International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 172-181
https://doi.org/10.70917/ijcisim-2026-5287
Article

A Hybrid Multilingual Sentiment Analysis Pipeline Using n8n Workflow Orchestration and LangChain Large Language Model Ensemble

Dr. Aniruddha Shelotkar[1], Dr. Deepali Bhamare[2], Dr. Pallavi Salunkhe[3], Dr. Vinodkumar Patil[4], Dr. Rajendra Pandit Desale[5], Dr. Machchhindra Garde[6]

[1]Cybage Software Pvt. Ltd., Pune, Email: anishelotkar@gmail.com
[2]Department of E&TC Engineering, S.S.V.P.S.B.S.D.'s COE Dhule, Email: dr.bhamare@ssvpsengg.ac.in
[3]Department of E&TC Engineering, S.S.V.P.S.B.S.D.'s COE Dhule, Email: pb.salunkhe@ssvpsengg.ac.in
[4]Department of E&TC Engineering, NES's Gangamai College of Engineering, Nagaon, Dhule, Email: vinod.patil293@gmail.com
[5]Associate Technology, Software Development, Synechron Technologies Pvt. Ltd., Email: desale.rajendra@gmail.com
[6]Department of Electronics Engineering, S.S.V.P.S.B.S.D.'s COE Dhule, Email: mchgarde@gmail.com

Abstract: Despite significant research into multilingual sentiment analysis, challenges persist since many of the best-performing sentiment classifiers are English-only, and non-English text poses difficulties due to errors in translation, cultural context, and a lack of training data. Current approaches involve either the use of pre-trained multilingual encoders (such as mBERT and XLM-R), which have problems with accuracy for informal or specialized text, or translating non-English text using online services and applying standard sentiment analysis models to the translated text. This paper proposes an approach that combines the n8n workflow automation service with LangChain large language model orchestration and uses a model ensemble of RoBERTa, BERTweet, and GPT-3 models for sentiment classification. Before performing sentiment classification, this pipeline involves dynamic text translation using Google Translate and LibreTranslate, depending on whether the text is formal or not. We conduct experiments with a pipeline involving five language datasets (Arabic, Chinese, French, Marathi, and Hindi) and compare the performance of our proposed approach with mBERT, XLM-R, RoBERTa, and GPT-3 models. Our model demonstrates the best performance in sentiment classification across all five languages (accuracy is from 85.5% to 87.6%) with an improvement in precision, recall, F1-score, and specificity metrics by 1.2–2.3 percentage points compared to the strongest baseline.
Keywords: multilingual sentiment analysis; workflow automation; LangChain; large language models; ensemble learning; machine translation; n8n

1. Introduction

Sentiment analysis is an essential component of natural language processing used widely on social media and other user-generated content platforms to identify public opinions and preferences. In general, sentiment analysis is performed using English-oriented pre-trained models such as BERT [1], RoBERTa, and GPT-3 [3], which demonstrate impressive performance when used for English text. However, their performance is much worse when the text is written in another language since it requires a different set of features [6], [11].

One of the ways of expanding the coverage to multiple languages is to apply the translation approach: the text is translated using a machine translator and then is passed to an English sentiment analysis model.

