

Sentiment Analysis of Government Mobile App Reviews: Validating Zero-Shot Classification with a Large Language Model

Lukáš Johanovský

lukas.johanovsky@ujep.cz

Jan Evangelista Purkyně University in Ústí nad Labem

Pavla Vrabcová

Jan Evangelista Purkyně University in Ústí nad Labem

Research Article

Keywords: citizen feedback, digital public services, Czech language, text classification, human annotation, classification stability

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11156504/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

User reviews of mobile applications provide a source of citizen feedback on digital public services, yet their automated analysis in non-English settings is constrained by the limited availability of domain-adapted classifiers. This study develops and evaluates a documented procedure for three-class sentiment classification of Czech reviews of government mobile applications using a large language model without task-specific fine-tuning. The corpus comprised 10,198 reviews of six applications. An adjudicated human reference was created for a stratified random sample of 500 reviews, and LLM performance was compared with a majority-class baseline, two lexicon-based methods, and two readily deployable encoder-based transfer models. The validation included five repeated runs under identical settings and an analysis of classification errors. Across five GPT-5.6 Sol validation runs, mean accuracy was 93.0%, mean macro-F1 was 0.827, and mean Cohen's kappa was 0.857. Macro-F1 ranged from 0.814 to 0.845, while agreement among the five runs was high (Fleiss' kappa = 0.918). The neutral class was the most challenging, with most disagreements occurring at its boundaries with negative and positive sentiment. In the full corpus, 65.9% of reviews were classified as negative, 7.9% as neutral, and 26.2% as positive. Textual sentiment was strongly associated with star ratings but was not identical to them. These findings provide corpus-specific evidence that a standardized LLM procedure can help analyze short-form non-english citizen feedback. They also underscore the need for human reference labels, stability checks, and precise documentation of the model, prompt, settings, and execution workflow.

1 Introduction and Theoretical Background

1.1 Government Mobile Applications and User Reviews

Mobile applications have become one of the channels through which public administrations provide digital services. Their significance, however, cannot be assessed solely in terms of technical availability; it also depends on whether citizens adopt them and, on the experiences, citizens have when using them. Heeks and Bailur [1] note that the benefits of e-government do not follow automatically from the introduction of technology and must be evaluated within specific institutional and user contexts. Research on government mobile applications therefore examines factors such as perceived usefulness, trust, and information quality [2], as well as usability, accessibility, interaction, privacy, and security [3]. User reviews can capture some of these experiences, but their sentiment cannot be equated with the objective quality of an application, the quality of the public service as a whole, or the satisfaction of all citizens [4, 5, 6, 7, 3].

1.2 Mobile App Reviews as Citizen Feedback

Mobile app reviews provide public administrations with a continuous, user-initiated record of experiences with digital services. Previous studies have used them to analyze citizen feedback, dimensions of satisfaction, and problems with government applications [8, 9, 10]. Unlike an aggregate rating, review text can indicate what the user's attitude concerns and can capture technical problems,

requests, or responses to application updates [7, 11]. Reviews are therefore a potentially useful supplementary source for monitoring citizens' experiences with a specific digital interface.

At the same time, the analytical value of reviews has clear limitations. Reviewers are self-selected, rating distributions tend to be polarized, and the relationship among review text, star ratings, and the objective quality of a service is not straightforward [4, 5, 6]. Review sentiment therefore cannot be interpreted as representative of satisfaction among all citizens. In this study, sentiment refers to the attitude expressed in the available review text, not to an overall assessment of the application or public service.

1.3 Sentiment Analysis in Non-English Settings

The challenge of non-English sentiment analysis lies not in a complete absence of methods but in their uneven applicability across languages, domains, and classification schemes. Cross-lingual transfer is motivated primarily by the scarcity of annotated data and specialized resources in target languages [12]. Research on mobile app reviews sometimes excludes non-English texts at the data-collection stage or translates them into English before analysis [11]. The availability of a general multilingual model therefore does not, by itself, resolve the mismatch among the target language, text type, and required classes.

Empirical studies have documented this problem across a range of non-English settings. A language-pretrained model for Arabizi outperformed multilingual BERT [13], combined English-Italian fine-tuning improved the classification of Italian reviews [14], and IndoBERT outperformed multilingual BERT on Indonesian mobile app reviews [15]. The latter study, however, derived sentiment labels from star ratings or a lexicon rather than from independent human annotation. These findings thus demonstrate the importance of linguistic and domain adaptation, while also highlighting the limited comparability of existing solutions.

Czech is treated here as an appropriate case-study setting rather than as representative of all non-English languages. It is a morphologically rich inflectional language with relatively free word order [16], examined here in the form of short, informal texts. In studies of Czech social media, annotation disagreements were concentrated particularly in neutral and bipolar cases and in texts involving irony, sarcasm, or context-dependent meaning [17]. These characteristics do not establish that Czech is generally more difficult than other languages, but they support its use as a relevant test case for classifying original non-English text.

1.4 Transferability of Classification Methods

Lexicon-based methods are transparent and do not require annotated training data. Language-specific or multilingual lexicons exist for a range of European languages [18, 19, 20, 21], but their transferability depends on translation quality, domain coverage, and text preprocessing [22, 19, 20]. An error analysis of a Czech lexicon-based approach identified particular difficulties with negation, multiword and domain-dependent expressions, adversative constructions, especially those involving the Czech equivalent of

'but,' and implicit evaluation [23, 21]. An available lexicon is therefore an appropriate transparent baseline, but not necessarily a readily transferable classifier for a new domain.

Encoder-based transformers, most notably BERT and RoBERTa, create contextual representations of text, but a pretrained encoder alone is not a complete sentiment classifier [24, 25]. Classification output is generally obtained only after task-specific fine-tuning; transfer to another language or domain therefore depends on the relationship among pretraining, the sentiment data, and the target text [13, 14, 26]. The weaker performance of a particular off-the-shelf checkpoint cannot therefore be generalized to the entire BERT or RoBERTa architecture, nor to a model fine-tuned directly on domain-matched data.

The performance of all methods is also conditioned by how the task itself is defined. Neutrality may include factual description, a question, a request, an absence of attitude, or a mixed evaluation with no clear predominance. In a meta-analysis of 217 studies and more than 1,100 experimental results, Hartmann et al. [27] found that moving from binary to three-class classification reduced accuracy by an average of 12.7 percentage points. In three-class zero-shot classification, Koto et al. [28] explicitly define the neutral class using a decision interval, illustrating the need for its operational specification. In a multilingual evaluation of LLMs, Kastrati et al. [29] found that the neutral class produced the weakest results across different languages. A credible comparison therefore cannot rely solely on overall accuracy; it must also examine class-specific performance and patterns of misclassification.

1.5 Large Language Models, Validity, and Reproducibility

Prompted large language models offer a different trade-off. Target classes and decision rules can be specified through instructions without creating a new task-annotated training set or conducting further fine-tuning. The terms zero-shot and 'without task-specific fine-tuning' refer to the mode of use, not to the origin of the model, which has already undergone extensive pretraining and instruction tuning. The practical advantage of an LLM therefore lies in making classification available for a new task, not in an assumption of universal superiority.

Existing findings in this area are mixed. Without new task-specific training, GPT achieved competitive performance in multilingual psychological text analysis [30] and in cross-lingual classification of Czech and French film reviews [31]. At the same time, strong performance on overall polarity does not transfer automatically to more complex sentiment tasks [32], and outcomes depend on the model, language, and prompt [29, 33]. An LLM must therefore be evaluated as a candidate method for a specific corpus rather than assumed in advance to be the best classifier.

