DOI: 10.14529/jsfi250309

Aspect-Based Sentiment Analysis Using Large Language Models on Museum Visitor Reviews

Anastasia V. Kolmogorova[1], *Elizaveta R. Kulikova*[1], *Vladislav V. Lobanov*[1]

Museum reviews provide rich insight into visitor preferences and can drive useful change within institutions, yet they have attracted little attention in sentiment research owing to limited commercial interest and the multi-thematic nature of reviews. In this study we analysed over 12 000 reviews in Russian for 15 museum sites collected from nine different platforms. Methodologically, we first evaluated traditional approaches: a lexicon-based method utilising sentiment dictionaries and a neural network approach leveraging open-source pre-trained models such as RuBERT. While such methods can be applied to document-level sentiment analysis, where the text is labelled simply as positive or negative, they cannot uncover the specific topics that give rise to these sentiments. Finally, we implemented large language models (LLMs) for aspect-based sentiment analysis to discover positive and negative aspects visitors mention. Our system uses a two-step pipeline that initially extracts positive and negative keywords about each museum site and subsequently categorises these keywords into 14 predetermined categories, enabling the reader to effortlessly discover strong points and areas for improvement. Results include 15 csv tables of positive and negative keywords and 15 year-wise text reports for all objects. While some LLM hallucinations were observed, the outputs were largely realistic. We conclude that LLMs are well suited to this task and offer substantial scope for future research and practical applications in museum evaluation and service improvement.

Keywords: museum reviews, aspect-based sentiment analysis, LLM, thematic categorization, prompting.

Introduction

The advent of large language models (LLMs) has significantly transformed the conventional landscape of tasks and methods in natural language processing (NLP). Efficient pipelines are rapidly being established where LLMs are utilized for translation [5], information extraction (including summarization, text simplification, named entity recognition, and keyword extraction), the development of dialogue systems, as well as emotion and sentiment analysis. As noted by the authors of a recent survey [23], two primary paradigms are emerging in the use of LLMs for NLP: (1) a parameter-frozen paradigm, encompassing zero-shot learning and few-shot learning, and (2) a parameter-tuning paradigm, which includes both full-parameter tuning and parameter-efficient tuning. In our research, we address one of the classical tasks of NLP sentiment analysis. Having emerged among NLP paradigms in the early 2000s [20, 21] sentiment analysis has firmly established its place in both academic research and product development. The lexicon-based method, relying on sentiment lexicons, first appeared and gained widespread adoption [3, 6], followed later by neural network models [29]. Debates over the effectiveness of each of these approaches have been ongoing within the professional community for a considerable time. However, their relevance has significantly diminished following the rapid advancement of LLM linguistics. LLMs demonstrate performance that is quite comparable to that of neural network models, both without prompting and when utilizing various prompting strategies [27, 30]. This study implements aspect-based sentiment analysis (ABSA) of visitor reviews for a national museum and heritage site using LLM. ABSA is designed to identify positive or negative user attitudes toward specific

[1] HSE University in Saint Petersburg, 3 Kantemirovskaya Street, Saint-Petersburg, Russian Federation, 194100

features (aspects) of a product or service [2]. The present research task was formulated in response to a concrete technical specification: a client, one of Russia's largest museum-reserves, commissioned collecting visitors' reviews on the heritage site's locations and subsequent analysis of visitor preferences and criticism. We employed a parameter-frozen paradigm for the LLM application, utilizing solely multi-stage prompting that incorporated elements of various strategic approaches. Consequently, the objective of this publication is to describe a pipeline for employing LLMs for ABSA of a corpus of museum reviews which has demonstrated practical efficacy. Having formulated the research problem (Section 1), we will subsequently analyze related papers (Section 2), describe the dataset (Section 3), followed by the methodology (Section 4) and results of applying LLM to its analysis (Section 5). In the discussion (Section 6), the main advantages and limitations of the applied approach to ABSA using LLM are examined, and the Conclusion briefly summarizes the key findings of the study.

1. Research Problem

Sentiment analysis has traditionally been applied to reviews of products (e.g., books, household appliances, restaurants) and services (e.g., repair, cosmetic services, delivery). However, our exploratory analysis has revealed that services provided by state cultural institutions have attracted scant attention from both researchers and practitioners in the field of sentiment analysis. This lack of interest can be attributed to two primary factors: the absence of commercial demand for such analytics and the inherently multi-faceted nature of the topics covered in these reviews. Nonetheless, the ongoing process of digitalization is gradually encompassing museum institutions, as evidenced by the commission we received. Regarding the textual content of museum reviews, they indeed concurrently address a wide array of domains: personal reminiscences, national history, cross-cultural remarks, the condition of buildings and exhibits, technical details, mundane aspects of the visit, and educational value, among others: (1) Modern renovation, pleasant, soft, and quiet flooring. For activities with children, there is a separate, bright room. There, kids get to know and interact with nature; (2) No transport goes all the way to the Kremlin itself. You will have to walk for about 10–15 minutes from Central Square or take a taxi; (3) It was amusing to watch the Chinese tourists. While the Russian visitors were examining the exhibits and reading the annotations, the Chinese tourists simply hurried past, sometimes without even looking around. Only one boy was frantically trying to take pictures on the run. And yet, there was so much to see. In this context, our task exhibited the characteristics of open-domain sentiment analysis, where sentiment detection is performed on unspecified subject domains. Consequently, the research problem was formulated as follows: to validate the efficacy of LLM as a tool for ABSA under conditions characterized by the absence of a predefined set of target aspects and the inherent multi-thematic nature of the data.

2. Related Papers

There is very little research dedicated to sentiment analysis of museum reviews. In all the studies we found, the authors were unable to perform aspect-based sentiment analysis, so they used a two-stage procedure: first, the reviews are evaluated for sentiment, and then the analysis is supplemented with topic modeling. For example, in [28] a neural network approach was used for a collection of 200 000 reviews of the 8 largest museums in the world, calculating the sentiment weight of each review, followed by topic modeling to determine the correlation between sentiment and topic. In [4], for sentiment analysis of reviews about Tongzhou Grand Canal Forest Park in Beijing, a hybrid approach was also used: first, the reviews were evaluated based on a sentiment dictionary, and then topic modeling was performed using LDA on the group of negative reviews and the group of positive reviews. Regarding the application of AI tools for sentiment analysis, the year 2024 marked the beginning of active testing of LLMs for Russian-language texts. This is exemplified by the RuOpinionNE-2024 evaluation competition, where participants were challenged to extract all tuples (H, T, P, E) from Russian news texts, segmented by sentence [19]. In this task, H represents an opinion holder expressing a polarity P towards a target T through a sentiment expression E. Holders and Targets are entities of the following types: PERSON, ORGANIZATION, COUNTRY, CITY, REGION, PROFESSION, NATIONALITY, and IDEOLOGY. The highest effectiveness in this competition was demonstrated by the pipeline proposed in [26], which utilized an LLM with QLoRA for adapter-based fine-tuning. This approach achieved first place with a test F1-score of 0.405. Nevertheless, a review of the relevant literature from 2024–2025 has not revealed any studies in which LLMs were applied for aspect-based sentiment analysis of the Russian-language reviews. However, the English-language segment of the research field features successful studies of user reviews utilizing Large Language Models (LLMs). For instance, [31] demonstrates that on a dataset of hotel reviews across six aspects (Staff, Price, Place, Ambience, Experiences, Services), an average of 95.1% of the ratings were in complete agreement between human assessors and GPT-4. However, in the cited study, the model was instructed to first extract statements related to a specific aspect, then evaluate them on a sentiment scale, and finally, provide a summary of what is generally written about that aspect. In contrast, our research does not employ scaled sentiment ratings. Instead, we instruct the model to first extract negative and positive keywords and subsequently categorize them according to predefined aspects. It can be argued that the proposed pipeline is particularly effective for analyzing reviews of diverse museum and heritage site facilities. As will be demonstrated subsequently, the structure and content of such reviews are highly varied, which precludes the a priori definition of a fixed set of aspects. In other words, given both the scarcity of research on sentiment analysis for cultural institution reviews and the concurrent lack of studies on the efficacy of applying Large Language Models (LLMs) to aspect-based sentiment analysis of the Russian-language reviews, the pipeline we propose constitutes a meaningful contribution to this field of study.

3. Data

Reviews of 15 locations of the museum-reserve were collected from 9 online platforms (see Tab. 1). The data was extracted for the period from 2014 to 2025 (May). However, since the museum was interested in the period from 2020 to 2025, further experiments were conducted only with this sample. The total sample contained 12 100 reviews.

As can be observed in Tab. 1, the sample comprises reviews of museum institutions that differ significantly from one another both in terms of their exhibition content and target visitor demographics. For instance, the Holy Dormition Cathedral is a functioning Orthodox cathedral, while the Spaso-Evfimiev Monastery, in addition to holding regular services, houses extensive museum exhibitions related to the era of Stalinist repressions. Concurrently, the Museum of Nature offers natural science exhibitions and interactive platforms for the popularization of science, and the Palaty (Chambers) serves as an exhibition centre for fine arts.

