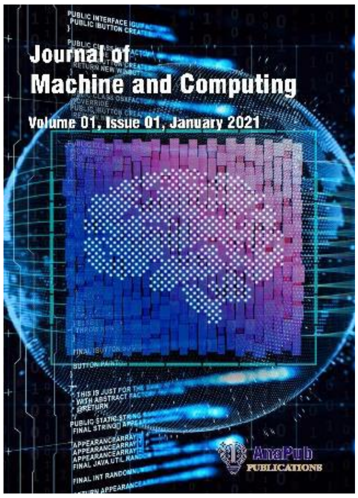
Journal Pre-proof

Comparing Multilingual Emoji-Enhanced Product Reviews: A Transformer-Based Approach for Language Pair and Emotion Detection



Priyanka Sharma and Ganesh Gopal Devarajan

DOI: 10.53759/7669/jmc202505063

Reference: JMC202505063

Journal: Journal of Machine and Computing.

Received 10 May 2024

Revised form 23 December 2024

Accepted 14 February 2025

Please cite this article as: Priyanka Sharma and Ganesh Gopal Devarajan, “Comparing Multilingual Emoji-Enhanced Product Reviews: A Transformer-Based Approach for Language Pair and Emotion Detection”, Journal of Machine and Computing. (2025). Doi: https://doi.org/10.53759/7669/jmc202505063

Comparing Multilingual Emoji-Enhanced Product Reviews: A Transformer-Based Approach for Language Pair and Emotion Detection

Priyanka Sharma[1], Ganesh Gopal Devarajan[2,*]

[1,2] Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Delhi-NCR Campus, Delhi – Meerut Road, Modinagar, Ghaziabad, Uttar Pradesh 201204, India.

*Corresponding author.
E-mail: ps9627@srmist.edu.in (P. Sharma),
dganeshgopal@gmail.com (G. Devarajan)

Abstract

This paper presents a multilingual sentiment analysis approach leveraging two transformer-based architectures—XLM-RoBERTa (base) and BERT-based multilingual cased—to classify sentiment across four language pairs (English–Spanish, English–French, English–Hindi, and English–Italian). We fine-tune XLM-RoBERTa by unfreezing only its last three layers to adapt the model to domain-specific sentiment cues while preserving its robust cross-lingual representations. Training over ten epochs yields a best validation accuracy of 0.95 and a test accuracy of 0.975, with an average F1-score around 0.92–0.97 across the four language pairs. The BERT-based multilingual cased model achieves a slightly higher test accuracy of 0.98, demonstrating comparable or improved performance in capturing sentiment nuances. These results confirm that selectively fine-tuning large-scale multilingual encoders is an effective strategy for cross-lingual sentiment classification, achieving high accuracy and strong generalization.

Keywords: Natural Language Processing, Language pair Identification, Transformers, XLM-RoBERTa, BERT, mT5, Emotion Detection.

1 Introduction

The increasing volume of multilingual user-generated content especially in e-commerce and social media domains has accelerated the need for robust natural language processing (NLP) systems. Users often interleave words from different languages (code-mixed text) and rely on emojis to express sentiment. Traditional monolingual models tend to underperform when confronted with these informal, mixed-lingua data streams.

Recent advances in natural language processing (NLP) have seen the rise of transformer-based architectures [1], offering state-of-the-art performance on various language tasks. In cross-lingual sentiment analysis, the ability to generalize across languages is crucial particularly in scenarios with limited labeled data for lower-resource languages. This paper investigates the efficacy of XLM-RoBERTa (base) and BERT-based [3] multilingual cased models in classifying sentiment [9] for multiple English-centric language pairs: English–Spanish (en-es), English–French (en-fr), English–Hindi (en-hi), and English–Italian (en-it).

In this paper, we propose a system that compares the fine-tuned versions of XLM-RoBERTa and BERT-based multilingual cased [13] models for accurate detection of language pairs in user reviews with base models like Roberta. We also integrate a simple but effective keyword and emoji-based sentiment classifier [18]. This presents new challenges for natural language understanding because code mix reviews involve interspersed use of languages, e.g. for English and Hindi language pairs (e.g., “Yeh product bohot accha hai ❤️, totally worth the price”).

Fig 1. illustrates the overall pipeline for handling multilingual product reviews containing emojis. Starting with the dataset (comprising English-Hindi, English-Spanish, English-French, and English-Italian reviews), the process begins with data pre-processing (e.g., cleaning, normalizing, and preserving emojis) followed by tokenization. From there, the pipeline branches into two paths: one prepares input for a classification model (such as XLM-RoBERTa or mBERT) by converting each review into tokens suitable for language-pair detection; the other produces tokens for a text-to-text model (mT5) using prompt-based fine-tuning, which can support additional tasks like rewriting or summarization. After fine-tuning on the respective models, the system integrates outputs into the proposed approach where emotion detection (based on emojis and keywords) and language pair identification (e.g., en-hi) are performed on each input review. Finally, both results feed into the system's output, offering sentiment [9] insights and language-pair labels for downstream applications.