Corpus-level validation must distinguish accuracy from stability. Repeated runs can show whether a model produces similar outputs under the same instructions, but a stable model can still be systematically wrong. Performance must therefore be compared against an independently created human ground truth [34, 35, 36]. Inter-annotator agreement alone does not guarantee content validity; Klie et al. [36] regard it as a necessary but insufficient condition and emphasize the importance of

instructions, annotation quality control, and adjudication. Validation of automated content analysis must therefore be tied to a specific purpose and corpus [35].

Proprietary LLMs also have reproducibility limitations arising from opaque training data and ongoing model updates [37]. Documenting the exact model designation, date of use, prompt, settings, and batch-processing procedure cannot guarantee identical replication in the future, but it is essential for evaluating and reproducing the procedure as closely as possible.

1.6 Research Gap

Studies in public administration have already used reviews of government mobile applications to analyze citizen feedback, usability and authentication problems, reliability, and the overall user experience [8, 10]. These studies generally draw on different linguistic, institutional, and domain contexts, and their conclusions cannot therefore be transferred directly to Czech reviews of government mobile applications without further validation [10, 38].

The literature on LLM-based sentiment analysis cautions that the performance of large language models cannot be assumed. LLMs may be particularly useful when annotated data are limited, but their results depend on the task, domain, prompt, and validation procedure [32, 39]. Studies of multilingual zero-shot sentiment analysis suggest that LLMs can be promising but may not match specialized or task-fine-tuned models, reinforcing the need for corpus-specific validation [32, 40].

Research on Czech sentiment analysis is developing concurrently, especially through transformer models and, more recently, large language models. These studies, however, primarily use other types of data, including film reviews, product reviews, social-media texts, and aspect-based datasets, rather than reviews of Czech government mobile applications [41, 42]. Recent work on Czech aspect-based sentiment analysis likewise indicates that the use of LLMs for Czech remains an area requiring task-, language-, and domain-specific evaluation [42, 43].

Existing research leaves open how a prompted LLM procedure performs on Czech reviews of government mobile applications under a three-class sentiment scheme. This study addresses that corpus-specific question by combining a real-world review corpus with evaluation against adjudicated human labels, comparison with selected lexicon-based and encoder-based transfer baselines, repeated-run stability checks, error analysis, and subsequent application to the full corpus.

The study addresses the following research questions:

RQ1: To what extent does sentiment classification using a standardized LLM procedure agree with adjudicated human ratings compared with selected lexicon-based and transfer-learning methods?

RQ2: How stable is LLM-based sentiment classification across independent runs under identical settings?

RQ3: In which sentiment classes and types of borderline cases do disagreements between the LLM and adjudicated human ratings occur most frequently?

RQ4: What is the distribution of sentiment in the corpus of reviews of digital public services, and how is it associated with star ratings and individual applications?

2 Methods

2.1 Research Design

The study combines methodological validation of a classification procedure with its application to a real-world corpus of citizen feedback. Its primary methodological objective is to determine whether a standardized LLM procedure without task-specific fine-tuning can serve as a sufficiently accurate and repeatable tool for three-class analysis of short non-English reviews. The Czech corpus provides a case-study setting that combines a morphologically rich language, short informal texts, and limited availability of readily deployable domain-specific classifiers. The findings therefore provide corpus-specific evidence on the utility of a standardized LLM procedure for three-class classification of short-form non-English citizen feedback when no domain-fine-tuned classifier is readily available, that is, for zero-shot sentiment classification.

2.2 Dataset and Selection of the Validation Sample

The initial corpus comprised 10,198 user reviews of six Czech government mobile applications. Reviews were collected from Google Play on 30 April 2026 using a Python notebook based on the google-play-scraper package. The notebook was run separately for each application using its Google Play package identifier, with lang='cs' and country='cz' in the scraper request. The six exported files were combined manually in Excel, and reviews with empty text were excluded. The exports also contained reviewer usernames and other review metadata; usernames were not used in the reported analyses. At the time of data collection, these six applications constituted the complete set of official nationwide, citizen-facing mobile eGovernment applications operated by or on behalf of Czech central public authorities and available through Google Play in the Czech Republic. The resulting corpus covered the following review periods: eRecept, 8 December 2017-29 April 2026; Klíč eGovernmentu, 27 June 2019-26 April 2026; EZKarta, 30 June 2021-29 April 2026; eDoklady, 18 January 2024-29 April 2026; Portál občana, 17 April 2024-29 April 2026; and JENDA, 30 May 2025-28 April 2026. All included records contained non-empty review text and a unique review identifier. Records with identical text but different identifiers were retained as distinct reviews. Because the applications were launched at different times, the periods covered differed across applications.

Annotated Python notebook (Online Resource 1) documenting the Google Play review collection procedure for one application. The same collection procedure was used for all six applications with the corresponding APP_ID; removal of empty review texts and merging of exports were performed separately in Excel.

A stratified random sample of 500 reviews was then selected for manual annotation and comparison of classification procedures. First, a proportional quota by application was established using the largest remainder method. Within each application, the quota was then allocated by star rating using the same method. Reviews were randomly selected from the final strata defined by the combination of application and star rating, after which their order was randomized. Stratification preserved the representation of both applications and rating categories. Star ratings were used solely to construct the sample and were not used in sentiment classification.

2.3 Manual Annotation and Construction of the Ground Truth

Each of the 500 reviews was independently classified by two annotators into one of three categories: positive, neutral, or negative. The assessment concerned the author's overall attitude toward the application, its functions, usability, or technical operation. The positive category indicated predominant satisfaction or praise, and the negative category indicated predominant dissatisfaction, criticism, or failure. The neutral category was operationalized broadly. It included factual or descriptive content, questions and requests, mixed evaluations without a clearly predominant polarity, text without discernible sentiment, and cases in which neither positive nor negative sentiment could be determined reliably. This definition was established before classification and was applied consistently in the instructions provided to both the human annotators and the LLMs.

The first two annotators worked independently to prevent their initial decisions from influencing one another. When their ratings agreed, that category was adopted directly. Disagreements and borderline cases were reviewed by a third annotator, who determined the final category. This adjudication procedure produced a reference ground truth for all 500 reviews. Agreement between the first two annotators was evaluated before adjudication.

2.4 Classification Methods Compared

The validation sample was classified using a simple majority-class baseline, two lexicon-based methods, two readily deployable encoder-based transformer sentiment checkpoints, and the prompted generative models GPT-5.5 and GPT-5.6 Sol (corresponding official API model identifiers: gpt-5.5 and gpt-5.6-sol). The majority-class baseline assigned every review to the most frequent category in the adjudicated ground truth, namely negative sentiment. Encoder inference was performed without additional fine-tuning using a batch size of 16, a maximum sequence length of 512 tokens, and truncation for longer sequences; each input string was limited to its first 2,000 characters before tokenization. The BERT implementation used `nlptown/bert-base-multilingual-uncased-sentiment`, with its one- and two-star outputs mapped to negative, three-star output to neutral, and four- and five-star outputs to positive. The XLM-RoBERTa implementation used `cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual`, whose negative, neutral, and positive outputs mapped directly to the target categories. The comparison was deliberately restricted to procedures that did not require the creation of a new task-annotated training set or additional model fine-tuning. Models adapted to other non-English languages or domains were

included in the theoretical comparison but were not used to classify the Czech corpus. BERT and XLM-RoBERTa therefore represent selected off-the-shelf checkpoints rather than the maximum attainable performance of encoder architectures after domain-specific fine-tuning.