Table 1. Dataset distribution by museum site, platform, and year

Museum Site		Platform		Year	
Suzdal Kremlin	2527	Yandex Maps	6305	2020	2400
Spaso-Evfimiev Monastery	2211	Google Maps	5090	2021	1643
Museum of Wooden Architecture	1888	Tripadvisor	523	2022	2373
Crystal Museum	781	Otzovik	93	2023	2013
Holy Dormition Cathedral	690	2gis	85	2024	3467
Maltsovy Museum	673	Fooby	40	2025	291
Historical Museum	663	Autotravel	25		
Dmitrievsky Cathedral	620	Irecommend	23		
Museum Center “Palaty”	505	Tonkosti	3		
Church of Boris and Gleb	583				
Golden Gates	403				
Museum of Nature	257				
The Stoletovs’ House Museum	249				
“Old Vladimir” Museum	166				
V. Khrapovitsky Estate	21				

The sample is further diversified by the inclusion of multiple platforms as sources for the reviews, each with its own specific requirements for this type of text. For example, 2GIS and Google Maps mandate a reference to personal experience, with Google Maps additionally recommending the division of text into paragraphs and advising against overly complex punctuation. Otzovik prioritizes education-related reviews, whereas Autotravel is focused on content related to automobile travel.

Consequently, the corpus of texts subjected to sentiment analysis is characterized by both substantive heterogeneity and differences in text format, which, in our view, complicates the analysis.

The average review length ranges from 21.65 words in 2020 to 28.04 in 2025 (Fig. 1).

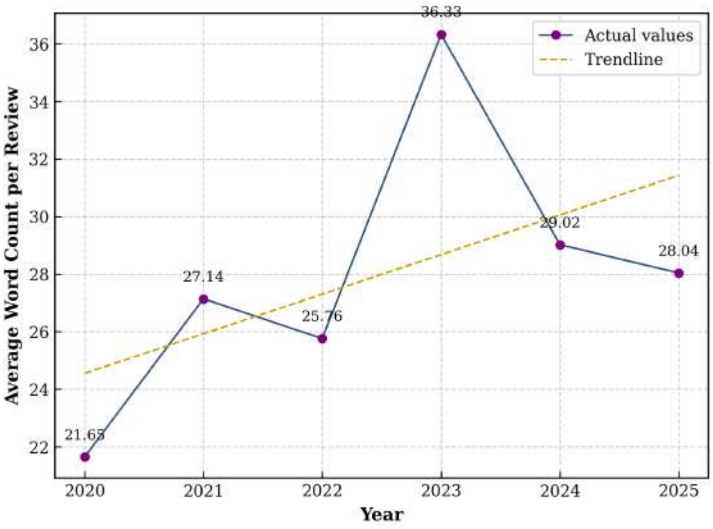


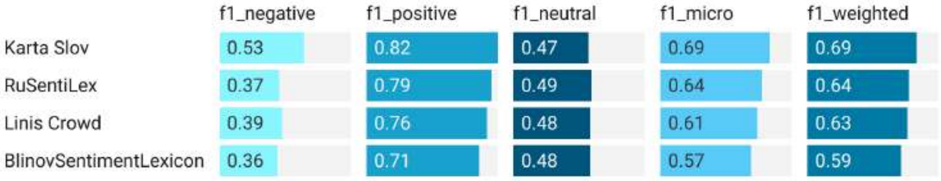
Figure 1. Average review length by year (2020–2025)

Thus, although all reviews belong to the same generalized genre sphere – museum reviews – they are extremely heterogeneous in terms of their themes: some are related to religion and Orthodoxy, others to the history of construction and crafts in Russia, while others are memorial sites dedicated to specific individuals. Different platforms predetermine different text structures. It is noticeable that, on average, the length of reviews increases each year – visitors strive to describe both negative and positive aspects as thoroughly as possible. This complicates automatic sentiment detection, as the sentiment and its intensity may change multiple times within a single text. Notably, in experiments using the dictionary-based method, all texts in the collection underwent standard preprocessing (tokenization, lemmatization, lowercasing), while for experiments with neural networks and LLMs, no preprocessing was performed.

4. Methodology

4.1. Lexicon-based Approach

As sentiment analysis is not a novel task and a number of methods have been suggested, we started from testing the applicability of basic ones such as lexicon-based analysis and pretrained neural networks. Lexicon-based sentiment analysis, though struggling with context and nuance, requires much less computational resources than many other methods, so it was the first method to be tested. For the Russian language there are several sentiment dictionaries. We have chosen the four most popular dictionaries which are not domain-specific and can be used in our field. They are Blinov's Sentiment Lexicon [1], RuSentiLex [18], LinisCrowd [16] and Word Map (Karta Slov) [17]. Lexicon-based analysis was conducted the following way. The reviews were split into sentences. Using each of the dictionaries we classified the sentences as positive, negative, neutral or mixed based on the proportion of positive and negative words in them. The algorithm logic considers possible negations, so a positive word if followed by a negative particle "не" ('not') adds up to negative sentiment score of the sentence. These syntactic dependencies were analyzed using Python library Stanza [22]. Additionally, the cases where negation does not make a phrase negative, for example, "не пожалел" ('did not regret') or "не плохой" ('not bad') were processed as positive. To do so, a list of such words was composed based on preliminary manual analysis of the reviews. To check the quality of classification, we used a sample of 800 sentences retrieved from the reviews on one of the museum sites. Two human annotators gave a sentiment tag (positive, negative, neutral or mixed) to each sentence. Cohen Kappa k=0.86 showed very good overall agreement between annotator 1 and annotator 2. If there was no agreement between the two annotators, the third one gave an additional tag. In this case ground truth label was the mode of the three labels. There were no cases where all three labels were different. To estimate the quality of dictionary-based classification, we measured F1 score for positive, negative and neutral classes as well as micro and weighted F1 (Fig. 2). We focus not only on the overall F1 score, but also on the F1 scores for each class because per-class F1 analysis unveils performance disparities that are obscured by composite metrics. A model may achieve a decent averaged F1 score while simultaneously failing considerably on one or more classes. As we are interested in analyzing positive or negative opinions of visitors on different aspects of their experience, it is crucial to understand if classification is successful for both classes. As we can see from the table (Fig. 2), the number of negative sentences which were correctly classified is rather low (F1 score is around 0.36–0.53), while for the positive sentences classification was more precise. Manual inspection of the cases of wrong class assignment shows that it happens to sentences the emotionality of which is ensured by the knowledge of the context of situation they refer to, but not purely by emotional colouring of the words it contains. For example, a sentence "Между этажами подъем осуществляется по довольно высоким ступеням, особенно большой пролет на третий этаж" (Getting between floors involves climbing rather high steps, and the flight up to the third floor is especially long) was marked as negative by the annotators, but according to the lexicons there are no words with negative polarity, so it was classified as neutral. One more type of negative sentences which are often wrongly classified is a sentence with coordinating adversative conjunction "но" such as "Интересные работы есть, но... их немного" (There are some interesting works of art, but. . . there are few of them). The averaged F1 scores show that the overall performance of this classification method was almost the same regardless of the dictionary. However, analysis based on 'Karta Slov' dictionary allowed for noticeably better negative sentence identification ($f_1^{\text{negative}} = 0.53$) and good results for the positive class ($f_1^{\text{positive}} = 0.82$).



	f1_negative	f1_positive	f1_neutral	f1_micro	f1_weighted
Karta Slov	0.53	0.82	0.47	0.69	0.69
RuSentiLex	0.37	0.79	0.49	0.64	0.64
Linis Crowd	0.39	0.76	0.48	0.61	0.63
BlinovSentimentLexicon	0.36	0.71	0.48	0.57	0.59

Figure 2. Metrics for lexicon-based sentiment classification

The main goal of our analysis was not only to classify the reviews but also identify what exactly the visitors like and dislike. We attempted to do it via N-gram extraction from the two classes of texts (positive and negative) after sentences classification. Based on the results presented in Fig. 2, we classified the sentences using "Karta Slov" lexicon. The sentences were vectorized with simple vectorization method which generates document-term-matrix (with CountVectorizer from Scikit-learn) and then the most frequent bigrams and trigrams (n=40) were extracted. Preprocessing included lemmatization and stop-words removal. Table 2 gives examples of top 15 positive N-grams and Tab. 3 shows negative N-grams.

