Multi-Model Framework for Sentiment Annotation: A Use Case of Multilingual Wikipedia

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

Abstract

In this work, we investigate whether data annotation can be automated using state-of-the-art large language models (LLMs), while maintaining high consistency and reliability. We propose a multi-model, multi-run framework in which several LLMs act as annotators and their outputs are aggregated using a strict consensus criterion to improve annotation robustness and reduce model-specific biases.

The proposed approach is evaluated on a multilingual sentiment analysis task using lead sections of Wikipedia articles covering diverse topics in five languages. The resulting annotations are assessed using intra-model consistency, cross-model agreement, and standard inter-annotator agreement metrics. Additionally, we conducted human evaluation.

The described framework enables scalable and reproducible data annotation, and the resulting dataset is publicly released to support further research.

1 Introduction

The rapid development of machine learning and artificial intelligence has led to an increasing demand for large-scale annotated datasets. However, manual annotation is time-consuming, costly, and difficult to expand. Although recent approaches have explored hybrid human-model annotation workflows, these methods still require significant human involvement and may not fully address scalability constraints. This motivates the need for automated annotation frameworks that can provide reliable labels with minimal human intervention while maintaining reproducibility and annotation consistency across languages and domains.

In this work, we propose a structured framework for automated annotation based on a multi-model, multi-run approach, in which multiple large language models (LLMs) act as annotators and their outputs are aggregated using a strict consensus criterion. We investigate whether the data annotation process can be automated using state-of-the-art LLMs, while maintaining high consistency of the resulting labels. Rather than treating individual LLM outputs as objective ground truth, we interpret cross-model consensus as a high-stability proxy for annotation reliability.

We apply our approach to multilingual Wikipedia text. Sentiment analysis was selected as a representative and challenging task, particularly in the context of Wikipedia, where maintaining a *neutral point of view* (NPOV) is a fundamental principle (English Wikipedia, 2026; Johnson et al., 2025). As one of the most widely used sources of general knowledge, Wikipedia requires high standards of neutrality. Wikipedia is also an open and continuously evolving knowledge resource, with articles regularly updated by a large community of contributors and available in many languages. This makes it a suitable setting for evaluating scalable and multilingual approaches to text analysis. Although sentiment analysis does not fully capture NPOV (which also involves source selection, due weight, and balanced representation of viewpoints), it provides a useful computational proxy for studying sentiment and tone-related neutrality in encyclopedic text. Importantly, sentiment labels in this study do not refer exclusively to subjective opinions or explicitly expressed attitudes. Rather, they represent the overall sentiment conveyed by the text, including the evaluative or affective tone that may arise when factual events, entities, or situations are described. In particular, systematically positive or negative sentiment may reflect evaluative or emotionally loaded language and thus provide a signal for identifying potential deviations from neutral presentation.

Wikipedia provides a challenging setting for sentiment analysis because its encyclopedic content is

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39079–39090
October 24-29, 2026 ©2026 Association for Computational Linguistics

predominantly factual and intended to avoid subjective or emotionally charged language. Consequently, sentiment is often subtle and may be expressed through nuanced wording or context-dependent cues rather than explicit polarity markers. At the same time, articles describing events such as wars or disasters may naturally contain negative sentiment without violating the NPOV principle. This makes Wikipedia a useful stress test for sentiment analysis models, requiring them to distinguish the sentiment conveyed by the text from the subjective intent of its authors.

In contrast to previous works (Al Khatib et al., 2012; Zhou et al., 2015; Stróżyna et al., 2025; Lewoniewski et al., 2026), this study explores the use of modern generative LLMs and extends the analysis to a multilingual setting. The goal of this work is therefore not to establish an objective ground truth for sentiment analysis, but to investigate whether agreement-based LLM annotation can provide sufficiently reliable labels to support multilingual sentiment analysis research at scale. The broad topical coverage and availability of Wikipedia across multiple languages further enable the assessment of the cross-lingual robustness of the proposed approach. Moreover, unlike many sentiment analysis datasets, which are primarily based on user reviews and comments (e.g., social media), our dataset consists of encyclopedic content spanning diverse domains. This allows us to evaluate the proposed annotation framework in a more general and challenging setting.

This paper makes the following contributions:

1. A consensus-based framework for automatic multilingual sentiment annotation using multiple state-of-the-art LLMs, designed to improve reliability through consensus-based aggregation;
2. The framework on a multilingual sentiment analysis task using Wikipedia data and assess annotation reliability using agreement metrics and comparison with human annotations;
3. Release of an annotated dataset to support further research and ensure reproducibility.

2 Related Work

2.1 Data Annotation Techniques

Data annotation plays a fundamental role in the development of machine learning systems, as the quality of labeled data directly impacts model performance. However, data annotation, especially manual, which is still considered a gold standard for quality, remains one of the most costly and time-consuming parts of the machine learning process (Alenzi et al., 2022). Additionally, the process of manual data annotation can be subjective, inconsistent, and may require domain expertise (Qureshi et al., 2022; Tan et al., 2024).