Table 1
Sentiment Classification Methods Tested

Group	Method	Implementation and source
Reference baseline	Majority class	Assigned every review to negative sentiment, the most frequent class in the adjudicated ground truth (329 of 500; 65.8%); no model fitting or textual features.
Lexicon-based	NRC	Automatically translated Czech version of the NRC Emotion Lexicon [44]; UDPipe lemmatization [45]; the sign of the difference between positive and negative matches determined the class.
Lexicon-based	Subjectivity	Czech Subjectivity Lexicon [21]; UDPipe lemmatization [45]; analogous sign-based rule.
Encoder-based transformer	multilingual BERT	Fine-tuned for sentiment on product reviews in six languages, excluding Czech. Outputs were mapped as follows: 1–2 stars = negative, 3 stars = neutral, and 4–5 stars = positive [24, 46].
Encoder-based transformer	XLM-RoBERTa / XLM-T	Fine-tuned for sentiment on multilingual tweets rather than Czech mobile app reviews [47].
Generative LLM	GPT-5.5 / GPT-5.6 Sol	ChatGPT interface; official model names GPT-5.5 and GPT-5.6 Sol (API documentation identifiers gpt-5.5 and gpt-5.6-sol); medium reasoning effort; identical prompt; classification based solely on review text; complete input file supplied in each run with an instruction for internal groups of no more than 100 records [48, 49].
Source: Authors' analysis		

2.5 Standardized LLM Procedure

The LLM procedure used a single English-language zero-shot prompt throughout the study (Appendix A). It specified the three sentiment categories and decision rules but provided no labeled examples. Before validation, this prompt was tried once solely to check whether the model could return one sentiment label for each of the 500 reviews in the supplied file. This was an output-completeness check, not a test of classification accuracy or a prompt-selection exercise; no alternative wording was tested. The same prompt was then used unchanged for the five reported validation runs per model and the subsequent full-corpus application. The human annotators received the same category definitions and decision rules. In each LLM run, the researchers supplied one complete input file. The prompt asked the model to process no more than 100 reviews at a time as an internal workflow instruction; the researchers did not split the file into batches or control how the model handled that instruction internally.

The LLM validation classifications were performed on 23 July 2026. GPT-5.5 and GPT-5.6 Sol were each evaluated in five runs, for ten reported validation runs in total. Each run took place in a separate, context-

free ChatGPT conversation. The unchanged prompt and one Excel file containing all 500 validation reviews were supplied at its start. The file contained review text and identifiers needed to preserve record order, but no star ratings, application names, developer responses, or other metadata. The prompt requested internal processing in groups of no more than 100 reviews and one consolidated output file. Whether or how the model internally divided the reviews could not be observed. No reported run was interrupted or incomplete, and no output from one run was supplied in another. Every output was checked against the input for completeness, unchanged review text, identifiers, and order.

All ten planned runs were included in the analysis. The final model was selected only after both models had completed all five runs, based on their agreement with the adjudicated ground truth and their stability across runs.

The selected model, prompt, and settings were then applied to a single file containing the full corpus of 10,198 reviews, again with the instruction for internal grouping. This was the final corpus application, not a sixth validation run. Classifications of the 500 validation reviews were extracted from its output by identifier and compared descriptively with the ground truth and the five preceding GPT-5.6 Sol runs. This check assessed consistency of the full-corpus application, not out-of-sample performance.

2.6 Evaluation and Statistical Analysis

The primary classification metric was macro-F1, defined as the unweighted mean of the F1 scores for all three classes. This metric limits the dominance of the largest class when sentiment categories are unevenly represented. Cohen's kappa and overall accuracy were also calculated. Precision, recall, and F1 were reported for each class, and confusion matrices were constructed.

For annotator agreement and the agreement of each method with the ground truth, 95% stratified bootstrap confidence intervals for the main metrics were calculated from 2,000 resamples. Strata were defined by the combination of application and star rating, thereby preserving the structure of the validation sample during resampling. The stability of the five LLM runs was assessed using pairwise agreement, pairwise Cohen's kappa, macro-F1, and Fleiss' kappa.

Error patterns were described by the directions of misclassification in the confusion matrices. A brief exploratory analysis of the type of linguistic difficulty was conducted for reviews that all five GPT-5.6 Sol runs classified consistently but incorrectly. This was not a new independent annotation exercise but an interpretive supplement to the quantitative analysis.

Across the full corpus, the association between textual sentiment and the original star rating was assessed using Spearman's correlation, Pearson's chi-square test, and Cramer's V. In a sensitivity analysis, ratings of 1–2 stars were converted to negative, 3 stars to neutral, and 4–5 stars to positive; exact agreement, Cohen's kappa, and linearly weighted kappa were then calculated. Differences in model-assigned sentiment distributions across applications were evaluated using an overall Pearson chi-square test and Cramer's V. To explore heterogeneity among applications, all 15 application pairs were subsequently compared using Pearson chi-square tests on 2 x 3 contingency tables. The resulting

p-values were adjusted using the Holm method to control the family-wise error rate, and Cramer's V was calculated for each pair. Given the size of the corpus, effect sizes and the observed distributions were considered alongside statistical significance. As a post hoc exploratory misclassification sensitivity analysis, the observed full-corpus class proportions were adjusted using the class-conditional error rates observed in the embedded 500-review validation sample from the final corpus run. The latent class distribution was estimated as the distribution that would produce the observed model-assigned proportions under the validation error matrix. This calculation assumed that the validation error matrix applied unchanged to the full corpus.

3 Results

3.1 Manual Annotation Agreement and Stability of LLM Runs

The two annotators agreed on 90.6% of cases in their initial independent ratings. Cohen's kappa was 0.820 (95% CI [0.781, 0.860]), and macro-F1 was 0.807 (95% CI [0.761, 0.852]). Agreement varied markedly across classes: F1 was 0.948 for both negative and positive sentiment, but only 0.525 for neutral sentiment. Disagreements were therefore concentrated primarily around the definition of the neutral category. The disputed cases were subsequently reviewed by a third annotator, and adjudication produced the final ground truth used to evaluate the classification methods in Table 3.

Stability across repeated runs differed substantially between the two GPT versions. For GPT-5.5, the mean pairwise agreement across five validation runs was 78.4%, with individual pairwise comparisons ranging from 66.6% to 95.8%. Fleiss' kappa was 0.621, and mean pairwise macro-F1 was 0.720, with a wide range of 0.572–0.913. Outputs from GPT-5.5 were therefore not highly stable across repeated runs. By contrast, GPT-5.6 Sol achieved a mean pairwise agreement of 96.0%, with a substantially narrower range of 94.6%–98.2%, Fleiss' kappa of 0.918, and a mean pairwise macro-F1 of 0.911 within a range of 0.879–0.958. The GPT-5.6 Sol final corpus run also remained in close agreement with the five validation runs: mean agreement was 95.6%, Cohen's kappa was 0.911, and macro-F1 was 0.903. These results indicate substantially greater between-run stability for the newer GPT-5.6 Sol model.