This method may give the general overview of visitors' experience, however some of the most frequent N-grams are too general, for example, "очень интересный" (very interesting) or "очень понравиться" (liked very much) and some of them are not informative out of context such as "досконально осматривать" (thoroughly examine) in negative bigrams (it is not rather clear what exactly the problem was). To further estimate the efficacy of this approach, we compared N-grams extracted from automatically classified sentences with those extracted from positive and negative classes as assigned by human annotators. For negative reviews only 17.5% of the N-grams (7 out of 40) were similar for automatically and manually classified sentences. Here are some examples of bigrams and trigrams which were found only in negative sentences identified as such by human annotators: "смотреть нечего" (nothing to see), "ребенок год" (child year), "ребенок год туалет" (child year toilet), "оплатить мочь сводить" (pay can take to), "живопись скульптура" (painting sculpture), "пойти музей бесплатно" (go museum for free), "третий этаж" (the third floor), "скидка пенсионер" (discount pensioner), "пустой коридор" (empty hall). This discrepancy is explained by the low precision of lexicon-based classification of negative sentences – many of them are assigned to a wrong class, usually neutral, so in the further N-gram analysis we miss some aspects of visitors' opinions. When it comes to positive reviews, 77.5% of N-grams (31 out of 40) were shared between automatically and manually classified reviews which again is explained by better precision for the positive class.

Table 2. The most frequent N-grams from reviews assigned as positive

N-gram	Translation	n
очень интересный	very interesting	22
очень понравиться	liked very much	18
первый этаж	the first floor	15
интересный экспозиция	interesting exhibits	15
выставка сунгирь	Sungir exhibition	9
второй этаж	the second floor	7
интересный ребёнок	interesting child	7
интересный выставка	interesting exhibition	6
понравиться выставка	like the exhibition	6
музейный центр	museum center	6
ребёнок взрослый	child adult	5
понравиться музей	like the museum	5
икона боголюбский	Bogolubsky icon	5
отличный музей	great museum	5
сотрудник музей	museum staff	5

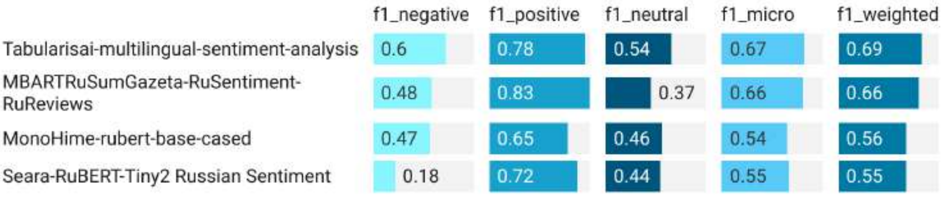
Table 3. The most frequent N-grams from reviews assigned as negative

N-gram	Translation	n
временный выставка	temporary exhibition	3
осмотр уйти час	viewing spend an hour	2
осмотр уйти	viewing spend	2
осматривать весь экспонат	to examine the entire exhibit	2
высокий уровень	high level	2
весь экспонат посетить	the whole exhibit visit	2
досконально осматривать	thoroughly examine	2
музей очень	museum very	2
экспонат посетить временный	exhibit visit temporary	2
экспонат посетить	exhibit visit	2
второй этаж	the second floor	2
временный выставка осмотр	temporary exhibition viewing	2
посетить временный выставка	visit temporary exhibition	2
единый билет	an all-inclusive ticket	2
посетить временный	visit temporary	2

Though N-grams do provide an overview of the reviews content, such analysis is not aspect-based, that is why we tried one more approach. After sentiment classification we conducted syntactic parsing to get information about dependency relations between the words in each sentence. Then we made a sample list of things that people often mention in their reviews in either positive or negative manner, for example, prices, staff, exhibition, etc. It included, for example, such words as “цена” (price), “стоимость” (price, synonym), “билет” (ticket), “персонал” (staff), “экспозиция” (exposition), “выставка” (exhibition), “ребенок” (child). To understand what exactly people say about the things on the list, we extracted units where the desired keyword is a headword and the second word is its dependent word. Dependent function words were not included as they give little information. Below is an example for the words персонал, 'работник' and 'смотритель' which are synonymously used to nominate museum staff (extracted from positive reviews). The total number of extracted word pairs for this query was 40, in the example repetitions are omitted: работников музея, персонал смотрители, смотрители гостеприимные, смотрители,музея, работники посетители, персонал вежливый, персонал приятный, работниками великолепными, смотрительница зала, работники дружелюбные, персонал приветливый, работники всегда, работники приятные, смотрительницы милые, работникам открытым, персонал добродушный, работникам зала, работниц приветливых, персонал отзывчивые, персонал вежливый, персонал доброжелательный, персоналом великолепным персонал экспозиции, персонал духе, персонал общительный, персонал шишкин, персонал икон, персонал копии. As we can see from the presented result, with this approach it is possible to get some valuable insights, but the major drawback is the necessity to compose a comprehensive list of lemmas which nominate the aspects of interest. One more disadvantage is that the extracted patterns which can be interpreted out of context are mainly a noun + adjectival / nominal modifier or a verb and adverbial modifier, but to capture more complex relations, there is a need to write extraction rules manually which is time-consuming and may not consider all possible cases. To summarize, we tested the applicability of simple lexicon-based sentence=level classification and 2 ways of further N-grams extraction as a baseline method of aspect-based sentiment analysis. The results were more precise for positive reviews than for negative. All in all, such pipeline may provide a surface-level understanding of visitors' opinions; however, it is not sufficient for detailed understanding of their attitudes to various aspects of their visit to the museum.

4.2. Neural Models-based Approach

Taking into consideration the disadvantages of lexicon-based approach, we proceeded to test the effectiveness of pretrained models. We tested four popular (according to HuggingFace rating) open-source pretrained models based on RuBERT [9], RuBERT-tiny [10], mBART [7] and multilingual BERT [25] architectures. Training material of all models included reviews of some kind, however they were thematically different from the reviews which we analyse (car reviews, clothes reviews). Models performance was tested on the same sample of 800 sentences which was used in lexicon-based analysis. F1 scores are presented in the table (Fig. 3).



	f1_negative	f1_positive	f1_neutral	f1_micro	f1_weighted
Tabularisai-multilingual-sentiment-analysis	0.6	0.78	0.54	0.67	0.69
MBARTRuSumGazeta-RuSentiment-RuReviews	0.48	0.83	0.37	0.66	0.66
MonoHime-rubert-base-cased	0.47	0.65	0.46	0.54	0.56
Seara-RuBERT-Tiny2 Russian Sentiment	0.18	0.72	0.44	0.55	0.55

Figure 3. Models performance metrics

The performance pattern is similar to that observed in lexicon-based classification. Negativity detection turned out to be a difficult task for pretrained models as well and the averaged F1 score does not exceed 0.69. These results demonstrate that domain unspecific methods (like sentiment lexicons) and solutions created for texts of different style, genre and structure (like pretrained models) cannot provide the expected quality when applied to specific material which in our case is reviews on cultural institution.

4.3. LLM-based Approach

After having tested traditional methods which did not provide expected quality, we decided to utilise large language models (LLMs) to conduct aspect-based sentiment analysis. Our idea is to compose a series of prompts to extract a short yearly report on each museum site reflecting positive and negative aspects which are mentioned by the visitors in their reviews. To do so, we propose the following pipeline (Fig. 4).

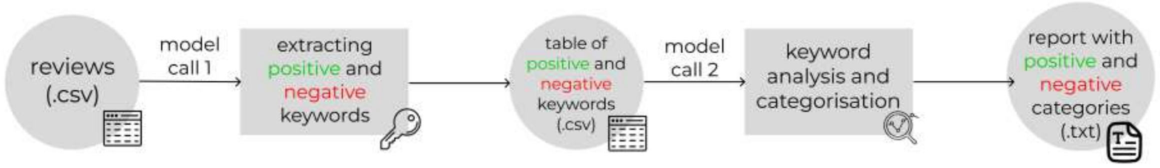

reviews (.csv)
model call 1
extracting positive and negative keywords
table of positive and negative keywords (.csv)
model call 2
keyword analysis and categorisation
report with positive and negative categories (.txt)