Recent advances in LLMs have introduced new possibilities for automating and supporting data annotation with the potential to significantly reduce costs while maintaining competitive performance and mitigating the challenges encountered with traditional annotation (Tan et al., 2024). Studies have shown that semi-automated and human-in-the-loop (HITL) annotation frameworks are commonly used to support annotation tasks (Alahmary and Al-Dossari, 2021; Quilio and Fioravanti, 2024; Wu et al., 2022). Such approaches combine machine learning models with human expertise in order to improve annotation efficiency and quality. Various tools have been applied within these frameworks, including SentiWordNet, BabelNet (Navigli et al., 2021), TextBlob, VADER (Alenzi et al., 2022) and also methods using such algorithms as Naive Bayes, CNN, LSTM, and SVM (Alahmary and Al-Dossari, 2021). Recent works on sentiment analysis of Wikipedia texts (Stróżyna et al., 2025; Lewoniewski et al., 2026) showed how smaller BERT-based and lexicon-based models can be used in similar task. However, these approaches still require substantial human involvement, limiting scalability and reproducibility, while remaining susceptible to annotation inconsistency (Mosqueira-Rey et al., 2023).

A less extensively explored direction is fully automated annotation approach, which relies on computational methods to label data without direct human involvement. One common strategy is to use language models to generate so-called "silver-standard" annotations (Liao et al., 2024), which are subsequently reviewed and refined by human experts to create gold-standard datasets (Akbaş et al., 2025).

In contrast to these approaches, our work investigates whether a consensus-based framework using multiple state-of-the-art LLMs can itself provide sufficiently reliable multilingual sentiment annotations without requiring manual correction at the annotation stage. Rather than using LLM outputs only as an intermediate step toward human annotation, we evaluate their agreement and consistency as a scalable alternative to manual labeling.

While fully automated methods enable rapid and scalable annotation, their quality depends strongly on the underlying models and may be less robust for complex or ambiguous cases. Nevertheless, recent studies suggest that modern LLM, such as GPT, can achieve annotation quality comparable to human annotators while providing greater consistency (Wu et al., 2022).

2.2 General-purpose LLMs for Sentiment Analysis

With the emergence of generative LLMs, such as the GPT family developed by OpenAI, Claude models from Anthropic, Gemini by Google, or open-weight models, these architectures have increasingly been adopted in NLP research and have shown strong potential for sentiment analysis (Gautam et al., 2025). Their advantage lies in processing texts holistically, capturing contextual and linguistic nuances important for sentiment classification.

Studies show that modern generative LLMs, including GPT- and LLaMA-based architectures, often match or surpass traditional supervised models in zero-shot and few-shot sentiment analysis settings (Almontashry et al., 2025). Recent work further indicates that advanced models, such as GPT-class systems, can achieve near state-of-the-art performance even without task-specific fine-tuning (Elmitwalli and Mehegan, 2024). The primary method for applying modern LLMs in sentiment analysis is prompt engineering, where models are instructed to classify text without retraining. Since performance is highly sensitive to prompt structure, multiple prompt variants are often tested to optimize accuracy (Kang and Choi, 2025).

Overall, recent studies support the suitability of generative LLMs as robust and versatile tools for sentiment analysis, motivating their use as annotators in automated dataset construction.

3 Dataset Description

To construct a multilingual dataset of encyclopedic text, we selected a set of high-importance Wikipedia articles from different domains based on two community curated lists (as of March 2026):

1. Wikipedia:Vital articles/Level/3[1]
2. List of articles every Wikipedia should have[2]

The first list contains approximately 1,000 articles considered essential for the English Wikipedia and is used to identify the most important topics across domains.

The second list is a complementary resource maintained on Meta-Wikimedia. It defines a set of 1,000 Wikidata items intended to serve as a minimal core of knowledge for all language editions of Wikipedia. Unlike the "Vital Articles" list, this resource is designed to support multilingual alignment by linking each concept to corresponding articles across languages via Wikidata identifiers.

We extracted article titles from both lists using their English Wikipedia (en) counterparts and removed duplicates, resulting in 1,394 unique titles. For each selected title, we additionally retrieved corresponding articles in four languages: German (de), Spanish (es), Polish (pl), Russian (ru). Cross-lingual mapping was performed using Wikipedia interlanguage links. It is important to note, that the resulting dataset is not strictly parallel, because Wikipedia articles are not guaranteed to be exact translations.

For each selected Wikipedia article in each considered language version, we extracted only the text from the lead section. This design choice was made to avoid sampling unrelated or highly heterogeneous article fragments and to focus on concise summaries of each topic. Since Wikipedia lead sections are expected to summarize the subject in neutral encyclopedic style, they are particularly suitable for studying tone-related aspects of neutrality.

3.1 Sentence Segmentation

To prepare the data for large-scale annotation using LLM APIs, we implemented a three-stage pipeline for sentence segmentation, normalization, and batching. The goal was to obtain linguistically meaningful sentence units suitable for reliable automated annotation, while ensuring robustness to Wikipedia-specific formatting artifacts.

In the first stage, we performed sentence segmentation using the Stanza NLP toolkit[3]. For each language, we initialized a language-specific tokenization pipeline.

Each detected sentence in the leading section was assigned a unique identifier composed of the Wikipedia page ID and a sentence index ("PageID-SentenceNumber"). Empty or whitespace-only sen-

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

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

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

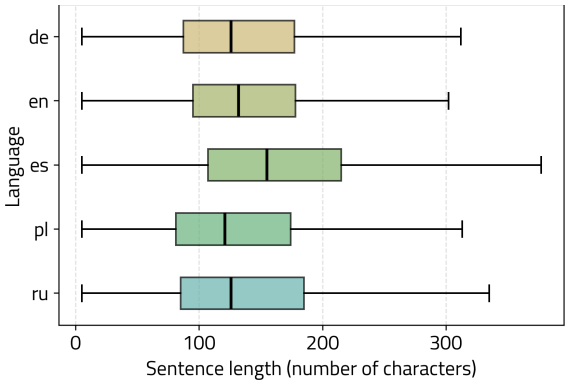

Figure 1: Sentence lengths across languages.

tences were discarded. The output of this stage was a collection of sentences stored in JSONL format.