Table 2
Reliability of Manual Annotation and Stability of the LLM

Comparison	Scope	Agreement	Kappa	Macro-F1
Manual annotators	500 reviews	0.906	0.820 [0.781, 0.860]	0.807 [0.761, 0.852]
GPT-5.5 validation runs 1–5	10 pairs	0.784 (0.666–0.958)	0.621 (Fleiss)	0.720 (0.572–0.913)
GPT-5.6 Sol validation runs 1–5	10 pairs	0.960 (0.946–0.982)	0.918 (Fleiss)	0.911 (0.879–0.958)
GPT-5.6 Sol final vs. validation runs	5 pairs	0.956 (0.948–0.964)	0.911 (0.893–0.928)	0.903 (0.887–0.918)
<p>Note: Square brackets indicate 95% stratified bootstrap confidence intervals for the manual-annotation statistics, calculated from 2,000 resamples. For the GPT comparisons, values reported before parentheses are means across the pairwise comparisons specified in the Scope column, and parentheses indicate the observed minimum-maximum range across those comparisons. Fleiss' kappa was calculated jointly across all five runs of each model and is therefore reported as a single coefficient.</p>				
Source: Authors' analysis				

3.2 Performance of the Classification Methods

Table 3
Performance of the Methods Against the Adjudicated Ground Truth

Method	Accuracy	Macro-F1	Cohen's kappa	Neutral-class F1
Majority-class baseline	0.658	0.265	0.000	0.000
NRC	0.308	0.282 [0.245, 0.317]	0.038 [-0.005, 0.082]	0.119
Subjectivity	0.298	0.281 [0.243, 0.316]	0.048 [0.002, 0.089]	0.129
BERT	0.782	0.621 [0.575, 0.670]	0.568 [0.504, 0.630]	0.247
XLM-RoBERTa	0.752	0.627 [0.583, 0.672]	0.548 [0.488, 0.607]	0.299
GPT-5.5 validation runs 1–5	0.823 (0.666–0.932)	0.725 (0.574–0.838)	0.684 (0.446–0.860)	0.438 (0.253–0.603)
GPT-5.6 Sol validation runs 1–5	0.930 (0.924–0.938)	0.827 (0.814–0.845)	0.857 (0.844–0.874)	0.563 (0.521–0.597)
GPT-5.6 Sol final corpus run	0.946	0.871 [0.824, 0.912]	0.892 [0.853, 0.926]	0.675
<p><i>Note: Square brackets indicate 95% stratified bootstrap confidence intervals based on 2,000 resamples. For rows summarizing the five repeated validation runs, the value reported before parentheses is the mean across the five runs, and parentheses indicate the observed minimum-maximum range across those runs. These ranges are descriptive measures of between-run variability and are not confidence intervals.</i></p>				
Source: Authors' analysis				

The majority-class baseline, which assigned every review to the negative class, achieved an accuracy of 0.658, a macro-F1 score of 0.265, Cohen's kappa of 0.000, and a neutral-class F1 score of 0.000. Its accuracy exceeded that of both lexicon-based methods because the ground truth was class-imbalanced, whereas its lower macro-F1 reflected its complete failure to identify the two minority classes. The lexicon-based methods achieved nearly identical macro-F1 scores (NRC, 0.282; Subjectivity, 0.281). Both transformer checkpoints achieved markedly higher macro-F1 scores than the lexicon-based methods (BERT, 0.621; XLM-RoBERTa, 0.627), and their point estimates were similar. Because no paired inferential comparison of the two checkpoints was conducted, the study does not support a conclusion that either checkpoint performed better than the other. Across the five GPT-5.6 Sol validation runs, macro-F1 ranged from 0.814 to 0.845 and Cohen's kappa from 0.844 to 0.874. The older GPT-5.5 model exhibited a markedly wider range across all reported indicators, confirming substantial instability across runs and suggesting dependence on the model version used. The final corpus run achieved a macro-F1 score of 0.871, Cohen's kappa of 0.892, and an overall accuracy of 94.6%.

Lexicon coverage was limited. NRC matched no lexicon entry in 165 of the 500 reviews (33.0%), and the Subjectivity Lexicon matched no entry in 138 reviews (27.6%); the median number of matched tokens per review was one for both methods. Because a total score of zero was assigned to the neutral class, at least 165 of the 199 neutral predictions produced by NRC (82.9%) and 138 of the 212 neutral predictions produced by the Subjectivity Lexicon (65.1%) resulted from the absence of any matched lexicon entry rather than from balanced positive and negative evidence.

Within the embedded 500-review validation subset, the GPT-5.6 Sol final corpus run achieved an accuracy of 94.6%, a macro-F1 score of 0.871, and Cohen's kappa of 0.892 against the adjudicated ground truth. Considered together with the between-run stability reported in Table 2, these findings support the deployment consistency of the full-corpus application but do not constitute an independent estimate of out-of-sample performance.

Neutral sentiment was the most difficult class for every procedure. Its F1 score was 0.119 for NRC, 0.129 for Subjectivity, 0.247 for BERT, and 0.299 for XLM-RoBERTa. Across the five GPT-5.6 Sol validation runs, neutral-class F1 ranged from 0.521 to 0.597, reaching 0.675 in the final run. In the final run, neutral-class precision was 0.609 and recall was 0.757. Of the 46 reviews classified as neutral, 18 (39.1%) were categorized as negative or positive in the adjudicated ground truth. Conversely, 9 of the 37 reviews adjudicated as neutral (24.3%) were classified as polarized. No direct confusions occurred between the positive and negative classes. The corresponding negative- and positive-class F1 scores were 0.972 and 0.966, respectively. The complete confusion matrix and class-specific performance measures are reported in Appendix B. These results must be interpreted in light of the small number of neutral reviews in the ground truth ($n = 37$). GPT-5.6 Sol achieved the highest observed performance among the directly deployable methods tested in this corpus

3.3 Error Patterns and LLM Stability

Table 4
Summary of Error Patterns and Stability in LLM Classification

Indicator	GPT-5.6 Sol validation runs 1–5	GPT-5.6 Sol final run
Errors between the neutral and negative classes	23.4 (19–26); 52.8%-77.4% of errors	18 of 27 (66.7%)
Errors between the neutral and positive classes	27.0% of all errors	9 of 27 (33.3%)
Direct positive-negative confusion	5.7% of all errors	0 cases
Error rate for reviews adjudicated as neutral	42.7% (35.1%-48.6%)	24.3%
All five runs identically incorrect	17 of 500 (3.4%)	-
<i>Note: The five validation runs classified the same reviews; the aggregate error counts therefore do not represent 2,500 independent observations.</i>		
Source: Authors' analysis		

Across the validation runs, 94.3% of disagreements occurred between the neutral class and one of the polarized classes; confusions between neutral and negative sentiment were the most frequent, accounting for 67.2% of errors. The same pattern appeared in the final run: all 27 disagreements involved the neutral category, and there were no direct confusions between positive and negative sentiment.

The five validation runs produced identical classifications for 456 reviews (91.2%), but they also classified 17 reviews identically and incorrectly. Of these, eight were factual questions, requests, or technical descriptions expressing only a weak attitude; six described mixed or updated experiences; and three involved indirect or context-dependent statements. This exploratory analysis clarifies that high stability does not imply error-free classification and that the boundary of neutral sentiment remains the principal difficulty.