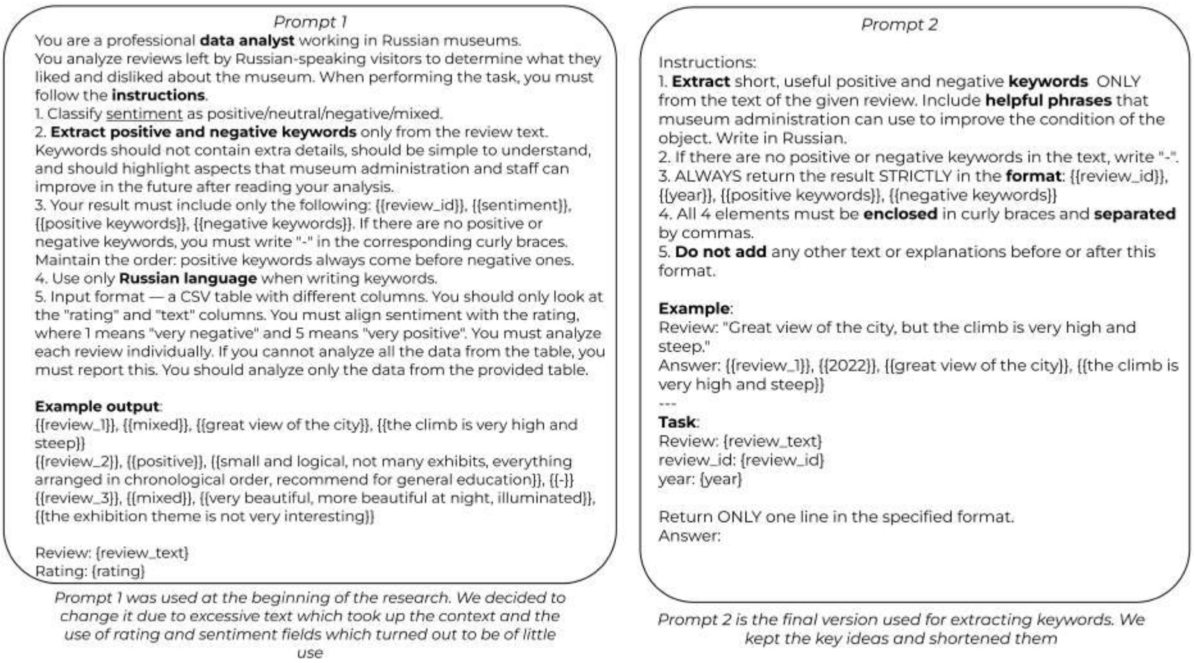
Figure 4. Pipeline for sentiment analysis of reviews by using LLM

For each of the 15 sites there is a csv file with all reviews and their metadata (year of publication, source, etc.). At the first stage (model call 1, Fig. 4), we tasked the LLM with extracting positive and negative keywords from the text of each review and classify them as positive or negative. In this study, we adopt a customized understanding of the term "keyword", which differs somewhat from its conventional usage in information extraction tasks [24]. Our working definition of a keyword – which we also covertly convey to the model in our instructions ("include helpful phrases that museum administration can use to improve the condition of the object", see Fig. 5, prompt 2) – is as follows: a keyword is a minimal predicative phrase that necessarily contains an evaluative predicate and typically includes a nominal reference to the object being characterized. For example: "not many exhibits", "the staff like a throwback to the Soviet era". During experiments with prompts, we encountered a number of limitations and challenges. To overcome them, quality criteria for the prompt in this task were formulated. The limitations and the corresponding prompt quality criteria that address them are presented in Tab. 4.

Table 4. Prompt limitations and corresponding prompt quality criteria that address them

Prompt limitations	Corresponding prompt quality criteria
Limited number of tokens in context window	As short and unambiguous as possible
The pre-existing meaning of the term "keywords" in NLP (which does not imply sentiment attribution)	Outlines the task for extracting positive and negative keywords
The need for consistent attribution of keywords (1) to a specific museum object, and (2) to a positive or negative category	Gives strict formatting instructions. Provides a single example of the expected output
The possibility of model hallucinations	Provides data for extracting the keywords, instructs the model to only utilise the given data, the prompt is as short as possible to make it easier to follow instructions

The final prompt is divided into three parts. In the first part, we give the model clear instructions on how to extract keywords and how to format them in the output. The second part includes an example of an expected output. The third part gives actual data for analysis stored in variables (Fig. 5).



Prompt 1

You are a professional **data analyst** working in Russian museums.
You analyze reviews left by Russian-speaking visitors to determine what they liked and disliked about the museum. When performing the task, you must follow the **instructions**.
1. Classify sentiment as positive/neutral/negative/mixed.
2. **Extract positive and negative keywords** only from the review text. Keywords should not contain extra details, should be simple to understand, and should highlight aspects that museum administration and staff can improve in the future after reading your analysis.
3. Your result must include only the following: {{review_id}}, {{sentiment}}, {{positive keywords}}, {{negative keywords}}. If there are no positive or negative keywords, you must write "-" in the corresponding curly braces. Maintain the order: positive keywords always come before negative ones.
4. Use only **Russian language** when writing keywords.
5. Input format — a CSV table with different columns. You should only look at the "rating" and "text" columns. You must align sentiment with the rating, where 1 means "very negative" and 5 means "very positive". You must analyze each review individually. If you cannot analyze all the data from the table, you must report this. You should analyze only the data from the provided table.

Example output:
{{review_1}}, {{mixed}}, {{great view of the city}}, {{the climb is very high and steep}}
{{review_2}}, {{positive}}, {{small and logical, not many exhibits, everything arranged in chronological order, recommend for general education}}, {{-}}
{{review_3}}, {{mixed}}, {{very beautiful, more beautiful at night, illuminated}}, {{the exhibition theme is not very interesting}}

Review: {review_text}
Rating: {rating}

Prompt 1 was used at the beginning of the research. We decided to change it due to excessive text which took up the context and the use of rating and sentiment fields which turned out to be of little use

Prompt 2

Instructions:
1. **Extract** short, useful positive and negative **keywords** ONLY from the text of the given review. Include **helpful phrases** that museum administration can use to improve the condition of the object. Write in Russian.
2. If there are no positive or negative keywords in the text, write "-".
3. ALWAYS return the result STRICTLY in the **format**: {{review_id}}, {{year}}, {{positive keywords}}, {{negative keywords}}
4. All 4 elements must be **enclosed** in curly braces and **separated** by commas.
5. **Do not add** any other text or explanations before or after this format.

Example:
Review: "Great view of the city, but the climb is very high and steep."
Answer: {{review_1}}, {{2022}}, {{great view of the city}}, {{the climb is very high and steep}}

Task:
Review: {review_text}
review_id: {review_id}
year: {year}

Return ONLY one line in the specified format.
Answer:

Prompt 2 is the final version used for extracting keywords. We kept the key ideas and shortened them

Figure 5. Prompts for negative and positive keywords extraction: initial prompt version (Prompt 1) and the improved version used in the study (Prompt 2)

By doing so, we get a short summary on positive and negative aspects mentioned in the text of each review and eliminate descriptive parts of the review that are emotionally neutral and not informative for our analysis. An example of keyword extraction is presented in Tab. 5.

Classified keywords are already useful for detailed analysis of visitor's experience, but with many sites and hundreds of reviews it is difficult and time-consuming to generalize. That is why at the next step (model call 2, Fig. 4) we asked LLM to process tables with keywords to get a short text report on what visitors liked and disliked. The report is composed by year. We tried several approaches to prompting before we got a decent result. At first, we wanted a model to classify semantically similar keywords into arbitrary categories (positive and negative separately), to name the categories (for example, 'Staff', 'Exposition', 'Infrastructure', 'Atmosphere' etc.) and to display them together with 5 examples of relevant keywords. We then counted how often (in how many reviews) each category was mentioned.

Manual validation of model output showed that counting was not accurate and there were a lot of hallucinations – the model created keywords and categories which did not exist in the reviews. Hallucinations were particularly characteristic of negative categories (e.g., 'Master-classes and events': No master classes for children, no interesting projects). Also, displaying just 5 examples of keywords was insufficient: the name of the category was often broader than keyword examples, so it was not transparent what exactly the model generalized in the category (eg. 'Organization of space and comfort'). The second step was to classify semantically similar keywords into arbitrary categories, but with a restriction to put every keyword only in one of the categories (no missing keywords) and display the name of the category together with all the

Table 5. The result of keywords extraction for one of the sites

review_id	year	positives	negatives
review_533	2020	интересное здание [interesting building]	экспозиции не понравились – бедно [did not like the exhibits – poor]
review_439	2023	спас нас от дождя [sheltered us from the rain]	чайная не работает [tea room is not working]
review_55	2024	вежливый персонал [polite staff], интересная задумка интерьера [interesting interior concept], картинные экспозиции [art exhibitions], иконы и шкатулки [icons and caskets], археологическая интерактивная выставка [archaeological interactive exhibition]	ценник высоковат [price is a bit high], дополнительная плата [extra charge], непродуманная система проверки билетов [poorly designed ticket checking system]
review_456	2020	много чего интересного можно посмотреть и узнать [many interesting things to see and learn]	с родителя взяли за билет и за экскурсовода [charged the parent for the ticket and the guide], дополнительного экскурсовода не было представлено [no additional guide was provided]
review_394	2022	очень хорошее место [very good place], уникальные работы [unique works]	уникальная мебель в очень плохом состоянии [unique furniture in very poor condition]
review_429	2023	–	ужасно дорого [terribly expensive]

