Evaluating Multilingual Sentiment Classifiers Using an LLM-Annotated Wikipedia Benchmark

Milena Stróżyna[1], **Włodzimierz Lewoniewski**[1], **Izabela Czumałowska**[1]
[1]Poznań University of Economics and Business,
Correspondence: milena.strozyna@ue.poznan.pl

Abstract

We present a multilingual study of sentiment evaluation on Wikipedia articles from various topics in five languages (German, English, Spanish, Polish, and Russian). In this paper, we compare three large language models (Gemini Pro 3.1, Claude Opus 4.6, and GPT 5.2), each queried three times per sentence, with two popular multilingual sentiment classifiers. This setup allows us to analyze not only inter-model differences but also intra-model stability as a proxy for confidence.

To support systematic evaluation, we construct a benchmark dataset based on strict consensus across evaluators and analyze sentiment distributions across topics and languages. We show substantial variation in sentiment distributions, agreement, and consistency across models and languages. Our results suggest that sentiment evaluation on encyclopedic text remains an underexplored challenge for multilingual NLP.

1 Introduction

Sentiment analysis is often framed as a classification task, but this perspective is inherently limiting, as sentiment is better understood as an evaluation problem, since opinions are typically gradual rather than strictly categorical (Borg and Boldt, 2020).

Wikipedia makes an especially interesting stress test for sentiment analysis, since all Wikipedia articles are explicitly designed to follow a Neutral Point of View (NPOV) rule (Matei and Dobrescu, 2011). This policy indicates that articles should avoid subjective or emotionally charged language, making general sentiment relatively rare and subtle. As a result, detecting sentiment in this domain requires models to handle implicit framing, nuanced wording, and context-dependent cues, rather than relying on obvious polarity markers.

Wikipedia's articles also offer a unique setting to motivate multilingual sentiment analysis, because it exists in hundreds of language editions, each shaped by distinct cultural, linguistic, and editorial communities (Van Dijk, 2009). All versions should aim to follow Neutral Point of View policy, but the way neutrality is expressed can vary across languages.

Despite substantial progress in sentiment analysis, important research gaps remain. First, there is a lack of multi-domain evaluation datasets that enable systematic assessment of model robustness across heterogeneous text sources; most existing benchmarks are domain-specific. Second, there is limited understanding of how to reliably aggregate sentence-level predictions into document-level sentiment, as simple heuristics often fail to capture discourse structure, contextual dependencies, and the varying importance of individual sentences. Addressing these gaps is crucial for developing sentiment analysis methods that are both transferable across domains and capable of producing coherent, document-level evaluations.

To do so, this paper introduces a multilingual, multi-domain sentiment evaluation benchmark derived from Wikipedia, constructed using consensus across state-of-the-art LLM evaluators. We evaluate the performance of sentiment analysis BERT-based models against this LLM-based benchmark. Additionally, we propose and compare two methods for aggregating sentence-level sentiment predictions into document-level representations. Finally, we analyze how aggregation strategies influence model agreement in a multilingual texts. This work builds upon and extends our prior research on sentiment analysis of Wikipedia content (Stróżyna et al., 2025; Lewoniewski et al., 2026b).

2 Related Work

Sentiment analysis is one of the areas of Natural Language Processing (NLP) that aims to determine the emotional tone of a text. Various methods are used for this purpose, including rule-based and

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 692–703
July 4, 2026 ©2026 Association for Computational Linguistics

lexicon-based approaches, as well as machine learning algorithms and deep learning techniques.

Although many earlier studies show that lexicon-based methods are among the most widely used (Baccianella et al., 2010; Wankhade et al., 2022; Ajik et al., 2023), it is now clear that transformer-based models and large language models (LLMs) are gradually dominating the sentiment analysis landscape (Gowda et al., 2025). Introduced by Vaswani et al. (2017), the transformer architecture addresses key limitations of lexical approaches, such as the need for constant lexicon updates, difficulty handling negation (Mao et al., 2024), reliance on domain-specific dictionaries (Wankhade et al., 2022), and challenges with capturing complex or nuanced emotions (Wankhade et al., 2022). By using self-attention, transformers model relationships between words, enable parallel processing, and capture dependencies across the input (Bashiri and Naderi, 2024).

2.1 BERTs in sentiment analysis

A widely known example of transformer-based models is BERT (Bidirectional Encoder Representations from Transformers). Its introduction marked a major breakthrough in NLP due to its bidirectional architecture. Unlike models that read text sequentially from left-to-right or right-to-left, BERT processes both contexts simultaneously, enabling a deeper understanding of text (Bashiri and Naderi, 2024).

In this study, we selected two BERT-like models based on their strong multilingual capabilities and recency. The first one is **modernBERT-base-multilingual-sentiment**[1], further referred to as **modBERT**, is a system built on top of the ModernBERT-base architecture[2]. It supports multilingual sentiment classification in 16+ languages. The model was fine-tuned using the Multilingual Sentiment dataset[3], which contains approximately 3.93 million annotated text samples. The dataset also covers 16+ languages, including English, German, Spanish, and Russian, with substantial variation in representation. English is by far the dominant language (over 1.5 million instances), while German, Spanish, and Russian are also well represented (each exceeding 200,000 samples). Importantly, Polish is not included among the supported languages. Instead, the dataset focuses on a set of high-resource and widely used languages (e.g., English, Chinese, French, Spanish, German, Russian), alongside several additional languages such as Arabic, Japanese, and Korean. Overall, the dataset exhibits a strong imbalance toward English and other major languages, while entirely omitting some European languages such as Polish, which may have implications for model performance. After fine-tuning, the model reaches an F1-score of 80.16 and has roughly 150 million parameters.