The second stage addressed noise and segmentation artifacts specific to Wikipedia content. In particular, we observed that raw segmentation often produces: fragments containing LaTeX-style expressions, very short sentence fragments (e.g., 1–2 tokens), formatting artifacts such as inline formulas or markup remnants. To mitigate these issues, we applied the following transformations:

1. Removal of brace-enclosed content: text within curly braces (e.g., `{...}`).
2. Removal of common LaTeX-like tokens (e.g., "`\displaystyle`") were removed.
3. Whitespace normalization by collapsing multiple spaces and tab characters.
4. Sentences of length 0 or 1 character and sentences containing fewer than 3 tokens were treated as short fragments.
5. Merging short fragments with subsequent sentences to form more coherent units.

After extraction and cleaning, the dataset contained the following number of sentences:

- German (de): 12,715 sentences
- English (en): 21,991 sentences
- Spanish (es): 17,021 sentences
- Polish (pl): 10,707 sentences
- Russian (ru): 12,739 sentences

Figure 1 presents the distribution of sentence lengths across languages, while Figure 2 illustrates the relative share of analyzed sentences across topics in the selected Wikipedia language versions. Overall, topic distributions are relatively similar across languages, with differences mainly resulting from variations in lead section length between Wikipedia language versions.

In the final stage, sentences were grouped into subsets suitable for submission to LLM APIs. Each sentence was formatted as JSON lines: `{"id": sentence_id, "text": sentence_text}`. Subsets were constructed to not exceed 25,000 tokens. Sentences were grouped sequentially according to their original order in the dataset, meaning that sentences originating from the same Wikipedia lead section were typically processed together within the same API request. However, the prompt itself did not explicitly indicate, that the sentences originated from the same article. Each language was processed independently and constituted a separate collection of subsets.

4 Annotation Procedure

To evaluate sentiment in the prepared datasets, each sentence was annotated using general-purpose LLMs. The sentences collections described in Section 3 served as input corpora, while sentiment labels were generated through a controlled multi-model prompting procedure described below.

To ensure methodological rigor and minimize model-specific bias, sentiment annotation was performed using three state-of-the-art general-purpose LLMs: GPT-5.2 (OpenAI), Gemini 3.1 Pro (Google), and Claude Opus 4.6 (Anthropic). At the time of the study, these three models ranked among the top-performing systems for text-based tasks on the LMarena.ai leaderboard[4], which justified their selection as high-capacity evaluators. As discussed in Section 2, modern generative LLMs demonstrate strong zero-shot sentiment classification capabilities across domains and languages, and this approach was adopted in our study.

All models were accessed via their respective APIs using batch processing, enabling consistent and scalable execution of the annotation procedure. Each batch corresponded to a single language (de, en, es, pl, ru) and contained multiple requests. Due to limitations in the GPT API, two batches were required for en, es, and ru, and one batch for pl and de. Each individual request included up to 500 sentences submitted within a single prompt in order to comply with the token limits imposed by the APIs. Consequently, sentiment labels were generated for multiple sentences simultaneously within the same input context. Each sentence within a request was assigned an individual sentiment label by the model.

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

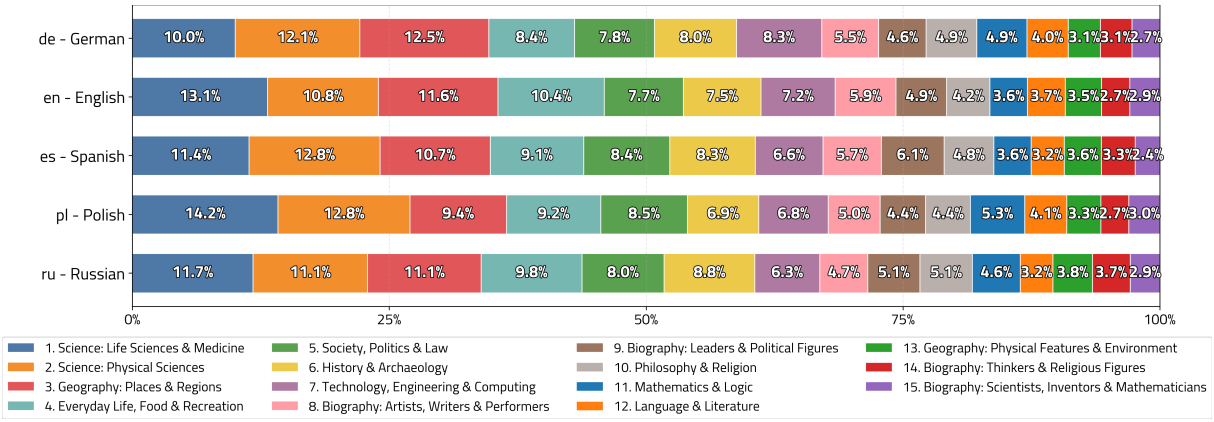

Figure 2: Distribution of analyzed sentences across topics in the selected Wikipedia language versions.

4.1 Annotation Execution

Sentiment was classified using a four-class scheme (positive, neutral, negative, n/a), where 'n/a' is assigned in cases where the text is empty, contains non-interpretable content (e.g., numbers or tables), is insufficient for reliable analysis, or the model is unable to determine the sentiment.