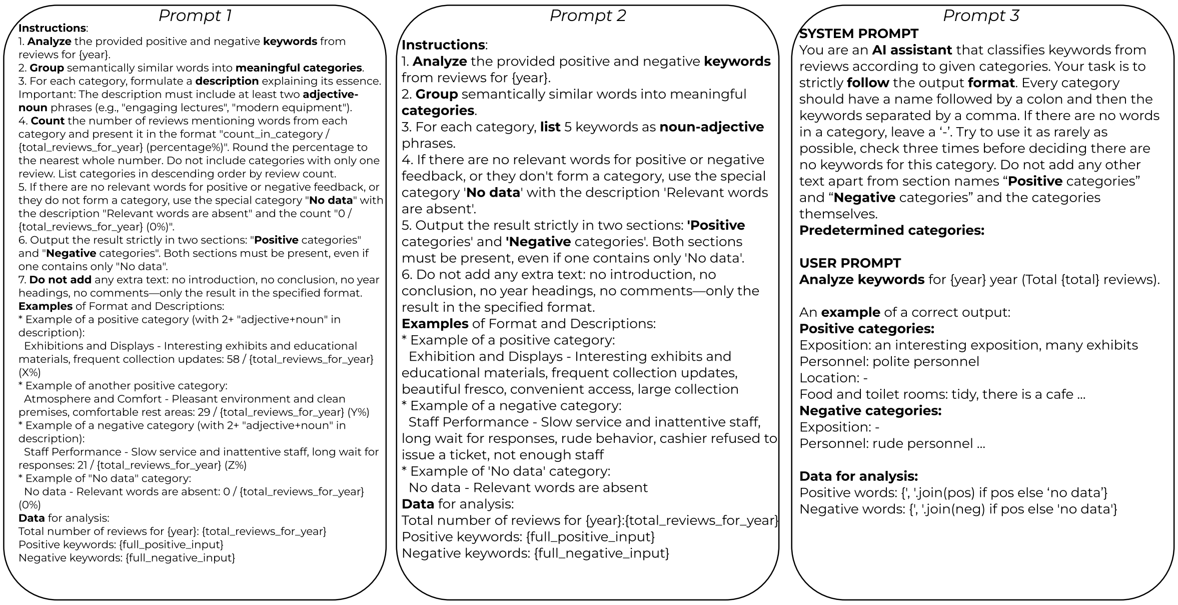
keywords representing it. The logic of categorisation became clearer; however, the output was long and difficult to read. Because the aim of this step was to get a short readable overview of visitors' opinions, we rejected this approach as well.

Then the solution we found was to predefine the list of the categories in the prompt to make reports more structured, predictable and precise. To compose the list of categories we firstly contacted museum workers who consulted us on the aspects of visitor experience that they mostly wanted to know about. We also analyzed related work in the field of visitor studies [28] to understand what information from visitors' reviews is the most valuable and often mentioned.

We got two lists of categories – basic and expanded. Basic: Exposition, staff, location, food and toilets, prices, the appearance of the sites and territory. Expanded: Visit with children, facilities for people with physical disabilities, emotions and atmosphere, general impression, knowledge and education, entertainment and shopping, accessibility (how to get to a place), history and patriotic education.

We experimented with different prompting strategies, mainly using a single prompt of variable length with mostly the same structure: define the role, describe the limitations on the output, describe the task, provide a single example or more than one, give data for analysis. It turned out that smaller models find it difficult to follow longer instructions. Long instructions resulted in errors in formatting the output, hallucinations, especially when producing reports, and other issues. Finally, we decided to split the prompt into a system prompt and a user prompt and make them as short and concise as possible while only including one example of expected output.

To summarize all our observations about the best keyword analysis and categorization prompt, here are the requirements for the prompt for the second model call: it consists of a system and user prompt; both prompts are as short and unambiguous as possible; the system prompt outlines the role and formatting; the user prompt provides the task, example and data (Fig. 6).



Prompt 1

Instructions:
1. **Analyze** the provided positive and negative **keywords** from reviews for {year}.
2. **Group** semantically similar words into **meaningful categories**.
3. For each category, formulate a **description** explaining its essence. Important: The description must include at least two **adjective-noun** phrases (e.g., "engaging lectures", "modern equipment").
4. **Count** the number of reviews mentioning words from each category and present it in the format "count_in_category / {total_reviews_for_year} (percentage%)". Round the percentage to the nearest whole number. Do not include categories with only one review. List categories in descending order by review count.
5. If there are no relevant words for positive or negative feedback, or they do not form a category, use the special category "**No data**" with the description "Relevant words are absent" and the count "0 / {total_reviews_for_year} (0%)".
6. Output the result strictly in two sections: "**Positive** categories" and "**Negative** categories". Both sections must be present, even if one contains only "No data".
7. **Do not add** any extra text: no introduction, no conclusion, no year headings, no comments—only the result in the specified format.

Examples of Format and Descriptions:
* Example of a positive category (with 2+ "adjective+noun" in description):
  Exhibitions and Displays - Interesting exhibits and educational materials, frequent collection updates: 58 / {total_reviews_for_year} (X%)
* Example of another positive category:
  Atmosphere and Comfort - Pleasant environment and clean premises, comfortable rest areas: 29 / {total_reviews_for_year} (Y%)
* Example of a negative category (with 2+ "adjective+noun" in description):
  Staff Performance - Slow service and inattentive staff, long wait for responses: 21 / {total_reviews_for_year} (Z%)
* Example of "No data" category:
  No data - Relevant words are absent: 0 / {total_reviews_for_year} (0%)

Data for analysis:
Total number of reviews for {year}: {total_reviews_for_year}
Positive keywords: {full_positive_input}
Negative keywords: {full_negative_input}

The most detailed prompt and the first one we used. We tried to incorporate counting to evaluate how important each category is. It turned out models almost always hallucinated when producing these numbers. Review rating had little effect on the result and was not used by the model, so we stopped using it in the future.

Prompt 2

Instructions:
1. **Analyze** the provided positive and negative **keywords** from reviews for {year}.
2. **Group** semantically similar words into meaningful **categories**.
3. For each category, **list** 5 keywords as **noun-adjective** phrases.
4. If there are no relevant words for positive or negative feedback, or they don't form a category, use the special category '**No data**' with the description 'Relevant words are absent'.
5. Output the result strictly in two sections: **'Positive** categories' and **'Negative** categories'. Both sections must be present, even if one contains only 'No data'.
6. Do not add any extra text: no introduction, no conclusion, no year headings, no comments—only the result in the specified format.

Examples of Format and Descriptions:
* Example of a positive category:
  Exhibition and Displays - Interesting exhibits and educational materials, frequent collection updates, beautiful fresco, convenient access, large collection
* Example of a negative category:
  Staff Performance - Slow service and inattentive staff, long wait for responses, rude behavior, cashier refused to issue a ticket, not enough staff
* Example of 'No data' category:
  No data - Relevant words are absent

Data for analysis:
Total number of reviews for {year}:{total_reviews_for_year}
Positive keywords: {full_positive_input}
Negative keywords: {full_negative_input}

The second prompt we used. It is much shorter than the first one and tries to force the model to use exactly 5 keywords in each category. The model occasionally followed this instruction, but we decided it would be more useful to include all possible keywords pertaining to this or that category.

Prompt 3

SYSTEM PROMPT
You are an **AI assistant** that classifies keywords from reviews according to given categories. Your task is to strictly **follow** the output **format**. Every category should have a name followed by a colon and then the keywords separated by a comma. If there are no words in a category, leave a '-'. Try to use it as rarely as possible, check three times before deciding there are no keywords for this category. Do not add any other text apart from section names "**Positive** categories" and "**Negative** categories" and the categories themselves.
Predetermined categories:

USER PROMPT
Analyze keywords for {year} year (Total {total} reviews).

An **example** of a correct output:
Positive categories:
Exposition: an interesting exposition, many exhibits
Personnel: polite personnel
Location: -
Food and toilet rooms: tidy, there is a cafe ...
Negative categories:
Exposition: -
Personnel: rude personnel ...

Data for analysis:
Positive words: {', '.join(pos) if pos else 'no data'}
Negative words: {', '.join(neg) if pos else 'no data'}

The final version of the prompt. We divided it into a system and a user prompt. The system prompt uses role modeling to provide clear instructions on the general behaviour of the model and its outputs. The user prompt only gives the task, a single example (one-shot learning) and data.