3.4 Association Between LLM-Derived Textual Sentiment and Star Ratings

Table 5
Statistical Comparison of Textual Sentiment and Star Ratings (n = 10,198)

Analysis	Statistic	Result	Interpretation
Ordinal association	Spearman's ρ	0.847	strong positive association
Test of independence	$\chi^2(8)$	9,348.54; $p < .001$	statistically significant association
Association strength	Cramer's V	0.677	strong association
Sensitivity analysis	agreement; κ ; weighted κ	87.4%; 0.748; 0.826	high to very high agreement
<i>Note: The sensitivity analysis converts 1–2 stars to negative, 3 stars to neutral, and 4–5 stars to positive. Star ratings were not provided to the model during classification and are not treated as the ground truth for textual sentiment.</i>			
Source: Authors' analysis			

Star ratings were strongly associated with textual sentiment. Exact agreement after conversion to three categories was 87.4%, but a direct polarity conflict - positive text paired with a 1- or 2-star rating, or negative text paired with a 4- or 5-star rating - occurred in 242 reviews (2.4%). This finding supports the convergent validity of the LLM classification but not the correctness of every individual label.

Source: Authors' analysis

The association was nearly unequivocal at the extremes: negative sentiment accounted for 96.8% of 1-star reviews, and positive sentiment accounted for 91.4% of 5-star reviews. By contrast, 3-star reviews were substantively heterogeneous: 52.4% were negative, 33.6% neutral, and 14.1% positive. Mechanically equating 3 stars with neutral sentiment would therefore discard a substantial share of the information contained in the text.

3.5 Sentiment in Reviews of Government Mobile Applications

Table 6
Sentiment in Reviews of Government Mobile Applications

Application	Reviews (n)	Negative n (%)	Neutral n (%)	Positive n (%)
eDoklady	2,635	1,641 (62.3%)	171 (6.5%)	823 (31.2%)
EZKarta	4,513	3,020 (66.9%)	456 (10.1%)	1,037 (23.0%)
eRecept	598	478 (79.9%)	34 (5.7%)	86 (14.4%)
JENDA	197	166 (84.3%)	12 (6.1%)	19 (9.6%)
Klíč eGovernmentu	1,397	784 (56.1%)	72 (5.2%)	541 (38.7%)
Portál občana	858	636 (74.1%)	60 (7.0%)	162 (18.9%)
Total	10,198	6,725 (65.9%)	805 (7.9%)	2,668 (26.2%)
Test of differences	$\chi^2(10) = 313.47$	$p < .001$	$V = 0.124$	modest overall association
<i>Note: The table includes text-bearing reviews available through Google Play up to the collection date. Sentiment was determined by the final corpus run using GPT-5.6 Sol.</i>				
Source: Authors' analysis				

Negative sentiment predominated among the model-assigned classifications in the full corpus: 6,725 reviews (65.9%) were classified as negative, 2,668 (26.2%) as positive, and 805 (7.9%) as neutral. Klíč eGovernmentu had the lowest model-assigned proportion of negative reviews and the highest proportion of positive reviews. JENDA had the highest model-assigned proportion of negative reviews, although this result was based on the smallest number of reviews, followed by eRecept. These findings describe the distribution of labels assigned to published reviews and do not constitute a ranking of the applications' overall quality.

The proportions in Table 6 represent model-assigned sentiment categories rather than error-free estimates of underlying sentiment prevalence. In the exploratory misclassification sensitivity analysis, application of the class-conditional error rates observed in the embedded validation sample yielded an estimated distribution of 67.9% negative, 5.6% neutral, and 26.5% positive, compared with the observed model-assigned proportions of 65.9%, 7.9%, and 26.2%, respectively. The substantive conclusion that negative sentiment predominated remained unchanged, although the estimated proportion of neutral reviews was sensitive to classification error. These adjusted proportions are exploratory because they assume that the validation error matrix applied unchanged to the full corpus.

Model-assigned sentiment distributions differed statistically across applications, $\chi^2(10) = 313.47$, $p < .001$. The global association was modest (Cramer's $V = 0.124$); however, this summary coefficient does not imply that every application-specific difference was small. For example, the proportion of reviews classified as negative ranged from 56.1% for Klíč eGovernmentu to 84.3% for JENDA. Because the applications differed in their numbers and composition of published reviews, these results describe

model-assigned sentiment distributions and do not justify a ranking of application quality. Table B3 in Appendix B reports descriptive classification performance by application for the 500 reviews in the validation subset extracted from the final corpus run. Because the application-specific groups were small ($n = 10\text{--}221$) and several contained few or no neutral reference cases, the observed differences in accuracy do not establish that the model is more reliable for some applications than for others.

Holm-adjusted pairwise comparisons indicated statistically detectable differences in model-assigned sentiment distributions for 13 of the 15 application pairs (Appendix B, Table B4). No statistically significant difference was observed between eRecept and JENDA (adjusted $p = .234$) or between eRecept and Portál občana (adjusted $p = .071$). Pairwise Cramer's V values ranged from 0.058 to 0.242, indicating that the magnitude of the association varied across application pairs. These pairwise comparisons concern model-assigned label distributions and do not establish differences in overall application quality.

4 Discussion

On the validation sample, the standardized GPT-5.6 Sol procedure achieved the highest agreement with the human ground truth among all methods tested and demonstrated high stability across repeated runs. This finding is consistent with multilingual evidence suggesting that prompted LLMs can achieve competitive sentiment classification without new task-specific fine-tuning [30, 31]. At the same time, it respects the limitations identified in that literature: performance varies by language, model, prompt, and task type [32, 29, 33]. The study therefore does not support the universal superiority of LLMs, but rather their use as a validated practical procedure when a suitable off-the-shelf domain classifier is unavailable.

The comparison of model versions reveals the most important methodological nuance. With the same prompt and declared settings, five GPT-5.5 runs exhibited a substantially wider range than five GPT-5.6 Sol runs: macro-F1 ranged from 0.574 to 0.838 versus 0.814 to 0.845, and Cohen's kappa ranged from 0.446 to 0.860 versus 0.844 to 0.874. The labels 'LLM' and 'ChatGPT' are therefore insufficient descriptions of a method; a reproducible procedure must identify the specific model version, prompt, settings, and date of use. Because the system is proprietary, the observed difference cannot be attributed causally to model version alone. Nevertheless, the result provides empirical support for the conclusion that performance and stability are properties of a specific model configuration rather than of the entire category of LLMs [29, 33, 37].

The weaker performance of the lexicon-based baselines cannot be generalized to all lexicon-based approaches. The majority-class baseline exceeded both lexicons in accuracy because of the predominance of negative cases, whereas the lexicons exceeded it only slightly in macro-F1. The Czech version of NRC used here was produced through automated translation, and both lexicons were applied using a simple aggregation rule without new domain adaptation. Their coverage was limited, with no matched entry for 33.0% of reviews using NRC and 27.6% using the Subjectivity Lexicon. Moreover, reviews with no matched entry and reviews with balanced positive and negative scores were both

assigned to the neutral class. The resulting neutral predictions therefore combined an absence of lexical evidence with balanced polarity evidence. Studies in other languages show that human review of translations, semantic transfer, lemmatization, and domain coverage can substantially alter performance [22, 19, 20]. The transformer baselines were similarly affected by a mismatch between the training domain and target classes: BERT mapped a five-star output to three categories, while XLM-T had been fine-tuned for sentiment on tweets in other languages. Linguistically and domain-adapted encoders can achieve substantially better results [13, 14, 26]. The ranking in this study therefore concerns the tested off-the-shelf implementations for government mobile applications and demonstrates the practical benefit of a task-specified LLM procedure under the study conditions, not the general superiority of one architecture over all possible alternatives.