The second selected model is **multilingual-sentiment-analysis**[4], further referred to as **mulBERT**. This is a DistilBERT-based sequence classifier fine-tuned for multilingual sentiment analysis. Model documentation stated that training was conducted only on synthetic multilingual data generated with large language models. The model supports 23 languages (incl. English, Spanish, German, Polish, and Russian) and its reported off-by-one accuracy up to 0.93.

Both models were selected as they support a wide range of languages, which makes them suitable for our experimental setup involving five different language versions of Wikipedia. Additionally, as recently released (2024) models fine-tuned for sentiment analysis, they provide a relevant and up-to-date baseline for evaluating sentiment classification performance.

2.2 LLMs as evaluators

With recent advancements in modern generative AI, large language models - such as OpenAI's GPT series, Anthropic's Claude, and Google's Gemini - have been increasingly applied in sentiment analysis research. This shift has moved the field beyond traditional discriminative BERT-like models toward more versatile, general-purpose approaches. Thanks to extensive pretraining on large datasets, these models excel at sentiment analysis, particularly in zero-shot and few-shot settings, allowing effective classification even with minimal task-specific data (Gautam et al., 2025).

In contrast to embedding-based methods, models like GPT and Gemini treat sentiment analysis as a text classification task, determining sentiment polarity through vectorized representations and contextual semantic understanding, without the need

[1] https://huggingface.co/clapAI/modernBERT-base-multilingual-sentiment

[2] https://huggingface.co/answerdotai/ModernBERT-base

[3] https://huggingface.co/datasets/clapAI/MultiLingualSentiment

[4] https://huggingface.co/tabularisai/multilingual-sentiment-analysis

for extensive task- or domain-specific fine-tuning (Xie et al., 2023). Modern LLMs also offer the capability to process and evaluate large volumes of text with greater speed and consistency than human annotators. They streamline assessment workflows and reduce the manual burden of data analysis (Sushil et al., 2024).

Among other methods, LLMs stands out in providing reasoning or explanation for classifications (Explainable AI). Models like ChatGPT can point out specific text premises that might influence sentiment label (Kocoń et al., 2023).

3 Benchmark Construction

3.1 Data preparation

In order to construct a multilingual benchmark dataset for sentiment analysis, we selected a set of high-importance Wikipedia articles from diverse domains based on two community-curated lists (as of March 2026): (1) *Wikipedia: Vital articles/Level/3*[5], comprising approximately 1,000 articles, and (2) *List of articles every Wikipedia should have*[6], which defines a core knowledge set for each language edition, consisting of approx. 1,000 Wikidata items. Article titles from both sources were extracted, merged, and deduplicated. For each selected article, corresponding versions in four additional languages: German (de), Spanish (es), Polish (pl), and Russian (ru) were retrieved. The cross-lingual alignment was established using Wikipedia interlanguage links.

A preprocessing pipeline was then applied to the selected articles. First, only the lead sections (abstracts) were retained to focus on the most representative content. The text was subsequently segmented into sentences using a language-specific tokenization pipeline based on the Stanza NLP toolkit[7]. Each sentence was assigned a unique identifier composed of the Wikipedia page ID and a sentence index (i.e., PageID–SentenceNumber).

Next, normalization procedures were applied, including the removal or standardization of special tokens and the merging of very short sentences to improve coherence. This step resulted in the following number of sentences: 21,991 in English; 12,715 in German; 17,021 in Spanish; 10,707 in Polish; 12,739 in Russian.

In the final step, sentences were grouped into batches suitable for LLM processing. Each sentence was formatted as a JSON line (`"id": sentence_id, "text": sentence_text`), and batches were constructed to respect API constraints, with a maximum size of 25 thousand tokens.

This processing pipeline enabled the construction of a multi-domain evaluation benchmark that goes beyond commonly used sentiment datasets based on user-generated reviews.

3.2 Sentiment Annotation Procedure

To assign sentiment labels to the previously constructed Wikipedia dataset, the corpus of sentences was annotated using general-purpose LLMs. The annotation process followed a controlled multi-model prompting procedure to ensure consistency and comparability across models.

Sentiment annotation was performed using three state-of-the-art LLMs: GPT-5.2 (OpenAI), Gemini 3.1 Pro (Google), and Claude Opus 4.6 (Anthropic). At the time of the study, these models ranked among the top-performing systems for text-based tasks on the LMarena.ai leaderboard[8]. Furthermore, as discussed in Section 2, general-purpose LLMs are suitable candidates for high-quality annotation due to their ability to capture contextual information, broad training data coverage, and strong zero-shot capabilities. All models were accessed via their official APIs, and the input data were processed in batches.

Each sentence was assigned a sentiment label using a four-class scheme: positive, neutral, negative, or `n/a`. To ensure procedural consistency, the same prompt was applied across all models (see Appendix A). The final version of the prompt was developed iteratively, based on pilot experiments conducted on a sample dataset.

The annotation procedure was repeated three times for each model, resulting in nine independent runs (three models × three repetitions) per language. Consequently, each sentence received multiple sentiment labels generated by different models and across repeated executions. This multi-run setup enabled a systematic analysis of annotation stability, including both intra-model consistency and inter-model agreement.

[5] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles/Level/3

[6] https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have

[7] https://stanfordnlp.github.io/stanza/

[8] https://arena.ai/en/leaderboard

3.3 Agreement Analysis and Benchmark Construction

In the next step, annotation quality was assessed using metrics for both intra-model and cross-model agreement. Intra-model consistency was evaluated by measuring whether each model assigned identical sentiment labels to the same sentence across its three independent runs. Cross-model agreement was defined using a strict criterion requiring unanimous agreement across all nine runs (see Table 2).