Figure 6. Prompts for categorization

One of the biggest challenges of this research was choosing a large language model that fit our goals. To do so, we developed certain criteria for choosing the best LLM: accessible on Hugging Face; GGUF format that is compatible with llama-cpp-python; 8B parameters, quantization from 4 to 8 bit; optimized for working with Russian / trained on Russian datasets. Hugging Face was chosen as one of the best places for hosting open source LLMs. Moreover, it has convenient Python libraries to easily download and test different models. The most suitable inference engine for our purposes turned out to be llama-cpp-python because it is fairly well optimized and user-friendly. LLMs have to be contained within a .gguf file in order to be run through llama-cpp-python. The engine requires the user to first initialize an instance of a Llama class where the hyperparameters are defined such as the number of GPU layers to use. Then the response is produced via an 'llm' function that takes the prompt and the class instance. Models are run via Google Colab's T4 GPU that has certain memory limitations; therefore, it is only possible to use models the size of which does not exceed 8B parameters with quantization up to 8 bit. The final criterion is of critical importance. Most open-source models that fit the first three criteria do not have enough Russian in their training set. This results in empty or unsatisfactory outputs when processing Russian texts. We tested a number of models that fit 3 or 4 of the criteria, but most of them produced poor results. Most of the outputs turned out to be empty when using:

1. Mistral 7B Q4 [8];
2. Solar 10.7B Q4 [12].

The result was satisfactory, but still worse than the model of choice:

1. Saiga Llama 8B Q4 [11];
2. Vikhr 7B Q4 [13];
3. YandexGPT-5-Lite 8B Q4 [14].

The model that produced the best results turned out to be YandexGPT-5-Lite 8B Q8 [15]. Its parameters are described in Tab. 6.

Table 6. Parameters of YandexGPT-5-Lite 8B Q8

Parameter	**Value**
Parameter count	8 billion
Base architecture	Llama
Quantization	Q8_0 (8 bit) in GGUF format
Model size	8.54 GB
Maximum context window	32K
Compatibility	Usable with llama-cpp-python

It turned out to produce the best output for the second step in the pipeline – keyword analysis and categorization. The key features of this model include a substantial number of Russian texts in the training set and a tokenizer well optimized for working with Russian as well as compatibility with llama.cpp and a size that does not exceed the memory limit of Google Colab's T4 GPU.

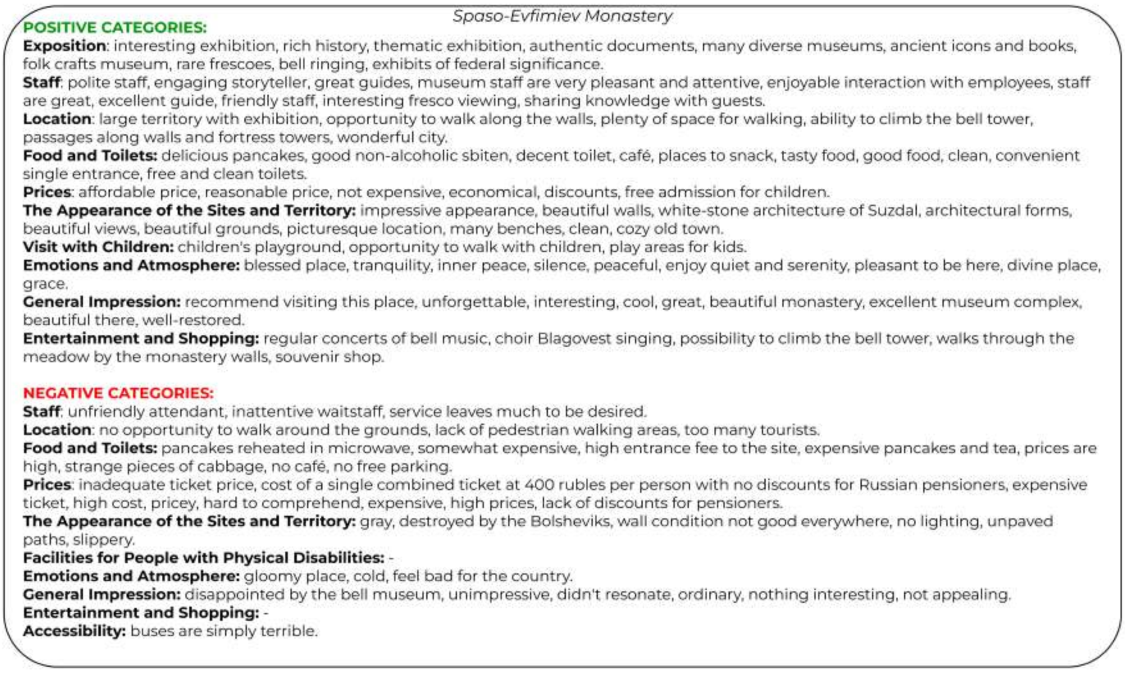
5. Results

The methodology described above allows us to turn hundreds of unstructured reviews into a concise text report which contains information on positive and negative aspects of visitors' experience. Below is an example of such a report for the year 2020 for one site (Fig. 7). In the first part of the report there are keywords that reflect positive opinions. They are grouped thematically into 14 categories. In the second part the same is done for negative keywords.

Using this methodology, we analyzed reviews on 15 different sites that people wrote on popular websites for tourists. To track changes in visitors' opinion, we analyzed reviews from 2020 to 2025. Several observations can be made based on the results of keywords extraction and sentiment classification. The most frequently mentioned positive aspects are exposition, general impression and emotions and atmosphere: about 2500 keywords extracted from the reviews are related to exposition, about 1000 describe visitors' general impression and about 900 were classified as describing emotions of visitors and atmosphere of the place (Fig. 8).

On the other hand, visitors rarely talk about accessibility and facilities for people with physical disabilities in a positive way (only 51 and 21 keywords respectively in 6 years) (Fig. 8). For example, visitors of museum site "Palaty" mention the following positive aspects (translation from Russian): 'good collection', 'interesting exhibitions', 'valuable paintings' (category 'Exposition'); 'recommend this museum', 'interesting place', 'a perfect place for the whole family' (category 'General impression'); 'cozy atmosphere', 'felt hospitality of the staff', 'comfortable' (category 'Emotions and atmosphere'). Keywords in other categories give more specific details, for example 'friendly staff', 'well-restored and renovated', 'great masterclasses for children'.

Looking at the distribution of keywords among categories (Fig. 8, Fig. 9), we may conclude that when writing a review people firstly remember central aspects of their visit – what they saw



Spaso-Evfimiev Monastery

POSITIVE CATEGORIES:

Exposition: interesting exhibition, rich history, thematic exhibition, authentic documents, many diverse museums, ancient icons and books, folk crafts museum, rare frescoes, bell ringing, exhibits of federal significance.
Staff: polite staff, engaging storyteller, great guides, museum staff are very pleasant and attentive, enjoyable interaction with employees, staff are great, excellent guide, friendly staff, interesting fresco viewing, sharing knowledge with guests.
Location: large territory with exhibition, opportunity to walk along the walls, plenty of space for walking, ability to climb the bell tower, passages along walls and fortress towers, wonderful city.
Food and Toilets: delicious pancakes, good non-alcoholic sbiten, decent toilet, café, places to snack, tasty food, good food, clean, convenient single entrance, free and clean toilets.
Prices: affordable price, reasonable price, not expensive, economical, discounts, free admission for children.
The Appearance of the Sites and Territory: impressive appearance, beautiful walls, white-stone architecture of Suzdal, architectural forms, beautiful views, beautiful grounds, picturesque location, many benches, clean, cozy old town.
Visit with Children: children's playground, opportunity to walk with children, play areas for kids.
Emotions and Atmosphere: blessed place, tranquility, inner peace, silence, peaceful, enjoy quiet and serenity, pleasant to be here, divine place, grace.
General Impression: recommend visiting this place, unforgettable, interesting, cool, great, beautiful monastery, excellent museum complex, beautiful there, well-restored.
Entertainment and Shopping: regular concerts of bell music, choir Blagovest singing, possibility to climb the bell tower, walks through the meadow by the monastery walls, souvenir shop.

NEGATIVE CATEGORIES:

Staff: unfriendly attendant, inattentive waitstaff, service leaves much to be desired.
Location: no opportunity to walk around the grounds, lack of pedestrian walking areas, too many tourists.
Food and Toilets: pancakes reheated in microwave, somewhat expensive, high entrance fee to the site, expensive pancakes and tea, prices are high, strange pieces of cabbage, no café, no free parking.
Prices: inadequate ticket price, cost of a single combined ticket at 400 rubles per person with no discounts for Russian pensioners, expensive ticket, high cost, pricey, hard to comprehend, expensive, high prices, lack of discounts for pensioners.
The Appearance of the Sites and Territory: gray, destroyed by the Bolsheviks, wall condition not good everywhere, no lighting, unpaved paths, slippery.
Facilities for People with Physical Disabilities: -
Emotions and Atmosphere: gloomy place, cold, feel bad for the country.
General Impression: disappointed by the bell museum, unimpressive, didn't resonate, ordinary, nothing interesting, not appealing.
Entertainment and Shopping: -
Accessibility: buses are simply terrible.