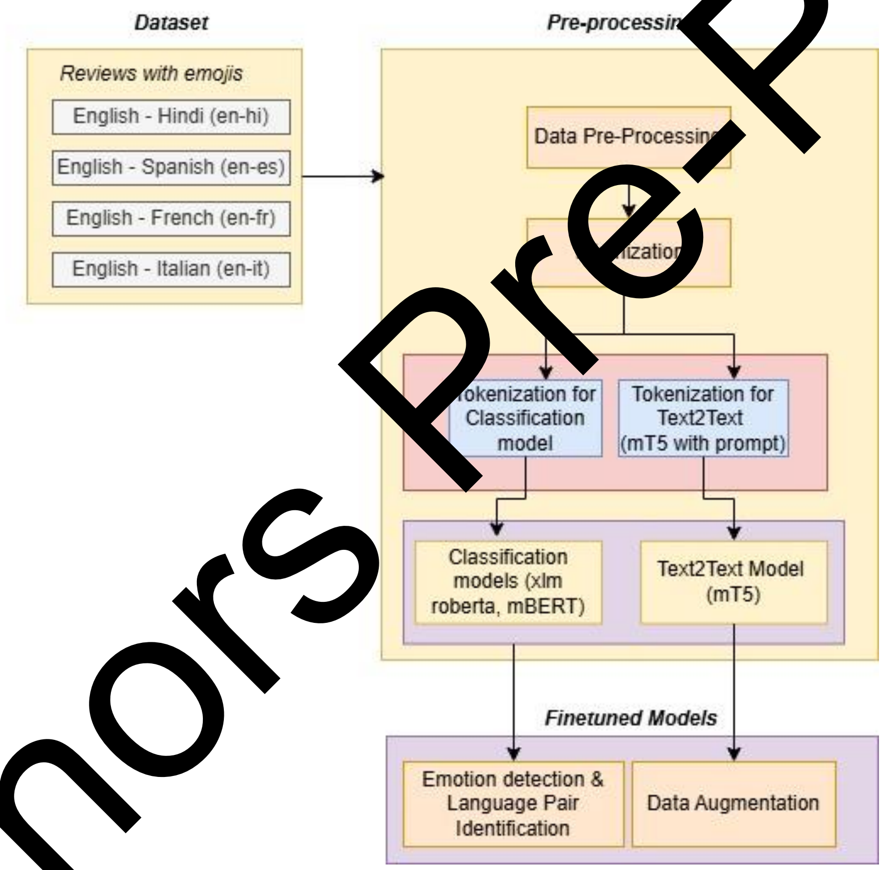


Fig. 1. High-level architecture for handling multilingual product reviews containing emojis.

1.1 Motivation

The primary motivation for this research arises from two observations: first, the increasing prevalence of multilingual, code-mixed reviews—often combining languages like English with Hindi, Spanish, French, or Italian—and second, the heightened use of emojis to express sentiment in informal text. Monolingual NLP models struggle to parse these blended linguistic forms, and ignoring emojis can lead to an incomplete understanding of user sentiment. To address these challenges, our system integrates language pair identification with emotion detection. The pipeline combines a classification-based approach for language recognition (e.g., XLM-RoBERTa or mBERT) with a text-to-text model (mT5) for more advanced processing [14]. By preserving emoji-based cues and refining code-mixed tokenization, the framework provides enriched context for accurate sentiment insights and robust language-specific analysis.

The remainder of this paper is organized as follows: Section 2 introduces the proposed system architecture and methodology. Section 3 outlines our experimental setup and implementation details. Section 4 presents the results and corresponding discussion, and Section 5 concludes the paper with potential future directions.

1.2 Literature Survey

Code-Mixed Text Analytics in which the author described that social media in multilingual regions often contains code-mixed text, leading to unique computational linguistic challenges. Hinglish (Hindi-English) is a prime example of code-mixing that uses either Latin or Devanagari scripts, sometimes both, for words and phrases. Similarly, language pairs like (English-Spanish), (English- French) and (English- Italian) has the morphological and syntactic irregularities of code-mixed[4] text complicate tasks such as part-of-speech tagging and sentiment analysis. Approaches using multilingual language models (mT5, XLM-R) have shown promise for handling such diversity in text data.

In the study Emoji [5] and sentiment analysis, Emojis add emotional or contextual nuance to textual content. Research has shown that ignoring emojis can lead to misinterpretations of sentiment, especially in informal communication. Models that incorporate emoji embeddings or treat emojis as separate tokens often perform better in sentiment classification tasks [6]. For instance, emoji-based features can boost sentiment detection [illegible] in social media data by capturing user emotions more accurately.

