Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCED MULTILINGUAL TEXT CLASSIFICATION WITH PRE-TRAINED LANGUAGE MODELS

[1*]KANDULA NARASIMHARAO, [2]ANGARA S. V. JAYASRI
[1*]Research Scholar, Department of CSE, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh, India.
[2]Department of CSE, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh, India.
[1*]Corresponding Author Email id: kandulanarasimharao@gmail.com

ABSTRACT

Introduction: Multilingual text classification is an important task in natural language processing (NLP), especially in the context of sentiment analysis in languages with different grammatical, semantic, and lexical properties. Labelled data sets are not available for many languages, and hence it is difficult to predict the sentiment. The use of pre-defined language models and translation-based methods is a promising solution to cross-lingual generalization.
Objectives: This research has the following research objectives: (1) multilingual sentiment classification, (2) neural machine translation to translate foreign language data, (3) neural machine translation using transformer models and large language models (LLMs) combined with ensemble learning, and (4) multilingual model performance evaluation.
Methods: Four languages- Arabic, Chinese, French, and Italian are translated into English using neural machine translation tools: DeepL Translator and Google Translate. The translated text was then fed into an ensemble of 3 pre-trained models: Twitter-RoBERTa-Base-Sentiment-Latest, BERT-Base-Multilingual-Uncased-Sentiment and GPT-4 from OpenAI. The sentiment predictions were aggregated output in a weighted voting mechanism, which was based on performance metrics, such as accuracy, precision, recall, and F1-score.
Results: The ensemble model suggested here obtained an overall accuracy of greater than 96%, which was improved when compared to individual models. The ensemble learning with translation enhanced the understanding of context, robust across languages and showed a consistent level of performance in multilingual sentiment classification.
Conclusion: The obtained results support the fact that translation-based ensemble learning using pretrained transformers and LLMs is a scalable and accurate multilingual sentiment analysis approach. This paper shows how neural translation and the best pre-trained models can overcome language barriers in NLP.

Keywords: *Multilingual Sentiment Analysis; Pre-Trained Language Models; Ensemble Learning; Machine Translation; Large Language Models (LLMs).*

1. INTRODUCTION

Multilingual text organization is one of the research hotspots in the field of natural language processing. The process of classifying free-text texts into predetermined categories is known as text classification. With the use of natural language processing (NLP). NLP describes the interaction between human languages and computers. Here, the classifier is fed a text, and it returns a category based on the content [1]. Multilingual LMs combine flows of multilingual text. Information independent of a specific task, acquiring general knowledge. As evidence of this in the landscape, we discover various multilingual versions emerging from well- recognized monolingual LMs, which have now established themselves as a norm. Within the NLP community [2]. Recently, pretrained models have garnered considerable interest because of their outstanding results in a variety of downstream NLP tasks. During the pretraining phase, it uses language modeling tasks and acquires contextualized word representations by leveraging a vast quantity of training text. Furthermore,

multilingual pretrained models facilitate various NLP applications for different languages through zero-shot translation between languages. Cross-lingual transfer with zero shots has demonstrated encouraging outcomes for quickly developing applications in low-resource languages [3]. Pretrained language models (LMs) can be optimized to perform well on many tasks related to natural language processing (NLP). These models and their generative variations are being utilized more frequently to solve problems through straightforward text production that requires no fine-tuning. [4,5] BERT, RoBERTa, and their multilingual offspring, mBERT and XLM-R, are examples of large-scale monolingual pre-trained language models (MonoLMs) that have demonstrated encouraging outcomes for both high-resource and low-resource languages, especially for text categorization. BERT makes use of a multi-layer bidirectional transformer encoder that may be optimised for many purposes, such as text classification, after learning deep bidirectional representations [6]. Numerous variables affect how well the pre-trained Multilingual Language Models (MultiLMs) work, including the quantity of pre-training language data, a language's relatedness to other languages for the pre-trained model, its features, typography, and the quantity of fine-tuning data used [7]. Multilingual pre-trained language models (PLMs) have been shown to transfer to other languages despite the restrictions of language coverage. A number of low-resource languages that were not used in pretraining. However, compared to pre-training languages, there is still a significant performance disparity. Language adaptive fine-tuning (LAFT) is among the most effective ways to learn a new language; it involves fine-tuning a multilingual PLM on monolingual literature with the same pre-training goal in the target language. This has been demonstrated to result in significant increases on numerous low-resource languages and tasks, including cross-lingual transfer. Low-resource languages in the pretraining data are an orthogonal strategy to increase their coverage [8]. Online communications frequently involve code-mixing and code-switching. It is difficult to classify such texts if one of the languages has limited resources. One interesting approach to code-mixed text categorization is to refine pre-trained multilingual language models. [9]. Several methods, such as transformer-based text categorization models, are available in a range of real-world monolingual and multilingual pretraining and fine-tuning contexts. [10].

Transformer language modelling using pretraining and word representations paired with transfer learning has significantly improved several natural language understanding (NLU) tasks, such as text categorisation and natural language inference. [11]. By learning shared information across several related tasks, multi-task learning (MTL) improves performance compared to learning each task separately. Multi-task learning has been widely employed in text classification tasks because it leverages possible correlations between related tasks to identify shared characteristics and improve performance. [12] Deep learning techniques were used to increase the classifiers' accuracy, while text feature conversion and fusion techniques were used to solve the classifier's domain adaptability in many languages [13]. The most popular transfer learning strategy in deep learning is pre-training using self-supervised learning on large amounts of unannotated data. Pre-training techniques differ in that they employ unannotated data for self-supervised training and can be applied to a variety of downstream tasks through few-shot learning or fine-tuning. [14]. Lastly, PLMs offer better training parameters, higher training performance, and the lowest training cost. This is particularly crucial for certain training data projects that are more difficult to find. The task of training their own data may not properly train neural networks with very big parameters. Consequently, the model can be examined using a better initial state through the pre-training method, which can result in superior performance [15].

Research Questions

(a) How effectively do NMT integration and pre-trained Transformer models aid multilingual sentiment analysis on datasets collected from Arabic, Chinese, French, and Italian?

(b) How well does our ensemble method utilizing GPT-4, BERT-Multilingual and Twitter-RoBERTa perform against individual pre-trained models?

(c) RQ3: How well can our framework remain cross-lingual adaptable and maintain contextual awareness for varied NMT services and exhibit improved metrics such as accuracy, precision, recall, F1 score, and specificity?

The important contribution of this research is as follows:

- Developed a translation-based multilingual sentiment classification framework using DeepL Translator and Google Translate.

- Proposed an ensemble model combining transformer-based models and an LLM.
- Integrated Twitter-RoBERTa, BERT-Multilingual, and GPT-4 to enhance accuracy and robustness.

This study is prepared into separate sections for clarity and consistency. Section 1 introduces the research focus, followed by Section 2 providing a thorough literature review. Section 3 outlines the chosen investigation organization, though Section 4 presents and discusses the obtained results. Finally, Section 5 encapsulates the study's results in the conclusion.