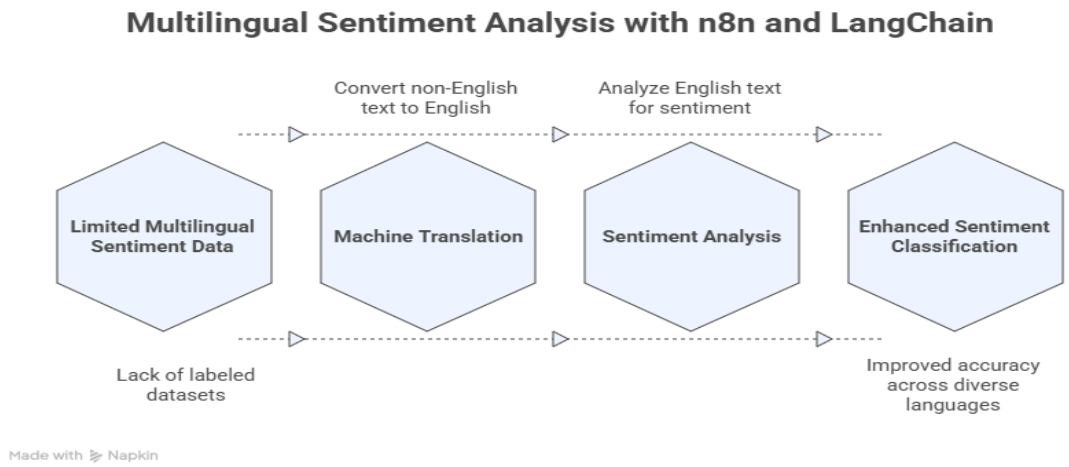

Fig. 1. Multilingual inputs are first converted to a common base language before running ensemble sentiment classification.

In contrast, this paper overcomes these issues by presenting a hybrid pipeline approach (Fig. 1) where (i) n8n, an open-source workflow automation engine, is used for collecting data, dynamic routing based on content and orchestration, and (ii) LangChain, a deep-learning-based LLM integration framework, is used for ensemble sentiment classification with RoBERTa, BERTweet, and GPT-3, with majority vote aggregation. Instead of relying on a pre-defined translation engine, the pipeline routes formal content (e.g., news) through machine translation and informal content (e.g., social media posts, reviews) through colloquially appropriate translation models to minimize loss of context prior to sentiment classification.

1.1 Contributions

- A hybrid architecture integrating workflow automation (n8n) and LLM ensemble orchestration (LangChain) for end-to-end multilingual sentiment analysis in real time.
- Content-aware dynamic translation that routes formal and informal text to appropriate translation engines to prevent sentiment loss.
- A majority-vote ensemble with RoBERTa, BERTweet, and GPT-3 that is more robust than each individual model.
- A study over five languages of different typological classes (Arabic, Chinese, French, Marathi, Hindi), over three domains, benchmarked on mBERT, XLM-R, RoBERTa, and GPT-3.

The rest of this paper is structured as follows: Section 2 surveys the relevant literature; Section 3 states the problem; Section 4 describes our proposed approach; Section 5 provides the experimental results; Section 6 outlines open problems; and Section 7 offers future directions.

2. Related Work

Earlier work on multilingual sentiment analysis involved feature engineering specific to the target language using classical classifiers (e.g., SVMs) [10], which poorly scaled up in performance across multiple languages. Multilingual transformer architectures such as mBERT and XLM-R [1], [2] have brought about improvement by learning masked language modeling from multilingual texts, but are still weak in handling slang and emoji-ridden social media language [6], [10].

Machine-translation pipelines first translate the non-English text into English and then apply English-centric classifiers [5]. Such approaches exploit mature technology for English but suffer due to the loss of context caused by machine-translation, especially when dealing with idioms, sarcasm, and domain-specific vocabulary [5], [6]. This problem becomes even more pronounced for low-resource languages [8].

Large-scale language models (GPT-3, T5) [3], [4] provide better contextual understanding and have been applied for sentiment classification either alone or combined with translation [5], [7]. Mixed solutions involving multilingual transformers, LLMs, and machine-translation ensembles [7] claim an increase in performance but still have trouble adapting to the domain-specific and culturally specific nuances. The state-of-the-art is summarized in Table 1.

Table 1. Comparative Analysis of Notable Recent Work on Multilingual Sentiment Classification

Study	Approach / Advantage	Key Limitation
Miah et al. (2024) [7]	Ensembles a fine-tuned transformer with an LLM over LibreTranslate/Google-Translate output for cross-lingual sentiment classification.	Translation-induced errors still affect accuracy on informal/domain-specific text.
Salameh et al. (2015) [5]	Machine translation bridges Arabic sentiment analysis.	Context loss from translation degrades sentiment fidelity.
Poria et al. (2020) [11]	Surveys open challenges in cross-lingual and context-dependent sentiment analysis.	Struggles with sarcasm and cultural nuance.
Wankhade et al. (2022) [10]	Surveys sentiment analysis methods, applications, and open challenges, including LLM-based approaches.	Broad survey; does not propose or evaluate a deployable pipeline.
Chan et al. (2023) [12]	Reviews sequential transfer-learning strategies for adapting multilingual models across domains.	Limited adaptation to informal, social-media language.