Product reviews provide valuable user feedback that influences consumer decisions. Text-based recommendation systems usually rely on star ratings and latent representation of review text via word embeddings, TF-IDF, or transformer-based embeddings). However, user opinions expressed in code-mixed scripts require specialized preprocessing to capture contextual meaning. Incorporating advanced deep learning techniques for sentiment [7] scoring can refine product ranking and lead to improved user satisfaction.

2 Method

2.1 Proposed System Methodology

Fig. 2 illustrates our approach, which enhances multilingual review analysis by integrating code-switching, emoji processing, and fine-tuning advanced NLP models. It begins with multilingual training data (en-hi, en-es, en-fr, en-it), which is used to generate a Parallel Seed Set. The mT5 model is fine-tuned to create code-switched [24] sentences, followed by data augmentation [19] to generate synthetic code-switched data. This is combined with the original dataset for improved representation. Simultaneously, emoji-rich text is incorporated, forming a Re-Combined Training Set. The dataset undergoes text normalization, transliteration, and emoji processing before classifier fine-tuning using XLM-Roberta and mBERT. These models are optimized for emotion detection and language pair identification.

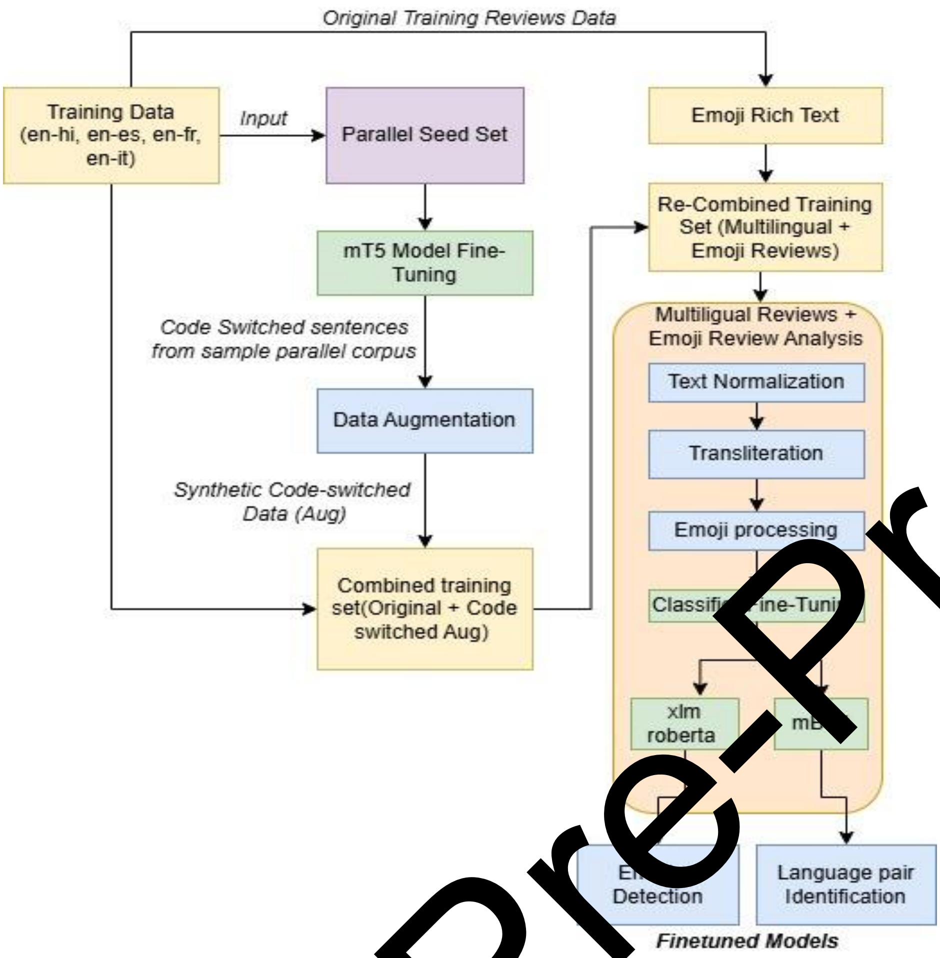


Fig. 2 Proposed Methodology for multilingual online reviews for en-hi, en-es, en-fr and en-it language pairs with emojis.

The parallel seed set provides a foundational collection of English–target language sentences labelled for sentiment or containing minimal annotations. Leveraging this small seed corpus, we employ mT5 fine-tuning to train the model to generate code-switched text [15]. Subsequently, the synthetic code-switched data (Aug) produced by the fine-tuned mT5 model significantly broadens the training pool by converting any additional monolingual data into code-mixed samples, thus addressing data gaps for lower-resource languages or language mixing scenarios.