The neutral class was the main source of uncertainty for both annotators and automated methods. The neutral category combined several substantively different types of text, including genuinely non-evaluative statements, balanced mixed evaluations, and cases in which a predominant polarity could not be assigned reliably. This heterogeneity may have contributed to the lower inter-annotator agreement and weaker classification performance observed for the neutral class. Because these subtypes were not annotated separately, their individual contributions to disagreement cannot be determined from the present data. This pattern is consistent with the meta-analytic decline in performance observed when moving from two to three classes [27]. Koto et al. [28] highlight the difficulty of defining the boundary of the neutral class in zero-shot classification, while Kastrati et al. [29] documented its weakest performance across six language families. The findings also show that repeatability cannot substitute for validation of correctness: all five runs made the same error on 17 reviews. Before classifying an entire new corpus, researchers should therefore construct at least a representative, manually annotated validation subsample, ideally using independent annotation and adjudication of disagreements. This subsample need not be used to train the model; its purpose is to evaluate a specific model version and prompt against human interpretations of the text, estimate class-specific errors, and determine whether the procedure is acceptable for the corpus concerned [35, 36].

The strong association between textual sentiment and star ratings provides supplementary support for the convergent validity of the selected procedure. At the same time, the heterogeneity of 3-star reviews illustrates why ratings cannot be treated mechanically as the ground truth for textual sentiment. Star ratings may express an overall assessment of an application, whereas the text may capture a specific problem, a change following an update, or a mixed experience.

The results suggest that reviews of government mobile applications can provide public administrations with a continuing and relatively low-cost supplementary source of citizen feedback. They are not a substitute for representative surveys or a direct measure of objective service quality, but rather a supplementary signal of problems that users themselves consider important. In a dataset of more than 100,000 reviews of Australian government applications, Aamir et al. [10], for example, identified problems involving stability, complexity, navigation, integration, and identity verification and described reviews as a useful and cost-effective source of continuous feedback.

Automated review analysis could therefore be incorporated into quality management for digital public services, particularly to monitor recurring dissatisfaction and prioritize areas requiring closer investigation. Sentiment alone, however, does not explain the cause of a problem and should be interpreted together with the textual content of reviews and other operational data, such as information about updates, error states, help-desk requests, or usability-testing results.

LLMs can facilitate the classification of large volumes of short texts in this process, but they should not operate as an uncontrolled substitute for human judgment. The results of this study demonstrate the importance of validation against human annotation and checks of the stability of a specific model version. Reproducible use requires documentation of the model, its version, the prompt, settings, classification date, batching procedure, and output checks. In public administration, requirements relating to transparency, accountability, data protection, risk management, and the preservation of public trust must also be considered [50, 51, 36, 37].

The predominance of negative reviews therefore does not reflect the satisfaction of all citizens and does not, in itself, measure the objective quality of the applications. Reviews are self-selected, and users with particularly strong positive or negative experiences may be more likely to post them, limiting the representativeness of the resulting distribution [6]. The applications also differ in how long they have been available and in the number of reviews collected. The findings thus constitute an analysis of published user feedback rather than a representative population survey.

Future research should validate the procedure on an independent dataset, in additional non-English languages, and against models fine-tuned directly on domain-matched reviews. A coordinated multilingual design using the same classes, annotation rules, metrics, and repeated-run protocol would be particularly valuable for separating the effects of language from those of domain and data. Substantive analysis could also be extended to aspect-based classification and longitudinal monitoring of updates, outages, or institutional changes.

Conclusion

This study provides empirical and methodological evidence on a documented procedure for three-class sentiment classification of Czech government mobile app reviews without task-specific fine-tuning. Its contribution lies in evaluating agreement with adjudicated human labels, comparing selected readily deployable baselines, and examining repeatability and classification errors in a specific public-service corpus.

On the stratified sample of 500 reviews, GPT-5.6 Sol achieved higher observed agreement with the adjudicated labels than the tested lexicon-based and encoder-based baselines. Across five repeated validation runs, mean accuracy was 93.0%, mean macro-F1 was 0.827, and mean Cohen's kappa was 0.857. Between-run agreement was high, with Fleiss' kappa of 0.918. The subsequent full-corpus application showed close agreement with the preceding classifications of the embedded validation reviews; this comparison was a consistency check, not an independent test of generalization.

In this corpus, disagreements were concentrated at the boundaries of neutral sentiment. As specified in the annotation rules, the neutral category combined non-evaluative content, mixed evaluations, and cases of uncertain polarity; these subtypes were not assessed separately. With only 37 reviews adjudicated as neutral, estimates for this class remain uncertain. Application of the selected procedure to 10,198 reviews yielded model-assigned proportions of 65.9% negative, 7.9% neutral, and 26.2% positive sentiment. These distributions describe published feedback and do not measure population-wide satisfaction or establish differences in application quality.

The same validation sample was used to select the model and estimate its performance, and the small application-specific subsamples do not establish equivalent classification accuracy across applications. The findings therefore support the potential utility of this configuration under the conditions examined, while its performance on new data remains untested. Further research should prioritize independent validation and comparison with stronger Czech or domain-adapted classifiers, followed by aspect-based and longitudinal analyses. Practical use requires corpus-specific human validation, transparent documentation of the execution workflow, and explicit treatment of classification uncertainty.

Declarations

Author Contribution

Both authors contributed to the conception and design of the study. Lukáš Johanovský carried out data collection, data preparation, analysis, and drafting of the manuscript. Pavla Vrabcová contributed to the methodology, supervision, interpretation of the results, and critical revision of the manuscript. Both authors read and approved the final manuscript.

Data Availability

The reviews examined in this study were publicly accessible on Google Play at the time of collection. Supplementary material 1 documents the collection procedure; however, a new download may not reproduce the dataset collected on 30 April 2026. The original exports included reviewer usernames and other review metadata, but usernames were not used in the reported analyses. The authors do not redistribute the original downloaded files. A cleaned analytical extract, including the adjudicated labels, can be made available from the corresponding author on reasonable request after removal of direct identifiers and screening of potentially identifying review text, subject to applicable platform terms and privacy safeguards.

References

1. Heeks R, Bailur S (2007) Analyzing e-government research: Perspectives, philosophies, theories, methods, and practice. *Government Inform Q* 24(2):243–265.
<https://doi.org/10.1016/j.giq.2006.06.005>