2. RELATED WORKS

In 2023, Çelik *et al.* [16] introduced a unified benchmark for zero-shot text organization in Turkish, evaluating three approaches: Next Sentence Prediction, Natural Language Interpretation, and a suggested model built on Masked Language Modelling and pre-trained word embeddings. Using Turkish transformer models that have already been trained, such as BERT, ConvBERT, DistilBERT, and mBERT, ConvBERT with the NLI method demonstrated superior results, achieving 79% accuracy and outperforming the before used polyglot XLM-RoBERTa framework by 19.6%.

In 2022, Zahra and Alami *et al.* [17] presented an actual method based on Semantics and contextual information in texts, are captured using Bidirectional Encoder Representations from Transformers (BERT). Using a bilingual dataset from the Semi-supervised Offensive Language Identification Dataset (SOLID), experimental results showed that the translation-based approach, when combined with Arabic BERT (AraBERT), achieved F1-score and accuracy rates exceeding 93% and 91%, respectively.

In 2022, Kit *et al*. [18] suggested a simpler system that requires low training time when employing the BERT framework without the need for feature reduction and fine-tuning in order to minimize sentence embeddings. A single mBERT model has also been used to demonstrate that the suggested method is effective for multilingual tasks.

In 2022, Alduailej *et al*. [19] introduced the primary XLNet-based Arabic verbal model, AraXLNet, and applied it to Arabic sentiment analysis to improve prediction accuracy. AraXLNet, coupled with the Farasa segmented dataset, achieved superior results with accuracy rates of 94.78%, 93.01%, and 85.77%, outstripping AraBERT, which gained 84.65%, 92.13%, and 85.05% for the corresponding datasets. In 2023, Badawi *et al*. [20] conducted an evaluation using the recently published medical corpus, employing both Bidirectional Encoder Representations from Transformers (BERT), the most recent NLP deep learning technique, and traditional machine learning. The evaluation revealed that BERT outperformed all machine learning classifiers, achieving a higher accuracy of 92%, surpassing them by two points.

In 2022, Kim *et al*. [21] addressed the challenges in analyzing Korean medical text due to agglutinative language characteristics and complex medical terminologies. A Korean medicinal verbal model was developed using deep learning NLP and BERT's pre-training framework. In 2022, zahra and Alami *et al*. [22] discovered the potential of the pre-trained Arabic BERT model for universal contextualized sentence illustrations in Arabic text multi-class categorization. AraBERT was utilized through transfer learning and as a feature extractor, combined with various classifiers.

In 2020, Elnagare et *al*. [23] presented a comprehensive comparison of different deep learning (DL) models for the classification of Arabic text, assessing their effectiveness on SANAD and NADiA datasets. Additionally, the research explored the influence of employing word2vec embedding models to enhance classification task performance.

In 2015, Romeo *et al*. [24] proposed a novel framework for transductive multilingual document classification, leveraging the extensive multilingual knowledge base, BabelNet. Utilizing a cutting-edge transductive learner yielded effective document classification, demonstrating the model's superiority over conventional multilingual and cross-lingual analysis approaches.

In 2021, Chen *et al.* [25] suggested using latent word-wise label information in a machine learning text arrangement model. The approach captured label correlations independently of predefined order or exposure bias, demonstrating significant advantages over state-of-the-art systems in extensive experimental results.

In 2025, Turunen et al. [34] used document classification performance in the public sector on IMP document sets. The effectiveness of AI/ML and NLP techniques for this HMC-TC task was measured. Results provided a proof-of-concept prototype classifier employing Zero-Shot Classification (ZSC) for efficient information classification, reduction of human effort, and to minimize incorrect data labelling. The classifier used is modular, explores the IMP ontology to

provide classification suggestions without training on any specialized language models. It generates a readily deployable system without special requirements (customizations, heavy computations, or dependencies on external tools).

In 2025, Heseltine et al. [35] assessed the alternative trade-offs that closed vs. Open-source LLMs offer for text annotation. We compare the performance of 41 models that are classified by text, cover a spectrum of topics, text topics and languages. Models are tested against seven texts across ten country use cases and reveal significant differences in performance. As such, consideration must be given to how model choice should support the planned classification study that centers around an LLM. Overall, both closed-source models with good relative performance and strong performance on all tasks, such as GPT-4 or Gemini3, and a set of open-source LLMs, like Llama3 and Qwen2.5, yield competitive results comparable with closed-source models, if not outperforming them.

2.1. Problem Statement

Sentiment analysis in more than one language is still a significant problem in NLP because of the linguistic variation, scarce labeled data, and discrepancies in cross-linguistic interpretation. The majority of sentiment analysis models are usually trained only on English data, which limits their capacity to be effectively generalized to other languages with different grammatical and semantic structures. Moreover, neural machine translation and multilingual pre-trained models have enhanced language comprehension, but when applied independently, they tend to cause translation errors and contextual sentiment clues. Thus, it is necessary to develop a powerful multilingual sentiment analysis system that combines translation-based methods with powerful transformers and large language models (LLMs). This study aims to come up with an ensemble-based model that integrates the advantages of neural translation and transformer models to improve accuracy and precision as well as contextual knowledge of sentiment classification across different languages. Although multilingual sentiment analysis has witnessed considerable progress through transformer-based language models and neural machine translation (NMT), several important challenges remain unresolved. Most existing approaches are optimized for individual languages or depend on language-specific annotated datasets, limiting their applicability to real-world multilingual environments. Translation-based approaches often introduce semantic inconsistencies, while standalone multilingual transformer models struggle to preserve contextual sentiment across linguistically diverse languages. Furthermore, existing studies rarely investigate whether combining multiple translation engines with multiple pre-trained language models can improve classification robustness across heterogeneous datasets.

Table 1 illustrates the aim and limitations of the previous research

Table 1: Aim and Limitation of the Previous Research

Author	Method	Aim	Limitation
zahra and Alami *et al*. [22]	Employed fine-tuned AraBERT for Arabic text categorization, generating contextual semantic embeddings	Enhanced multi-class categorization accuracy	Model performance subject to corpus specificity and fine-tuning data quality.
Elnagar *et al*. [23]	Implemented Arabic text classification using deep learning models.	Enhanced accuracy in Arabic text categorization tasks.	Model performance may vary with dataset characteristics and size.
Romeo *et al*. [24]	Developed a knowledge-based representation for transductive classification of multilingual documents.	Improved accuracy in categorizing multilingual documents.	Dependent on knowledge base quality and coverage
Chen *et al.* [25]	Utilized latent word-wise label information for multi-label text classification.	Improved accuracy in assigning multiple labels.	Dependent on label representation quality.
Turunen et al. [34]	Developed a Zero-Shot Classification (ZSC) prototype for Hierarchical Multi-Class Text Classification (HMC-TC) on IMP ontology	- No need for labeled training data. - Reduce human effort in manual annotation.	- Hard to update the model on changing or new domain terms if not reflected in ontology. - May not cope with Ambiguous or complex documents.
Heseltine et al. [35]	Compared a sample of 41 closed-source and open-source Large	- Assistance for an informed choice of LLMs	- Heavily tied open-source LLMs need more infrastructure and

	Language Models (LLMs) for text annotation / text classification over numerous Languages, Topics, and Country specific datasets.	for your classification project. - Evaluates the models based on multilingual and cross-domain settings.	computational expertize for hosting. - Model Choice is to problem, dataset, context and task at hands.