We then consolidate the newly created examples with the original dataset, yielding a combined training set that blends authentic and augmented code-switched data. Finally, we fine-tune a sentiment [10] classifier, such as XLM-RoBERTa (with only its final three layers unfrozen) or a fully fine-tuned BERT-based multilingual cased model, by adding a classification head to produce sentiment scores. This arrangement ensures the classifier learns robust representations across diverse inputs, especially where language mixing is prevalent. The synergy between mT5-based data augmentation [20] and partial or full fine-tuning of these transformer[3] architectures boosts sentiment accuracy across multiple language pairs, facilitating more effective handling of code-switched inputs and reducing reliance on extensive fully labelled corpora.

2.2 Data Collection and Processing

Data is collected from different e-commerce sources, combining multilingual text reviews that contain emojis and star ratings into a unified structure. Each product has a unique product ID along with manually labelled language pair (en-hi, en-es, en-fr, en-it).

2.2.1 Text Normalization: First, raw text undergoes a normalization process that removes extraneous punctuation, corrects spacing inconsistencies, and standardizes letter cases. This helps reduce noise and improve downstream processing.

2.2.2 Transliteration: Since Hinglish typically blends Hindi words (written in the Devanagari script) with English words (Roman script), or may present Hindi words in Roman transliteration, the pipeline applies a transliteration module. This module systematically converts any Hindi words from Roman script back into standardized Devanagari script (or vice versa, depending on the target representation). This ensures that linguistic features of Hindi are preserved and recognized correctly in subsequent steps. For other language pairs, this step is not required.

2.2.3 Language pair Text Normalization

Cross lingual [2] text is highly variable; for instance, “Dekh yaar, yeh product bahut accha hai!” may appear as “Dekh yaar, yeh product bohot acha hai” or “Dekh yarr, yeh product bahut accha hey!”. To address these inconsistencies, we employ:

Spell Correction: Identify and correct commonly misspelled Hindi transliterations (e.g., “bohot” → “bahut”).

Stopword Removal: Remove or reduce influence of common filler words (e.g., “yaar,” “hey,” etc.).

Emoji Tokenization: Convert emojis to specific tokens (e.g., “❤️” → “<red_heart_emoji>” or embedded as a special token.

2.2.3 Data Format

Fig. 3. shows a sample of the parallel seed training data. This dataset contains user-generated product reviews that exhibit code-mixing and multilingual phrases (e.g., English, Hindi, Spanish, Italian, and French), along with emojis and informal expressions. Each row pairs the original text with its translated or rewritten output in another language, capturing the nuances of style and sentiment [10] across cultures

Input	Output
"ye t-shirt acchi hai, Full paisa wasool ❤️	ये टी-शर्ट अच्छी है, पूरा पैसा वसूल ❤️
"product ki installation difficult hai, regret it 😞"	"उत्पाद की स्थापना कठिन है, इसका पछतावा है 😞"
"amazing design, quality bhi top-notch!" 😍	"अद्भुत डिजाइन, गुणवत्ता भी श्रेष्ठतम!" 😍
"Este producto es amazing, quality is top-notch!" 😍	"Este producto es increíble, ¡la calidad es de primera categoría!" 😍
"Muy good product, worth every penny!" 😊	"Muy good product, worth every penny!" 😊
"I love this product, exceeded my expectations!" ❤️	"Me encanta este producto, ¡superó mis expectativas!" ❤️
"Questo prodotto è eccezionale, totally worth it!" 😊	"Ce produit est exceptionnel, ça vaut vraiment le coup !" 😊
"I love this prodotto, quality is amazing!" 😍❤️	"J'adore ce produit, la qualité est incroyable !" 😍❤️
"Ottima performance, and the design è fantastico!" ❤️	"Ottima performance, et le design è fantastico !" ❤️
"Ce produit est incroyable, totally worth it!" 😊	"Questo prodotto è fantastico, ne vale assolutamente la pena!" 😊
"I love this produit, quality is top-notch!" 😍	"Adoro questo prodotto, la qualità è di prim'ordine!" 😍
"Great performance, et le design est magnifique!" ❤️	"Ottime prestazioni e il design è bellissimo!" ❤️

Fig. 3 Sample Parallel seed Training data for mT5-xl

Fig 4. shows the short user-generated reviews in various language pairs, complete with user IDs, product IDs, star ratings, sentiment scores, and emojis. The Review Text column provides code-mixed content, informal expressions, and emojis that reflect users' emotions and opinions, while the Language Pair column explicitly labels which languages are being used.

