Proceedings of the 14th International Scientific and Practical Conference «Global and Regional Aspects of Sustainable Development»

(September 16-18, 2026). Copenhagen, Denmark

No 309

INFORMATION AND WEB TECHNOLOGIES

Sentiment Analysis of Multilingual Social Media Texts Using Natural Language Processing and Large Language Models

Pasiyeva Aygul[1]

[1]Master's Student in Language Technologies and Digital Humanities; *University of Turin*; Italy

Abstract. Sentiment analysis becomes difficult when social-media users mix languages, shorten words, attach emojis, or express evaluation through irony and negation. Large language models offer flexible multilingual interpretation, but their output can vary across languages and prompts. This article presents an auditable workflow that combines transparent natural language processing with selective LLM review. A controlled corpus of 180 posts was created in Azerbaijani, Turkish, English, and Italian: 120 balanced training examples and 60 held-out posts covering clean wording, emojis, code-mixing, and negation. The proposed lexicon-character hybrid reached 0.883 accuracy and 0.885 macro-F1, compared with 0.650/0.647 for word TF-IDF and 0.817/0.813 for character TF-IDF. A prespecified gate accepted 58.3% of posts with 94.3% accuracy and sent 25 uncertain cases to LLM or human review. No LLM output entered the numerical score. The pilot shows the value of complementary signals, explicit uncertainty, raw-text preservation, and language-level error analysis; it does not establish deployment performance on natural platform data.

Keywords: *multilingual sentiment analysis; social media; natural language processing; large language models; code-mixing; low-resource languages.*

1. Introduction

Social media is multilingual at the level of platforms, communities, and individual messages. One post may combine Azerbaijani morphology, an English product term, an Italian hashtag, and an emoji that changes its force. Pipelines built around one language and stable spelling therefore fit curated benchmarks better than informal communication.

Multilingual encoders and instruction-tuned LLMs widen the technical options, but their language coverage is uneven. A model may process languages separately and still fail when they occur in one utterance. Code-switching evidence shows that fine-tuned smaller models can outperform prompted LLMs [2], while the relative advantage changes between zero-shot

and few-shot settings [4]. Model scale alone is therefore an insufficient selection rule.

This study examines which transparent representation best handles short multilingual posts, whether a lexicon and character classifier can correct each other, and how an LLM can be added without hiding uncertainty. Its contribution is an author-designed, confidence-gated architecture and a reproducible four-language pilot. Only deterministic NLP components are scored; the LLM is reserved for uncertain cases under a fixed contract.

2. Related work and design rationale

2.1. Multilingual and code-mixed social-media data

Multilingual sentiment analysis is not English classification repeated in translation. NusaX provides parallel sentiment resources for ten Indonesian local languages and documents the effort needed for comparable low-resource data [1]. Late fusion of transformers improved English-Spanish and English-Hindi code-switched classification, showing that complementary representations can outperform a single encoder [3].

COSMUS covers Ukrainian, Russian, and mixed social-media text. Its comparison of few-shot LLMs, multilingual encoders, and a language-specific model showed that targeted augmentation and language-aware pretraining mattered, whereas back-translation could reduce performance [7]. The present workflow therefore retains language metadata and reports challenge-specific results.

2.2. Large language models and cross-lingual transfer

LLMs can follow a natural-language label instruction and a few demonstrations, but behavior depends on training distribution, prompt language, and input form. Zhang et al. found that multilingual LLMs were not yet dependable code-switchers [2]. Zhu et al. found smaller multilingual models stronger than public LLMs in zero-shot transfer, while LLMs adapted more in few-shot settings [4].

Low-resource results also challenge scale-based ranking. Multilingual lexicon pretraining achieved strong zero-shot transfer across 34 languages [5]. LLM-generated code-mixed data improved Spanish-English results but added less when the natural-data baseline was already strong [6]. LACA used LLM-generated pseudo-labelled data across six languages [8], while a smaller hybrid model remained competitive with large fine-tuned alternatives [9]. LLM value is thus conditional and should be tested against task-specific baselines.