Based on this criterion, the final benchmark dataset was constructed by retaining only those sentences for which all models produced identical sentiment labels in all nine repetitions. This strict filtering resulted in removal of 6,088 sentences in English, 3,144 in German, 9,071 in Spanish, 2,757 in Polish, and 3,104 in Russian. This rule was adopted to maximize label reliability and reduce the influence of model-specific biases and stochastic variation. This corresponds to retaining approximately 72% of the original sentences in English, 75% in German, 71% in Spanish, 74% in Polish, and 76% in Russian. The relatively high number of filtered sentences reflects the strictness of the consensus criterion, which prioritizes label reliability and consistency.

For the evaluation framework presented in the following section, the benchmark dataset derived from a strict consensus procedure was fixed, ensuring reproducibility. The number of sentences before applying the strict agreement (source) and in the benchmark dataset after agreement (filtered) across languages is presented in Table 1. The retained sentences are subsequently named as the sentence-level benchmark.

3.4 Topic Classification

In order to enable analysis across different topics, we constructed a unified, flat categorization scheme from two independently curated lists (which were mentioned before in subsection 3.1). The objective was to eliminate hierarchical dependencies, ensure full coverage of all articles, and constrain category sizes to a manageable and comparable range (between 40 and 200 items per category). The first list (”Vital Articles”, level 3) organized by domains (e.g., People, History, Science) with multiple levels of subcategories. A second list containing Wikidata-linked entries grouped under thematic headings (e.g., Biography, Philosophy, Technology), also partially hierarchical.

As a result, a set of 15 unified categories was defined, and each Wikipedia article, along with all sentences it contains, was assigned to exactly one category through the following rule-based mapping process:

1. Primary mapping: articles inherited the unified category corresponding to their original top-level domain (e.g., “Science” → physical sciences).
2. Subcategory refinement: when necessary, subsection headings were used to refine assignment (e.g., “Biology” → life sciences; “Physics” → physical sciences).
3. Biographical classification: articles under “People” or “Biography” were classified into subgroups based on their original subcategory (e.g., “Writers” → artists; “Scientists” → scientific biography).
4. Conflict resolution: in cases where an article could belong to multiple domains, precedence rules were applied:
    - Biography overrides topical domain for person entities.
    - Domain-specific categories (e.g., mathematics, language) override general ones.
    - Remaining ambiguities were resolved manually.

4 Evaluation Framework

In this section, we present the evaluation framework designed to assess the performance of two BERT-based models, modBERT and mulBERT (see Section 2), in the task of sentiment analysis on general, multi-domain Wikipedia texts. The evaluated BERT-based models were not fine-tuned on the constructed benchmark dataset.

The framework compares their sentiment predictions with a benchmark dataset derived from general-purpose LLMs. In this setup, annotations produced by Gemini, GPT, and Opus serve as evaluators, and the predictions of the BERT-based models are assessed against this benchmark.

4.1 Sentence-Level Evaluation

In the first step, the sentence-level benchmark was compared with sentiment predictions generated by the BERT-based models. The analysis was conducted separately for each model and jointly across both models and the benchmark evaluators.

We first examined the distribution of sentiment labels for the benchmark and the BERT models across different languages. In addition, we analyzed sentence-level predictions within documents belonging to different topics (see Section 3.1 for topics definition), enabling a more fine-grained, domain-aware comparison.

To assess agreement between models, inter-model compliance was evaluated for each language. Agreement was measured both as simple percent-

age agreement (cross-model agreement) and using standard inter-annotator agreement metrics:

- Cohen's κ (Cohen, 1960) – measures pairwise agreement between two annotators while correcting for chance agreement.
- Fleiss' κ (Fleiss, 1971) – extends Cohen's κ to multiple annotators.

Cohen's κ was calculated pairwise between the benchmark and each BERT model. Additionally, it was computed for sentences grouped by topic, resulting in a matrix of agreement scores. Fleiss' κ was used to assess overall agreement at the sentence level across multiple annotators.

We additionally counted macro F1, which was calculated between the benchmark labels and each BERT model separately. For each comparison, class-wise F1 scores were computed for the positive, neutral, and negative labels, treating each label as a one-vs-rest classification problem.

4.2 Document-Level Evaluation

Following the sentence-level evaluation, the next step of the analysis involves aggregating sentiment predictions to the document level. Therefore, we propose two methods of aggregation to assess the overall sentiment of each document.

Majority Voting (MV) method – each sentence is assigned a single sentiment label: positive (pos), neutral (neu), or negative (neg). Each sentence contributes one vote. The overall sentiment of the document is determined by the class that receives the highest number of votes.

Weighted Majority Voting (WMV) method – each sentence is assigned a single sentiment label: positive (pos), neutral (neu), or negative (neg). The weight of each sentence's vote is proportional to its length, based on the assumption that longer sentences convey more semantic information and should therefore have a greater impact on the final decision. For each document, the weights of all sentences within each class are summed, and the overall sentiment of the document is assigned to the class with the highest cumulative weight.

The results of aggregation step are summarized in Table 1. The number of documents differs between aggregation methods due to differences in how sentence-level sentiment labels were combined and filtered. In MV method, sentiment labels were first aggregated from the sentence level to the document level across all nine annotation runs (i.e., three independent repetitions for each of the three LLM evaluators). Subsequently, only those documents were retained for which the aggregated sentiment label remains identical across all nine runs. The final number of documents therefore reflects a strict consensus at the document level. An analogous procedure was applied using WMV method. As a result, differences in aggregation lead to variations in the number of documents that satisfy the strict agreement criterion. These documents are subsequently treated as the document-level benchmark.