User ID	Product ID	Language Pair	Review Text	Star Rating	Sentiment Score	Extracted Emojis
U111	P100	en-hi	"Yeh shoe mast hai, totally pasand ❤️"	5	0.90	❤️
U222	P200	en-es	"Me encanta este vestido, super guay 😍"	4	0.75	😍
U333	P300	en-fr	"Ce produit est vraiment cool, I love it 😍"	3	0.50	😍
U444	P400	en-it	"Questo cappello is amazing, perfect fit 😍"	5	0.85	😍
U555	P999	en-hi	"Yeh product bekar hai 😞, not recommended."	2	-0.70	😞

Fig. 4 Multilingual Language paired Reviews with emojis and sentiment score

2.3 Feature Extraction

2.3.1 Text Embedding Generation

To handle code-mixed languages and emojis, a multilingual T5 (mT5-XL) model is fine-tuned. First with sub-word Tokenization where the model's tokenizer splits text into subwords, capturing Hindi, Spanish, Italian, French and English words. Emojis are tokenized as separate tokens. Then, Contextual Embeddings which generate embeddings for each token, aggregated (e.g., by taking the [CLS] token representation or the average of all token embeddings) into a fixed-length vector. Last Fine-Tuning for Sentiment where a model is fine-tuned on a labeled sentiment dataset to produce better sentiment [12] embeddings that incorporate emoji-based emotional cues.

2.3.2 Sentiment Scoring

We fine-tune XLM-RoBERTa on parallel English-to-Hindi, English-to-Spanish, English-to-French, and English-to-Italian datasets to generate sentiment scores in the range [-1, +1]. A dedicated sentiment classification head is added on top of the XLM-RoBERTa encoder and trained using standard cross-entropy loss for multiple epochs with a moderate learning rate (e.g., 3e-5) and batch size (16). Early encoder layers can be partially frozen to preserve multilingual representations, while upper layers adapt to domain-specific sentiment cues.

Additionally, an mBERT-based sentiment [16] classifier is fine-tuned on the same curated dataset, which includes language mixing and emojis, to improve sentiment prediction accuracy. This fine-tuning process ensures that mBERT effectively captures multilingual sentiment patterns and emoji-based expressions, allowing for a comparative evaluation against the XLM-RoBERTa-based sentiment model.

2.3.3 Emoji Processing

Given that emoji can carry sentiment and semantic cues, the pipeline treats them as meaningful tokens rather than discarding or ignoring them. Each emoji is mapped to a standardized label or short textual description, allowing the model to incorporate its emotional or contextual significance into the text representation.

2.3.4 Emotion Detection

Emotion detection is applied together with the main language modeling process. Instead of categorizing each review as simply positive, negative, or neutral, the pipeline classifies the text into more specific emotion categories, such as happy, sad, disappointed, angry, good, bad, etc. These emotions are then appended as additional features or integrated into the representation itself, enabling the system to better capture and utilize user attitudes when generating recommendations or summaries.

2.4 Model Architecture and Recommendation

For each product in the database, we aggregate the textual embeddings of its reviews into a final product-level textual embedding vector:

$$ProductEmbedding_{text} = \frac{1}{N}\sum_{i=1}^{N}(E(Review_i).S(Review_i))$$

where $E(Review_i)$ is the embedding of review i and $S(Review_i)$ is its sentiment score. Multiplying embedding vectors by sentiment scores weights the contribution of each review by its positivity/negativity magnitude.

We also have a product-level emoji embedding:

$$ProductEmbedding_{emoji} = \frac{1}{M}\sum_{i=1}^{M}(E(ProductEmoji_j)$$

where $E(ProductEmoji)$ is the emoji embedding.

The final product embedding (ProductEmbed) is a concatenation or fusion (e.g., element-wise addition) of ProductEmbedding_text and ProductEmbedding_emoji

$$ProductEmbed = Fuse(ProductEmbedding_{text} + ProductEmbedding_{emoji})$$

3 Experimental Setup and Implementation

3.1 Implementation Details

The implementation of the proposed multilingual [17] language paired model necessitates the utilization of advanced programming languages and libraries optimized for deep learning and natural language processing tasks. Python serves as the primary language, leveraging frameworks such as PyTorch for the development and deployment of deep learning models. The text processing pipeline incorporates the Transformers [22] library by Hugging Face, specifically employing mT5-xl based embeddings for enhanced textual representation. To ensure computational efficiency, particularly for large-scale model training was conducted using both CPUs and 20 GB of GPU RAM Nvidia A100 machine is employed, significantly accelerating processing speeds and optimizing system performance.

3.2 Implementation Steps

The following Implementation steps are followed:

3.2.1 Data Loading: Load the review dataset into memory.

3.2.2. Review Preprocessing:

- Code-mixed text normalization.
- Emoji tokenization.

3.2.3. Embedding Generation: Generate or load pretrained mT5-xl embeddings for each review.

3.2.4 Aggregation: For each product, combine textual embeddings of reviews (weighted by sentiment) into a single vector. Similarly, average reference image embeddings for each product.

3.2.5 Fusion: Fuse textual and emojis embeddings to obtain a single product-level embedding.

3.3 Dataset Description

As seen in Table 1, we curated a dataset consisting of four language pair with emojis. Each language pair has around 500 user reviews, for a combined corpus of 2,000 reviews. The textual content of these reviews primarily includes code-mixed en-hi, en-es, en-fr, en-it with an average of 8 tokens per review, and approximately 80% of them contain emojis. This composition reflects a realistic e-commerce setting where informal, code-mixed language is prevalent, capturing the nuances of user feedback in a multilingual environment.