3. PROPOSED METHODOLOGY

In this study, a multi-step method is presented to accomplish multilingual sentiment classification through translation of non-English texts into a base language, English. The methodology is divided into five major stages, namely data gathering, data preprocessing, translation, sentiment classification and performance evaluation. To begin with, the multilingual text data are gathered using various sources of information, including social media, online reviews, and news articles in four target languages: Arabic, Chinese, French, and Italian. The data is then cleaned and preprocessed, such as tokenized, lowercased, stop words removed, stemmed and lemmatized to guarantee quality and consistency. Two neural machine translators, DeepL Translator and Google Translate, are then used to translate the preprocessed texts into English. The translated texts are then processed with a collection of pre-trained sentiment analysis models, such as OpenAI's GPT-4, Twitter-RoBERTa-Base-Sentiment-Latest, and BERT-Base-Multilingual-Uncased-Sentiment. Lastly, performance measures of precision, accuracy, recall, and F1-score are used to analyze the obtained results to determine the effectiveness and strength of the model. The general workflow in the proposed methodology is shown in Figure 1.

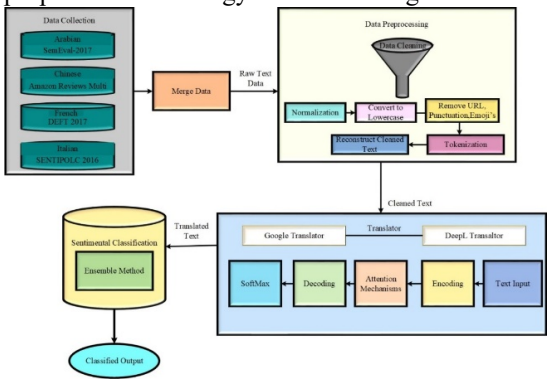

Figure 1. Architecture of the proposed methodology

3.1. Data Acquisition

During the primary step of this study, multilingual text data are gathered using various publicly available benchmark datasets in order to cover a variety of languages and areas. The datasets utilized in this paper are SemEval-2017 Task 4: Sentiment Analysis in Twitter [26], Amazon Reviews Multi [27], DEFT 2017 [28] and SENTIPOLC 2016 [29]. The four languages used in social media platforms like Twitter, namely Chinese (zh), Arabic (ar), French (fr), and Italian (it), are chosen to be analyses due to their high usage as well as their international importance in the linguistic world. All the datasets are already annotated with three categories of sentiment, including positive, negative, or neutral. Human evaluators or crowd-sourced contributors are used to manually validate the annotations in order to guarantee accuracy in labelling. The datasets are followed by the compilation of the obtained datasets to create a multilingual corpus that is used to classify the sentiment based on translation.

3.2. Pre-processing

The next stage of the method involved a methodical approach to pre-processing and data cleansing of the collected multilingual data to make sure it has high-quality input for translation and sentiment classification. The steps that are used in the process are tokenization, lowercasing, stop-word removal, punctuation and emoji filtering, and URL elimination. The process of tokenization is done to divide the text into separate tokens or words to allow uniform input between languages. Stop words like "and", "the" and "in" are eliminated to remove words that do not add much to the sentiment discrimination. Let the raw multilingual dataset be denoted in Eqn (1),

$$D = \{d_1, d_2, d_3, \ldots, d_n\} \quad (1)$$

where each d_i represents a text sample in one of the four languages: Arabic, Chinese, French, or Italian. Each text $d_i$is normalized through several transformation functions and is represented in Eqn (2),

$$d_i^{(1)} = Lowercase(d_i) \quad (2)$$

Hyperlinks and hashtags are often non-semantic and introduce bias; hence, they are removed as in Eqn (3). This eliminates hyperlinks and hashtags that do not contribute to semantic meaning.

$$d_i^{(2)} = d_i^{(1)} - URLs\left(d_i^{(1)}\right) - Hashtags\left(d_i^{(1)}\right) \quad (3)$$

Since this paper uses Transformer-based pre-trained models, no stemming and lemmatization are used as the models already encode contextual and morphological differences in their embeddings. The removal of emojis and special characters is done to reduce noise that

biases the translation, and sentiment prediction is represented in Eqn. (4), whereas URLs and hashtags are omitted because they are not semantically relevant.

$$d_i^{(3)} = RemovePunctEmoji(d_i^{(2)}) \quad (4)$$

$$T_i = Tokenize(d_i^{(3)}) = \{t_{i1}, t_{i2}, \dots, t_{im}\} \quad (5)$$

$$T_i^{(1)} = \{t_{ij} \epsilon\, T_i | t_{ij} \notin S\} \quad (6)$$

Each cleaned text is split into tokens (words/subwords) t_{ij} is expressed in Eqn (5). Let S denote the predefined stop-word set (language-specific) is specified in Eqn (6). This filters out frequent but semantically weak words such as *and*, *the*, *in*. Finally, the cleaned tokens are concatenated back to form the cleaned sentence is represented in Eqn (7). After processing all samples, which is expressed in Eqn (8), the overall pre-processing transformation is represented in Eqn (9),

$$d_1' = Join(T_i^{(1)}) \quad (7)$$

$$D' = \{d_1', d_2', \dots., d_n'\} \quad (8)$$

$$D' = \{Join(RemoveStopWords(Tokenize(Remove\ URLs(Lowercase(di)))) \quad (9)$$

Where D represents the raw multilingual dataset, d_i specifies an individual text sample, D' denotes the Cleaned and pre-Where, D represents Raw multilingual dataset, d_i specifies an individual text sample, D' denotes the cleaned and pre-processed dataset t_{ij} is a token in the sample i, S represents the set of stop Words, F_{clean} refers to the complete pre-processing transformation function. The entire process of pre-processing is done in a similar manner in the four languages, which are Arabic, Chinese, French, and Italian, prior to the translation process. This would provide uniformity and less ambiguity in the translation during the next stage. The entire data scrubbing and pre-processing algorithm is outlined in Algorithm 1.