The above-mentioned literature analysis reveals three common shortcomings, which are addressed by this research: (1) decoupling and static nature of translation vs. classification; (2) insufficient robustness of one-model classifiers; (3) absence of pipelines for automation and large-scale deployment in production environment.

3. Problem Formulation

Multilingual sentiment classification faces five challenges connected to each other: (1) linguistic differences in the structure and morphology that limit cross-lingual generalization capabilities [12]; (2) translation mistakes distorting emotionally meaningful words, such as slang or irony [6]; (3) insufficient availability of labeled data for the target language [8], [10]; (4) cultural differences in expressing the sentiment, e.g., indirect expressions [7] or context-based expressions [11]; and (5) domain-specific vocabularies where the sentiment polarity depends on context, e.g. "cheap" as negative for quality, but positive for price [7].

Problem statement. The problem is in construction of a robust to translation errors, domain shifts, and cultural peculiarities pipeline that will output a sentiment label $\hat{y} \in$ {positive, negative, neutral} based on the multilingual and multi-domain text corpus. This research proposes a solution in the form of a dynamic translation-based classifier ensemble, automated via workflow management system.

4. Proposed Methodology

The pipeline consists of four stages – data acquisition, pre-processing, dynamic translation, and sentiment classification using ensemble of LLMs, orchestrated end-to-end by n8n, where LangChain is responsible for invoking the LLMs for classification purposes (Figure 1).

4.1 Data Collection

Data collection occurs through social media, customer reviews, and news from platforms in Arabic, Chinese, French, Marathi, and Hindi using n8n APIs and web-scraping nodes, thus allowing for ongoing, real-time data collection. Model evaluation uses five public, sentiment-labeled corpora, one per language: ASAD, a 95,000-tweet Arabic Twitter sentiment corpus [13]; ChnSentiCorp, a Chinese product- and service-review sentiment corpus [14]; the French subset of Webis-CLS-10, a cross-lingual Amazon product-review sentiment corpus [15]; L3Cube-MahaSent-MD, a multi-domain Marathi sentiment dataset covering movie reviews and tweets [16]; and the IIT Patna Hindi movie- and product-review sentiment dataset [17]. Because these public resources are review- and tweet-based rather than news-based, the French and Marathi domains are reported as reviews (not news articles) throughout this paper (Section 5.1, Table 2) to accurately reflect the data actually used; exact version numbers, access dates, and license terms for each corpus are to be reported by the authors from their experimental records before submission (Section 6).

4.2 Preprocessing

Before translation, raw text undergoes cleansing: URLs are removed, emojis are retained as Unicode symbols so that potentially sentiment-bearing information is not discarded, text is converted to lowercase, stopwords are stripped off, and tokenization followed by lemmatization is performed (Algorithm 1). Figure 2 demonstrates sample input and output at this step.

Algorithm 1: Text Cleaning

Input: Unprocessed text d. Output: Cleaned tokens.

1. Delete all web addresses from d.
2. Preserve emojis as Unicode symbols rather than deleting them.
3. Convert the remaining text to lowercase.
4. Split the text into word tokens and remove stop words.
5. Lemmatize the remaining tokens.