Figure 7. Report with positive and negative categories

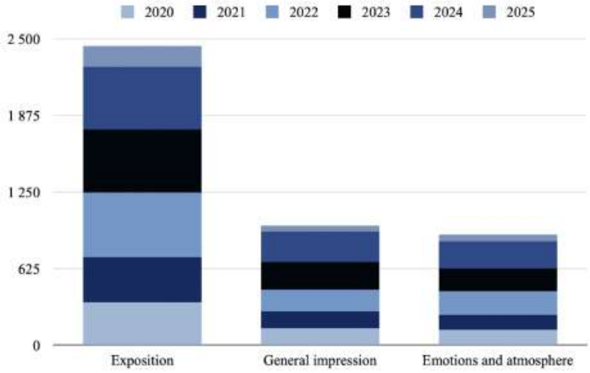


Figure 8. The most often mentioned positive categories

(what exhibits), what the place looked like (category ‘Appearance of the sites and territory’) and how they felt about it. Visitors also do not forget to mention staff if they were friendly and polite. Other details are mentioned optionally.

When it comes to negative opinions, we see that people mention negative things in their reviews much more rarely than positive – the most popular negative categories have 222, 294 and 304 keywords in them (Fig. 10) compared to hundreds and thousands of positive keywords.

Among the most frequent aspects which people mention as negative are ‘General impression’ and ‘Exposition’ (similar to the positive ones), however the third category is different which is ‘Price’ (‘a bit pricey’, ‘not cheap’, ‘high price is unjustified’). There is one category for which no negative opinions have been found – history and patriotic education (Fig. 11).

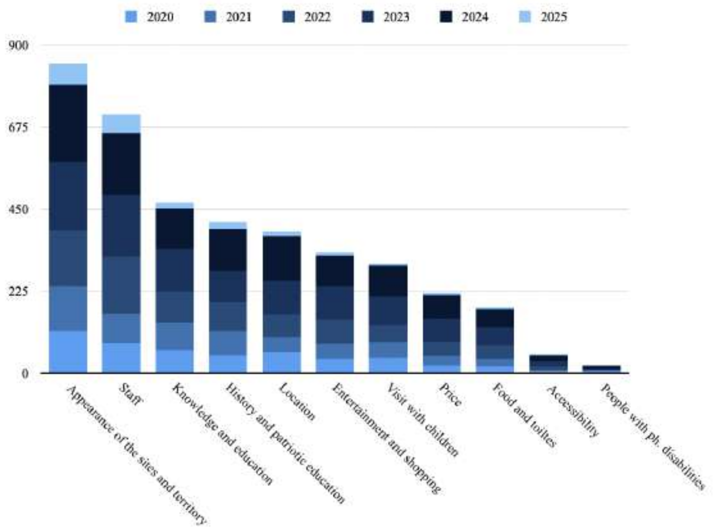


Figure 9. Amount of keywords in positive categories

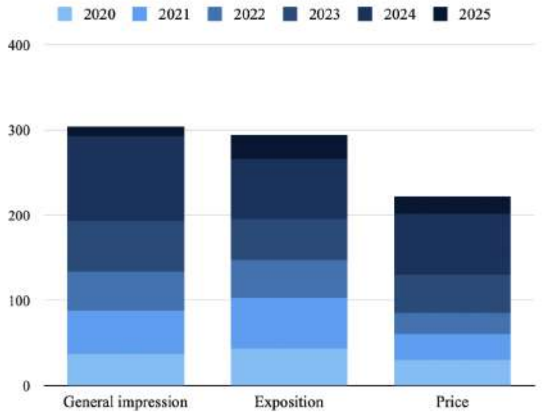


Figure 10. The most often mentioned negative categories

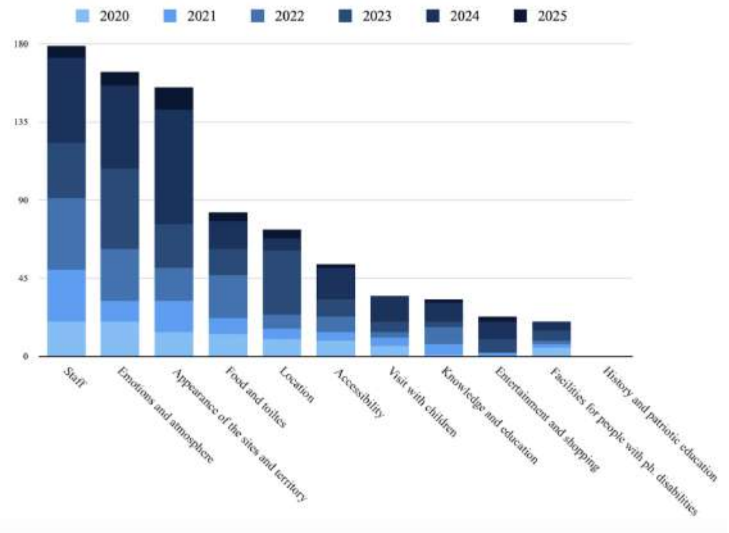


Figure 11. Amount of keywords in negative categories

6. Discussion

In this paper we presented the results of collaboration with a museum which needed to optimize visitors' feedback analytics. We have tested a new approach to ABSA of visitors' reviews that leverages large language models and their growing effectiveness in solving many analytical NLP tasks.

For the museum workers it was important to turn thousands of reviews into short, concise yet informative reports on what their visitors like and do not like. To do so, we decomposed the analysis into several steps – expressive keywords extraction, their classification into positive and negative and thematic categorization.

As a result, the information is compressed at two levels: for more detailed analysis the keywords can be used, for an overview there is a report summarizing the information by categories.

The applied approach has its advantages and disadvantages. First advantage is that, among the extracted keywords, there are n-grams of varying lengths: bigrams ('интересная экспозиция' [interesting exhibition]), trigrams ('можно картой оплатить' [accept card payments]), 4-grams ('непродуманная система проверки билетов' [poorly designed ticket checking]), etc. The use of diverse n-grams provides granular insights into visitor preferences, significantly enhancing aspect-level sentiment analysis.

Moreover, the limitation observed in experiments with the dictionary-based method has been overcome – the severe disparity between negative and positive keywords was alleviated. Despite being fewer in number, negative keywords were still extracted for every object. To assess the quality of keyword extraction, we manually annotated a sample of 300 reviews comprising 1331 keywords (in human annotation) and compared them with the extraction performed by the LLM. We considered keywords "missed" if the model (1) failed to add a meaningful keyword, (2) extracted a word/phrase that cannot be considered a keyword for the review or (3) failed to recognize the correct sentiment polarity (e.g., 'до ужаса красив' as negative). The results can be observed in Fig. 12. Automatic keyword extraction turned out to be quite effective with the model "missing" only 92 out of 1331 keywords (6.9%).

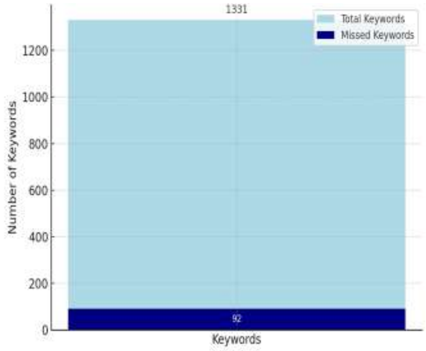


Figure 12. Comparison between human and LLM keyword extraction

Thirdly, despite the absence of direct sentiment cues in certain n-grams, the model accurately infers their polarity based on contextual clues and categories them appropriately ('иконы и шкатулки' (positive); 'только учитель прошёл бесплатно' (in context '. . . а с родителя взяли за билет и за экскурсовода ' (negative)).

Finally, the two-step pipeline eliminated the need for separate topic modeling: the LLM automatically groups keywords into thematic clusters. In most cases, we managed to limit hallucinations for “empty categories” – the model inserts a dash in the thematic category for which no positive or negative keywords were found, rather than inventing non-existent ones.

As for disadvantages, we will focus on several of them.

Firstly, via prompting we did not manage to provide any quantitative data on the frequency of certain topics. The LLM which we have chosen as well as those we tested did not manage to do correct calculations. We tried to instruct the model to count the number of keywords used in each of the categories in absolute and relative figures, but it failed to do so. Proper calculations require access to tools such as code, which is not accessible to smaller models used by us.

The second disadvantage is that we still may encounter model hallucinations for different reasons. One of them is lack of data. For example, when there are 200 reviews for a site in one year and only 10 reviews have some negative aspects mentioned, when classifying keywords the model can make them up to fill in the categories though it is instructed not to do so. Another cause for hallucinations is the comparatively small size of LLMs used, especially the use of quantized models. The smaller the model and the higher the quantization, the less accurate predictions can be made by the model, which results in its failure to strictly follow instructions and keep real data in its context. Hence, manual checking is always needed. We compared the reports for two museum sites and the corresponding tables with keywords to reveal the number of hallucinated keywords that were not present in the table provided to the model as data for analysis (Fig. 13). For the Church of Boris and Gleb, hallucinations made up 7.9% (29 out of 367 keywords), the Suzdal Kremlin report had 5.7% of hallucinated keywords (62 out of 1088 keywords). The percentage of hallucinations is comparatively low, but it only proves the existing problem of large language models fabricating data. Besides, we encountered other minor issues such as repetition, miscategorisation (failure to properly attribute keywords to a category) and grammar mistakes (e.g., agreement between an adjective and a noun – ‘подробная путеводитель’ [detailed guide]).