Algorithm 1: Data Cleaning and Pre-Processing

Input: Raw multilingual dataset D = {d_1, d_2, …, d_n}
Output: Cleaned and pre-processed dataset D'
1: Initialize empty list $D' \leftarrow \emptyset$
2: For each text sample $d_i \in$ D do
3: # Step 1: Text Normalization
4: $d_i \leftarrow$ Convert to Lowercase (d_i)
5: $d_i \leftarrow$ Remove URLs (d_i)
6: $d_i \leftarrow$ Remove Punctuation (d_i)
7: $d_i \leftarrow$ Remove Emojis and Special Chars (d_i)
8: # Step 2: Tokenization
9: tokens ← Tokenize (d_i)
10: # Step 3: Stop-Word Removal
11: tokens ← Remove Stop Words (tokens)
12: # Step 4: Reconstruct Cleaned Text
13: clean text ← Join(tokens)
14: Append clean_ text to D'
15: End For
16: Return D'

3.3. Translation Phase

The third step in the methodology involved translating the cleansed and pre-processed multilingual information into English with two Neural Machine Translation (NMT) systems, DeepL Translator and Google Translate. This step of translation is necessary to be able to perform uniform sentiment analysis with English-based pre-trained models. DeepL Translator is a free API for machine translation that translates text using pre-trained NMT models from one language to another. It has an encoder-decoder neural network structure, with the input text being tokenized and converted to a mathematical form first. This coded message is then decoded to translate into the target language sentence. Google Translate, in its turn, is a cloud-based NMT engine, which uses large-scale deep learning architectures trained on large bilingual datasets. It encodes input text as linguistic units and uses a multi-layer neural network to encode semantic and contextual interrelationships between languages. This enables the system to easily deal with idiomatic phrases, sentence structure of the sentences and subtle language differences.

The four chosen languages of Chinese, Arabic, Italian, and French were translated using both systems to guarantee the accuracy of translations and cross-linguistic consistency. The results of the two translators were contrasted, and the English translation that was contextually apt was used in further sentiment analysis. Each cleaned text d_i' is translated into English using both NMT systems, as given in Eqn (10) and (11),

$$t_i^{(1)} = MT_1(d_i', lang(d_i') \rightarrow English) \quad (10)$$

$$t_i^{(2)} = MT_2(d_i', lang(d_i') \rightarrow English) \quad (11)$$

Where, MT_1 represents Google Translate (cloud-based NMT), MT_2 denotes DeepL Translator (deep neural NMT). To measure translation quality, the semantic similarity between the source and translated texts is computed in Eqn (12),

$$S_1 = Sim(d_i', t_i^{(1)}),\ S_2 = Sim(d_i', t_i^{(2)}) \quad (12)$$

Where $Sim(\cdot)$ denotes a semantic similarity function such as cosine similarity between sentence embeddings. The translation with the higher similarity score is selected in Eqn (13) and more compactly in Eqn (14),

$$t_i - \begin{cases} t_i^{(1)}, & if\ S_1 \geq S_2 \\ t_i^{(2)}, & otherwise \end{cases} \quad (13)$$

$$t_i = argmax_{t_i^{(k)}} Sim\left(d_i', t_i^{(k)}\right), \quad k\epsilon\{1,2\} \tag{14}$$

After processing all n samples, the final English-translated dataset is represented in Eqn (15). The complete translation phase is represented in Eqn (16),

$$T = \{t_1, t_2, \ldots, t_n\} \tag{15}$$

$$T = F_{trans}(D') = \{argmax_{t_i^{(k)}} Sim(d_i', MT_k(d_i'))\mid d_i' \epsilon D', k\epsilon\{1,2\}\} \tag{16}$$

Here, D' represents the cleaned and pre-processed dataset, d_i' denotes the cleaned text sample, MT1 and MT2 specify translation systems (Google Translate, DeepL Translator), $t_i^{(1)}$, $t_i^{(2)}$ refers to translated texts by each system, S_1,S_2 is the semantic similarity scores, T denotes the final translated English dataset, F_{trans} represents the Translation function. This translation step is used to bridge the gap between multilingual data, English-based transformers, and LLM models to be able to classify sentiments cross-linguistically. The workflow of translation used in this study is described in Algorithm 2, and the following section explains the sentiment analysis flow that is applied to the translated data.

Algorithm 2: Translation Process

Input: Cleaned multilingual dataset D' ={d_1, d_2, …, d_n}
Source languages $L = \{Arabic, Chinese, French, Italian\}$
Output: English-translated dataset $T = \{t_1, t_2, \ldots, t_n\}$
1: Initialize empty list T ← ∅
2: Load translation systems:
MT1 ← Google Translate (cloud-based NMT)
MT2 ← DeepL Translator (deep neural NMT)
3: For each text sample $d_i \in D'$ do
4: Identify source language lang (d_i) ∈ L
5: # Step 1: Translation using multiple NMT systems
6: trans1 ← MT1.translate(d_i, source=lang (d_i), target="en")
7: trans2 ← MT2.translate(d_i, source=lang (d_i), target="en")
8: # Step 2: Translation Quality Comparison
9: Compute semantic similarity scores:
10: sim1, sim2← SemanticSimilarity(d_i, trans1, trans2)
11: # Step 3: Select Best Translation
12: best translation ← Select Translation ({trans1, trans2}, argmax(sim))
13: Append best_translation to T
14: End For
15: Return T

3.4. Sentiment Analysis

Sentiment analysis of the translated text was the fourth step in the technique. English text with both pre-trained transformer-based models, as well as an LLM within an ensemble learning model. The purpose of the ensemble model is to improve the accuracy and robustness of classification across languages through the complementary benefits of different architectures. The three models are defined in Eqn (17), (18) and (19),

$$M_1 = Twitter\ RoBERTa\ Base\ Sentiment\ Latest \tag{17}$$

$$M_2 = BERTBase\ Multilingual\ Uncased\ Sentiment \tag{18}$$

$$M_3 = GPT-4 \tag{19}$$

Each model M_k(for $k \in \{1,2,3\}$) outputs a probability distribution over classes for sample t_i represented in Eqn. (20). Here $p_i^{(k)}(c)$ is the predicted probability model $k$that model t_i belongs to class c.

$$P_i^{(k)} = (p_i^{(k)}(c))_{c\epsilon C}, \quad \sum_{c\epsilon C} p_i^{(k)}(c) - 1 \tag{20}$$

For each sample t_i and class c, compute the weighted ensemble score in Eqn. (21). This is the aggregated confidence for class $c$after combining the three models.

$$S_i(c) = \sum_{k=1}^{3} w_k\, p_i^{(k)}(c) \tag{21}$$

Assign a non-negative weight $w_k$to each model representing its reliability derived from validation performance, such as F1 or accuracy. Enforce is derived in Eqn (22). A practical choice is to set weights proportional to a performance metric m_k I is expressed in Eqn (23). The final predicted label choose the class with the maximum aggregated score, is represented in Eqn (24),

$$w_k \geq 0, \sum_{k=1}^{3} w_{k-1} \tag{22}$$

$$w_k = \frac{m_k}{\sum_{j=1}^{3} m_j}, \quad m_k > 0 \tag{23}$$

$$\hat{y}_i = argmax_{c\epsilon C} S_i(c) \tag{24}$$

This paper uses three pre-trained models, including GPT-4, BERT-Base-Multilingual-Uncased-Sentiment, and Twitter-RoBERTa-Base-Sentiment-Latest, created by OpenAI. The first two patterns are retrieved from the Hugging Face model repository, which offers a large selection of state-of-the-art transformer patterns in NLP tasks. The OpenAI GPT-4 model is a large language model capable of capturing more contextual and semantic data than sentiment patterns on the surface.