Table 1: Benchmark dataset statistics: number of sentences and resulting documents after aggregation

Lang.	Source Sent.	Filtered Sent.	Documents MV	Documents WMV
de	12,715	9,571	1,123	1,104
en	21,991	15,910	1,063	1,023
es	17,021	12,097	1,056	996
pl	10,707	7,950	1,089	1,045
ru	12,739	9,635	1,147	1,094

In the next step, we evaluated sentiment predictions at the document level by comparing the document-level benchmark with the outputs of the BERT-based models. The analysis was conducted separately for each aggregation method. Similarly to the sentence-level evaluation, we assessed cross-model agreement, expressed as percentage agreement, as well as inter-annotator agreement using Cohen's κ and Fleiss' κ. These metrics were analyzed across multiple dimensions, including different languages, aggregation methods (MV and WMV), and document groups defined by their associated topics. Overall, the evaluation framework compares predicted sentiment labels with benchmark annotations without additional post-processing. The benchmark itself is fixed and derived from a strict consensus procedure, ensuring reproducibility. The selected metrics capture both exact agreement and agreement corrected for chance, enabling a comprehensive assessment of model performance at both sentence and document levels.

4.3 Dataset Publication

We publicly release the constructed datasets[9] - for each of two document-level sentiment aggregation methods described in this work (Majority Voting and Weighted Majority Voting). Each dataset contains sentiment labels derived from the benchmark and from the evaluated models at the level of

[9] https://www.kaggle.com/datasets/lewoniewski/wikipedia-sentiment-benchmark

Table 2: Comparison of sentiment annotation agreement across languages between benchmark and BERT models

Lang.	Agreement		Cohen's κ		Fleiss' κ
	modBERT	mulBERT	modBERT	mulBERT	
Sentence-level					
de	71%	81%	0.23	0.09	0.12
en	81%	76%	0.40	0.14	0.23
es	69%	77%	0.35	0.15	0.18
pl	55%	81%	0.08	0.12	0.02
ru	23%	82%	0.06	0.15	-0.13
Document-level – Majority Voting (MV)					
de	90%	95%	0.16	0.00	0.07
en	93%	93%	0.23	0.00	0.1
es	87%	91%	0.39	0.00	0.16
pl	68%	93%	0.07	0.03	-0.02
ru	9%	94%	0.02	0.00	-0.31
Document-level – Weighted Majority Voting (WMV)					
de	84%	92%	0.19	0.00	0.07
en	92%	90%	0.33	0.00	0.15
es	82%	90%	0.32	0.02	0.13
pl	55%	92%	0.06	0.04	-0.06
ru	7%	93%	0.02	0.00	-0.32

• **Agreement** between benchmark and respective BERT models; values represent % of instances with identical sentiment.
• **Cohen's** κ calculated between benchmark and respective BERT models.
• **Fleiss'** κ computed across all models (benchmark + BERT models), to measure overall agreement corrected for chance.

Wikipedia lead sections. Additionally, the dataset includes annotations produced by each evaluator (LLM) for both document-level sentiment aggregation methods applied to the texts of Wikipedia articles in 5 language versions.

5 Results

Table 1 summarizes the number of sentences in the benchmark dataset across five languages. After applying the strict inter-model agreement criterion, the largest number of sentences was obtained for English, while the smallest for Polish. The table also reports the number of benchmark documents for each language after aggregation using MV and WMV. In all cases, MV yields a slightly higher number of documents.Figure 1 presents the distribution of sentiment labels across languages for the benchmark dataset and the two BERT models. The distributions are shown separately for each language and further segmented by topic. This allows for a detailed comparison of sentiment label distributions across models, languages, and topical categories. The bar charts illustrate the proportion of negative, neutral, and positive labels.

Table 2 presents agreement metrics across languages between the benchmark evaluators and the BERT-based models at both sentence and document levels. In addition, pairwise inter-model agreement between the benchmark and each BERT model is presented using Cohen's κ. Finally, Fleiss' κ values are reported to capture overall agreement across all models, i.e., the benchmark evaluators and both BERT-based models. Figure 2 presents Cohen's κ between the benchmark dataset and the BERT-based models across languages and topics at both sentence and document levels. The values are reported separately for each language, with columns corresponding to sentence-level and document-level results, and rows representing different topic categories. Overall, the agreement results confirm the patterns already observed in the label distributions. Cohen's κ was also computed between the two BERT-based models; however, the agreement between them was consistently lower than that observed between the benchmark and each individual model. Therefore, these results are not reported in the figure.

Model	Lang.	Pos	Neu	Neg	Macro F1
modBERT	de	0.37	0.83	0.22	0.47
modBERT	en	0.52	0.89	0.40	0.60
modBERT	es	0.52	0.79	0.38	0.56
modBERT	pl	0.23	0.71	0.13	0.36
modBERT	ru	0.26	0.23	0.17	0.22
mulBERT	de	0.10	0.90	0.14	0.38
mulBERT	en	0.11	0.86	0.25	0.41
mulBERT	es	0.16	0.87	0.23	0.42
mulBERT	pl	0.17	0.90	0.14	0.40
mulBERT	ru	0.18	0.90	0.21	0.43

Table 3: Macro F1 scores by language version between the benchmark and each BERT model.

Table 3 presents class-wise F1 scores for the sentence-level sentiment classification task, computed separately for each language and model.