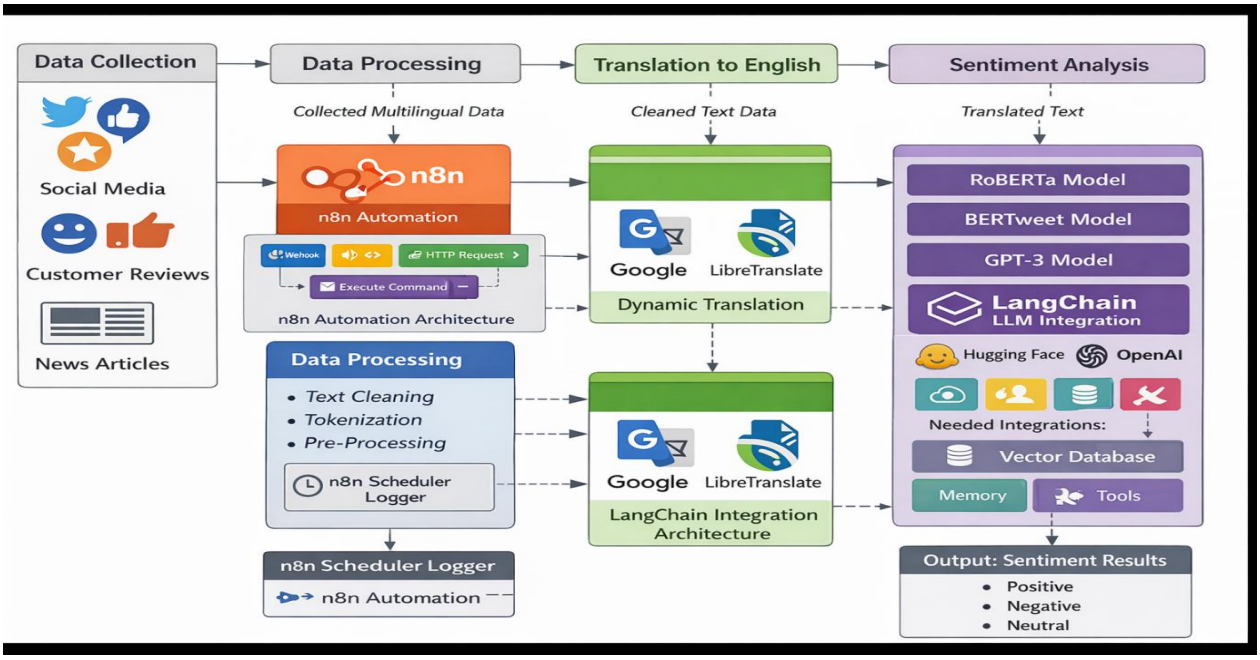

Figure 2. Outputs from the pre-processing stage on different multilingual input samples.

4.3 Dynamic, Content-Aware Translation

The processed text is converted to English — the language used by the pipeline for classification — using Google Translate or LibreTranslate (Algorithm 2). Instead of using a single translation engine for every input, the routing stage first determines whether the content is formal (for example, news) or informal (for example, social-media text or reviews) and then selects the corresponding translation engine. The exact classifier or rule set used for this formal/informal decision is an implementation detail that must be documented from the n8n workflow configuration before final submission; it is not inferred or fabricated here.

Algorithm 2: Content-Aware Translation

6. Identify the source language of the cleansed document.
7. Determine whether the content style is formal or informal, using the content-style routing classifier/rule set configured in the workflow [exact implementation to be reported from the experimental workflow].
8. Send to the respective translator and translate into English.
9. In case of uncertainty about the translated text, try using another engine or send for human inspection.

Let d_i be the cleansed document i in the source language L_i, then the translation can be represented as follows:

$$d_i^{translated} = T(\, d_i^{cleaned}\,) \quad (1)$$

4.4 Ensemble Sentiment Classification via LangChain

The LangChain framework uses parallel calling of three pre-trained transformer models, namely RoBERTa, BERTweet, and GPT-3, each of which produces an output in the form of a label from the set {positive, negative, neutral} based on the translated text (Eq. 2). The final label is generated through the majority voting scheme (Eq. 3, Algorithm 3), which accounts for possible misclassification errors by the individual models.

$$\hat{y}_i^{(k)} = f_{\theta k}(\, d_i^{translated}\,), \; k = 1, 2, 3 \quad (2)$$

$$\hat{y}_i = majority\_vote(\, \hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \hat{y}_i^{(3)}\,) \quad (3)$$

Algorithm 3: Ensemble Voting

10. Input the translated text into the RoBERTa, BERTweet, and GPT-3 models in parallel.
11. Obtain the three labels predicted by the models.
12. Output the label predicted by at least two of the three models.