Both models analyzed the translated text separately and gave sentiment polarity predictions as either favorable, negative or neutral. The results of the three models are then added through a weighted voting system where the ultimate sentiment tagging is achieved through the summation of the confidence scores of the three models. Such an ensemble approach improved generalization and reduced the bias of individual models.

Contextual embedding features of transformer-based models like RoBERTa and BERT are

selected because they are known to be effective in working with social media data and multilingual text. GPT-4 is added to enhance interpretability and contextual consistency in the determination of sentiment. Figure 2 displays the block diagram for the suggested Ensemble model.

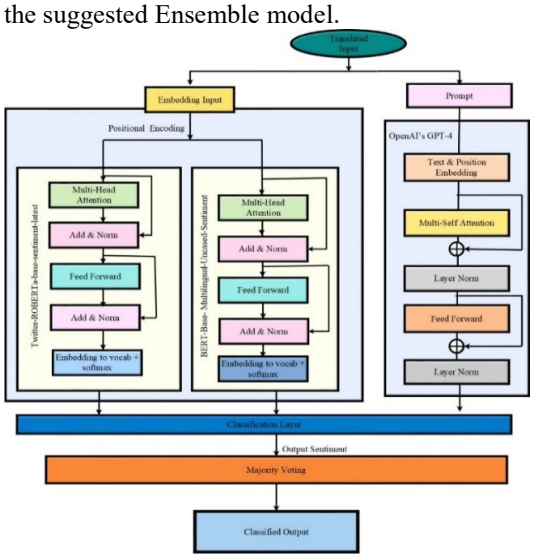

Figure. 2. Block diagram of proposed Ensemble model

3.5. Evaluation Phase

At the last stage of the suggested methodology, the work of the ensemble-based multilingual sentiment analysis framework to understand its accuracy and its overall performance is tested. The results of the evaluation process are compared with the predicted sentiment labels of the ensemble model using the actual data sentiment labels of the innovative datasets. To measure the accuracy, the standard evaluation measures that are commonly utilized in sentiment classification are used, such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These measures are used to measure the accuracy of the model in recognizing sentiment categories. In particular, TP and TN are correctly predicted and rejected opposite sentiments, respectively; FP and FN stand for false positive and false negative predictions, respectively.

Since three classes of feelings, including neutral, negative, and positive, are considered in this study, each of the measures is calculated individually. The assessment model employed precision, recall, accuracy, specificity, and F1-score to provide a thorough analysis of the model's performance. Precision determines the percentage of instances that are correctly predicted among all instances that are classified as a specific sentiment. The rate at which all true instances of a sentiment are accurately identified is known as recall. By calculating the harmonic mean of precision and recall, the F1-score offers a fair evaluation of the correctness of the model. The overall percentage of correctly identified cases is called accuracy. Specificity describes the model in its capacity to accurately detect negative sentiment cases.

The evaluation process guarantees that the quality of translation of Google Translate, DeepL Translator and the sentiment predictions of the ensemble of RoBERTa, BERT-Multilingual, and GPT-4 are strictly tested based on the cross-linguistic consistency of sentiment and the strength of the results.

3.6 Study Validity

There are several methodological and experimental features contributing to the validity of this study, ensuring that the reported findings are robust, repeatable, and generalize well for multilingual sentiment classification.

Internal Validity

• Standardization of the proposed approach is used across two multilingual sentiment datasets and involving two different translation approaches.

• The use of state-of-the-art pre-trained language models: Twitter-RoBERTa, BERT-Multilingual, and GPT-4 are used with their fine-tuned version on. common evaluation environment is used for all translation systems, and a single set of experimental settings are employed throughout the study.

Construct Validity

• The proposed model's performance is assessed using standard sentiment classification measures: Accuracy, Precision, Recall, F1-score, and Specificity. This comprehensive suite ensures that the model's classification effectiveness and stability across different metrics are well accounted for. External Validity

• The validation process is performed across diverse multilingual settings, encompassing Arabic, Chinese, French and Italian to represent varying linguistic structures and writing systems.

• Experiments are conducted with multiple translation systems to generalize the model beyond a single translator's limitations and ensure broad Applicability.

• Standardized multilingual sentiment analysis benchmark datasets are adopted, enabling replicability and comparison to future work.

Comparative Validity

• Performance of the proposed approach is evaluated against strong existing baseline models including Twitter-RoBERTa, BERT-Multilingual, GPT-4, mBERT and XLNet, thereby providing concrete evidence for its relative strengths and applicability.

4. RESULTS AND DISCUSSIONS

4.1 Experimental findings of the Positive Metrics

The proposed multilingual sentiment analysis framework is implemented using Python 3.10 in the Google Colab environment, leveraging libraries such as Transformers (Hugging Face), Scikit-learn, Pandas, NumPy, and Matplotlib for model implementation, evaluation, and visualization. Two NMT systems, such as Google Translate API and DeepL Translator, are employed to translate non-English texts (Arabic, Chinese, French, and Italian) into English. For sentiment analysis, three pre-trained transformer and large language models are utilized. This segment contains the presentation and discussion of the investigational outcomes of our multilingual sentiment analysis framework. The experiment consisted of sentiment analysis of foreign language texts that had been initially translated into English with two NMT systems, Google Translate and DeepL Translator. The languages chosen to be evaluated were Arabic, Chinese, French, and Italian because they represent different linguistic constructions and are common in digital communication channels. [30] The sentiment analysis of the translated English texts is performed with three pre-trained deep learning models: BERT-Base-Multilingual-Uncased-Sentiment and Twitter-RoBERTa-Base-Sentiment-Latest [31], ZSC [34], GPT-3.5 [35] and OpenAI GPT-4 [32], as well as the proposed ensemble model that integrates the strengths of the three. Each language underwent 8 experiments, which led to 32 experimental runs that comprised the combinations of various translation tools and sentiment analysis models.

Table 2: Experimental Findings for Various Combinations of the Sentiment Analyser Model and Translator.