All three LLMs were prompted using the same instruction to ensure consistency across models (see Fig. 3). In the API calls, we specified the maximum number of output tokens[5], model identifiers ('gpt-5.2', 'gemini-3.1-pro-preview', 'claude-opus-4-6'), inference level (set only for Gemini to 'low', as this parameter was mandatory), and instructions in English to maintain a consistent prompting strategy across languages. We did not modify the temperature parameter, as not all APIs provided the option to control this setting. To ensure methodological consistency across all evaluated models, we conducted all experiments using their default configurations. For Opus, the instructions (see Fig. 3) were provided as a system prompt and input data as a user message, whereas for GPT and Gemini the same instructions were included directly within the user message together with the input data.

To ensure structural consistency and facilitate automated post-processing, strict output constraints were imposed. The models were instructed to return only valid TSV-formatted outputs, one sentence per line, following the schema: `[id]\t[r]`, where `[r]` denotes the sentiment label.

Prior to the main annotation process, a pilot run was conducted on smaller subsets for each language. These subsets were used exclusively to validate the initial prompt and verify whether the models returned outputs consistent with the expected format and label set. During this stage, we observed that GPT occasionally produced labels outside the predefined schema (e.g., "mix" or "mixed"), which led to a refinement of the prompt by explicitly instructing the model to return 'n/a' in cases of uncertainty. The pilot subsets were not included in the final annotation process or in the final dataset, ensuring that prompt refinement was performed independently.

The final prompting procedure was repeated three times per model, resulting in nine independent annotation runs (3 models × 3 repetitions) for each language. This resulted in a multi-layer annotation structure, where each sentence has multiple sentiment annotations from different annotators. This setup enabled us to analyse annotation stability both intra-model consistency and inter-model agreement of the assigned sentiment labels.

4.2 Agreement Analysis

Several metrics were used to analyze annotation quality separately for each language. First, intra-model consistency was evaluated by measuring whether each model assigned identical sentiment labels to the same sentence across its three independent runs (see Table 1). Next, a strict cross-model agreement criterion was applied. A label was considered valid only if all three models assigned the identical sentiment in all three repetitions, i.e., unanimous agreement across all nine runs (see Cross-Model Agreement in Table 1).

Finally, the inter-model agreement was measured using standard inter-annotator agreement metrics: Cohen's κ (Cohen, 1960), that measures

[5] 65536, corresponding to the maximum limit supported by the Gemini API and selected as a common value compatible with all evaluated models; this parameter was explicitly required only for Opus.

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")

Figure 3: Prompt used for sentiment classification.

pairwise agreement between two annotators while correcting for chance agreement, and Fleiss' κ (Fleiss, 1971), that extends Cohen's κ to multiple annotators. Fleiss' κ was computed across all models and runs, separately for each language. For Cohen's κ was calculated pairwise, resulting in a matrix of agreement scores for each language.

For Cohen's κ and Fleiss' κ, we followed interpretation ranges commonly used in NLP and sentiment analysis research (Artstein and Poesio, 2008; Hassan et al., 2021; Guzmán-Monteza, 2023), where values above 0.60 generally indicate substantial agreement, and values above 0.70 indicate strong annotation consistency.

4.3 Human Evaluation

To verify the reliability of the proposed LLM-based annotation framework, an additional manual annotation of a subset of the data was conducted. Texts in both Polish and English were selected for the evaluation. For the Polish dataset, 10% of the corpus was sampled (1,071 sentences), while for the English dataset, 1% was selected (220 sentences). Sentences were sampled randomly without preserving the original article context, ensuring that human annotations were assigned independently of the broader text.

The annotation process involved three human annotators. Each annotator assigned every sentence to one of three classes: positive, negative, or neutral. If a sentence was deemed insufficient for sentiment determination, the 'n/a' label was assigned, consistent with the prompting strategy applied to the LLMs evaluators. When all human annotators assigned the same class, that label was retained as the final annotation. Otherwise, the majority label was selected. This set is further referred to as test set.

4.4 Dataset Release

The dataset with 75,173 annotated sentences was released publicly on Hugging Face[6]. Each instance corresponds to a single sentence extracted from a Wikipedia lead section and includes sentiment annotations, agreement information, and metadata describing the source article and language. For downstream applications, we recommend using unanimously classified instances, as the strict agreement criterion maximizes label reliability and reduces the influence of model-specific biases and stochastic variation.

5 Results and Discussion

Figure 4 shows the distribution of sentiment labels across languages for each LLM annotator and run. The results indicate that neutral sentiment is dominant across all models and languages, what is expected for vital Wikipedia articles that follow the Neutral Point of View principle. For English (en) and Spanish (es), a slightly higher proportion of both positive and negative labels is observed.

5.1 Annotation Agreement

Table 1 summarizes agreement metrics across models and languages. The analysis of intra-model

[6] https://huggingface.co/datasets/lewoniewski/wikipedia-sentences-sentiment

Table 1: Intra-model and cross-model agreement for sentiment classification across languages.