Table 1 Dataset Statistics

Statistics	Value
No. of language pair	4
Total Reviews	2000
Emoji Occurrence Rate	~80% of reviews
Avg. Tokens/Review	10
Language Mix (Hindi/English)	50% Eng, 50% Hindi
Language Mix (French/English)	60& Eng 40 % French
Language Mix (Spanish/English)	60& Eng 40 % Spanish
Language Mix (Italian/English)	60& Eng 40 % Italian

3.4 Finetuning and Hyper-parameter details

Given computational constraints and the strong multilingual features of mT5, XLM-RoBERTa, and mBERT, we unfreeze only the last three layers for fine-tuning while keeping the lower layers frozen [2, 3, 14]. This approach retains cross-lingual knowledge in the frozen layers while enabling domain adaptation in the final layers.

3.4.1 Text Encoder

Preprocessing:
An 80:20 train-validation split is applied to ensure a balanced distribution of language pairs in both subsets. Emojis within the text are replaced with special tokens, such as <heart_emoji>, to standardize input representation.

Forward Pass:
The tokenized sequences are processed through a partially unfrozen model. The [CLS] (or pooled) embedding vector is extracted and passed to a classification head, which maps the representation to the corresponding number of language pairs.

Loss and Optimization:

Cross-Entropy Loss is computed between the predicted logits and the true language pair labels. The AdamW optimizer is employed to update the parameters of the last three layers and the classification head, ensuring effective learning.

Hyperparameters:
The model is trained with a learning rate of 3e-5 and a batch size of 8. Training is conducted for 10 epochs, with a weight decay of 0.01 to prevent overfitting. The optimizer used is AdamW, maintaining a learning rate of 3e-5. Early stopping is implemented by monitoring validation accuracy, with the best-performing checkpoint restored to ensure optimal model performance.

3.4.2 Integration of Emoji Embeddings

Extend the base tokenizer's vocabulary with frequently used emojis. Align with the base model dimension. The sentiment head is explicitly trained to interpret these emoji tokens as carrying emotional signals.

4 Results

We evaluated the system on a held-out test set (with 4 language pairs 2,000 reviews). We measure recommendation accuracy using precision, recall and F1 score. We also separately analyze the correctness of sentiment scoring for code-mixed language reviews.

Table 2 Comparison of performance among models with proposed methodology.

Models	precision	recall	f1score
RoBERTa-base	0.68	0.85	0.72
XLM-RoBERTa	0.74	0.87	0.80
mBERT	0.78	0.91	0.81
XLM-RoBERTa-base (en-hi + Emoji)	0.91	0.92	0.87
XLM-RoBERTa-base (en-es + Emoji)	0.93	0.92	0.89
XLM-RoBERTa-base (en-fr +Emoji)	0.94	0.90	0.85
XLM-RoBERTa-base (en-it + Emoji)	0.96	0.89	0.82
bert-base-multilingual-cased (en-hi + Emoji)	0.90	0.94	0.92
bert-base-multilingual-cased (en-es + Emoji)	0.96	0.95	0.97
bert-base-multilingual-cased (en-fr +Emoji)	0.94	0.92	0.93
bert-base-multilingual-cased (en-it + Emoji)	0.92	0.91	0.91

Table 2 compares the performance of several multilingual Transformer-based model on a classification task using precision, recall, and F1 score. The baseline models (RoBERTa-base, XLM-RoBERTa, and mBERT) yield moderate performance, with precision scores ranging from 0.68 to 0.78 and corresponding F1 scores between 0.72 and 0.81.

When additional language pairs and emoji are introduced (e.g., XLM-RoBERTa-base (en-hi + Emoji)), performance improves substantially in both precision and recall, resulting in higher F1 scores (up to 0.89). Notably, bert-base-multilingual-cased models with bilingual plus emoji inputs generally achieve the best overall results. For instance, bert-base-multilingual-cased (en-es + Emoji) obtains a precision of 0.96, recall of 0.95, and an F1 score of 0.97, which is the highest among all evaluated systems. These findings suggest that expanding the training data to include multilingual text and emojis consistently enhances model robustness and classification accuracy [5, 6].

4.1 Language Pair identification for multilingual reviews with Emojis

We subset of 2,000 labelled multilingual reviews (en-hi, en-es, en-fr, en-it) to measure language pair prediction accuracy. Fig. 5 presents sample predictions from our proposed approach on user reviews that mix Spanish and English (labeled "en-es"), often with emoji indicating sentiment.