Translator	Sentiment Analyzer	Accuracy	Precision	Recall	F1-score	Specificity
Google Translate	Twitter-RoBERTa-Base-Sentiment-Latest	0.9215	0.9281	0.9198	0.9239	0.9124
	BERT-Base-Multilingual-Uncased	0.9324	0.9356	0.9287	0.9321	0.9189
	GPT-4 (OpenAI)	0.9458	0.9421	0.9383	0.9402	0.9265
	GPT-3.5	0.9438	0.9402	0.9369	0.9385	0.9251
	ZSC	0.9477	0.9462	0.9401	0.9402	0.9278
	Proposed Ensemble Model	0.9582	0.9543	0.9518	0.953	0.9357
DeepL Translator	Twitter-RoBERTa-Base-Sentiment-Latest	0.9289	0.9317	0.9262	0.9289	0.9163
	BERT-Base-Multilingual-Uncased	0.9385	0.9414	0.9342	0.9378	0.9227
	GPT-4 (OpenAI)	0.9503	0.9471	0.9425	0.9448	0.9309
	GPT-3.5	0.9497	0.9465	0.9428	0.9446	0.9318
	ZSC	0.9501	0.9494	0.9463	0.9486	0.9345
	suggested Ensemble Model	0.9624	0.9587	0.9562	0.9574	0.9405

Table 2 shows the performance comparison of various combinations of the models for sentiment analysis and translation in the tasks using sentiment analysis, which are measured in five main metrics: F1-score, recall, accuracy, precision, and Specificity. The research compares two large machine translators, Google Translate and DeepL Translator, to four models for sentiment analysis: BERT-Base-Multilingual-Uncased, Twitter-RoBERTa-Base-Sentiment-Latest, GPT-4 (OpenAI), and the Proposed Ensemble Model.

Based on the results, it is seen that both translators have performed well in all models, and DeepL Translator has shown slightly better overall results than Google Translate. When coupled with DeepL Translator, with an accuracy of 0.9624, precision of 0.9587, recall of 0.9562, F1-score of 0.9574, and specificity of 0.9405, the Proposed Ensemble Model outperformed all other models. This shows the usefulness of the ensemble strategy, which combines several model predictions to enhance the overall reliability and robustness of sentiment categorization.

GPT-4 (OpenAI) model also showed high performance in both translation systems and this confirmed that the model had a robust generalization capacity in cross-lingual sentiment analysis. In the meantime, BERT-Base-Multilingual-Uncased and Twitter-RoBERTa-Base-Sentiment-Latest models yielded a little bit lower but consistent scores, which indicates that they are also competent in multilingual sentiment comprehension. The results show that the Deep Learning Translator Using the Suggested Ensemble Framework provides the greatest correct and balanced sentiment predictions. This indicates that the combination of a high-quality neural translation system and an ensemble-based sentiment analysis method is boosting cross-lingual sentiment detection substantially, as discussed further in Figure 3 by visualizing performance comparison graphs and sentiment analysis models under important evaluation metrics, including specificity, F1-score, precision, accuracy, and recall.

A visual comparison is shown in the figure. of the performances of individual transformer-based models, using the suggested ensemble model, GPT-4, BERT-Base-Multilingual, ZSC, GPT-3.5 and Twitter-RoBERTa-Base in the environment of Google Translate and DeepL Translator. The findings are clear that the proposed ensemble model is always more effective in all metrics than the individual models, which means that it is more effective in grasping more subtle sentiment information and is more robust across languages. Moreover, the figure shows that DeepL Translator with the ensemble model is the most effective overall, which underlines the advantage of using high-quality translation with ensemble-based sentiment classification in multilingual analysis.

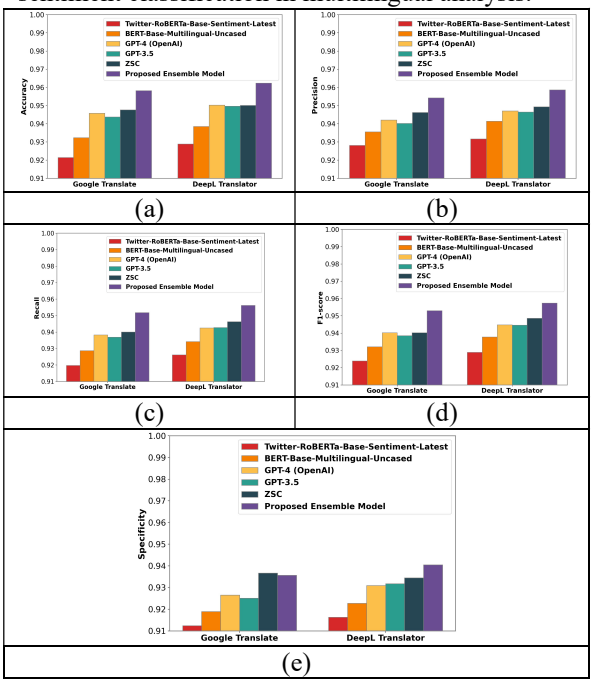

Figure 3. Results of experiments demonstrating the effects of various assessment metrics

Figure 4 shows the confusion matrices resulting from the different experimental settings, presenting the presentation of sentiment analysis representations for different translators and model combinations. Each confusion matrix gives a graphic illustration of the ability of the model to distinguish between generally positive and negative sentiment categories by comparing the predicted and real sentiment labels. The values across the diagonal are indicative of correctly classified instances, while the off-diagonal elements show incorrect categories. From the figure, it follows that the proposed ensemble model produces higher values along the diagonal, which means that it is more accurate with fewer misclassifications when compared to individual models such as BERT and RoBERTa. This reflects the enhanced generalization capability of the ensemble model towards sentiment prediction for different languages and translation systems, hence making the predictions of sentiment classification more reliable and consistent.

Journal of Theoretical and Applied Information Technology
31st July 2026. Vol.104. No.14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