6 Discussion

6.1 Sentiment Distribution

The analysis of Figure 1 shows a clear pattern across all five languages in the benchmark set – neutral sentiment dominates in nearly all topical categories. In most science-, technology-, and knowledge-oriented domains, the proportion of neutral labels exceeds 90%, which is in line with the encyclopedic and normatively neutral character of Wikipedia. At the same time, the benchmark reveals meaningful topical variation. Categories related to biographies are substantially less neutral and contain considerably higher proportions of positive labels - positive sentiment often reaches or exceeds 45%, and in some cases surpasses 50%. An-

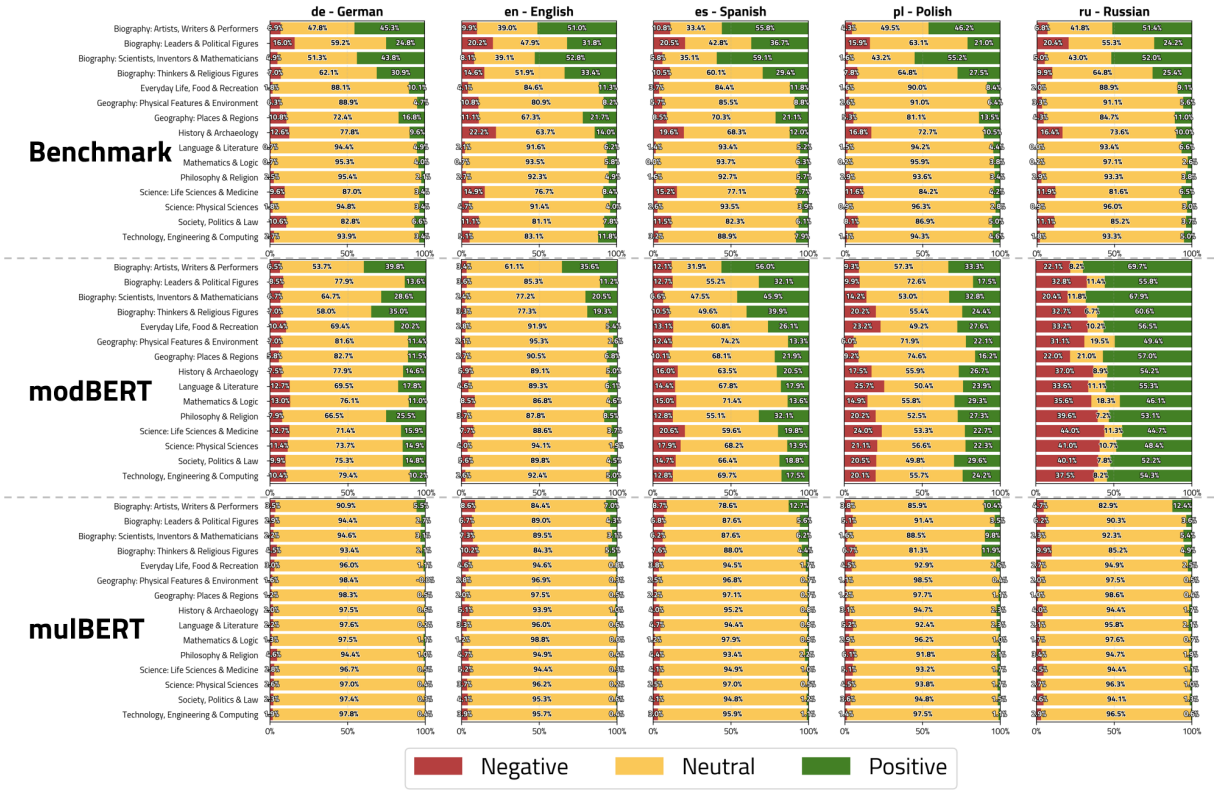

Figure 1: Sentiment label distribution across languages in sentences for benchmark, modBERT and mulBERT.

other important pattern concerns geography-related topics - a relatively high proportion of positive labels across all languages, reaching 21.7% in English and 21.1% in Spanish. This topic appears more positively framed. Nevertheless, neutral labels remain dominant in all languages.

The results for modBERT differ substantially from the benchmark distributions. In general, modBERT produces far fewer neutral and more polarized predictions, although the direction and magnitude of this effect vary considerably across languages. The most extreme behavior is observed for Russian. In this language, modBERT produces highly polarized outputs in nearly all categories, with neutrality often collapsing to very low values. This suggests severe language-specific instability or miscalibration in the Russian setting.

In contrast to modBERT, mulBERT produces a much more uniform prediction pattern across languages and categories. It shows a very strong preference for the neutral label. As a result, mulBERT preserves the overall neutral orientation of Wikipedia lead sections, but it also appears to suppress much of the topic-specific variation present in the benchmark. Unlike modBERT, mulBERT behaves consistently across all five languages and tends to treat Wikipedia lead sections as overwhelmingly neutral regardless of topic.

6.2 Sentence-Level Agreement Analysis

The results presented in Table 2 also reveal several consistent patterns in agreement between the benchmark evaluators and the BERT models. At the sentence level, mulBERT generally achieves higher percentage agreement with the benchmark than modBERT across most languages. This indicates that it aligns more closely with benchmark annotations at the level of individual sentences.