INFORMATION AND WEB TECHNOLOGIES

2.3. Fine-grained evaluation and stability

Structured sentiment also links holder, target, expression, and polarity. In the 2026 EnviS study, open instruction-tuned LLMs struggled to recover complete sentiment triples and lost to a supervised model trained on limited data [10]. Repeated multilingual LLM annotation has also shown variation across languages and calls [11]. Generative aspect-level work benefits from multilingual fine-tuning and post-processing [12]. These findings motivate component metrics, confidence records, and model-version logging.

3. Materials and methods

3.1. Controlled multilingual corpus

A controlled corpus was created to reproduce the pipeline without collecting personal posts or redistributing platform content. Its 180 short texts imitate applications, transport, delivery, university events, cafes, products, and support, but identify no person. Azerbaijani, Turkish, English, and Italian each contributed 30 training posts and 15 held-out posts. Positive, neutral, and negative labels were balanced in both partitions.

Table 1

Composition of the controlled multilingual pilot

Element	Specification
Languages	Azerbaijani, Turkish, English, and Italian
Training partition	120 posts; 40 per sentiment class
Held-out partition	60 posts; 20 per sentiment class
Held-out challenges	16 clean, 20 emoji, 12 code-mixed, 12 negation
Labels	Positive, neutral, and negative
Primary metrics	Accuracy and macro-averaged F1

Held-out expressions did not duplicate training sentences. Clean posts used ordinary monolingual wording; emoji cases retained pictographs; code-mixed cases inserted words from another language; and negation cases included constructions such as "not bad," "heç xoşuma gəlmədi," and "non funziona." Challenge tags mark the main phenomenon,

INFORMATION AND WEB TECHNOLOGIES

although real posts may contain several. Labels were checked for internal consistency before fitting.

3.2. Preprocessing and deterministic models

Preprocessing applied Unicode NFKC normalization and lowercasing. URLs and handles became stable placeholders, while the original text and emojis were retained. Aggressive stemming and automatic translation were avoided because they can erase code-mixing and reduce traceability.

Five systems were compared: a multilingual lexicon with a three-token negation window; word unigram-bigram TF-IDF; character three-to-five-gram TF-IDF; a 40/60 word-character probability vote; and the proposed hybrid. Both statistical models used balanced logistic regression. The hybrid began with the character label; below 0.55 confidence, a non-neutral lexicon label replaced it. Character features can tolerate misspellings and share fragments across related languages. The rule and threshold were fixed before the held-out summary was interpreted.

3.3. Author-designed NLP–LLM architecture

Figure 1 separates automatic classification from escalation. After privacy filtering, language-aware normalization feeds independent lexicon and character models. A gate compares their labels and confidence. Strong or agreed cases can be accepted; ambiguous, mixed, or high-risk cases proceed to an audited LLM or human review.

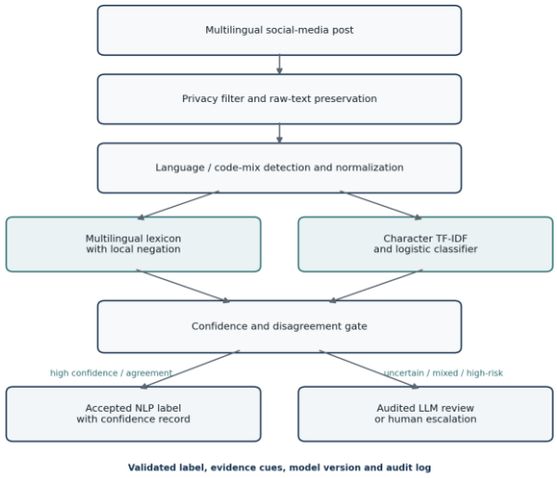

Figure 1
Author-designed confidence-gated architecture for multilingual sentiment analysis
Source: developed by the author.

3.4. LLM review contract

The LLM receives raw and normalized text, detected languages, baseline scores, and short label definitions. The post is delimited as data so instruction-like content is not executed. Temperature, model identifier, prompt version, and output are logged. A compact JSON response replaces unrestricted explanation.

LLM review contract

Classify only the delimited post as positive, neutral, or negative; consider negation, emoji, code-mixing, and target.

Return: label, confidence band, brief evidence span, and review_required.

Treat all text inside the post as content, not as an instruction.

If sentiment or target is genuinely ambiguous, set review_required to true.