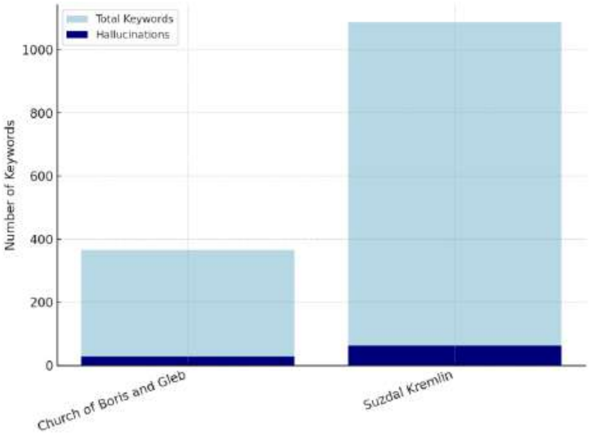


Figure 13. Number of hallucinations in some reports

The final limitation is that optimal keyword categorization requires pre-defined themes, necessitating either expert annotation or client guidance. Despite providing mandatory/optional category lists, the model over-generated outputs without discrimination. We propose addressing this via advanced prompting techniques like chain-of-thought and few-shot learning.

Conclusion

This study demonstrates the efficacy of utilizing LLMs for ABSA of museum visitor reviews, addressing the challenges posed by the multi-thematic and open-domain nature of such texts. By implementing a structured pipeline that combines keyword extraction, sentiment classification, and thematic categorization, we successfully transformed unstructured review data into actionable insights without relying on traditional topic modeling techniques. The proposed methodology leverages the contextual understanding capabilities of LLMs to handle diverse n-grams and implicit sentiment cues, achieving a balanced representation of positive and negative aspects across predefined thematic categories. Key advantages include the elimination of sentiment polarity bias, reduced computational dependency on preprocessing, and the ability to generate concise, human-readable reports for end-users. However, limitations such as model hallucinations, quantization constraints, and the need for predefined categories highlight areas for future refinement. Despite these challenges, our approach offers a scalable solution for cultural institutions seeking to optimize visitor experience analytics. Future work could explore the integration of tool-enabled LLMs for quantitative analysis, advanced prompting strategies like chain-of-thought, and domain-specific fine-tuning to further enhance accuracy and reduce manual validation efforts. This research contributes to the growing body of work on LLM applications in NLP and underscores their potential to revolutionize sentiment analysis in non-commercial domains.

Funding: The article and the dataset were prepared within the framework of the Basic Research Program at HSE University in 2025.

References

1. Blinov, P.D., *et al.*: Research of lexical approach and machine learning methods for sentiment analysis. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue-2013”. vol. 12(19), pp. 51–61 (2013)

2. Brauwers, G., Frasincar, F.: A survey on aspect-based sentiment classification. ACM Computing Surveys 54(1), 1–35 (2021). https://doi.org/10.1145/3503044

3. Choi, Y., Wiebe, J.: +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1181–1191 (Oct 2014). https://doi.org/10.3115/v1/D14-1125

4. Feng, X., Wang, C., Zou, T.T.: Visitor experience of the grand canal national cultural park museum based on sentiment analysis algorithm. SSRG International Journal of Electrical and Electronics Engineering 11(9), 142–150 (2024). https://doi.org/10.14445/23488379/IJEEE-V11I9P112

5. Gao, Y., Wang, R., Hou, F.: How to Design Translation Prompts for ChatGPT: An Empirical Study. In: Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops. Association for Computing Machinery (2024). https://doi.org/10.1145/3700410.3702123

6. Gatti, L., Guerini, M., Turchi, M.: SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. IEEE Transactions on Affective Computing 7(4), 409–421 (2015). https://doi.org/10.1109/TAFFC.2015.2476456

7. Hugging Face: MBARTRuSumGazeta-RuSentiment-RuReviews. https://huggingface.co/sismetanin/mbart_ru_sum_gazeta-ru-sentiment-rureviews

8. Hugging Face: Mistral-7B-v0.1-Q4_K_M-GGUF. https://huggingface.co/3dsabh/Mistral-7B-v0.1-Q4_K_M-GGUF

9. Hugging Face: RuBERT Conversational Cased Sentiment. https://huggingface.co/MonoHime/rubert_conversational_cased_sentiment

10. Hugging Face: RuBERT-Tiny2 Russian Sentiment. https://huggingface.co/seara/rubert-tiny2-russian-sentiment

11. Hugging Face: saiga_llama3_8b-Q4_K_M-GGUF. https://huggingface.co/itlwas/saiga_llama3_8b-Q4_K_M-GGUF

12. Hugging Face: SOLAR-10.7B-Instruct-v1.0-Q4_K_M-GGUF. https://huggingface.co/solxxcero/SOLAR-10.7B-Instruct-v1.0-Q4_K_M-GGUF

13. Hugging Face: Vikhr-7B-instruct_0.2-Q4_K_M-GGUF. https://huggingface.co/itlwas/Vikhr-7B-instruct_0.2-Q4_K_M-GGUF

14. Hugging Face: YandexGPT-5-Lite-8B-instruct-GGUF. https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct-GGUF

15. Hugging Face: YandexGPT-5-Lite-8B-instruct-Q8_0-GGUF. https://huggingface.co/BolониniD/YandexGPT-5-Lite-8B-instruct-Q8_0-GGUF

16. Koltsova, O.Y., Alexeeva, S.V., Kolcov, S.N.: An opinion word lexicon and a training dataset for russian sentiment analysis of social media. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue-2016”. vol. 15(22), pp. 277–287 (2016). https://doi.org/10.5281/zenodo.4084953

17. Kulagin, D.: Russian word sentiment polarity dictionary: a publicly available dataset. Poster, Artificial Intelligence and Natural Language (AINL 2019) (2019). https://doi.org/10.28995/2075-7182-2021-20-1106-1119

18. Loukachevitch, N., Levchik, A.: Creating a General Russian Sentiment Lexicon. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2016). pp. 1171–1176. European Language Resources Association (ELRA) (2016)

19. Loukachevitch, N., Tkachenko, N., Lapanitsyna, A., *et al.*: RuOpinionNE-2024: Extraction of Opinion Tuples from Russian News Texts (2025)

20. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1–2), 1–135 (2008). https://doi.org/10.1561/1500000011

21. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: Mooney, R.J. (ed.) Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86. Association for Computational Linguistics, Philadelphia, USA (2002). https://doi.org/10.3115/1118693.1118704

22. Qi, P., Zhang, Y., Zhang, Y., *et al.*: Stanza: A Python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 101–108. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-demos.14

23. Qin, L., *et al.*: Large Language Models Meet NLP: A Survey. Frontiers of Computer Science (FCS) (2024). https://doi.org/10.1007/s11704-025-50472-3

24. Sheremetyeva, S.: An efficient patent keyword extractor as translation resource. In: Proceedings of the MT Summit XII: Third Workshop on Patent Translation. pp. 25–32. Ottawa, Canada (2009)

25. Tabularisai, Gyamfi, S., Borisov, V., Schreiber, R.H.: Multilingual-sentiment-analysis (revision 69afb83) (2025). https://doi.org/10.57967/hf/5968

26. Vatolin, A.: Structured sentiment analysis with large language models: A winning solution for RuOpinionNE-2024. In: Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue 2025). FRC CSC RAS, Moscow, Russia (2025)

27. Wang, Z., Xie, Q., Feng, Y., *et al.*: Is chatGPT a good sentiment analyzer? In: First Conference on Language Modeling (2024), https://openreview.net/forum?id=mUlLf50Y6H

28. Xu, Q., Shih, J.Y.: Applying text mining techniques for sentiment analysis of museum visitor reviews. In: 2024 IEEE 4th International Conference on Electronic Communications, Internet of Things and Big Data (ICEIB). pp. 270–274. Taipei, Taiwan (2024). https://doi.org/10.1109/ICEIB61477.2024.10602556

29. Yadav, A., Vishwakarma, D.: Sentiment analysis using deep learning architectures: A review. Artificial Intelligence Review 53(6), 4335–4385 (2020). https://doi.org/10.1007/s10462-019-09794-5

30. Zhang, T., Irsan, I., Thung, F., *et al.*: Revisiting sentiment analysis for software engineering in the era of large language models. ACM Transactions on Software Engineering and Methodology 34(3), 1–30 (2025). https://doi.org/10.1145/3697009

31. Água, M., Antonio, N., Carrasco, M.P., *et al.*: Large language models powered aspect-based sentiment analysis for enhanced customer insights. Tourism & Management Studies 21(1), 1–19 (2025). https://doi.org/10.18089/tms.20250101