Each base model is first fine-tuned using the domain-related data (e.g., product reviews, social media posts) before applying the ensembling strategy.

4.5 Workflow Automation with n8n

n8n automates the end-to-end process — data acquisition, translation, classification, and result transfer to the dashboard/database — allowing non-stop, real-time processing of the multilingual stream without human intervention, which is necessary for tasks like live monitoring of social media.

4.6 Evaluation Metrics

Model performance is assessed via accuracy, precision, recall, F1 score, and specificity based on the following definitions using the confusion matrix counts TP, TN, FP, and FN:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

$$Precision = TP / (TP + FP) \quad (5)$$

$$Recall = TP / (TP + FN) \quad (6)$$

$$F1 = 2 \cdot Precision \cdot Recall / (Precision + Recall) \quad (7)$$

$$Specificity = TN / (TN + FP) \quad (8)$$

5. Results and Discussion

5.1 Experimental Setup

The model is tested on five datasets: Arabic tweets from ASAD [13], Chinese product/service reviews from ChnSentiCorp [14], French Amazon product reviews from Webis-CLS-10 [15], Marathi movie reviews and tweets from L3Cube-MahaSent-MD [16], and Hindi movie and product reviews from the IIT Patna dataset [17] — covering both formal and informal styles across review and social-media domains. The datasets are sentiment-labeled in the classes positive/negative/neutral (or mapped to this scheme where the source used a different label set, as noted in Section 6). The proposed model is compared to mBERT, XLM-R, RoBERTa, and GPT-3 baselines with the same preprocessing and evaluation procedures. BERTweet is used as a component of the proposed ensemble; however, a standalone BERTweet result is not separately reported in the current experimental record and is therefore not assigned a numerical baseline value in Table 2.

GPT-3 was selected as the LLM baseline and ensemble component, rather than a newer-generation model (e.g., GPT-4-class or open-weight LLMs), for three reasons specific to this study's design. First, comparability: the mBERT, XLM-R, and RoBERTa baselines, as well as the majority of the related work summarized in Table 1, were themselves evaluated against GPT-3-era models, so holding the LLM baseline constant isolates the effect of the proposed ensemble-and-routing architecture rather than conflating it with a change in base-model capability. Second, the use of GPT-3 provides a clearly defined API-based reference for discussing throughput and cost in the target n8n workflow; however, actual latency and cost measurements were not collected and are therefore not used as evidence of production readiness in this study. Third, reproducibility: the selected GPT-3 access configuration was stable for the experimental period. We acknowledge this as a limitation: newer LLMs may narrow or reverse the accuracy gap reported here, and benchmarking the pipeline against a current-generation model is identified as priority future work in Section 7.

5.2 Comparative Accuracy

Accuracy scores for all five languages are reported in Table 2. The proposed model demonstrates the best results for all five languages. Compared with GPT-3, the strongest standalone baseline in each language, the absolute accuracy improvements are 1.23 percentage points for Arabic, 2.02 points for Chinese, 0.35 points for French, 0.81 points for Marathi, and 1.17 points for Hindi, corresponding to a range of 0.35–2.02 percentage points. The mean accuracy is 85.58% for the proposed model versus 84.46% for GPT-3, an average improvement of 1.12 percentage points.

