure 3), the analysis of scientific abstracts up to the end of 2022 reveals a prevalence of terms associated with significant challenges. To generate this visualization, the extracted opinion expressions (polar expressions) associated with the target "Russia" were first lemmatized using pymorphy2 (Korobov, 2015) to group different word forms and then translated using Google Translate for representation in English. Negative expressions like "sanction", "corruption", and "problem" are most frequent, often linked to economic issues ("economic sanction", "lag behind") and security concerns ("threat", "information war", "threat of national security"). This reflects the intense focus within academic discourse during this period on Russia's geopolitical and socio-economic difficulties. However, positive terms also appear, highlighting perceived strengths and potential avenues for devel-

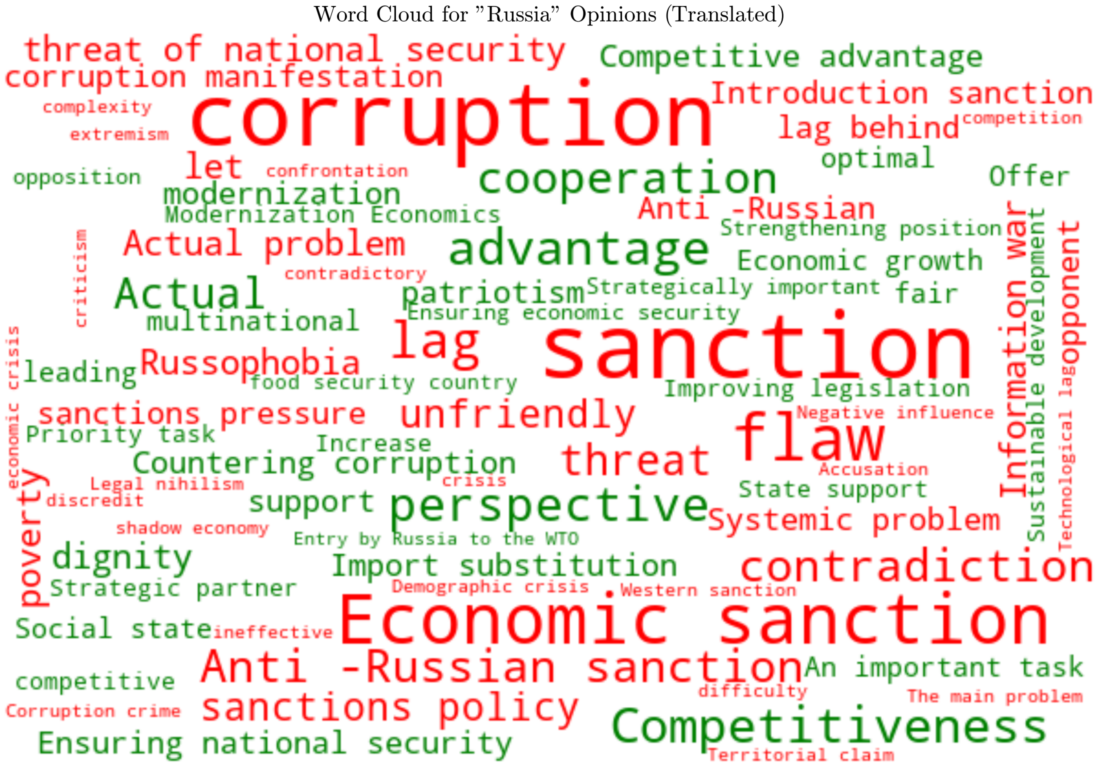

Figure 3: Word Cloud for Opinions Targeting "Russia" in Scientific Abstracts. Red color indicates negative sentiment, Green indicates positive.

opment. Words like "competitiveness", "advantage", "perspective", "cooperation", and "modernization" suggest discussions around national capabilities, strategic partnerships, and efforts towards improvement, such as "countering corruption" and "ensuring national security".

These exploratory applications suggest broader potential uses for the developed model. For instance, it could be integrated into specialized search engines to find articles mentioning specific entities (e.g., countries, organizations, individuals) in a positive or negative light. Furthermore, it could enable large-scale analytics, such as tracking sentiment trends towards certain topics or examining correlations in sentiment (e.g., identifying which countries are frequently mentioned negatively in articles where Russia, for example, is mentioned positively). These possibilities extend beyond the scope of this paper and represent promising avenues for future research.

7 Conclusion

Our work demonstrates that large language models can be effectively adapted for structured sentiment analysis in morphologically rich languages through targeted fine-tuning strategies. The combination of QLoRA-efficient adaptation and robust post-processing techniques enabled state-of-the-art performance on the RuOpinionNE-2024 task, achieving a 0.405 F1 score without task-specific architectural modifications.

The results establish several important principles for applying LLMs to structured information extraction tasks. First, adapter-based fine-tuning with 4-bit quantization enables efficient specialization of billion-parameter models to complex annotation schemes. Second, open-source models can surpass proprietary alternatives when properly adapted to domain-specific requirements.

This research provides practical guidelines for adapting LLMs to opinion mining tasks in resource-rich language scenarios. Future work should investigate the generalizability of these techniques to other structured prediction tasks and low-resource language settings. The success of our approach suggests that continued development of efficient fine-tuning methods will further unlock the potential of large language models for complex information extraction challenges.

References

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: Structured sentiment analysis. // *International Workshop on Semantic Evaluation*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. 2021. Nonautoregressive encoder–decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34:5544–5556.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. // Lun-Wei Ku, Andre Martins, and Vivek Srikumar, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 9920–9948, Bangkok, Thailand, August. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models, July.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. // Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, P 320–332. Springer International Publishing.