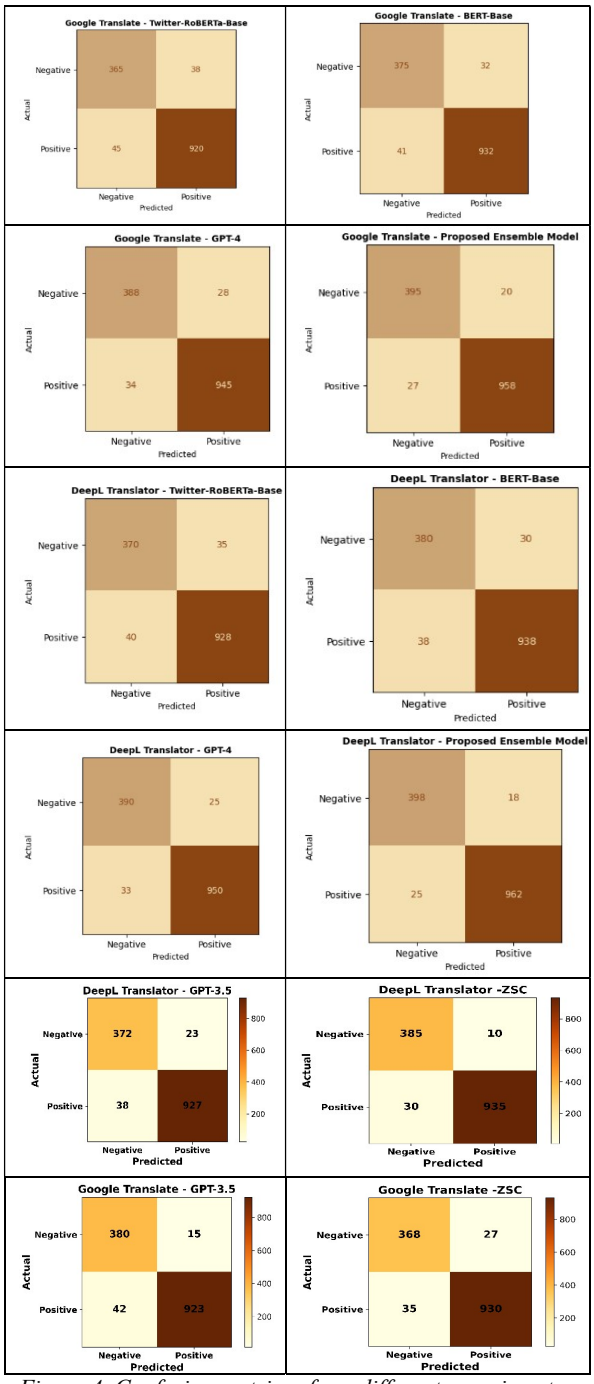

Figure 4. Confusion matrices from different experiments.

Table 3: Language-Wise Accuracy Scores for Different Translator and Sentiment Analyser Model Combinations

Translator	Model Sentiment Analyser	Arabic	Chinese	French	Italian
DeepL Translator	Twitter-RoBERTa-Base	0.86	0.86	0.87	0.86
	BERT-Base-Multilingual-Uncased-Sentiment	0.87	0.86	0.88	0.87
	GPT-4 (OpenAI)	0.89	0.88	0.89	0.89
	ZSC	0.88	0.89	0.87	0.86
	GPT-3.5	0.89	0.86	0.86	0.85
	Proposed Ensemble Model	0.91	0.9	0.91	0.9
Google Translate	Twitter-RoBERTa-Base	0.87	0.86	0.88	0.87
	BERT-Base-Multilingual-Uncased-Sentiment	0.88	0.87	0.89	0.88
	GPT-4 (OpenAI)	0.9	0.89	0.91	0.9
	ZSC	0.84	0.88	0.86	0.87
	GPT-3.5	0.85	0.88	0.86	0.87
	Proposed Ensemble Model	0.91	0.9	0.91	0.91

Table 3 shows the language-wise accuracy scores obtained using Arabic, Chinese, French, and Italian as the four target languages for which there are various combinations of translator and emotion analyser models. Results are provided showing the comparison of the performance of DeepL Translator and Four distinct sentiment analysis models integrated with Google Translate: Twitter-RoBERTa-Base, BERT-Base-Multilingual-Uncased-Sentiment, GPT-4 (OpenAI), ZSC, GPT-3.5 and the Proposed Ensemble Model.

The performances in different languages show quite consistent results, with values ranging from 0.86 to 0.91, which supports the fact that cross-lingual sentiment detection is strong and stable across different languages. Among them, out of all the languages, the Proposed Ensemble Model achieved the greatest accuracy value. For both translators, going up to a maximum value of 0.91 in most of the cases. This again proves the superior generalization capability of the model and effectively interprets multilingual sentiment expressions.

GPT-4 also performed on par with the model presented here, only falling back a little in terms of accuracy. This underlines its robustness and ability to adapt to various linguistic contexts. On the other end of the spectrum, Twitter-RoBERTa-Base and

BERT-Base-Multilingual-Uncased-Sentiment recorded relatively lower but consistent accuracy in their use in the handling of multilingual sentiment, despite being optimized, primarily for English-based content.

Overall, the results in Table 3 confirm that using DeepL Translator or Google Translate with the Proposed Ensemble Model produces reliable performance among all languages tested. The present study underlines the potential for ensemble approaches to enhance multilingual sentiment analysis by effectively combining insights from a range of linguistic representations.

Table 4: Language-Wise Precision Scores for Different Translator and Sentiment Analyser Model Combinations

Translator	**Sentiment Analyzer Model**	**Arabic**	**Chinese**	**French**	**Italian**
DeepL Translator	Twitter-RoBERTa-Base	0.87	0.86	0.85	0.86
	BERT-Base-Multilingual-Uncased-Sentiment	0.88	0.86	0.84	0.86
	GPT-4 (OpenAI)	0.89	0.90	0.89	0.91
	ZSC	0.87	0.86	0.86	0.85
	GPT-3.5	0.85	0.87	0.88	0.87
	Proposed Ensemble Model	0.91	0.92	0.93	0.91
Google Translate	Twitter-RoBERTa-Base	0.87	0.86	0.87	0.88
	BERT-Base-Multilingual-Uncased-Sentiment	0.89	0.88	0.89	0.87
	GPT-4 (OpenAI)	0.9	0.91	0.92	0.9
	ZSC	0.87	0.86	0.86	0.85
	GPT-3.5	0.85	0.87	0.88	0.87
	Proposed Ensemble Model	0.91	0.92	0.93	0.91

Table 4 provides the Arabic, Chinese, French, and Italian language-wise precision scores attained by various translator and emotion analyser model combinations. Four sentiment analysis models are used to compare the performance of DeepL Translator with Google Translate: GPT-4 (OpenAI), Twitter-RoBERTa-Base, BERT-Base-Multilingual Uncased Sentiment, ZSC, GPT-3.5 and the Suggested Ensemble Model.

Precision, which is the percentage of accurately predicted optimistic sentiments out of the total number of positive predictions, has been a vital indicator of how well a model differentiates between positive and negative sentiments. It is noticed that the precision scores remain consistently high for all combinations, ranging between 0.84 and 0.93, reflecting the reliability of these models in doing accurate sentiment classification.

The precision for all languages and translators using the Proposed Ensemble Model is the highest among the combinations tested, reaching as high as 0.93 for French when using Google Translate and 0.92 using DeepL Translator for multiple languages. This indicates that the ensemble model has a high ability to minimize false positives and provide more confident sentiment predictions.

The GPT-4 (OpenAI) model also yielded excellent precision, mainly for Chinese and French, at 0.90-0.92, which testifies to its advanced understanding of linguistic subtleties in multilingual tasks. In contrast, the Twitter-RoBERTa-Base and BERT-Base-Multilingual-Uncased-Sentiment models reported moderate but stable values for precision, indicating dependability at a comparatively lower level of discriminative ability.

Overall, Table 4 demonstrates that both DeepL Translator and Google Translate provide translation quality that is generally adequate for the task of sentiment analysis, while the Proposed Ensemble example exhibits the highest degree of accuracy for all the languages tested, further establishing its efficacy in relation to tasks involving sentiment analysis across languages.