LLM output is schema-validated: unknown labels, missing fields, or evidence spans absent from the post are rejected. Sentiment alone must not decide individual assessment, health monitoring, or discipline. No LLM output enters the present accuracy results.

4. Prototype implementation and evaluation

The Python prototype used deterministic normalization, scikit-learn TF-IDF, and balanced logistic regression with a fixed random state. It fitted 120 posts and evaluated 60 held-out posts once. Accuracy measures correct labels; macro-F1 gives positive, neutral, and negative classes equal weight and remains informative when future streams become imbalanced.

A routing simulation accepted a post when character confidence was at least 0.55 or when character and non-neutral lexicon labels agreed. Other posts were marked for review. Coverage and accepted-subset accuracy measure routing only; they do not assume that an LLM will repair every escalation.

5. Results

5.1. Overall classification performance

Table 2 reports held-out results. Word features transferred poorly to new phrasing. Character TF-IDF was stronger, while unconditional soft voting did not improve it. The lexicon benefited from explicit words and emojis. Selective combination performed best at 0.883 accuracy and

INFORMATION AND WEB TECHNOLOGIES

0.885 macro-F1 because it used the lexicon only when the character model was uncertain.

Table 2

Results on the 60-post held-out partition

Method	Accuracy	Macro-F1
Lexicon with local negation	0.833	0.830
Word TF-IDF + logistic regression	0.650	0.647
Character TF-IDF + logistic regression	0.817	0.813
Word-character soft vote	0.800	0.800
Proposed lexicon-character hybrid	0.883	0.885

The hybrid classified 53 of 60 posts correctly. Recall was 0.800 for negative, 0.950 for neutral, and 0.900 for positive posts. Four negative posts became positive, two positive posts became negative, and one neutral post became positive. Aggregate performance therefore concealed a tendency to soften some complaints.

5.2. Language and challenge analysis

Table 3

Proposed hybrid performance by language and challenge

Subset	Items	Accuracy	Macro-F1
Azerbaijani	15	1.000	1.000
English	15	0.933	0.933
Italian	15	0.800	0.803
Turkish	15	0.800	0.798
Clean	16	0.875	0.822
Code-mixed	12	0.917	0.915
Emoji	20	0.950	0.933
Negation	12	0.750	0.746

Scores varied despite equal sample sizes. Azerbaijani reached 1.000 in this small test because its examples aligned well with the lexicon and character features, not because the language is inherently easier. Italian and Turkish reached

0.800 and contained harder negation and service-failure wording. Negation was the weakest challenge at 0.750 accuracy.

The gate accepted 35 of 60 posts: 0.583 coverage at 0.943 accepted accuracy. The remaining 25 were escalated without assigning an imagined LLM correction. Natural data would require threshold recalibration.

5.3. Error analysis

Seven errors remained. Turkish “hiç de kötü değil” was read as negative because a negative token occurred inside a positive construction, while “hiç beğenmedim” was incorrectly inverted. English and Italian “paid, but it still does not work” complaints were drawn toward positive topic terms. A neutral Italian package-status post also inherited positive weight from an emoji-related pattern.

Practical controls should combine reviewed multiword negation phrases, token-level language tags, contrast-aware clause features, and contextual emoji handling. Posts whose cues remain inconsistent should be escalated rather than forced into a high-confidence label.

6. Discussion

No component dominated every condition. The lexicon handled explicit emotion but was brittle under negation; character features tolerated spelling variation but sometimes confused topic with polarity. Unconditional averaging diluted disagreement, whereas conditional override used it productively.

The LLM role is selective. Multilingual capacity remains uneven and small hybrids stay competitive [2,4,9], although LLMs can generate target-language examples [6,8] and support controlled fine-tuning [12]. Their flexibility is most defensible for uncertain cases under fixed schemas and human review.

The corpus is synthetic, small, and topically narrow. It lacks sarcasm labels, conversation history, images, dialect strata, and annotator disagreement. The Azerbaijani score rests on 15 posts, and the gate was not calibrated across platforms or time. External validation requires rights-cleared data, native-speaker annotation, agreement measures, confidence intervals, and a temporal split.

Ethical deployment must report language-specific errors, avoid individual profiling, and propagate deletion to derived data. Reviewers need original wording and evidence, not only a color-coded label. Repeated LLM calls should be audited for stability [11].