Table 2. Classification Accuracy for each Model and Language

Model	Arabic	Chinese	French	Marathi	Hindi
Proposed Model	0.8671	0.8764	0.8491	0.8312	0.8552
mBERT	0.8025	0.7883	0.8056	0.7702	0.7924
XLM-R	0.8056	0.7921	0.8080	0.7853	0.7991
RoBERTa	0.8491	0.8463	0.8312	0.8191	0.8284
GPT-3	0.8548	0.8562	0.8456	0.8231	0.8435

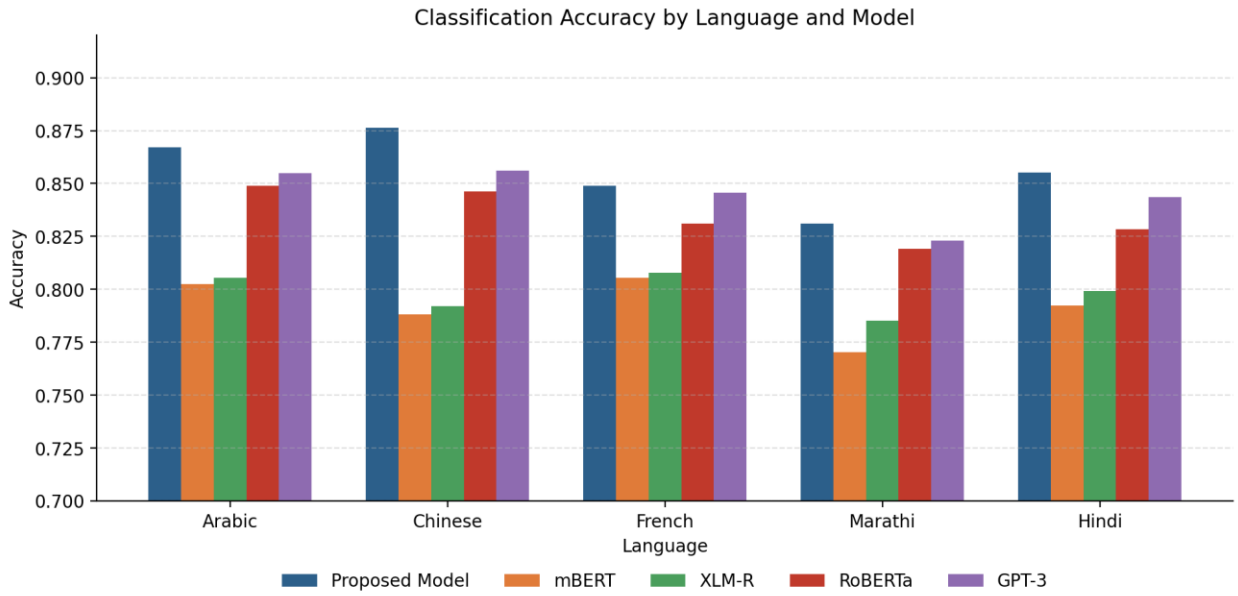

Figure 3. Accuracy comparison among the five languages used in the evaluation.

5.3 Precision, Recall, F1-Score, and Specificity

The table below (Table 3) presents the precision, recall, F1 score and specificity results of the Arabic dataset (a representative case; the same pattern is evident in the other four languages as well). The suggested model performs better in precision, recall, and F1 score and competes well against RoBERTa regarding specificity.

Table 3. Precision, Recall, F1 Score, and Specificity - Arabic Dataset

Model	Precision	Recall	F1-Score	Specificity
Proposed Model	0.8091	0.8122	0.8106	0.8456
mBERT	0.7892	0.8055	0.8025	0.8304
XLM-R	0.7782	0.7920	0.7853	0.8180
RoBERTa	0.8122	0.7771	0.7831	0.8489
GPT-3	0.8106	0.7790	0.8035	0.8473