Agreement varies substantially across languages. For modBERT, the highest agreement is observed for English, with relatively strong results also for German and Spanish, and lower agreement for Polish and Russian. Similar pattern is visible also when analysing Macro F1 values. This suggests that modBERT performs better for high-resource languages, while its performance degrades for languages that may be less represented in fine-tuning data. In contrast, mulBERT exhibits a different pattern, achieving higher agreement for German, Polish, and Russian, while showing comparatively lower agreement for English and Spanish. This may reflect the effect of multilingual train-

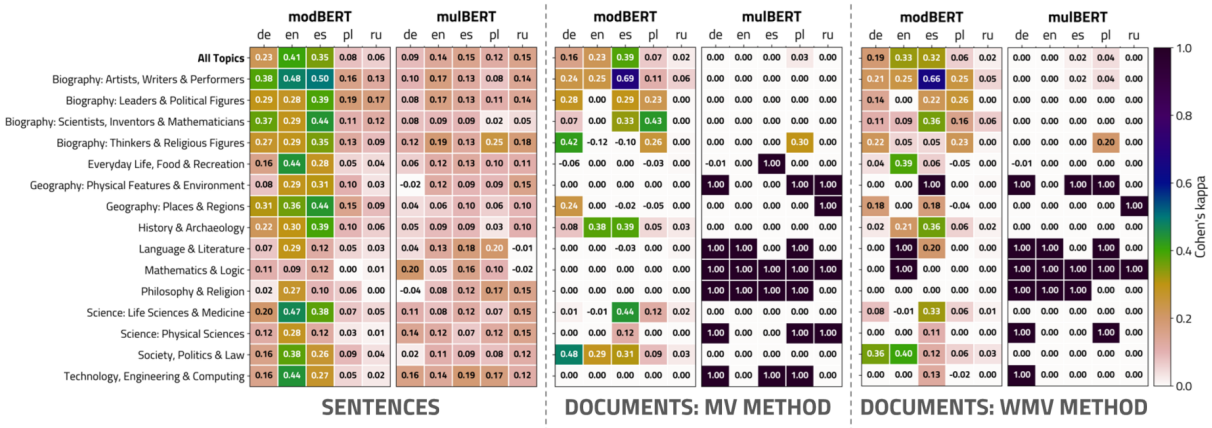

Figure 2: Cohen's κ between benchmark dataset and BERT-based models (modBERT, mulBERT) across topics and languages at sentence and document levels.

ing (23 languageas in contrast to 16 languages for modBERT), which promotes more balanced cross-lingual performance.

These observations are reflected in Cohen's κ values, which remain relatively low across languages, indicating limited agreement beyond chance. A more detailed view of these patterns is provided in Figure 2. Overall, in all topics, modBERT achieves moderate agreement with the benchmark in English ($\kappa = 0.41$) and Spanish ($\kappa = 0.35$), and lower but still positive agreement in German (0.23). In contrast, agreement remains weak in Polish (0.08) and Russian (0.06). Such a low level of agreement for Polish may stem from the lack of adequate training data, as the dataset did not include the Polish language. The case of Russian is somewhat more complex, as Russian was among the most highly represented languages in the dataset.

In case of mulBERT, a consistently low agreement is observed across all languages, with κ values ranging from 0.09 (German) to 0.15 (Russian). The strongest agreement between modBERT and the benchmark is observed in biography-related categories, where sentiment is more explicitly expressed. Overall, agreement is highest for English and Spanish, while Polish and Russian remain the most challenging languages. For mulBERT, agreement remains more uniform but consistently low across topics and languages. Most values fall between 0.08 and 0.15 across languages and categories. In general, mulBERT fails to capture the benchmark's sentiment structure, particularly in categories where non-neutral sentiment is more prominent.

The values of F1 metric (Table 3) show that modBERT obtains higher F1 scores for the positive class in all five languages, with the strongest results for English and Spanish (0.52 in both cases). For the negative class, modBERT also performs better in German, English, and Spanish, whereas mulBERT is slightly stronger in Polish and Russian.

In case of neutral label, we can observe a different patterns. For instance, mulBERT achieves higher F1 scores in German, Spanish, Polish, and Russian, reaching 0.90 in several languages. English is the only language where modBERT obtains a slightly higher neutral F1 score (0.89 compared with 0.86 for mulBERT). Overall, the comparison again suggests that modBERT is more effective for detecting polar sentiment, especially positive sentiment, while mulBERT tends to favor the neutral class.

6.3 Document-Level Agreement Analysis

We now turn to the document-level analysis, where sentiment labels are aggregated at the level of Wikipedia lead sections. This setting reflects a more practical evaluation scenario but introduces additional complexity, as sentence-level predictions are combined into a single document-level representation.

When analysing results in Table 2, similar trends can be observed as for the sentences, although agreement values generally increase after aggregation. This pattern persists at the document level, where mulBERT also shows higher percentage agreement with the benchmark than modBERT.

Also language-specific differences remain visible. For modBERT, the highest agreement is achieved for English, while performance for Pol-

ish and Russian remains substantially lower. For mulBERT, again relatively stronger agreement is visible for German, Polish, and Russian, resulting in a similar agreement values across all languages independent of the aggregation method.

Cohen's κ values at the document level follow patterns similar to those observed at the sentence level. The persistence of near-zero values for mulBERT suggests that aggregation on the document-level does not fully resolve discrepancies in label distributions between model predictions and the benchmark. Despite higher percentage agreement, mulBERT shows substantially lower agreement beyond chance, particularly at the document level.

The comparison of aggregation methods indicates that MV method generally yields slightly higher agreement values than WMV method, particularly for modBERT. However, differences between methods are relatively small, suggesting that both aggregation strategies produce comparable document-level representations. These effects are further illustrated in Figure 2, which presents Cohen's κ values also at the document level. If we look at aggregation using MV method, agreement with the benchmark generally decreases for both models, with the effect being particularly pronounced for mulBERT. While modBERT retains moderate agreement in some languages (e.g., English and Spanish), many topic-language combinations show κ values close to zero or even negative. For mulBERT, performance is especially poor at the document level, with κ values often equal to zero across topics and languages. This reflects its strong bias toward neutral predictions. An additional characteristic of document-level results is the occurrence of extreme κ values (e.g., $\kappa = 1.0$ in some cases). These are typically associated with subsets dominated by a single label, most often neutral, and therefore do not necessarily reflect meaningful agreement but rather trivial consistency in homogeneous subsets.

7 Conclusion

This paper investigated the use of a multilingual, multi-domain benchmark dataset derived from LLM consensus for evaluating sentiment analysis models. In particular, we examined how smaller, fine-tuned BERT-based models perform in comparison to this benchmark, and how sentence-level predictions can be aggregated into document-level sentiment representations. The evaluation framework allowed us to analyze model behavior across languages, topics, and levels of aggregation, providing a comprehensive view of their strengths and limitations.

The two evaluated models exhibit markedly different error profiles relative to the benchmark. These results suggest that neither model fully reproduces the benchmark distribution. After performing the Cohen's κ analysis, we observe that modBERT is generally closer to the benchmark than mulBERT, but still does not achieve robust agreement across all topic-language subsets. Other model, mulBERT appears more "conservative" but less faithful to the benchmark, because its strong neutral bias leads to near-zero agreement in most non-trivial settings. As a result, neither model can be regarded as a fully reliable substitute for the benchmark. The comparison between sentence-level and document-level agreement reveals an important pattern: aggregation amplifies model biases. While sentence-level evaluation captures partial agreement and local alignment with the benchmark, document-level aggregation tends to smooth out this variation and emphasize systematic deviations, particularly for models with strong class biases. In particular, mulBERT's tendency toward neutral predictions leads to near-zero agreement at the document level, even when some alignment is present at the sentence level.

These findings highlight the importance of carefully designed benchmarks and aggregation strategies for robust multilingual sentiment evaluation.

Ethics Statement

This work involves the automated annotation and analysis of publicly available Wikipedia content. As such, no personal or sensitive user data were collected or processed, and the study does not involve human subjects or private information.

All textual data used in this study are derived from Wikipedia, which is openly accessible and distributed under permissive licenses. The dataset construction follows standard research practices for using publicly available corpora.

The sentiment annotations in this work are produced automatically using large language models and machine learning classifiers. Such models may reflect biases present in their training data, including cultural, linguistic, or societal biases. Consequently, the resulting annotations may not be fully neutral or objective, particularly in sensitive

domains such as politics, history, or biographies. These limitations should be considered when interpreting the results or using the dataset in downstream applications.

The released dataset is intended for research purposes, particularly for the evaluation of sentiment analysis methods and language models. Users of the dataset should be aware of its limitations and avoid using it in contexts where misinterpretation of sentiment could lead to harm, such as decision-making systems affecting individuals or groups.

Limitations

Despite the contributions of this work, several limitations should be acknowledged.

First of all, the benchmark lacks an external ground truth beyond LLM consensus. The dataset is constructed based on strict agreement across evaluators, which provides high-confidence labels but does not necessarily represent an objective ground truth. In particular, sentiment in encyclopedic text is inherently subtle and context-dependent, and even full agreement does not guarantee correctness. As a result, the benchmark should be interpreted as a reliable reference rather than an absolute standard. Furthermore, the use of a strict consensus criterion reflects a deliberate trade-off between reliability and coverage. While this approach prioritizes label consistency and reduces the influence of model-specific biases and stochastic variation, it may also exclude more ambiguous or nuanced instances that are inherently more difficult to classify. In particular, the requirement of unanimous agreement across all nine LLM runs may lead to filtering out borderline cases, which are often the most informative and challenging in encyclopedic text, where sentiment is frequently expressed through subtle framing rather than explicit polarity markers. Consequently, the resulting dataset might be biased toward clearer sentiment expressions, while more nuanced, difficult, or contested cases are underrepresented. This filtering may also affect the evaluation results. Models are assessed primarily on sentences with high annotator agreement, which may lead to an overestimation of performance in simpler cases and does not fully reflect their behavior on more challenging inputs. In particular, disagreement patterns between models may be more pronounced in the excluded subset, which is not captured in the current benchmark. Future work could address this limitation by incorporating alternative annotation strategies, such as soft-labeling approaches, relaxed consensus thresholds, or human-annotated subsets, to better capture the full spectrum of sentiment variation.

Another limitation - label granularity. We adopt a coarse-grained three-class sentiment scheme (*negative*, *neutral*, *positive*). While this simplifies analysis and enables consistent comparisons across models and languages, it may obscure more nuanced distinctions such as weak polarity, mixed sentiment, or factual framing with implicit evaluative cues. This limitation is particularly relevant in the context of Wikipedia, where sentiment is often implicit rather than explicitly expressed.

The study focuses exclusively on Wikipedia lead sections, which are designed to follow a neutral point of view (NPOV). While this makes the dataset suitable for studying low-subjectivity text, it also limits the generalizability of the findings to other domains, such as opinion-rich content (e.g., reviews or social media), where sentiment is more explicit and diverse.

Although the analysis includes five languages (German, English, Spanish, Polish, and Russian), these represent only a subset of Wikipedia's linguistic diversity. The observed cross-lingual differences may not generalize to other languages, especially those with different typological properties, writing systems, or editorial conventions.

We evaluate a limited set of models, including three large language models and two multilingual BERT-based classifiers. While these models are representative of current approaches, the results may not extend to other architectures, especially newer or specialized sentiment models. In addition, LLM outputs depend on prompting strategies, which may influence both consistency and agreement.

Although we mitigate variability by querying LLMs multiple times per sentence, the results remain sensitive to prompt design and stochastic generation. The chosen prompting strategy may affect both label distributions and intra-model consistency, and alternative prompts could lead to different outcomes.

We rely primarily on agreement-based metrics such as Cohen's κ. While these metrics are widely used, they have known limitations, particularly in settings with class imbalance or low label diversity, which are common in neutral-dominated datasets. Consequently, agreement scores should be interpreted with caution (especially in cases with

$\kappa = 1.0$, where models returns only the "neutral" label for each sentence).

The study highlights differences between sentence-level and document-level evaluation, but the aggregation methods used for document-level sentiment may influence the results. Alternative aggregation strategies could yield different agreement patterns and should be explored in future work.

The benchmark dataset is based on curated lists of important Wikipedia articles. While this ensures broad topical coverage, it may introduce selection bias toward well-developed or widely edited articles, which could differ systematically from less prominent content.

Each article is assigned to a single high-level topic, which simplifies analysis but may not fully capture the multidimensional nature of many Wikipedia entries. Some articles span multiple domains, and this simplification may affect the interpretation of topic-specific results. In future work, we also plan to increase the number of topics at different levels of generalization, including science fiction and fantasy (Lewoniewski et al., 2026a).

While we analyze agreement, distributional differences, and consistency, the study does not provide a detailed explanation of why models behave differently across languages and topics. Understanding the underlying causes of these differences, including linguistic, cultural, or training data factors, remains an open challenge.

The use of LLMs as annotators introduces an additional layer of uncertainty. Although multiple runs are used to estimate stability, LLMs may exhibit hidden biases, inconsistencies, or sensitivity to subtle input variations. Their behavior as evaluators should therefore be interpreted critically.

While we provide datasets and methodological details, exact replication of LLM-based annotations may be affected by changes in model versions, API behavior, or underlying system updates over time.

References

Emmy Danny Ajik, Aminu Bashir Suleiman, and Muhammad Ibrahim. 2023. Enhancing user experience through sentiment analysis for katsina state transport agency: A textblob approach. *FUDMA Journal of Sciences*, 7(6):117–122.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Hadis Bashiri and Hassan Naderi. 2024. Comprehensive review and comparative analysis of transformer models in sentiment analysis. *Knowledge and Information Systems*, 66(12):7305–7361.

Anton Borg and Martin Boldt. 2020. Using vader sentiment and svm for predicting customer response sentiment. *Expert Systems with Applications*, 162:113746.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Himanshu Gautam, Abhishek Gaur, and Dharmendra Kumar Yadav. 2025. A survey on the impact of pre-trained language models in sentiment classification task. *International Journal of Data Science and Analytics*, 20(6):5197–5235.

K Puttaswamy Gowda, R Porwal, C Ramesh, SS Tiwari, K Srivastava, R Rambabu, and SG Rao. 2025. Transformers in sentiment analysis: A paradigm shift in deep learning research. *J. Inf. Syst. Eng. Manag.*, 10(5):2468–4376.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Włodzimierz Lewoniewski, Milena Stróżyna, Izabela Czumałowska, and Elżbieta Lewańska. 2026a. Science fiction and fantasy in wikipedia: Exploring structural and semantic cues. *arXiv preprint arXiv:2602.24229*.

Włodzimierz Lewoniewski, Milena Stróżyna, Izabela Czumałowska, Aleksandra Wojewoda, and Krzysztof Węcel. 2026b. Cross-Topic Sentiment Analysis of Wikipedia Articles: A Comparative Study of AI Models. In *Artificial Intelligence for Knowledge Acquisition and Management. AI4KAM 2025. IFIP Advances in Information and Communication Technology*, volume 776, Cham. Springer Nature Switzerland.

Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36.

Sorin Adam Matei and Caius Dobrescu. 2011. Wikipedia's "neutral point of view": Settling conflict through ambiguity. *The Information Society*, 27(1):40–51.

Milena Stróżyna, Włodzimierz Lewoniewski, Izabela Czumałowska, and Aleksandra Wojewoda. 2025. Sentiment analysis of wikipedia articles about companies: A comparison of different models. In *International Conference on Business Information Systems*, pages 101–115. Springer.

Madhumita Sushil, Travis Zack, Divneet Mandair, Zhiwei Zheng, Ahmed Wali, Yan-Ning Yu, Yuwei Quan, Dmytro S. Lituiev, and Atul J. Butte. 2024. A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports. *Journal of the American Medical Informatics Association*, 31(10):2315–2327.

Ziko Van Dijk. 2009. Wikipedia and lesser-resourced languages. *Language Problems and Language Planning*, 33(3):234–250.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot ner with chatgpt. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7935–7956.

A Appendix: Prompt Used for LLM Evaluators

Role: You are an expert linguistic analyst specializing in sentiment analysis and Natural Language Processing.

Task: Perform high-precision sentiment analysis on multilingual datasets and convert structured JSON input into a raw TSV (Tab-Separated Values) format.

Instructions:

1. **Sensitivity:** Be highly sensitive to subtle nuances, tone shifts, and evaluative language. Detect underlying bias or descriptive framing (positive/negative).

2. **Evaluation Scale:** Categorize sentiment as "positive", "negative", or "neutral".

3. **Handling Edge Cases:** If the text is empty, contains only non-interpretable data (numbers/tables), or is insufficient for analysis, or you cannot decide, return "n/a" for sentiment.

Input Structure: A JSON array of objects, each containing "id" (string) and "text" (string). All articles are provided in a single message.

Output Requirements (Strict):

- Format: Output ONLY the result in a flat TSV (Tab-Separated Values) structure.

- Do not include any header row.

- Structure: Each line must contain exactly: [id][TAB][r] (where r is ONLY "pos", "neu", "neg", or "n/a").

- Separation: Each record must be on a new line.

- No Markdown: Do not wrap the output in markdown code blocks or any other formatting. Return raw text only.

Output Schema per line: Return ONLY a valid TSV object.

[id] [r]

(where [r] is "pos" | "neu" | "neg" | "n/a")