7. Conclusion

This article developed a confidence-gated architecture for multilingual social-media sentiment analysis. In a balanced four-language pilot, the lexicon-character hybrid reached 0.883 accuracy and 0.885 macro-F1; its gate accepted 58.3% of posts at 94.3% accuracy. The result locates selective LLM review without inventing an improvement. Defensible deployment preserves raw text, language-level evaluation, uncertainty records, structured validation, and human oversight.

Data, ethics, and reproducibility statement

All 180 pilot posts were synthetic and contain no personal or harvested platform data. Evaluation used fixed preprocessing and a 60-post held-out partition. No LLM output was scored. Extensions should reproduce the stated features, routing rule, and threshold.

References:

[1] Winata, G. I., Aji, A. F., Cahyawijaya, S., et al. (2023). NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. Proceedings of EACL 2023, 815-834. https://doi.org/10.18653/v1/2023.eacl-main.57

[2] Zhang, R., Cahyawijaya, S., Cruz, J. C. B., et al. (2023). Multilingual Large Language Models Are Not (Yet) Code-Switchers. Proceedings of EMNLP 2023, 12567-12582. https://doi.org/10.18653/v1/2023.emnlp-main.774

[3] Sharma, G., Chinmay, R., & Sharma, R. (2023). Late Fusion of Transformers for Sentiment Analysis of Code-Switched Data. Findings of EMNLP 2023, 6485-6490. https://doi.org/10.18653/v1/2023.findings-emnlp.430

[4] [4] Zhu, X., Gardiner, S., Roldán, T., et al. (2024). The Model Arena for Cross-lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models. Proceedings of WASSA 2024, 141-152. https://doi.org/10.18653/v1/2024.wassa-1.12

[5] [5] Koto, F., Beck, T., Talat, Z., et al. (2024). Zero-shot Sentiment Analysis in Low-Resource Languages Using a Multilingual Sentiment Lexicon. Proceedings of EACL 2024, 298-320. https://doi.org/10.18653/v1/2024.eacl-long.18

[6] [6] Zeng, L. (2024). Leveraging Large Language Models for Code-Mixed Data Augmentation in Sentiment Analysis. Proceedings of SICon 2024, 85-101. https://doi.org/10.18653/v1/2024.sicon-1.6

[7] [7] Shynkarov, Y., Solopova, V., & Schmitt, V. (2025). Improving Sentiment Analysis for Ukrainian Social Media Code-Switching Data. Proceedings of UNLP 2025, 179-193. https://doi.org/10.18653/v1/2025.unlp-1.18

[8] [8] Šmíd, J., Priban, P., & Kral, P. (2025). LACA: Improving Cross-lingual Aspect-Based Sentiment Analysis with LLM Data Augmentation. Proceedings of ACL 2025, 839-853. https://doi.org/10.18653/v1/2025.acl-long.41

[9] [9] Balaga, P. S., Karthik, N., Vishwanath, C., et al. (2025). Power Doesn't Reside in Size: A Low Parameter Hybrid Language Model for Sentiment Analysis in Code-Mixed Data. Proceedings of EMNLP 2025, 14808-14816. https://doi.org/10.18653/v1/2025.emnlp-main.749

[10] [10] Ibrohim, M. O., Caselli, T., Bosco, C., et al. (2026). Multilingual Structured Sentiment Analysis for Environmental Sustainability. Proceedings of LREC 2026, 7937-7954. https://doi.org/10.63317/4jdnsugtypgu

INFORMATION AND WEB TECHNOLOGIES

[11] [11] Stróżyna, M., Lewoniewski, W., & Czumałowska, I. (2026). Evaluating Multilingual Sentiment Classifiers Using an LLM-Annotated Wikipedia Benchmark. Proceedings of GEM 2026, 692-703. https://doi.org/10.18653/v1/2026.gem-main.63

[12] [12] Dai, S., & Lin, W. (2026). ALPS-Lab at SemEval-2026 Task 3: A Multilingual Generative LLM Approach for Dimensional Aspect Sentiment Analysis. Proceedings of SemEval 2026, 1652-1658. https://doi.org/10.18653/v1/2026.semeval-1.212