Language	Gemini 3.1 Pro	GPT 5.2	Opus 4.6	Strict Cross-Model Agreement	Fleiss' κ
German (de)	11,052 (86.9%)	11,799 (92.8%)	12,440 (97.8%)	9,571 (75.3%)	0.74
English (en)	18,592 (84.5%)	20,318 (92.4%)	21,266 (96.7%)	15,910 (72.3%)	0.76
Spanish (es)	14,740 (86.6%)	15,699 (92.2%)	16,575 (97.4%)	12,097 (71.1%)	0.74
Polish (pl)	9,153 (85.5%)	10,006 (93.5%)	10,458 (97.7%)	7,950 (74.3%)	0.72
Russian (ru)	11,164 (87.6%)	11,843 (93.0%)	12,438 (97.6%)	9,635 (75.6%)	0.75

For each model and for the strict cross-model agreement, values represent the absolute value and percentage of instances with identical sentiment labels across repeated runs.

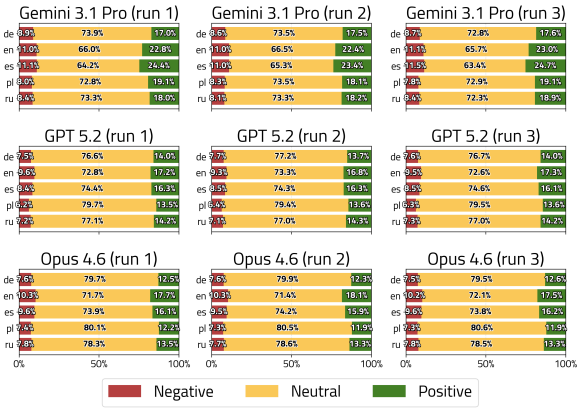

Figure 4: Sentiment label distribution across languages for each LLM annotator and run.

consistency shows that Claude Opus achieved the highest stability (approx. 97%), indicating a high level of stability in its predictions. GPT models also demonstrated strong consistency, with values ranging between 92% and 95%. In contrast, Gemini exhibited lower consistency (84-87%, depending on the language), suggesting a higher tendency to change classifications across repeated runs. Only minor differences in intra-model agreement were observed across languages, which likely reflects the strong multilingual capabilities of modern LLMs.

Cross-model agreement, defined as strict unanimous agreement across all nine runs, was lower than intra-model consistency but remained relatively high, ranging from 71.1% to 75.6% depending on the language. This indicates that, even after applying a strict consensus criterion, at least 71% of the original dataset can be retained, supporting the feasibility of automated labeling.

Finally, inter-annotator agreement metrics showed consistent results across languages. Fleiss' κ scores ranged from 0.72 (Polish) to 0.76 (English), which can be interpreted as good agreement for an NLP task.

Figure 5 presents pairwise Cohen's κ for the nine LLM annotation runs across the five Wikipedia language versions and for the combined dataset. The results again show that agreement is consistently highest within repeated runs of the same model, indicating stronger intra-model than inter-model consistency. Among the evaluated models, Claude Opus demonstrates the highest intra-model agreement across languages, while Gemini exhibits the lowest stability. In terms of inter-model agreement, higher consistency is observed between GPT and Opus, whereas the lowest agreement occurs between GPT and Gemini. These patterns are consistent across languages, suggesting that differences in agreement are primarily driven by model characteristics rather than language-specific factors.

Overall, these findings indicate that while LLMs can serve as reliable annotators, their outputs are model-dependent, similarly to how human annotations may vary across evaluators. Therefore, combining multiple models and applying consensus-based filtering can improve annotation robustness and mitigate individual model biases.

5.2 Annotation Error Analysis

During the annotation process, several types of inconsistencies in LLM outputs, including missing annotations, spurious labels assigned to non-existent identifiers, duplicate outputs, and labels outside the predefined schema (e.g., "mix", "mixed"). However, it concerned just a few cases.

To ensure annotation consistency, missing outputs were completed through additional queries, while invalid or non-compliant labels were normalized to 'n/a' during post-processing. These observations highlight the importance of validation and post-processing steps in large-scale LLM-based annotation pipelines.