Each example contains a blend of Spanish words (“Este reloj es muy estilizado,” “La durabilidad es dudosa,” etc.) interspersed with English phrases (“simply fantastic,” “not worth it”) and emojis expressing enthusiasm or dissatisfaction. Despite the linguistic code-switching and the additional visual cues provided by the emojis, the model consistently predicts the correct bilingual label ("en-es"), showing it can accurately detect language boundaries and sentiment-laden symbols in realistic review data [15, 19].

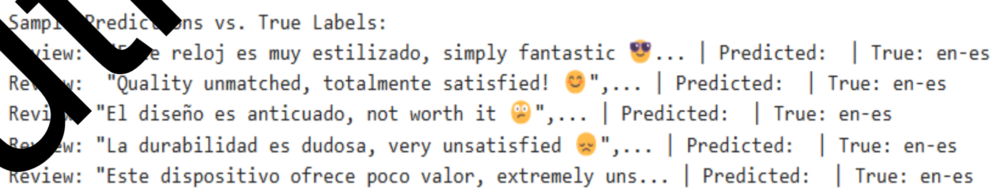

Sample Predictions vs. True Labels:
Review: "Este reloj es muy estilizado, simply fantastic ... | Predicted: | True: en-es
Review: "Quality unmatched, totalmente satisfied! ",... | Predicted: | True: en-es
Review: "El diseño es anticuado, not worth it ",... | Predicted: | True: en-es
Review: "La durabilidad es dudosa, very unsatisfied ",... | Predicted: | True: en-es
Review: "Este dispositivo ofrece poco valor, extremely uns... | Predicted: | True: en-es

Fig. 5 Predicted output for en-es review sample as input from proposed approach

4.2 Qualitative Analysis

Our qualitative analysis revealed interesting patterns in how language mixing and emoji usage correlated with sentiment scores. The given network graph[11] in Fig 6 illustrates the relationship between multilingual product reviews, sentiment categories, and product attributes. At the center of the graph, the "Unknown Product" node is linked to multiple language pairs (e.g., en-hi, en-es, en-fr, en-it), indicating that reviews for this product exist across different linguistic contexts. Additionally, the "T-shirt" node suggests that this visualization may be related to product-specific sentiment analysis.

Sentiment labels (happy, sad, anger, disappointed, bad, neutral) are connected to the central nodes, showing the range of emotions expressed in customer reviews. The edges between the language pairs and sentiment nodes imply that users from different linguistic backgrounds associate distinct sentiments with the product. For instance, English-Hindi (en-hi) and English-French (en-fr) might have more neutral or negative sentiment connections, while English-Spanish (en-es) and English-Italian (en-it) may lean toward a more positive perception.

This visualization helps in understanding sentiment distribution across multilingual feedback, which can be used to tailor product recommendations, enhance marketing strategies, and improve customer satisfaction based on localized sentiments.

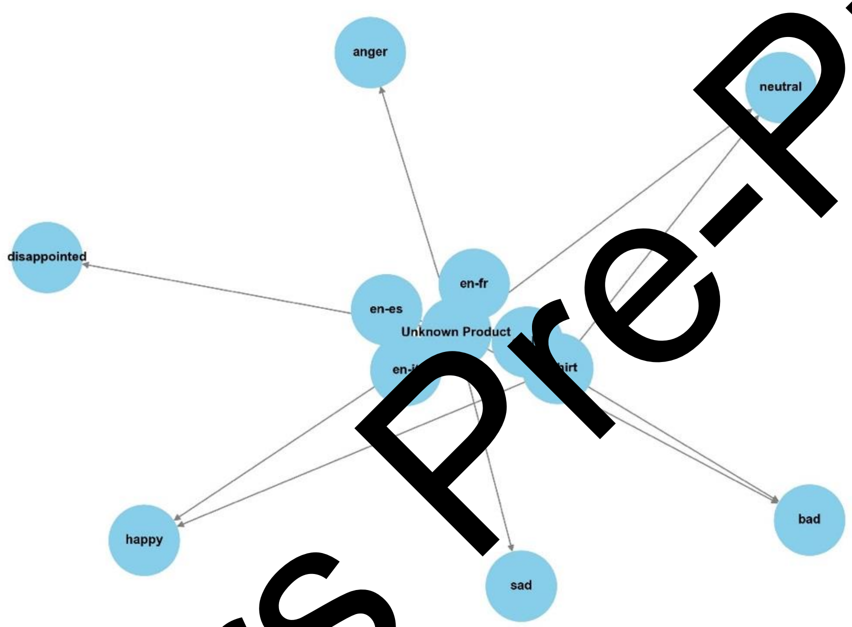


Fig. 6 Network graph showing multilingual product reviews linked to sentiment categories and product attributes

5 Conclusion and Future work

This paper introduces a multilingual pipeline leveraging transformer-based models—mT5 for data augmentation [24], XLM-RoBERTa and mBERT for fine-grained text classification—and incorporates specialized emoji processing and emotion detection to effectively handle multilingual, code-switched, and emoji-rich content. Experimental results confirm the pipeline's efficacy, especially with robust preprocessing and data augmentation [21].