Natalia Loukachevitch, Natalia Tkachenko, Anna Lapanitsyna, Mikhail Tikhomirov, and Nicolay Rusnachenko. 2025. Ruopinionne-2024: Extraction of opinion tuples from russian news texts. // *Proceedings of International Conference on Computational Linguistics and Intelligent Technologies Dialogue 2025*.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. // *Proceedings of the 4rd Workshop on Multilingual Representation Learning (MRL) @ EMNLP-2024*. Association for Computational Linguistics.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, Aj Ostrow, Akila Welihinda, Alan Hayes, and et al. 2024. Gpt-4o system card, October.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2019. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. // *AAAI Conference on Artificial Intelligence*.

Team Qwen. 2024. Qwen2.5: A party of foundation models, September.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. // Kevin Duh, Helena Gomez, and Steven Bethard, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, P 720–736, Mexico City, Mexico, June. Association for Computational Linguistics.

T-bank. 2024. T-pro-it-1.0.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. // *International Conference on Computational Linguistics*.

Mikhail Tikhomirov and Daniil Chernyshev. 2023. Impact of tokenization on llama russian adaptation. // *2023 Ivannikov Ispras Open Conference (ISPRAS)*, P 163–168.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. // *Annual Meeting of the Association for Computational Linguistics*.

Aleksei Vatolin, Nikolai Gerasimenko, Anastasia Ianina, and Konstantin Vorotsov. 2024. Ruscibench: Open benchmark for russian and english scientific document representations. *Doklady Mathematics, 2024, Vol. 110, No. 4, pp. 249–258*.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *ArXiv*, abs/2307.09702.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. // *Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. // Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, P 504–510, Online, August. Association for Computational Linguistics.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. // Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 3239–3248, Online, July. Association for Computational Linguistics.

A Prompt for unsupervised approach

Инструкция по разметке данных для задачи определения мнений

Общие правила

Разметка выполняется на уровне предложений. Каждое предложение может содержать одну или несколько пар "мнение-отношение"(или не содержать их вовсе). Для каждой пары создается объект, включающий следующие ключи: source, target, polar_expression, polarity.

Ключи и их значения

1. **source (источник)**
   - **Определение**: Источник мнения — лицо или организация, выражающие мнение в предложении.
   - **Формат**: Строка или NULL, если источник не указан явно.
   - **Требования**:
     - Указывайте только конкретные именованные сущности (например, "Муртаза Рахимов "Единая Россия").
     - Если в предложении источник не указан явно или используется местоимение, но его можно вывести из контекста, источник отмечается как NULL.
     - Автор текста (AUTHOR) может быть источником, если мнение выражено в форме редакционного заявления, например, при оценке событий.
2. **target (цель)**
   - **Определение**: Цель мнения — лицо, организация или объект, на который направлено мнение.
   - **Формат**: Строка
   - **Требования**:
     - Указывайте только конкретные именованные сущности (например, "главу администрации "Монти").
     - Цель НЕ может быть отмечена как NULL! Если вы не можете найти цель, значит это мнение не подходит и его не нужно писать.
3. **polar_expression (эмоциональное выражение)**
   - **Определение**: Слово или фраза, выражающие отношение источника к цели.
   - **Формат**: Строка (например, "раскритиковал "восстановившая").
   - **Требования**:
     - Указывайте фразу в точности так, как она представлена в тексте.
     - Если эмоциональное выражение отсутствует, пара не создается.
     - Учитывайте фразы, выражающие скрытую оценку (например, "значительно ослабла").
     - Не может быть NULL!

4. **polarity (полярность)**
   - **Определение**: Тип эмоциональной окраски выражения.
   - **Допустимые значения**:
     - NEG — негативное мнение (например, "раскритиковал "уволил").
     - POS — позитивное мнение (например, "восстановившая "приглашен").
   - **Требования**:
     - Полярность определяется исходя из контекста предложения.
     - Если эмоциональное выражение нейтральное, пара не создается.

ВАЖНО! Требования ко полям source, target, polar_expression: Если значение не равно NULL или AUTHOR, то это должна быть подстрока исходного текста. Будь вниматален, подстрока должна состоять из подряд идущих слов и в точности совпадать с исходным текстом.

Примеры и уточнения

1. **Пример предложения с источником и целью:**
   Входной текст: Муртаза Рахимов подписал указ об отставке главы своей администрации
   Твой ответ:

```
{
    "opinions": [
        {
            "source": "Муртаза Рахимов",
            "target": "главы своей администрации",
            "polar_expression": "отставке",
            "polarity": "NEG"
        }
    ]
}
```

2. **Пример с** AUTHOR **как источником:**
   Входной текст: Правительство технократов проводило непопулярную экономическую политику.
   Твой ответ:

```
{
    "opinions": [
        {
            "source": "AUTHOR",
            "target": "правительство технократов",
            "polar_expression": "непопулярную экономическую политику",
            "polarity": "NEG"
        }
    ]
```

}

3. **Пример нейтрального предложения:**
   Входной текст: Указ о назначении нового главы своей администрации пока не подписан.
   Твой ответ:
   {"opinions": []}

Запрещенные значения

- source **и** target:
  - Не используйте местоимения (например, "он "она").
  - Не используйте обобщенные термины, если конкретная сущность не упомянута.
- polar_expression:
  - Не добавляйте субъективные интерпретации, отсутствующие в тексте.
- polarity:
  - Не отмечайте нейтральные фразы как NEG или POS.

Дополнительно

- Если вы сомневаетесь в полярности или источнике, добавляйте комментарии к разметке для последующего обсуждения.
- Проверяйте каждую размеченную пару на соответствие контексту предложения.
- Если в тексте нет мнений, на выходе должен быть пустой массив "{"opinions": []}".

Теперь твоя задача порассуждать (рассуждения должны быть не больше 1000 слов) и разметить следующий текст: