Benchmarking LLMs for Aspect-Based Sentiment Classification in Slovene Historical Periodicals

Tina Munda[2], **Filip Dobranić**[1], **Uroš Šmajdek**[3], **Oliver Pejić**[1], **Ciril Bohak**[3], **Vojko Gorjanc**[1,2], **Darja Fišer**[1]

[1] Institute of Contemporary History
Privoz 11, SI-1000 Ljubljana
{filip.dobranic, oliver.pejic, darja.fiser}@inz.si

[2] Faculty of Arts, University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
{tina.munda, vojko.gorjanc}@ff-uni-lj.si

[3] Faculty of Computer Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
{uros.smajdek, ciril.bohak}@fri-uni-lj.si

Abstract

Historical newspapers present substantial challenges for computational sentiment analysis due to OCR errors, archaic linguistic features, and the absence of domain-specific labelled training data. This paper investigates whether instruction-following LLMs can facilitate aspect-level sentiment inference under such conditions. We benchmark four instruction-following LLMs on a manually annotated sample of collective-identity mentions drawn from Slovene historical newspapers. The results provide a benchmark for targeted sentiment classification in OCR-degraded historical Slovene and offer an empirically grounded assessment of the capabilities and limitations of an instruction-tuned LLM in digital humanities research.

Keywords: historical newspapers, digital humanities, aspect-level sentiment analysis, ALSA, ABSA, LLMs, benchmark, GaMS, sPeriodika

1. Introduction

Computational approaches have long been central to Digital Humanities (DH), enabling large-scale analyses of textual corpora through rule-based methods (e.g., lexicon-driven sentiment analysis), unsupervised modelling techniques such as topic modelling, and supervised machine-learning approaches. While these approaches have expanded the analytical scope of DH research, many rely on curated linguistic resources or supervised training data, while others require extensive pre-processing and normalisation. Such requirements pose substantial challenges for analysing historically noisy corpora, where annotated datasets and normalisation tools are limited.

The emergence of instruction-following large language models (LLMs) has fundamentally altered this landscape. Unlike traditional supervised approaches that require substantial amounts of task-specific annotated data, instruction-tuned LLMs can perform complex linguistic tasks in zero-shot settings, including sentiment inference, potentially lowering the barrier for computational analysis in DH contexts. This shift is particularly relevant for historical Slovene, where labelled sentiment datasets are scarce and OCR degradation further complicates large-scale annotation and model training. In such contexts, building and maintaining domain-specific training resources is costly and often infeasible.

At the same time, the robustness of instruction-following LLMs under historically degraded conditions remains insufficiently examined. It is unclear how instruction-following models behave in OCR-extracted text material, in morphologically rich languages, and in tasks that require fine-grained attribution of sentiment to specific lexical targets rather than to entire sentences or documents. Moreover, the extent to which language-adapted models outperform widely used general-purpose instruction-tuned LLMs in such settings has not been systematically evaluated for Slovene historical data.

This study addresses these questions by benchmarking instruction-following LLMs on targeted sentiment classification in late 19th–early 20th century Slovene newspapers from the *sPeriodika* corpus (Dobranić et al., 2023). The task involves classifying sentiment toward explicitly marked collective identity mentions—both nominal and adjectival realisations—in OCR-extracted historical text.

We evaluate GaMS3-12B-Instruct, a publicly available instruction-tuned LLM based on the Gemma 3 family and continually pretrained with a focus on Slovene-language data (Vreš et al., 2024), and compare it with other publicly available general-purpose instruction-tuned models of comparable scale: Gemma-3-12B-IT (Gemma Team, 2025),

Proceedings of Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026, pages 103–113

LLaMA 3.1 (Grattafiori and Dubey et al., 2024), and DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025). To probe the effect of model size within two model families, we additionally include the smaller Gemma-3-4B-IT (Gemma Team, 2025) and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI, 2025) variants. We restrict the comparison to publicly available downloadable models that can be run and versioned in a controlled, reproducible evaluation setup. Each model is assessed on a manually annotated sample of 371 mentions using standard classification metrics. Beyond overall performance, we examine variation across grammatical realisation and referential type, identifying systematic asymmetries that affect dataset-scale inference.

Finally, we investigate whether the best-performing model can be applied at dataset scale by aggregating mention-level predictions across newspapers. This aggregation functions as a methodological demonstration of scalability rather than as a substantive historical interpretation.

Taken together, this study contributes (i) a controlled benchmark for aspect-level sentiment classification in OCR-degraded historical Slovene text; (ii) a systematic comparison of aspect-level sentiment classification in a few-shot setting between a Slovene-adapted instruction-tuned LLM and general-purpose instruction-tuned LLMs of comparable scale; and (iii) a diagnostic analysis of performance variation across grammatical realisation and referential type, highlighting systematic asymmetries that must be considered in dataset-scale sentiment aggregation. By situating LLM evaluation within a historically degraded and morphologically complex dataset, the paper provides empirical evidence on the reliability and limitations of instruction-tuned models as analytical instruments in digital humanities research.

2. Related Work

Sentiment analysis has been a longstanding research area with early approaches being lexicon-based, utilising tools like VADER (Hutto and Gilbert, 2014) and the Liu-Hu lexicon (Hu and Liu, 2004) to assign sentiment scores based on predefined sentiment lexica. While simple and interpretable, these methods were limited by lexicon coverage and struggled to capture linguistic nuances such as sarcasm and idiomatic expressions (Cambria et al., 2017). Machine learning methods, such as Support Vector Machines (Cortes and Vapnik, 1995) and Naive Bayes classifiers (Pang et al., 2002), improved sentiment classification by leveraging features like n-grams (Cavnar et al., 1994) and part-of-speech tags (Marcus et al., 1993). However, these approaches relied heavily on manual feature engineering and lacked the ability to model contextual information, which limited their performance on more complex sentiment tasks.

In recent years, language models have set new benchmarks in sentiment analysis. The introduction of transformer-based models, such as BERT (Devlin et al., 2019), has reshaped the field. BERT's bidirectional context modelling enables it to consider both preceding and succeeding words in a sentence, significantly boosting performance on sentiment classification tasks. When fine-tuned for ABSA, BERT has demonstrated remarkable improvements in identifying and classifying aspect-specific sentiment (Sun et al., 2019).

In a comprehensive experiment testing capabilities of LLMs in performing various sentiment analysis tasks, Zhang et al. (2024) highlight the strengths and limitations of LLMs. LLMs excel in simpler tasks, such as binary or trinary sentiment classification, even in zero-shot and few-shot settings, often matching or surpassing fine-tuned smaller language models. This makes them particularly effective when training resources are limited.

For Slovene, sentiment analysis research has developed largely around contemporary, non-historical text types and has produced several widely used datasets. Early work includes sentiment annotation of user-generated content in the Janes corpus, where Fišer et al. (2016) applied an SVM-based three-class classifier (POS/NEG/NEU) trained on a large manually labelled tweet collection and used it to automatically assign sentiment metadata across heterogeneous Slovene genres (tweets, forums, blogs, news comments, and Wikipedia talk pages). A major step toward target- and entity-oriented sentiment is SentiNews 1.0, a manually annotated news corpus (10,427 articles; five-point Likert scale; document/paragraph/sentence levels) and its derivative SentiCoref 1.0, which enriches 837 selected news articles with named entities, coreference chains, and target-level sentiment labels for entities in context. Žitnik et al. (2022) operationalise this setup as target-level sentiment analysis over entity-based document representations and show strong gains from BERT-based approaches over traditional feature-based methods. Recent work continues to explore ABSA pipelines for Slovene on SentiCoref, comparing lexicon-based and tree-based feature extraction with neural embedding approaches (Adhikari et al., 2024). Finally, Slovene sentiment research has also expanded into specialized domains such as finance, where recent benchmarking work evaluates LLMs for target-based financial sentiment in news (Muhammad et al., 2025). In contrast to these largely contemporary, clean-text settings, the present study targets OCR-degraded historical newspapers and frames the

task as aspect-level targeted sentiment classification without supervised fine-tuning.

3. Data

The dataset on which we evaluate instruction-following LLMs for targeted sentiment classification in historical Slovene comprises three Slovene-language newspapers: *Slovenec* [The Slovene] (1873–1945), *Slovenski narod* [The Slovene Nation] (1868–1943), and *Slovenka* [The *Slovene* Woman] (1897–1902), sourced from the *sPeriodika* collection Dobranić et al. (2023). For copyright reasons, the collection only includes issues of newspapers published until 31 December 1914.

This dataset spans distinct political orientations and readership profiles, providing a heterogeneous discursive environment for evaluating targeted sentiment classification.

The newspapers differ substantially in size (Table 1). *Slovenski narod* contains approximately 183 million tokens, *Slovenec* 137 million tokens, and *Slovenka* 1.6 million tokens, amounting to over 320 million tokens in total. This volume enables dataset-level aggregation of model predictions while preserving variation across newspapers.

All texts are derived from historical scans using OCR of heterogeneous quality. Spelling inconsistencies, segmentation errors, and recognition noise introduce ambiguity at the token and sentence level. In addition, the data reflect morphologically rich and historically evolving language use, including orthographic variation and derivational complexity, affecting collective-identity expressions. The data were thus further refined through cleaning and preprocessing before undergoing automatic linguistic annotation (Dobranić et al., 2024).

Newspaper	Number of tokens
Slovenski narod	183,294,799
Slovenec	137,506,802
Slovenka	1,632,695

Table 1: Size of newspaper data

4. Methodology

4.1. Collective-Identity Extraction

In this study, collective identity refers to nouns and adjectives derived from ethno-national, regional/provincial, or other geography-based designations. Identity mentions therefore include: (i) ethno-national denominations (e.g., Španec [Spaniard], nemški [German], Jud [Jew]), (ii) regional or provincial identities (e.g., Istrijan [Istrian], Moravka [Moravian]), and (iii) other geography-based identities (e.g., evropski [European], južnoameriški [South American]).

Collective identity mentions were extracted using a manually constructed lexicon of nationality- and identity-denoting lemmas.

For nominal references, we manually inspected *Slovenka* and compiled a list of lemma types referring to national, regional, continental, or ethnic groups. This noun-lemma inventory was then applied unchanged to *Slovenec* and *Slovenski narod* to retrieve all matching nominal mentions across newspapers.

For adjectival references, lemma candidates were first extracted automatically based on Slovene derivational suffixes (e.g., -ski, -ški, including historical orthographic variants such as -zki and -žki). In *Slovenka*, all candidate adjective lemmas were manually reviewed. In the larger corpora (*Slovenec* and *Slovenski narod*), candidates were first filtered by frequency (minimum 90 occurrences) and subsequently manually inspected. The validated adjective lists from all three newspapers were merged, deduplicated, and consolidated into a single lexicon used for dataset-wide extraction.

To enable unified analysis, adjectival lemmas were mapped to their corresponding nominal identity categories (e.g., nemški [German] → Nemci [Germans], italijanski [Italian] → Italijani [Italians]), allowing nominal and adjectival realisations to be grouped under shared identity labels during aggregation.

No distinction was made at the extraction stage between adjectival mentions modifying collective actors (e.g., German army) and those modifying non-agentive entities (e.g., Slovene bread). This referential distinction is introduced later during manual evaluation (cf. Section 4.5).

4.2. Task Definition

The task is formulated as aspect-level sentiment classification. For each extracted collective identity mention (cf. Section 4.1), the model predicts the sentiment expressed toward that specific mention within its local context.

For adjectival realisations, sentiment is annotated with respect to the collective-identity adjective, even though the nominal head it modifies remains visible in context. The model must therefore attribute sentiment specifically to the marked identity expression rather than to the broader sentence or topic.

Each instance is represented as a structured entry containing: (i) a unique mention identifier, (ii) the target identity mention explicitly marked using XML-style tags, and (iii) a context window consisting of the sentence containing the mention together with the two preceding and two following sentences.

Given this input, the model assigns one of three labels to the marked mention: positive (POS), neutral (NEU), or negative (NEG).

4.3. Evaluation Dataset Construction

To evaluate model performance, we initially sampled 400 collective identity mentions from the three newspapers. The dataset was stratified at sampling time to ensure equal representation across newspapers and grammatical realisations: each newspaper contributed an equal number of instances, with 50% nominal and 50% adjectival mentions per outlet.

Mentions were randomly sampled within these constraints. To prevent over-representation of highly frequent identities, a frequency cap was applied during sampling: if a single identity accounted for more than 15% of a newspaper-specific subset (i.e., more than 20 instances within approximately 133 samples), excess instances were replaced via random resampling within the same newspaper and mention-type category. This threshold was set heuristically, as several identities were extremely prevalent in the dataset.

During manual annotation, annotators could assign the label *unknown* in cases where sentiment could not be determined due to pre-processing errors. A total of 29 instances received this label and were excluded from evaluation. The final evaluation dataset therefore contains 371 annotated mentions, which form the basis for all reported performance metrics.

The resulting dataset[1] spans a broad range of identity categories across newspapers and grammatical forms, providing a balanced and linguistically varied benchmark for mention-level sentiment classification under historical OCR noise.

4.4. Annotation Procedure

The evaluation set was divided evenly among three annotators, all trained linguists, with each mention annotated independently by a single annotator. Annotators consulted the sentence containing the target mention and, when necessary, up to two preceding and two following sentences, using the smallest sufficient context required to determine sentiment.

In addition to sentiment labelling, annotators recorded the referential type (group vs. non-group) for adjectival mentions. This variable was later used to analyse differences in model performance across referential contexts.

Inter-annotator agreement was not assessed, as each mention was annotated by only one annotator.

[1]Available for download through the GitLab repository at https://dihur.si/muki/llm4dh/.

4.5. Annotation Guidelines

Sentiment was defined as the evaluative stance expressed toward the referenced collective identity within its immediate context. Annotators were instructed to rely only on linguistic cues and other pragmatic signals available in the local context, and not on broader historical background knowledge. Mentions were labelled as positive or negative only when a linguistically explicit or clearly implied evaluative judgment was present, including cases of irony, patronizing stance, and related pragmatic effects where these could be inferred from the available context. In the absence of such evaluation, mentions were labelled as neutral. All interpretable mentions were assigned a sentiment label; the rest were marked as *unknown*.

Referential type was annotated independently of sentiment. All nominal identity mentions were treated as group references, as they inherently denote collective actors (e.g., Nemci [Germans], Slovenci [Slovenes]). For adjectival identity expressions, referential type was determined based on syntactic context. Adjectives modifying collective actors (e.g., German army) were classified as group references, whereas adjectives modifying inanimate or abstract entities (e.g., Slovene bread, German politics) were classified as non-group references.

Sentiment labelling was performed strictly with respect to the marked identity expression and not the broader topic or event described in the passage.

4.6. LLM Sentiment Inference Setup

Sentiment classification was performed in a few-shot setting using four openly available instruction-following LLMs representing a range of architectures and training regimes: GaMS3-12B-Instruct (Slovene-adapted), Gemma-3-12B-IT, Llama-3.1-8B-Instruct, and DeepSeek-R1-Distill-Qwen-14B. These models were selected to compare a language-adapted model against widely used general-purpose instruction-following models of comparable scale. We additionally included two smaller model variants, Gemma-3-4B-IT and DeepSeek-R1-Distill-Qwen-7B, to assess the impact of model size on performance and whether this impact generalises across model families.

Each collective-identity mention was processed independently using the structured input described in Section 4.2.

Models were prompted in Slovene with explicit instructions to perform targeted sentiment classification of the marked mention only. The prompt required that: (1) sentiment be evaluated exclusively for the tagged identity expression; (2) output be returned as valid JSON; and (3) the assigned label be one of the predefined categories: POS,

NEU, or NEG.

Short illustrative examples of positive, negative, and neutral classifications were included in the prompt to clarify labelling criteria.

Model responses were required to conform to a predefined JSON schema containing the mention text and predicted sentiment label. No recalibration, filtering, or manual correction of sentiment labels was applied; all reported results reflect the models' raw predictions.

4.7. Dataset-Level Aggregation of Sentiment Predictions

Sentiment predictions were generated at the level of individual collective-identity mentions. To examine whether the selected model can be applied on the entire dataset, we aggregated these mention-level predictions by collective identity and newspaper. For each identity within each newspaper, we counted the number of POS, NEU, and NEG predictions and computed their relative proportions.

The resulting class distributions are presented in 5.2. Given the class-specific performance differences observed in Section 5.1—particularly lower recall for positive sentiment—the aggregated proportions should be interpreted as reflecting the model's relative classification tendencies rather than exact estimates of historical evaluative stance.

5. Results

This section presents a comparison of multiple instruction-following LLMs, then analyses the performance profile of the top-scoring model in greater detail. Finally, we demonstrate dataset-level sentiment distributions to assess the scalability and interpretative plausibility of large-scale inference.

5.1. Model Evaluation

5.1.1. Overall Performance Across Models

We evaluated four cutting-edge instruction-following LLMs in a few-shot setting: GaMS3-12B-Instruct (GaMS3), Gemma-3-12B-IT (Gemma-3-12B), Llama-3.1-8B-Instruct (Llama 3.1), DeepSeek-R1-Distill-Qwen-14B (DeepSeek-R1-14B). To assess the effect of model size within two model families, we additionally included two smaller variants: Gemma-3-4B-IT (Gemma-3-4B) and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-R1-7B). All models were evaluated using the same prompt template, decoding parameters, and manually annotated evaluation sample (N=371 mentions). Table 2 reports overall accuracy, macro-averaged F1, weighted F1, and per-class F1 scores for each model.

Model	Acc	M-F1	W-F1	$F1_{POS}$	$F1_{NEU}$	$F1_{NEG}$
GaMS3	**0.693**	0.538	**0.708**	0.343	**0.794**	**0.478**
Gemma-3-12B	0.679	**0.555**	0.701	0.411	0.776	0.477
LLaMA 3.1	0.485	0.459	0.515	0.427	0.545	0.406
DeepSeek-R1-14B	0.580	0.519	0.616	**0.442**	0.665	0.450
Gemma-3-4B	0.620	0.428	0.637	0.154	0.742	0.388
DeepSeek-R1-7B	0.253	0.244	0.260	0.197	0.267	0.269

Table 2: Cross-model performance on mention-level sentiment classification (N=371). M-F1 denotes macro-averaged F1; W-F1 denotes support-weighted F1. Class-specific scores are reported as $F1_{POS}$, $F1_{NEU}$, and $F1_{NEG}$, corresponding to POS, NEU, and NEG sentiment labels.

GaMS3 achieves the highest overall accuracy (0.693) and the highest weighted F1 (0.708), indicating the strongest performance when class frequency is taken into account. Gemma-3-12B attains the highest macro-averaged F1 (0.555), suggesting slightly more balanced performance across sentiment classes. The difference between GaMS3 and Gemma-3-12B is small in overall accuracy ($\Delta \approx 0.014$), and in macro F1 ($\Delta \approx 0.017$), but the models differ slightly in their performance profile: GaMS3 performs better under class imbalance, whereas Gemma-3-12B shows slightly more balanced behaviour across sentiment classes. In relation to our research questions, these results suggest that Slovene adaptation yields a measurable but limited advantage over a strong general-purpose instruction-tuned model from the same family (Gemma-3-12B).

Comparing the smaller models, DeepSeek-R1-7B performs notably worse than its larger counterpart (0.253 vs. 0.580 accuracy). By contrast, Gemma-3-4B shows a more limited performance drop relative to Gemma-3-12B (0.620 vs. 0.679 accuracy), although its performance declines substantially on positive sentiment detection (0.154 vs. 0.411 $F1_{POS}$). This asymmetry may be related to the distillation process used to train DeepSeek-R1-7B, which can reduce sensitivity to low-frequency linguistic features important for handling Slovene. As a morphologically rich and relatively low-resource language, Slovene may be particularly sensitive to such degradation.

At the class level, all larger model variants perform best on neutral sentiment, though the magnitude differs substantially. GaMS3 achieves the highest F1 for neutral instances (0.794), followed closely by Gemma-3-12B (0.776). In contrast, both DeepSeek-R1-14B (0.665) and LLaMA 3.1 (0.545) have greater difficulties distinguishing descriptive from evaluative contexts in this domain.

Performance on negative sentiment is low for GaMS3 (0.478), Gemma-3-12B (0.477) and DeepSeek-R1-14B (0.450), and even lower for LLaMA 3.1 (0.406). Notably LLaMA 3.1 exhibits ex-

tremely high negative recall (0.950), paired with substantially lower precision, suggesting a tendency to overpredict negative sentiment. This pattern indicates reduced calibration rather than stronger discrimination.

Positive sentiment is consistently the most difficult class across models, especially smaller variants. GaMS3 yields an F1 of 0.343, Gemma-3-12B improves to 0.411, LLaMA 3.1 reaches 0.427, while DeepSeek-R1-14B achieves the best result (0.442). On the other hand, smaller models (Gemma-3-4B and DeepSeek-R1-7B) under-perform considerably. These results confirm that explicit positive evaluation in historical newspaper discourse is both less frequent and more challenging for models to detect reliably.

Taken together, these results indicate that no single model uniformly dominates across all evaluation criteria. GaMS3 provides the strongest overall performance under realistic class imbalance, while Gemma-3-12B demonstrates competitive class-balanced behaviour. The divergence between models—despite identical prompts and evaluation data—highlights the importance of empirical validation when deploying LLMs for sentiment inference in historically noisy corpora. Model performance in such DH settings appears sensitive not only to general instruction tuning, but also to language adaptation and domain robustness. These findings underscore the need for cross-model comparison when LLMs are used as analytical instruments in humanities research.

Given these results, subsequent dataset-level analyses are conducted using GaMS3. Although Gemma-3-12B yields slightly higher macro-averaged F1, GaMS3 provides the highest overall accuracy and weighted F1, as well as the strongest performance on neutral detection, which dominates the historical dataset distribution. This combination of robustness under class imbalance and stable neutral classification makes GaMS3 suitable for large-scale aggregation. At the same time, the cross-model variation observed above underscores that polarity estimates should be interpreted comparatively rather than as absolute measures of evaluative intensity.

5.1.2. Performance Profile of GaMS

Table 3 summarizes the performance of GaMS3-12B-Instruct on the manually annotated evaluation set. GaMS's overall accuracy reaches 0.693, with a weighted F1 of 0.708 and a macro-averaged F1 of 0.538. The gap between weighted and macro F1 reflects class imbalance and uneven performance across sentiment categories.

Neutral sentiment is detected most reliably (F1=0.794, P=0.858, R=0.739; support=287). Both precision and recall are comparatively high, indicating stable behaviour with relatively few false positives and false negatives in neutral contexts.

Sentiment	M	Ska	Snec	SN	**Total**
POS	P	0.400	0.500	0.200	**0.400**
	R	0.261	0.625	0.111	**0.300**
	F1	0.316	0.556	0.143	**0.343**
NEU	P	0.812	0.865	0.909	**0.858**
	R	0.812	0.703	0.700	**0.739**
	F1	0.812	0.776	0.791	**0.794**
NEG	P	0.154	0.511	0.222	**0.351**
	R	0.400	0.767	0.889	**0.750**
	F1	0.222	0.613	0.356	**0.478**
Macro	P	0.455	0.625	0.444	**0.536**
	R	0.491	0.698	0.567	**0.596**
	F1	0.450	0.648	0.430	**0.538**
Weighted	P	0.709	0.760	0.803	**0.749**
	R	0.694	0.713	0.669	**0.693**
	F1	0.697	0.724	0.708	**0.708**

Table 3: Performance of GaMS3 on the manually annotated sample for mention-level sentiment classification for *Slovenka* (Ska), *Slovenec* (Snec), *Slovenski narod* (SN), and the full dataset (Total). M denotes evaluation metric (P precision, R recall, F1-score). Macro and weighted averages are computed across classes.

Negative sentiment achieves an F1 of 0.478 (P=0.351, R=0.750; support=44). The substantially higher recall than precision indicates that the model captures most annotated negative instances but over-assigns negative labels in some neutral or positive contexts. In other words, negative sentiment is detected readily, though not always selectively.

Positive sentiment proves more challenging (F1=0.343, P=0.400, R=0.300; support=40). The relatively low recall suggests that a considerable proportion of positive instances are not recognized and are instead classified as neutral or negative. Compared to the negative class, the model appears more conservative in assigning positive sentiment.

Taken together, the results show a clear asymmetry in class behaviour. Neutral sentiment is most stable. Negative sentiment is detected with high sensitivity but reduced precision. Positive sentiment is under-detected relative to its annotated frequency. For downstream aggregation, this implies that dataset-level summaries are more likely to under-represent positive evaluations than negative ones, while neutral proportions remain comparatively robust.

5.1.3. GaMS Performance by Grammatical Category

We analyse GaMS's performance separately for nominal identity mentions (e.g., Nemci [Germans], Slovenci [Slovenes]) and adjectival modifiers (e.g., nemški [German], slovenski [Slovene]) to assess whether grammatical form affects classification behaviour (see Table 4).

Sentiment	M	Ska	Snec	SN	**Total**
Nouns					
POS	P	0.333	0.000	0.000	**0.222**
	R	0.143	0.000	0.000	**0.100**
	F1	0.200	0.000	0.000	**0.138**
NEU	P	0.702	0.714	0.919	**0.773**
	R	0.767	0.641	0.723	**0.713**
	F1	0.733	0.676	0.810	**0.742**
NEG	P	0.000	0.519	0.333	**0.385**
	R	0.000	0.667	0.857	**0.645**
	F1	0.000	0.583	0.480	**0.482**
Macro	P	0.345	0.411	0.417	**0.460**
	R	0.303	0.436	0.527	**0.486**
	F1	0.311	0.420	0.430	**0.454**
Weighted	P	0.581	0.615	0.799	**0.645**
	R	0.583	0.619	0.702	**0.633**
	F1	0.572	0.613	0.726	**0.630**
Adjectives					
POS	P	0.444	0.556	0.333	**0.476**
	R	0.444	1.000	0.167	**0.500**
	F1	0.444	0.714	0.222	**0.488**
NEU	P	0.918	1.000	0.900	**0.938**
	R	0.849	0.750	0.679	**0.759**
	F1	0.882	0.857	0.774	**0.839**
NEG	P	0.333	0.500	0.111	**0.310**
	R	1.000	1.000	1.000	**1.000**
	F1	0.500	0.667	0.200	**0.473**
Macro	P	0.565	0.685	0.448	**0.574**
	R	0.765	0.917	0.615	**0.753**
	F1	0.609	0.746	0.399	**0.600**
Weighted	P	0.833	0.898	0.818	**0.846**
	R	0.797	0.803	0.639	**0.749**
	F1	0.809	0.820	0.701	**0.777**

Table 4: Performance of GaMS3 on manually annotated nominal (N=180) and adjectival (N=191) mentions. M denotes evaluation metric (P precision, R recall, F1 score). Macro and weighted averages are computed across classes.

For noun mentions (N=180), GaMS achieves an accuracy of 0.633 and a weighted F1 of 0.630, indicating moderate overall performance. Neutral sentiment is identified most reliably (F1=0.742, P=0.773, R=0.713), showing relatively balanced behaviour. Negative sentiment yields an F1 of 0.482, with higher recall (0.645) than precision (0.385), suggesting a tendency to over-assign negative labels in ambiguous contexts.

Positive sentiment proves particularly difficult in the nominal subset. Although 20 positive instances are present, recall drops to 0.1, resulting in a very low F1 of 0.138. This indicates that the model rarely detects positive sentiment when it is expressed through nominal identity references, frequently defaulting instead to neutral or negative predictions.

For adjectival mentions (N=191), performance improves substantially. Accuracy increases to 0.749 and weighted F1 to 0.777, indicating greater overall stability. Neutral sentiment again achieves the highest F1 (0.839, P=0.938, R=0.759), reflecting strong precision and fewer false positives.

Positive sentiment shows marked improvement compared to nouns (F1=0.488, R=0.500), suggesting that evaluative meaning is more readily captured when embedded in adjectival modification rather than nominal reference.

Negative sentiment for adjectives displays a different error profile: recall reaches 1.00, but precision drops to 0.310, indicating systematic overprediction of negative labels in this subset. In other words, the model successfully captures all annotated negative adjectival instances but at the cost of labelling a substantial number of neutral contexts as negative.

Taken together, the results demonstrate that grammatical form influences classification behaviour not only in overall performance but in the precision–recall balance of individual classes. Nominal mentions are associated with missed positive evaluations, whereas adjectival mentions are more prone to over-attributing negative sentiment. This asymmetry is important for dataset-scale analysis, where mention-level predictions are aggregated into identity-level sentiment distributions, combining both grammatical realisations and therefore inheriting their respective error tendencies.

5.1.4. GaMS Performance by Referential Type

We next assess whether GaMS's performance varies according to referential type, distinguishing between group-referential mentions (direct references to collective actors, e.g., Nemci [Germans], or adjectival expressions modifying group-denoting heads, e.g., nemška vojska [German army]) and non-group mentions (typically adjectival nationality markers modifying inanimate or abstract heads, e.g., nemška politika [German politics]; see Table 5).

For group-referential mentions (N=245), GaMS achieves an accuracy of 0.645 and a weighted F1 of 0.660. Neutral sentiment is detected most reliably (F1=0.749, P=0.827, R=0.685), indicating reasonably balanced behaviour. Negative sentiment reaches an F1 of 0.491, with higher recall (0.711)

Sentiment	M	Ska	Snec	SN	**Total**
Group Modifiers					
POS	P	0.333	0.286	0.250	**0.304**
	R	0.235	0.500	0.200	**0.269**
	F1	0.276	0.364	0.222	**0.286**
NEU	P	0.741	0.816	0.936	**0.827**
	R	0.727	0.635	0.698	**0.685**
	F1	0.734	0.714	0.800	**0.749**
NEG	P	0.000	0.526	0.280	**0.375**
	R	0.000	0.741	0.875	**0.711**
	F1	0.000	0.615	0.424	**0.491**
Macro	P	0.358	0.543	0.489	**0.502**
	R	0.321	0.625	0.591	**0.555**
	F1	0.337	0.564	0.482	**0.509**
Weighted	P	0.619	0.710	0.822	**0.701**
	R	0.587	0.660	0.684	**0.645**
	F1	0.601	0.671	0.722	**0.660**
Non-Group Modifiers					
POS	P	0.667	1.000	0.000	**0.714**
	R	0.333	0.750	0.000	**0.357**
	F1	0.444	0.857	0.000	**0.476**
NEU	P	0.905	0.960	0.867	**0.907**
	R	0.927	0.857	0.703	**0.830**
	F1	0.916	0.906	0.776	**0.867**
NEG	P	0.500	0.429	0.091	**0.273**
	R	1.000	1.000	1.000	**1.000**
	F1	0.667	0.600	0.167	**0.429**
Macro	P	0.690	0.796	0.319	**0.631**
	R	0.753	0.869	0.568	**0.729**
	F1	0.676	0.788	0.314	**0.591**
Weighted	P	0.859	0.919	0.766	**0.856**
	R	0.857	0.857	0.643	**0.786**
	F1	0.848	0.874	0.688	**0.803**

Table 5: Performance of GaMS3 on manually annotated group-referential (N=245) and non-group mentions (N=126). M denotes evaluation metric (P precision, R recall, F1 score). Macro and weighted averages are computed across classes.

than precision (0.375), suggesting that the model captures most negative group evaluations but also assigns negative labels to a notable number of non-negative instances. Positive sentiment performs more weakly (F1=0.286, P=0.304, R=0.269), indicating that positive evaluations directed toward collective actors are frequently missed.

For non-group mentions (N=126), overall performance improves. Accuracy increases to 0.786 and weighted F1 to 0.803. Neutral sentiment again shows strong performance (F1=0.867, P=0.907, R=0.830), reflecting both low false-positive and low false-negative rates in descriptive contexts. Positive sentiment also improves compared to the group subset (F1=0.476, P=0.714, R=0.357). While precision is relatively high, recall remains moderate, indicating that some positive cases are still not detected. Negative sentiment achieves perfect recall (1.00); however, this result must be interpreted cautiously due to the very small number of negative instances in this subset (support=6), which limits stability and inflates recall.

Overall, the comparison indicates that direct group-referential mentions are more difficult for the model than non-group contexts. In group contexts, negative sentiment is more readily identified than positive sentiment, and positive evaluations are disproportionately missed. In non-group contexts, classification is more stable across classes, particularly for neutral and positive sentiment. This distinction is relevant for dataset-level aggregation, as summaries of sentiment toward collective actors may under-estimate positive polarity more than negative polarity.

5.2. Sentiment Distribution by Identity and Newspaper

Following the evaluation and diagnostic analysis above, we applied GaMS3-12B-Instruct to all identity mentions in the dataset (2.65 million instances) and aggregated predicted sentiment labels by identity and newspaper. Figure 1 shows the proportional distribution of POS, NEU, and NEG predictions for the five most frequently mentioned collective identities in the dataset. For each identity, three stacked bars correspond to the historical newspapers in Slovene: *Slovenka*, *Slovenec*, and *Slovenski narod*.

The distributions differ across identity categories. Nemci [Germans] show a consistently higher proportion of negative predictions across all three newspapers, with relatively smaller positive shares. In contrast, references to Slovenci [Slovenians] display a more balanced or mixed distribution, including a visibly larger proportion of positive predictions, particularly in *Slovenka*. Identities such as Avstrijci [Austrians] and Rusi [Russians] are dominated by neutral predictions across newspapers, with only limited positive or negative shares. The distribution for Čehi [Czechs] appears more mixed, with moderate levels of both positive and negative labels depending on the newspaper.

These patterns suggest that the model differentiates between identity categories rather than assigning sentiment uniformly. On the one hand, some of our preliminary findings largely align with mainstream historiographical narratives surrounding turn-of-the-century Slovene history. For example, the overwhelmingly negative sentiment that all three journals show towards Germans is indicative of contemporary Slovene-German nationalist political conflict (Čuček, 2016). Likewise, the pre-

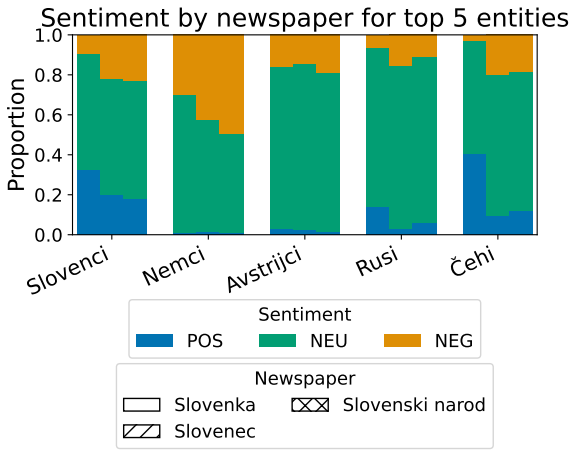

Figure 1: Sentiment class composition (POS/NEU/NEG) for the five most frequent collective identities in the dataset. For each identity, three stacked bars show the predicted class proportions in *Slovenka* (no hatch), *Slovenec* (//), and *Slovenski narod* (xx).

dominantly neutral sentiment expressed towards Austrians is expected given that Slovene political discourse of the time was mostly supportive of the Austrian state and Austrian political identity (Luthar et al., 2008).

Conversely, the results that we have gained by analysing the lemma Čehi [Czechs] demonstrate some of the interpretative complexities faced when compiling data using such models. While the overwhelmingly positive sentiment shown towards this group in *Slovenka* is not surprising, it is curious to see an overwhelmingly negative assessment of Czechs in the other two journals knowing that Slovene society was overwhelmingly sympathetic to Czechs during this period (Keršič-Svetel, 1996). While our interpretation remains speculative, we assume that the negative sentiment in the two journals was connected to the intensity of German-Czech nationalist conflict in the multiethnic crownland of Bohemia [Češka] — a province that is otherwise referred to using the same adjective [češki] as the Czech nation.

This analysis is intended as a plausibility check rather than a substantive historical argument. As shown in Section 5.1, model performance varies by class, with the highest reliability for neutral sentiment and lower performance for positive instances.

Figure 1 demonstrates that mention-level predictions can be aggregated at dataset scale while preserving identity-specific variation, provided that model performance characteristics are taken into account.

6. Conclusion

This study evaluated whether instruction-following LLMs can reliably perform targeted, mention-level sentiment classification in OCR-extracted text from historical Slovene newspapers and whether these predictions can support large-scale historical analyses when aggregated across the dataset. Using a manually annotated sample of 371 collective-identity mentions, we benchmarked four instruction-tuned LLMs and selected the Slovene-adapted GaMS3-12B-Instruct model for large-scale application. The comparison further shows that Slovene adaptation is beneficial, but that its advantage over a strong general-purpose model from the same family remains modest in a few-shot environment.

We show that the best-performing model is usable for this task, but its performance is class-dependent and varies across grammatical realisation and referential type. Neutral sentiment is detected most reliably. Negative sentiment is captured with relatively high recall but lower precision, indicating a tendency toward over-attribution, while positive sentiment is systematically under-detected, especially in nominal and group-referential contexts. These asymmetries directly affect dataset-level interpretation and require that aggregated polarity patterns be read as directional tendencies rather than exact measurements of historical evaluative stance.

This study also highlights an often overlooked disconnect between technical benchmarks and scholarly needs. For DH workflows, a model's F1 score is ultimately less important than its interpretative validity: how it behaves when used to map complex and often ambiguous human expression.

More broadly, the study provides a benchmark for targeted sentiment classification in OCR-degraded historical Slovene, highlights the necessity of cross-model comparison in DH settings, and offers an empirically grounded assessment of both the reliability and limitations of instruction-following LLMs as analytical instruments in DH research.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the Slovenian Research and Innovation Agency research programme "Digital Humanities: resources, tools and methods" (2022–2027) [grant number P6-0436], by the DARIAH-SI research infrastructure, by the Slovene Common Language Resources and Technology Infrastructure (CLARIN.SI), and by the project "Large Language Models for Digital Humanities" (2024–2027) [grant number GC-0002].

7. Bibliographical References

Swoichha Adhikari, Manan Gangwani, and Adithi Varadarajan. 2024. Aspect-based sentiment analysis for slovene texts: Models, lexicons, and embeddings. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, volume 2, pages 1–6.

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2017. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175, page 14. Las Vegas, NV.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186.

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2024. A lightweight approach to a giga-corpus of historical periodicals: The story of a slovenian historical newspaper collection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 695–703.

Darja Fišer, Jasmina Smailovic, Tomaž Erjavec, Igor Mozetic, and Miha Grcar. 2016. Sentiment annotation of slovene user-generated content. In *Proceedings of the 2016 conference language technologies and digital humanities (JTDH 2016)*, pages 65–70.

Gemma Team. 2025. Gemma 3.

Aaron Grattafiori and Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, pages 216–225.

Marjeta Keršič-Svetel. 1996. *Češko-slovenski stiki med svetovnima vojnama*. Zbirka Zgodovinskega časopisa. Zveza zgodovinskih društev Slovenije, Ljubljana.

Oto Luthar, Igor Grdina, Marjeta Šašel Kos, Petra Svoljšak, Peter Kos, Dušan Kos, Peter Štih, Alja Brglez, and Martin Pogačar. 2008. The land between: a history of slovenia.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Iftikhar Muhammad, Marco Rospocher, Timotej Knez, and Slavko Žitnik. 2025. Benchmarking large language models for target-based financial sentiment analysis. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 785–795, Cagliari, Italy. CEUR Workshop Proceedings.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 380–385.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. *arXiv preprint arXiv:2410.06898*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906. Association for Computational Linguistics.

Filip Čuček. 2016. Svoji k svojim: na poti k dokončni nacionalni razmejitvi na spodnjem Štajerskem v 19. stoletju.

Slavko Žitnik, Neli Blagus, and Marko Bajec. 2022. Target-level sentiment analysis for news articles. *Knowledge-Based Systems*, 249:108939.

8. Language Resource References

Filip Dobranić, Bojan Evkoski, and Nikola Ljubešić. 2023. Corpus of slovenian periodicals (1771-1914) sPeriodika 1.0. Slovenian language resource repository CLARIN.SI.