2. Sharma SK, Al-Badi A, Rana NP, Al-Azizi L (2018) Mobile applications in government services (mG-App) from user's perspectives: A predictive modelling approach. *Government Inform Q* 35(4):557–568. <https://doi.org/10.1016/j.giq.2018.07.002>
3. Desmal AJ, Hamid S, Othman MK, Zolait A (2022) A user satisfaction model for mobile government services: A literature review. *PeerJ Comput Sci* 8:e1074. <https://doi.org/10.7717/peerj-cs.1074>
4. Hu N, Zhang J, Pavlou PA (2009) Overcoming the J-shaped distribution of product reviews. *Commun ACM* 52(10):144–147. <https://doi.org/10.1145/1562764.1562800>
5. de Langhe B, Fernbach PM, Lichtenstein DR (2016) Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *J Consum Res* 42(6):817–833. <https://doi.org/10.1093/jcr/ucv047>
6. Schoenmueller V, Netzer O, Stahl F (2020) The Polarity of Online Reviews: Prevalence, Drivers and Implications. *J Mark Res* 57(5):853–877. <https://doi.org/10.1177/0022243720941832>
7. Dąbrowski J, Letier E, Perini A, Susi A (2022) Analysing app reviews for software engineering: A systematic literature review. *Empir Softw Eng* 27(2):43. <https://doi.org/10.1007/s10664-021-10065-7>
8. Alqaryouti O, Siyam N, Abdel Monem A, Shaalan K (2024) Aspect-based sentiment analysis using smart government review data. *Appl Comput Inf* 20(1/2):142–161. <https://doi.org/10.1016/j.aci.2019.11.003>
9. Hadwan M, Al-Hagery MA, Al-Sarem M, Saeed F, *Computers* (2022) *Mater Continua* 72(3):4675–4689. <https://doi.org/10.32604/cmc.2022.027311>
10. Aamir T, Chhetri MB, Chamikara MAP, Grobler M (2023) Government Mobile Apps: Analysing Citizen Feedback via App Reviews. In: 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp 1858–1863. <https://doi.org/10.1109/ASE56229.2023.00190>
11. Wang X, Zhang T, Tan Y, Shang W, Li Y (2024) How to effectively mine app reviews concerning software ecosystem? A survey of review characteristics. *J Syst Softw* 213:112040. <https://doi.org/10.1016/j.jss.2024.112040>
12. Xu Y, Cao H, Du W, Wang W (2022) A Survey of Cross-lingual Sentiment Analysis: Methodologies, Models and Evaluations. *Data Sci Eng* 7(3):279–299. <https://doi.org/10.1007/s41019-022-00187-3>
13. Baert G, Gahbiche S, Gadek G, Pauchet A (2020) Arabizi Language Models for Sentiment Analysis. In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp 592–603. <https://doi.org/10.18653/v1/2020.coling-main.51>
14. Catelli R, Bevilacqua L, Mariniello N, Di Scotto V, Magaldi M, Fujita H, De Pietro G, Esposito M (2022) Cross lingual transfer learning for sentiment analysis of Italian TripAdvisor reviews. *Expert Syst Appl* 209:118246. <https://doi.org/10.1016/j.eswa.2022.118246>
15. Nugroho KS, Sukmadewa AY, Wuswilahaken Dw H, Bachtiar FA, Yudistira N (2021) BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews. In: *6th International Conference on Sustainable Information Engineering and Technology 2021*, pp 258–264. <https://doi.org/10.1145/3479645.3479679>

16. Bojar O (2005) Problems of inducing large coverage constraint-based dependency grammar for Czech. In: Christiansen H, Skadhaug PR, Villadsen J (eds) *Constraint Solving and Language Processing*. Lecture Notes in Computer Science, vol 3438. Springer, Berlin, pp 90–103. https://doi.org/10.1007/11424574_6
17. Habernal I, Ptáček T, Steinberger J (2014) Supervised sentiment analysis in Czech social media. *Inf Process Manag* 50(5):693–707. <https://doi.org/10.1016/j.ipm.2014.05.001>
18. Cruz FL, Troyano JA, Pontes B, Ortega FJ (2014) Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Syst Appl* 41(13):5984–5994. <https://doi.org/10.1016/j.eswa.2014.04.005>
19. Abdaoui A, Azé J, Bringay S, Poncelet P (2017) FEEL: A French Expanded Emotion Lexicon. *Lang Resour Evaluation* 51(3):833–855. <https://doi.org/10.1007/s10579-016-9364-5>
20. Fehle J, Schmidt T, Wolff C (2021) Lexicon-based sentiment analysis in German: Systematic evaluation of resources and preprocessing techniques. In: *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pp 86–103. <https://doi.org/10.5283/EPUB.50833>
21. Veselovská K, Hajic J, Šindlerová J (2014) Subjectivity Lexicon for Czech: Implementation and Improvements. *J Lang Technol Comput Linguistics* 29(1):47–61. <https://doi.org/10.21248/jlcl.29.2014.183>
22. Hogenboom A, Heerschop B, Frasincar F, Kaymak U, De Jong F (2014) Multi-lingual support for lexicon-based sentiment analysis guided by semantics. *Decis Support Syst* 62:43–53. <https://doi.org/10.1016/j.dss.2014.03.004>
23. Veselovská K, Hajič J Jr (2013) Why words alone are not enough: Error analysis of lexicon-based polarity classifier for Czech. In: *Proceedings of the 3rd Workshop on Sentiment Analysis where AI meets Psychology*, pp 1–5. Asian Federation of Natural Language Processing. <https://aclanthology.org/W13-4101/>
24. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
25. Facebook AI (2026) Research (n.d.) RoBERTa: A Robustly Optimized BERT Pretraining Approach. fairseq documentation. GitHub. Accessed 24 Sep <https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md>
26. Negru V-A, Suci V, Lăpușan A-M, Lemnaru C, Dîșoreanu M, Potolea R (2025) Assessing language models' task and language transfer capabilities for sentiment analysis in dialog data. *Comput Speech Lang* 89:101704. <https://doi.org/10.1016/j.csl.2024.101704>
27. Hartmann J, Heitmann M, Siebert C, Schamp C (2023) More than a Feeling: Accuracy and Application of Sentiment Analysis. *Int J Res Mark* 40(1):75–87. <https://doi.org/10.1016/j.ijresmar.2022.05.005>

28. Koto F, Beck T, Talat Z, Gurevych I, Baldwin T (2024) Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp 298–320. <https://doi.org/10.18653/v1/2024.eacl-long.18>
29. Kastrati M, Manaf A, Imran AS, Kastrati Z, Daudpota SM, Biba M (2026) Do language families matter? Evaluating LLMs for sentiment analysis through a hierarchical cross-lingual lens. *Front Artif Intell* 9:1854873. <https://doi.org/10.3389/frai.2026.1854873>
30. Rathje S, Mirea D-M, Sucholutsky I, Marjeh R, Robertson CE, Van Bavel JJ (2024) GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences* 121(34):e2308950121. <https://doi.org/10.1073/pnas.2308950121>
31. Přibáň P, Šmíd J, Steinberger J, Mištera A (2024) A comparative study of cross-lingual sentiment analysis. *Expert Syst Appl* 247:123247. <https://doi.org/10.1016/j.eswa.2024.123247>
32. Zhang W, Deng Y, Liu B, Pan S, Bing L (2024) Sentiment Analysis in the Era of Large Language Models: A Reality Check. In: Findings of the Association for Computational Linguistics: NAACL 2024, pp 3881–3906. <https://doi.org/10.18653/v1/2024.findings-naacl.246>
33. Lima JVRJ, Pinheiro V, Caminha C (2026) Portuguese sentiment analysis with open-source LLMs: Models, prompts, and efficient deployment. In: Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) – Vol. 1, pp 212–221. Association for Computational Linguistics. <https://aclanthology.org/2026.propor-1.21/>
34. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D (2024) Can Large Language Models Transform Computational Social Science? *Comput Linguistics* 50(1):237–291. https://doi.org/10.1162/coli_a_00502
35. Grimmer J, Stewart BM (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Anal* 21(3):267–297. <https://doi.org/10.1093/pan/mps028>
36. Klie J-C, Castilho RED, Gurevych I (2024) Analyzing Dataset Annotation Quality Management in the Wild. *Comput Linguistics* 50(3):817–866. https://doi.org/10.1162/coli_a_00516
37. Ollion É, Shen R, Macanovic A, Chatelain A (2024) The dangers of using proprietary LLMs for research. *Nat Mach Intell* 6(1):4–5. <https://doi.org/10.1038/s42256-023-00783-6>
38. Xu N, Zhang W (2025) User satisfaction with Chinese government apps: Topic mining and sentiment analysis of user reviews. *Lex Localis – Journal of Local Self-Government* 23(3):95–124. [https://doi.org/10.52152/23.3.95-124\(2025\)](https://doi.org/10.52152/23.3.95-124(2025))
39. Krugmann JO, Hartmann J (2024) Sentiment Analysis in the Age of Generative AI. *Customer Needs Solutions* 11(1):3. <https://doi.org/10.1007/s40547-024-00143-4>
40. Wu C, Ma B, Zhang Z, Deng N, He Y, Xue Y (2025) Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models. *Int J Mach Learn Cybernet* 16(10):8079–8101. <https://doi.org/10.1007/s13042-025-02711-z>
41. Přibáň P, Steinberger J (2021) Are the multilingual models better? Improving Czech sentiment with transformers. In: Proceedings of the International Conference on Recent Advances in Natural