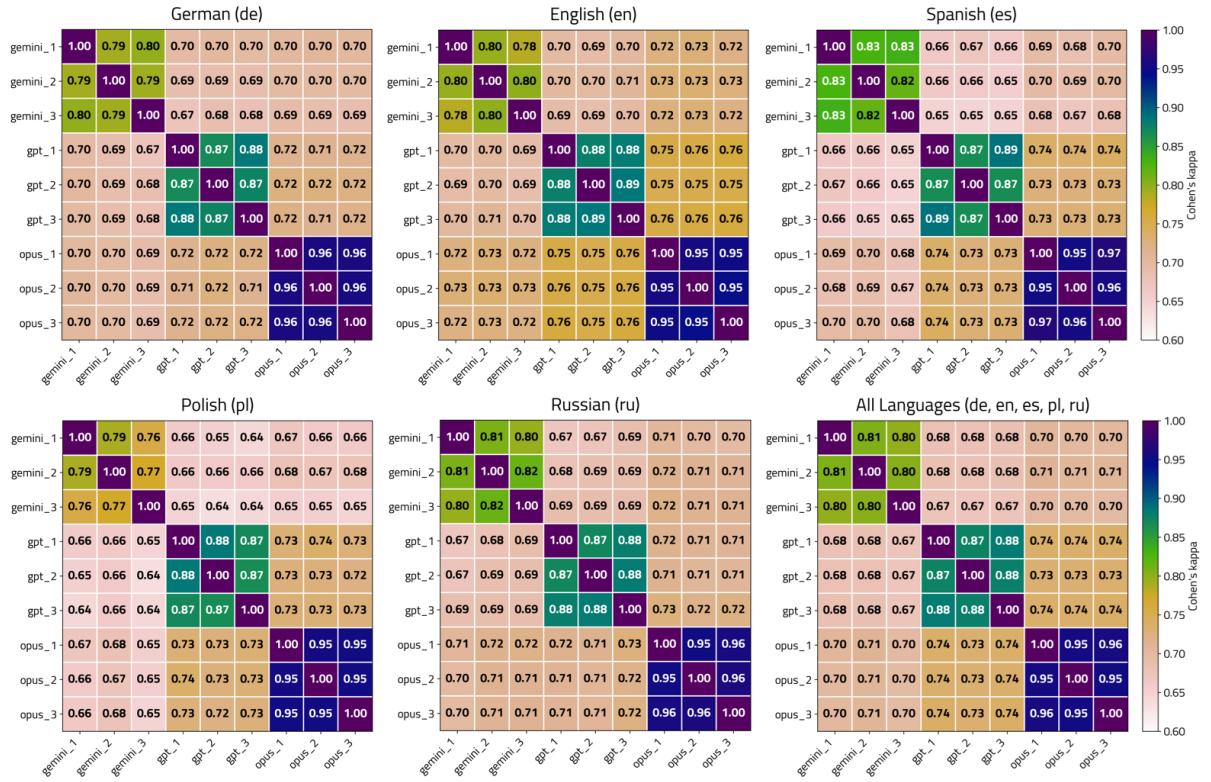

Figure 5: Pairwise Cohen's κ for nine LLM annotators across Wikipedia language versions.

5.3 Evaluation with Human Judgments

Table 2 summarizes the agreement between LLM annotations and human judgments using the test set described in Section 4.3. The evaluation was conducted for two scenarios: (1) sentences for which a given model produced identical predictions across all repeated runs (intra-model agreement), and (2) sentences satisfying the strict cross-model consensus criterion. This approach ensured that only stable and reproducible LLM outputs were included in the comparison.

For the first scenario, the highest agreement with human annotations was consistently observed for the neutral class across both languages, whereas substantially lower agreement was obtained for negative sentiment. In the Polish dataset, the F1 scores for the negative class ranged from 0.24 to 0.33, while in the English dataset they were 0.1 for all three models. Among the evaluated models, GPT demonstrated the most balanced overall performance, followed by Opus and Gemini.

The substantial difference between class-specific F1 scores and the overall metrics should be interpreted in the context of class imbalance in the evaluation data. The human-evaluated test subsets were randomly sampled from the underlying datasets, in which neutral sentiment was the dominant class. Consequently, accuracy and weighted F1, which reflect the underlying class distribution, remained relatively high (0.75-0.88), despite the substantially lower performance for the negative class. In contrast, Macro-F1 gives equal weight to all sentiment classes and is therefore more strongly affected by the poor detection of the minority negative class, resulting in values ranging from 0.36 to 0.65. The particularly low F1 scores for negative sentiment indicate that this class remains difficult for LLMs to identify in predominantly neutral encyclopedic text.

This pattern is consistent with the characteristics of Wikipedia content, where factual and neutral descriptions predominate and explicit negative sentiment is relatively uncommon. Negative sentiment may also be expressed in a more subtle way, making it more difficult to identify than the dominant neutral class.

For the second scenario, strict cross-model consensus reduced the dataset to 692 Polish sentences (65% of the test set) and 131 English sentences (60%). Within this subsets, agreement between the consensus LLM annotations and human labels was perfect for all evaluated metrics. However, these results should be interpreted in light of the selection criterion: only sentences on which all three

LLMs agreed were retained, and neutral sentences constituted the majority of the retained instances. Thus, the perfect agreement reflects the high consistency of the selected consensus cases rather than demonstrating equally high performance across the full, class-imbalanced test set.

Overall, the human evaluation demonstrates strong agreement between LLM and human annotations for cases selected through strict consensus. These results suggest that consensus-based annotation using multiple LLMs can provide reliable sentiment labels that closely align with human judgments. In particular, the strict agreement criterion substantially improves annotation consistency and reduces the influence of individual model biases (see "All LLMs" in Table 2).

Table 2: F1 scores and agreement metrics between LLM annotations and human judgments

Sentences in Polish:

	Gemini 3.1 Pro (910)	GPT 5.2 (1012)	Opus 4.6 (1039)	All LLMs (692)
Negative	0.24	0.26	0.33	1.00
Neutral	0.88	0.91	0.92	1.00
Positive	0.41	0.43	0.46	1.00
Macro avg	0.38	0.65	0.59	1.00
Weighted avg	0.85	0.88	0.88	1.00
Accuracy	0.80	0.84	0.85	1.00
Fleiss' κ	0.27	0.33	0.36	1.00

Sentences in English:

	Gemini 3.1 Pro (195)	GPT 5.2 (208)	Opus 4.6 (211)	All LLMs (131)
Negative	0.10	0.10	0.10	1.00
Neutral	0.85	0.88	0.87	1.00
Positive	0.49	0.50	0.54	1.00
Macro avg	0.36	0.49	0.38	1.00
Weighted avg	0.82	0.85	0.84	1.00
Accuracy	0.75	0.79	0.79	1.00
Fleiss' κ	0.28	0.28	0.31	1.00

Values in brackets correspond to the number of sentences for each scenario.

6 Conclusions

This paper presented a framework for automated multilingual sentiment annotation using multiple state-of-the-art LLMs. The proposed approach combines repeated prompting with strict cross-model consensus in order to improve annotation reliability and reduce model-specific biases. The framework was evaluated on a multilingual dataset of Wikipedia lead sections.

The results showed high intra-model consistency for all evaluated LLMs and relatively strong cross-model agreement across languages. Pairwise agreement analysis confirmed that annotations remain model-dependent, with higher consistency observed within repeated runs of the same model than between different models. These findings indicate that combining multiple LLM evaluators with strict consensus-based filtering provides a robust approach to automated sentiment annotation.

Evaluation against human judgments further confirmed the reliability of the proposed framework. In particular, sentences satisfying the strict consensus criterion showed perfect agreement with human annotations, suggesting that consensus-based LLM annotation can provide high-quality labels suitable for multilingual sentiment analysis tasks.

Despite the limitations indicated in the following section, the dataset was constructed using a transparent and reproducible procedure involving multiple independent annotation runs and agreement-based filtering. The resulting dataset has been publicly released to support further research.

Future work will focus on analyzing model-specific biases, alternative prompting strategies, and extensions of the framework to additional NLP tasks and document-level annotation settings. We also plan to incorporate additional metadata describing the analyzed Wikipedia articles, including indicators of article quality and popularity, such as those provided by WikiRank[7]. Furthermore, future extensions will consider the provenance of information contained in Wikipedia articles by analyzing their cited sources. For example, BestRef[8] provides importance assessments of individual web information sources across different language versions of Wikipedia.

7 Limitations

The adopted procedure is subject to several limitations. First, although general-purpose LLMs are strong general-purpose language processors, they may underperform specialized supervised models in some subjective or fine-grained classification tasks (Kocoń et al., 2023). In addition, LLMs may inherit, reproduce, or amplify biases present in their training data and model development pipelines (Inoshita and Zhou, 2024). These limitations are particularly relevant for sentiment annotation, where judgments may depend on linguistic, cultural and contextual cues.

A second limitation concerns the construction

[7] https://wikirank.net
[8] https://bestref.net

of API calls and the inference configuration of the evaluated LLMs. In the current study, the temperature parameter was not explicitly controlled and default API settings were used to ensure comparability across models as its effective range may vary between models. As a result, differences in default inference behavior across models may have influenced annotation consistency and agreement patterns. Future work could systematically investigate the impact of temperature and other inference parameters on annotation stability, sentiment distributions, and cross-model agreement. Lower temperature settings may improve determinism and reproducibility, whereas higher values may increase sensitivity to nuanced or ambiguous sentiment expressions.

The applied batching strategy used during annotation is also subject to several limitations. Multiple sentences were submitted together within a single API request, meaning that the LLM evaluators processed several inputs simultaneously within a shared context window. Moreover, the construction of the sentence identifier ("PageID-SentenceNumber") and the sharing of the same "PageID" across sentences originating from the same article, might have suggested to the model that these sentences were taken from the same article. As a result, individual predictions may have been influenced by neighboring sentences, potentially introducing contextual dependencies or consistency biases between annotations. This issue is especially relevant for encyclopedic texts, where consecutive sentences often concern the same topic or article section.

Another limitation concerns the prompting strategy used during the annotation process. A single English-language system prompt was applied across all evaluated languages, while each API batch contained sentences from only one language. Although modern LLMs are generally capable of cross-lingual understanding, this setup may have influenced how sentiment was interpreted across different linguistic and cultural contexts. In particular, subtle language-specific framing patterns or idiomatic expressions may not have been captured equally well when instructions were provided exclusively in English. As a result, the current setup should be interpreted as cross-lingual prompting rather than fully localized multilingual prompting.

The selected prompt persona, namely an "expert linguistic analyst", also reflects a particular interpretation of sentiment analysis as a linguistically grounded annotation task. Other prompt roles or application-oriented perspectives, such as customer-facing, domain-specific, or reader-oriented evaluators, could lead to different interpretations of sentiment. Future work could therefore examine how alternative prompt roles, task framings, and language-localized prompts affect annotation consistency, agreement patterns, and benchmark composition.

The analysis focuses primarily on the final sentiment labels returned by the LLM evaluators rather than on the explanations underlying these decisions. As a result, the study provides limited insight into why individual models assigned a given sentiment label. Future research could extend the annotation framework by collecting (short) model-generated justifications together with each prediction. However, such explanations should be interpreted cautiously, since generated rationales may not always faithfully reflect the internal decision process of the model.

Although strict agreement across three state-of-the-art LLMs increases label stability, it does not guarantee alignment with human judgment. It may also reinforce shared biases arising from similar training data, model design choices, or evaluation practices. A subset of the dataset was compared against human annotations, but this validation was limited in scope and does not constitute a comprehensive gold-standard evaluation of the entire dataset. Manual validation was conducted only for Polish and English and covered relatively small subsets of the data. Therefore, the reported agreement levels may not generalize to all evaluated languages, topics, or domains.

The human-evaluation subsets also retained the natural class distribution of the underlying data, resulting in a substantial imbalance between sentiment classes, with dominating neutral instances. This imbalance particularly affects the interpretation of Macro-F1 and class-specific performance, as the relatively high accuracy and weighted F1 scores may coexist with substantially lower performance for minority classes, especially negative sentiment. Future studies should investigate whether class-balancing strategies or alternative prompting approaches can improve the identification of minority sentiment classes.

Restricting the dataset to instances with unanimous agreement may favor clearer and less ambiguous cases, potentially limiting the representation of complex, implicit, or context-dependent sentiment.

Therefore, this filtering strategy may exclude exactly the nuanced or borderline examples that are especially relevant in encyclopedic text, where sentiment is often expressed through subtle framing rather than explicit emotional language.

The multilingual comparison is also affected by the structure of Wikipedia itself. Lead sections across languages are not strictly parallel, and different language versions of the same article may vary in content, framing, length, and editorial emphasis. Consequently, sentiment differences across languages may reflect both linguistic variation and genuine differences in article content. Furthermore, the predominantly factual and encyclopedic nature of Wikipedia limits the generalizability of the findings to other text genres, such as opinion-oriented or user-generated content, where sentiment is typically expressed more explicitly.

Another limitation concerns the linguistic diversity of the evaluated datasets. Although the study covers five languages, they are all European languages. The robustness of multi-model consensus may differ for languages with substantially different characteristics. Future research should therefore extend the evaluation to typologically diverse languages, including languages using different writing systems, such as Chinese, Japanese, or Arabic, to further assess the generalizability of the proposed approach.

The current framework operates primarily at the sentence level and does not explicitly account for broader discourse structure or document-level context. This may affect sentiment interpretation in encyclopedic texts, where evaluative meaning can emerge from contrast, framing, attribution or the relation between sentences.

Finally, the use of coarse sentiment categories, namely positive, negative, and neutral, simplifies more complex discourse phenomena. Future work could address this limitation through more fine-grained approaches, such as aspect-based sentiment analysis, emotion classification, stance detection or framing analysis.

References

Cem Emre Akbaş, Vladimír Ulman, Martin Maška, and Michal Kozubek. 2025. Deepfuse: A multi-rater fusion and refinement network for computing silver-standard annotations. *Computers in biology and medicine*, 192:110186.

Khalid Al Khatib, Hinrich Schütze, and Cathleen Kantner. 2012. Automatic detection of point of view differences in wikipedia. In *Proceedings of COLING 2012*, pages 33–50.

Rahma Alahmary and Hmood Al-Dossari. 2021. A semiautomatic annotation approach for sentiment analysis. *Journal of Information Science*, 49(2):398–410.

Badriya Murdhi Alenzi, Muhammad Badruddin Khan, Mozaherul Hoque Abul Hasanat, Abdul Khader Jilani Saudagar, Mohammed AlKhathami, and Abdullah AlTameem. 2022. Automatic annotation performance of textblob and vader on covid vaccination dataset. *Intelligent Automation & Soft Computing*, 34(2).

Nouf Almontashry, Siba Ziadi, Khadeja Njaai, and Rabab Alomairy. 2025. Evaluating large language models for sentiment analysis: A comparative study of accuracy and scalability. In *2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS)*, pages 413–419.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

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

Sherif Elmitwalli and John Mehegan. 2024. Sentiment analysis of COP9-related tweets: a comparative study of pre-trained models and traditional techniques. *Frontiers in Big Data*, Volume 7 - 2024.

English Wikipedia. 2026. Wikipedia: Neutral point of view.

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.

Yudi Guzmán-Monteza. 2023. Assessment of an annotation method for the detection of spanish argumentative, non-argumentative, and their components. *Telematics and Informatics Reports*, 11:100068.

Saeed-Ul Hassan, Aneela Saleem, Saira Hanif Soroya, Iqra Safder, Sehrish Iqbal, Saqib Jamil, Faisal Bukhari, Naif Radi Aljohani, and Raheel Nawaz. 2021. Sentiment analysis of tweets through altmetrics: A machine learning approach. *Journal of Information Science*, 47(6):712–726.

Keito Inoshita and Xiaokang Zhou. 2024. Sentiment bias and security analysis in training datasets of large language models. In *2024 IEEE International Conference on Big Data and Cloud Computing (BDCloud)*, pages 1–8.

Isaac Johnson, Yu-Ming Liou, Jacob Rogers, Aaron Shaw, and Leila Zia. 2025. Recommended practices for NPOV research on wikipedia.

Ji-Won Kang and Sun-Yong Choi. 2025. Comparative investigation of GPT and FinBERT's sentiment analysis performance in news across different sectors. *Electronics*, 14(6). Number: 1090.

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, Aleksandra Wojewoda, and Krzysztof Węcel. 2026. 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.

Yuxiang Liao, Hantao Liu, and Irena Spasić. 2024. Fine-tuning coreference resolution for different styles of clinical narratives. *Journal of Biomedical Informatics*, 149:104578.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of babelnet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Loris Di Quilio and Fabio Fioravanti. 2024. A comprehensive framework for aspect-category sentiment analysis. In *Proceedings of the Eight Workshop on Natural Language for Artificial Intelligence (NL4AI 2024)*, volume 3877, pages paper 13, Bolzano, Italy. CEUR Workshop Proceedings.

Muhammad Aasim Qureshi, Muhammad Asif, Mohd Fadzil Hassan, Ghulam Mustafa, Muhammad Khurram Ehsan, Aasim Ali, and Unaza Sajid. 2022. A novel auto-annotation technique for aspect level sentiment analysis. *Computers, Materials, & Continua*, 70(3):4987.

Fitrah Rumaisa, Halizah Basiron, Zurina Saaya, and Yoki Muchsam. 2020. Bmbi: A development of a special corpus on homonyms for multi-lingual sentiment analysis. *International Journal*, 9(4).

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.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Yiwei Zhou, Alexandra Cristea, and Zachary Roberts. 2015. Is wikipedia really neutral? a sentiment perspective study of war-related wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 160–168.