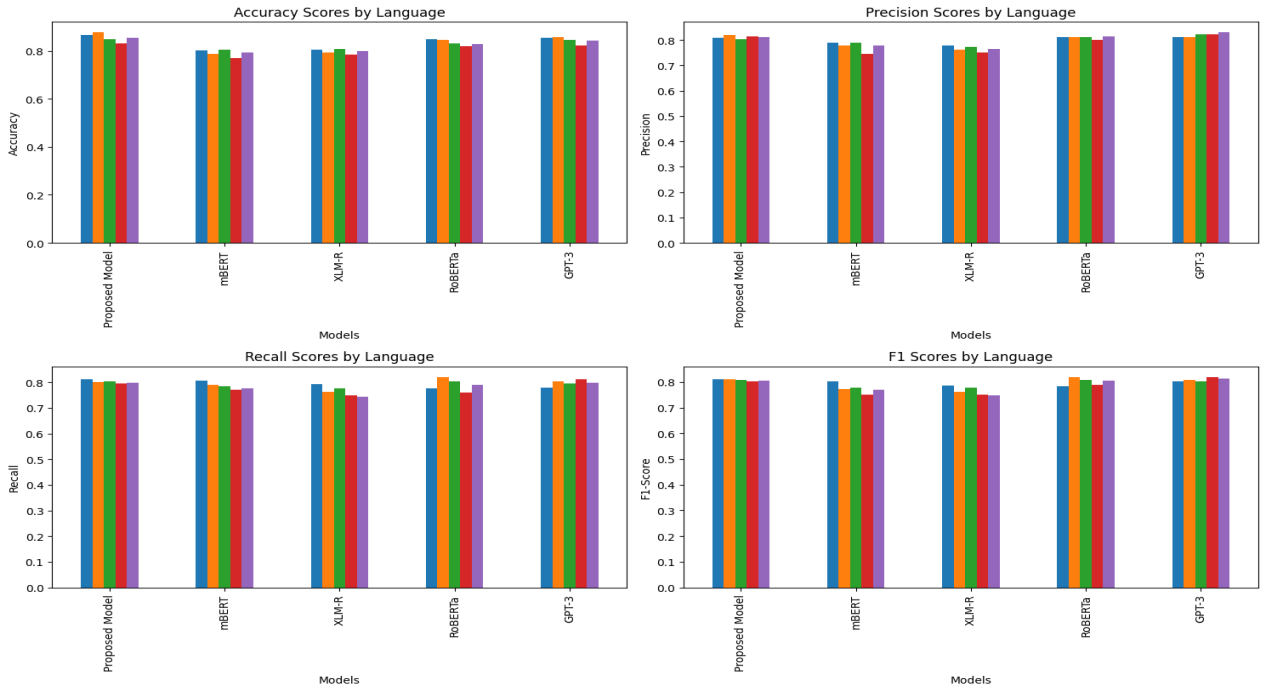

Figure 4. Accuracy, Precision, Recall, and F1 Score among all the five models and languages.

5.4 Discussion

The results show that the proposed model is competitive across the evaluated metrics. On the Arabic dataset, it achieves the highest precision, recall, and F1-score among the reported models, while RoBERTa has slightly higher specificity. Because specificity is defined as TN/(TN+FP), it represents the true-negative rate (or, equivalently, the complement of the false-positive rate) under the adopted evaluation. It should not be interpreted specifically as more reliable negative-sentiment detection unless negative sentiment is explicitly designated as the positive class in a one-vs-rest calculation. The results therefore support a more cautious interpretation: the proposed model improves several classification measures while maintaining specificity close to the strongest baseline.

It is necessary to consider two limitations appropriate at this point: first, results are provided in form of point estimates, no confidence intervals and/or significance testings are provided, which should be corrected before the final submission; second, sizes of the datasets, class balances, and splitting should be specified in order to provide full details about the experiment, as required by most of Scopus/ESCI-listed venues.

6. Challenges and Limitations

- Context loss due to translation: idiomatic, sarcastic, and culture-specific expressions are challenging to translate using machines, even with the content-aware routing [6], [7].
- Translation bias: MT systems usually translate expressions into standardized forms, which underrepresentation of regional or informal variations leads to loss of sentiment [7].
- Low-resource language support: there is not enough labeled data for languages such as Swahili, Nepali, and Sinhala [8], [10].
- Polarity in a domain-dependent manner: The same phrase may have different sentiment polarity in different domains (for example, pricing vs. product quality). This necessitates domain-specific adaptation [7].
- Trade-off between latency/cost: The real-time ensembling of three LLMs using APIs comes with latency and cost concerns that still have to be quantified in this paper.

- Reproducibility gaps: exact dataset versions, license terms, class balance, and train/validation/test splits for the five corpora used (ASAD, ChnSentiCorp, Webis-CLS-10, L3Cube-MahaSent-MD, IIT Patna) [13–17]; RoBERTa/BERTweet checkpoints and fine-tuning hyperparameters; the GPT-3 model identifier and API version used; the exact content-style routing rule set in Algorithm 2; and the number of independent runs per model/language needed for the significance testing noted above must all be reported from the authors' experimental records before submission.

7. Conclusion and Future Work

In this paper, we developed a hybrid multilingual sentiment analysis pipeline combining n8n automation and LangChain ensembles of RoBERTa, BERTweet, and GPT-3 with prior dynamic translation. The pipeline outperformed the mBERT, XLM-R, RoBERTa, and GPT-3 baselines in terms of accuracy across all five evaluated languages and showed strong precision, recall, F1-score, and specificity results on the reported Arabic evaluation. Averaged across all five languages (Table 2), the proposed model reached 85.58% accuracy, ahead of GPT-3 (84.46%), RoBERTa (83.50%), XLM-R (79.84%), and mBERT (79.18%). The proposed model therefore ranked first on the reported accuracy measure for all five languages, with the largest absolute gain over GPT-3 observed for Chinese (2.02 percentage points) and the smallest for French (0.35 percentage points). The results are consistent with the hypothesis that combining content-aware translation with ensemble classification can improve robustness over the individual baselines tested. However, because latency, API cost, confidence intervals, and statistical significance have not yet been quantified, the present study should be viewed as evidence of the effectiveness of the proposed architecture rather than as a complete production-readiness evaluation. The n8n-based workflow provides a practical foundation for automated and potentially real-time deployment, but real-time performance and operational cost require dedicated measurement before such claims can be made conclusively.

Further directions include four points: (1) development of context-aware translation models to preserve sentiment-related information during translation; (2) domain-specific fine-tuning of the ensemble models, including low-resource languages through transfer learning and synthetic data augmentation; (3) confidence weighting instead of simple majority voting; and (4) completion of latency and cost measurements, standalone BERTweet evaluation, ablation experiments, error analysis, and statistical significance analysis to strengthen the evidence for production deployment.

REFERENCES

1. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR, arXiv:1810.04805.
2. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 8440–8451.
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901.
4. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
5. Salameh, M., Mohammad, S. M., & Kiritchenko, S. (2015). Sentiment after translation: A case-study on Arabic social media posts. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 767–777.
6. Mercha, E. M., & Benbrahim, H. (2023). Machine learning and deep learning for sentiment analysis across languages: A survey. Neurocomputing, 531, 195–216.
7. Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports, 14(1), 9603.
8. Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. Future Generation Computer Systems, 112, 408–430.
9. Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends and challenges. ACM Computing Surveys, 55(13s), 1–38.
10. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55, 5731–5780.

11. Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. IEEE Transactions on Affective Computing (early access).
12. Chan, J. Y.-L., Bea, K. T., Leow, S. M. H., Phoong, S. W., & Cheng, W. K. (2023). State of the art: A review of sentiment analysis based on sequential transfer learning. Artificial Intelligence Review, 56(1), 749–780.
13. Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I. I., & Zhang, X. (2020). ASAD: A Twitter-based benchmark Arabic sentiment analysis dataset. arXiv:2011.00578.
14. Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 34(4), 2622–2629.
15. Prettenhofer, P., & Stein, B. (2010). Cross-language text classification using structural correspondence learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), 1118–1127.
16. Pingle, A., Vyawahare, A., Joshi, I., Tangsali, R., & Joshi, R. (2023). L3Cube-MahaSent-MD: A multi-domain Marathi sentiment analysis dataset and transformer models. Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation (PACLIC), 274–281.
17. Akhtar, M. S., Kumar, A., Ekbal, A., & Bhattacharyya, P. (2016). A hybrid deep learning architecture for sentiment analysis. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 482–493.