- Language Processing (RANLP 2021), pp 1138–1149. https://doi.org/10.26615/978-954-452-072-4_128
42. Šmíd J, Příbáň P, Král P (2026) Large language models for Czech aspect-based sentiment analysis. In: Ekštejn K, Konopík M, Pražák O, Pártl F (eds) Text, Speech, and Dialogue. Lecture Notes in Computer Science, vol 16030. Springer, Cham, pp 15–26. https://doi.org/10.1007/978-3-032-02551-7_3
43. Šmíd J, Příbáň P, Král P (2026) Extending Czech aspect-based sentiment analysis with opinion terms: Dataset and LLM benchmarks. In: Proceedings of the Fifteenth Language Resources and Evaluation Conference, pp 7973–7984. ELRA Language Resource Association. <https://doi.org/10.63317/4hkzdnwftkz>
44. Mohammad SM, Turney PD (2013) NRC emotion lexicon. Technical report, National Research Council Canada. <https://doi.org/10.4224/21270984>
45. Straka M, Straková J (2017) Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp 88–99. <https://doi.org/10.18653/v1/K17-3009>
46. NLP Town (2023) bert-base-multilingual-uncased-sentiment (version edd66ab) [machine learning model]. Hugging Face. <https://doi.org/10.57967/hf/1515>
47. Camacho-Collados J, Rezaee K, Riahi T, Ushio A, Loureiro D, Antypas D, Boisson J, Espinosa-Anke L, Liu F, Martínez-Cámara E, Medina G, Buhrmann T, Neves L, Barbieri F (2022) TweetNLP: Cutting-edge natural language processing for social media. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp 38–49. <https://doi.org/10.18653/v1/2022.emnlp-demos.5>
48. OpenAI (n.d.) GPT-5.5 model. API documentation. Accessed 13 Sep 2026. <https://developers.openai.com/api/docs/models/gpt-5.5>
49. OpenAI (n.d.) GPT-5.6 Sol model. API documentation. Accessed 13 Sep 2026. <https://developers.openai.com/api/docs/models/gpt-5.6-sol>
50. European Commission (2024) EU study calls for strategic AI adoption to transform public sector services. Shaping Europe’s digital future. <https://digital-strategy.ec.europa.eu/en/library/eu-study-calls-strategic-ai-adoption-transform-public-sector-services>
51. Vatamanu AF, Tofan M (2025) Integrating artificial intelligence into public administration: Challenges and vulnerabilities. Administrative Sci 15(4):149. <https://doi.org/10.3390/admsci15040149>

Figures

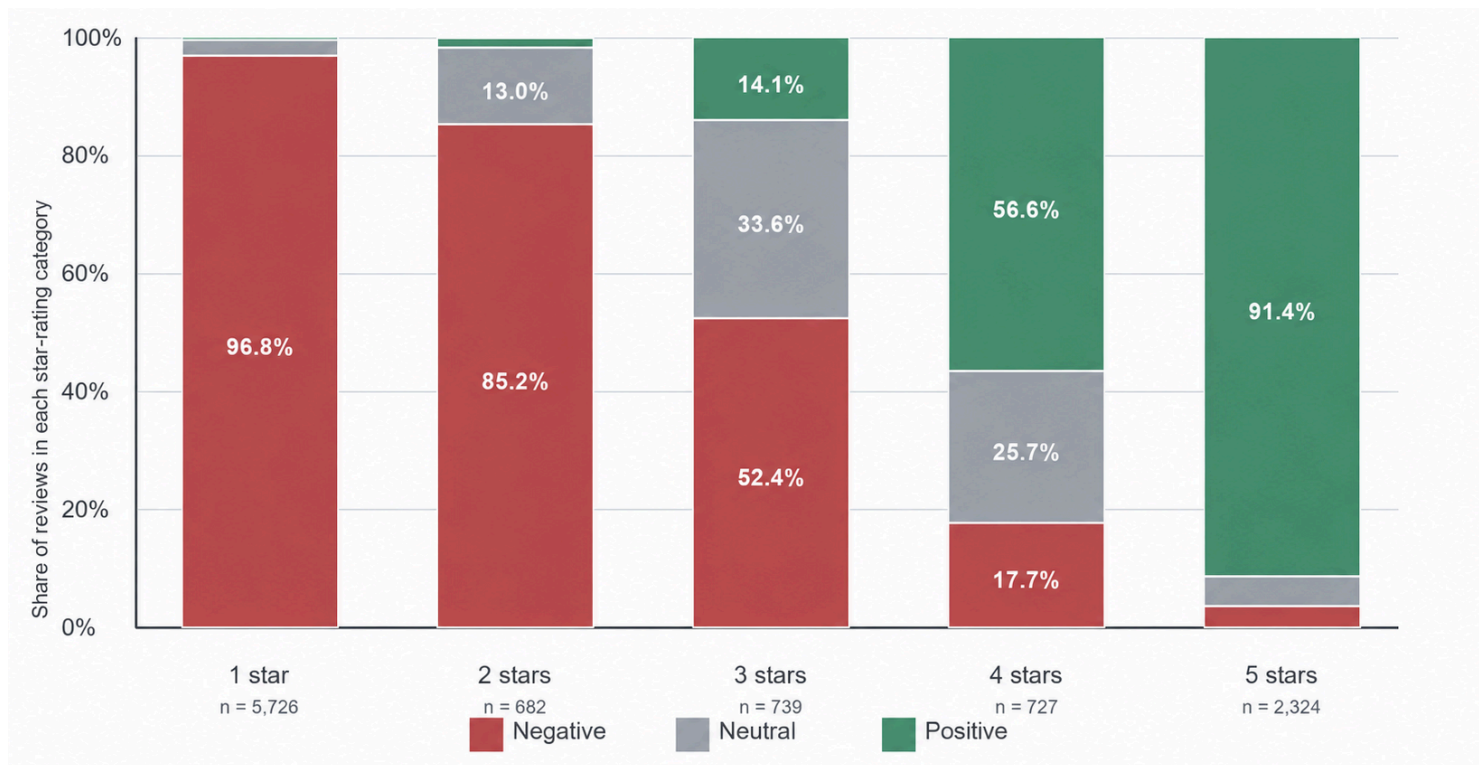


Figure 1

Distribution of Textual Sentiment Within Star-Rating Categories

Source: Authors' analysis

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial1.ipynb](#)
- [Appendices.docx](#)