Future research directions include developing more nuanced, transformer-based emotion detection, expanding language and dialect [23] coverage, implementing advanced domain adaptation and error correction techniques, and achieving real-time inference capabilities within industry-grade systems. This unified approach shows promise for managing complex multilingual data at scale with high interpretability.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Wang, Cindy, and Michele Banko. "Practical transformer-based multilingual text classification." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 121-129. 2021.

[2] A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised cross-lingual representation learning at scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451

[3] Puttaswamy, B. S., and N. Thillaiarasu. "Fine DenseNet based human personality recognition using english hand writing of non-native speakers." Biomedical Signal Processing and Control 99 (2025): 106910.

[4] S. Bhat, P. Y. Lakshmi, K. Bali, and M. Choudhury, "Code Mixing: A Challenge for Language Identification in the Indian Perspective," in *Proceedings of the 9th Workshop on Asian Language Resources (ALR9)*, 2014.

[5] G. Barbieri, F. Ronzano, M. Saggion, and H. Wanner, "What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis," in *LREC*, 2016.

[6] A. Wijeratne, Q. Balasuriya, D. Y. Sheth, and M. D. Goodman, "EmojiNet: Building a Machine Readable Sense Inventory for Emoji," in *Proceedings of the [illegible] International Conference on Social Informatics*, 2016.

[7] S. Liu, J. He, and B. Zhou, "Transformer-based joint deep learning for multilingual sentiment analysis," *IEEE Access*, vol. 8, 2020, pp. 301[illegible]0149.

[8] Singh, G.V., Ghosh, S., Firdaus, M. *et al.* Predicting multi-label emojis, emotions, and sentiments in code-mixed texts using an emojifying sentiments framework. *Sci Rep* 14, 12204 (2024). https://doi.org/10.1038/s41598-024-58944-5.

[9] Iseal, Sheed, et al. "Cross-Lingual Sentiment Analysis of E-Commerce Product Review." (2024).

[10] Ipa, Atia Shahnaz, et al. "BdSentiLLM: A Novel LLM Approach to Sentiment Analysis of Product Reviews." *IEEE Access* (2024).

[11] Khemani, B., Patil, S., Kotecha, K. *et al.* A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions". *J Big Data* **11**, 18 (2024). https://doi.org/10.1186/s40537-023-00876-4

[12] Tang, Tiancheng & Tang, Xinhuai & Yuan, Tianyi. (2020). "Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text". IEEE Access. 8. 193248-19[illegible]256. 10.1109/ACCESS.2020.3030468.

[13] Ashwin Shenoy, M., and N. Thillaiarasu. "Enhancing temple surveillance through human activity recognition: A novel dataset and YOLOv4-ConvLSTM approach." Journal of Intelligent & Fuzzy Systems Preprint (2023): 1-16.

[14] Khanuja, S., Dandapat, S., Sitaram, S., & Bali, K. (2020). GLUECoS: *An Evaluation Benchmark for Code-Switched NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL).*

[15] Akhter, [illegible] & S[illegible]a, S. (2021). Sentiment *Analysis on Code-Mixed Social Media Text in a Low-Resource Language. IEEE Transactions on Affective Computing.*

[16] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning *Transferable Visual Models from Natural Language Supervision (CLIP). International Conference on Machine Learning (ICML).*

[17] Schouten, K., & Frasincar, F. (2021). *Finding Emotion in Emojis: A Study of Emoji Use in Multilingual Sentiment Analysis. Information Processing & Management, [illegible]8(4).*

[18] Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). *TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[19] Zhou, X., Li, J., & Liu, Y. (2022). *Improving Low-Resource Question Answering with Cross-Lingual Data Augmentation.* ***.*** *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers)*, 3245–3257.*

[20] Kumar, R., & Singh, A. (2021). *Multi-Speaker TTS System for Low-Resource Languages Using Cross-Lingual Transfer Learning and Data Augmentation. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 6598–6602.

[21] Chen, J., Tam, D., Raffel, C., Bansal, M., & Yang, D. (2023). *An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. Transactions of the*

Association for Computational Linguistics, 11, 191–211.
[22] Bhat, S., Chandu, K. M., Java, S., & Black, A. W. (2021). *Evaluating Neural Approaches for Text Normalization in Low-Resource Scenarios. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).*
[23] Sitaram, S., Choudhury, M., & Bali, K. (2020). *A Survey of Code-Switched Speech and Language Processing. Frontiers in Artificial Intelligence*, 3, 17.
[24] Zhang B., Nakatani T., Hussey D. V., Walter S., & Tan L. (2024). *Don't Just Translate, Summarize Too: Cross-lingual Product Title Generation in E-commerce.* In Proceedings of the Seventh Workshop on e-Commerce and NLP @ LREC-COLING 2024, pages 58–64, Torino, Italia. ELRA and ICCL.