Table 5: Language-wise recall scores between different groupings of sentiment and translator analyzer model.

Translator	**Sentiment Analyzer Model**	**Arabic**	**Chinese**	**French**	**Italian**
DeepL Translator	Twitter-RoBERTa-Base	0.87	0.86	0.85	0.86
	BERT-Base-Multilingual-Uncased-Sentiment	0.88	0.87	0.86	0.88
	GPT-4 (OpenAI)	0.89	0.88	0.89	0.90
	ZSC	0.89	0.86	0.86	0.85

	GPT-3.5	0.85	0.87	0.88	0.87
	Proposed Ensemble Model	0.91	0.92	0.90	0.91
Google Translate	Twitter-RoBERTa-Base	0.88	0.87	0.89	0.87
	BERT-Base-Multilingual-Uncased-Sentiment	0.87	0.88	0.89	0.88
	GPT-4 (OpenAI)	0.9	0.89	0.90	0.9
	Proposed Ensemble Model	0.92	0.9	0.91	0.91

Table 5 shows the language-wise recall scores obtained from several sets of sentiment analysis models and translators for Arabic, Chinese, French, and Italian. Recall gauges how well the model can recognise all pertinent positive examples. Making it a key indicator of how effectively a sentiment analyzer captures true positive sentiments without omission.

This table compares two translation services-DeepL Translator and Google Translate each combined with four Models for sentiment analysis: GPT-4, BERT-Base-Multilingual-Uncased-Sentiment, and Twitter-RoBERTa-Base (OpenAI), and the suggested ensemble model. The recall ratings for each and every combination range between 0.85 and 0.92, which shows that all models are performing well in identifying positive sentiments across languages.

The Proposed Ensemble Model has demonstrated the highest recall among the models that were compared. The values consistently, with as high as 0.92 for Chinese with DeepL Translator and Arabic with Google Translate. The high value of recall signifies that the ensemble model has the highest capability to detect true positive sentiments, hence minimizing the chances of missing relevant sentiment cues.

However, the GPT-4 model stands out in showing strong recall performance, ranging from 0.88 to 0.90 across all languages and translators. This is indicative of the robustness and adaptability of GPT-4 in multilingual sentiment understanding. On the other hand, BERT-Base, ZSC, GPT-3.5 and RoBERTa-Base on Twitter: Multilingual, Uncased Sentiment **have** slightly lower recall scores but maintain consistency across different linguistic contexts. Overall, Table 5 shows that the two translation systems, DeepL and Google Translate, maintain most of the subtlety in sentiments when translating. However, the suggested ensemble model performs better than any alternative models, confirming its superior capability in identifying true positive sentiments across multilingual and translation platforms, thus becoming the most reliable for multilingual sentiment classification.

Table 6: Language-wise F1 scores for various combinations of the Sentiment Analyser Model and Translator.

Translator	**Sentiment Analyzer Model**	**Arabic**	**Chinese**	**French**	**Italian**
DeepL Translator	Twitter-RoBERTa-Base	0.85	0.85	0.87	0.86
	BERT-Base-Multilingual-Uncased-Sentiment	0.86	0.87	0.88	0.87
	GPT-4 (OpenAI)	0.89	0.90	0.89	0.89
	ZSC	0.89	0.86	0.86	0.85
	GPT-3.5	0.85	0.86	0.88	0.84
	Proposed Ensemble Model	0.90	0.91	0.90	0.91
Google Translate	Twitter-RoBERTa-Base	0.87	0.86	0.88	0.87
	BERT-Base-Multilingual-Uncased-Sentiment	0.88	0.87	0.89	0.88
	GPT-4 (OpenAI)	0.9	0.89	0.91	0.9
	ZSC	0.87	0.86	0.86	0.85
	GPT-3.5	0.85	0.87	0.88	0.87
	Proposed Ensemble Model	0.91	0.9	0.91	0.91

The F1-scores by language for various translator groupings and emotion analyser models in four languages: French, Italian, Chinese, and Arabic, are displayed in Table 6. To provide a fair assessment of both false positives and false negatives, the harmonic mean of recall and accuracy is known as the F1-score. It is especially helpful in the field of sentiment analysis, where the distribution of classes is uneven and more accuracy fails to reveal the actual model performance. Two translation

systems, DeepL Translator and Google Translate, are compared, each combined with Twitter-RoBERTa-Base, BERT-Base-Multilingual-Uncased-Sentiment, GPT-4(OpenAI), and the Proposed Ensemble Model, which are the four sentiment analysis models. The DeepL Translator resulted in the highest F1-scores of the Proposed Ensemble Model, amounting to as high as 0.91 for Chinese and Italian, indicating that the proposed ensemble methodology retains a strong balance between precision and recall in recognizing sentiments with minimal misclassifications. The GPT-4 model from OpenAI also had high F1-scores ranging from 0.89 to 0.90, reflecting its stability and generalization capability for different languages. Similarly, similar to Google Translate, the Proposed Ensemble Model had the greatest results in all four languages, rising from 0.90 to 0.91, while Twitter-RoBERTa-Base and BERT-Base-Multilingual-Uncased-Sentiare a little behind but consistently yielded results within the range of 0.86–0.88. On the whole, Table 6 indicates that the Proposed Ensemble Model has shown the best F1 performance consistently across languages and translation systems. This ascertains that the ensemble approach enhances the overall robustness in multilingual sentiment classification by effectively balancing precision and recall, ensuring accurate sentiment detection in a variety of language settings.

Table 7: Scores For Language-Specificity Across Various Translator and Sentiment Analyser Model Combinations

Translator	**Sentiment Analyzer Model**	**Arabic**	**Chinese**	**French**	**Italian**
DeepL Translator	Twitter-RoBERTa-Base	0.87	0.85	0.87	0.85
	BERT-Base-Multilingual-Uncased-Sentiment	0.86	0.86	0.88	0.86
	GPT-4 (OpenAI)	0.88	0.87	0.89	0.88
	ZSC	0.84	0.86	0.86	0.85
	GPT-3.5	0.85	0.85	0.88	0.87
	Proposed Ensemble Model	0.90	0.9	0.91	0.9
Google Translate	Twitter-RoBERTa-Base	0.86	0.85	0.87	0.86
	BERT-Base-Multilingual-Uncased-Sentiment	0.87	0.86	0.89	0.89
	GPT-4 (OpenAI)	0.89	0.88	0.90	0.90
	ZSC	0.86	0.86	0.86	0.85
	GPT-3.5	0.85	0.87	0.89	0.87
	Proposed Ensemble Model	0.90	0.91	0.92	0.91

Table 7 demonstrates the language-wise specificity scores obtained from Arabic, Chinese, French, and Italian, which are the four target languages for various combinations of translators and sentiment analysis models. Specificity basically calculates the ratio of negativity that has been correctly identified; it shows the model's ability to differentiate nonpositive sentiments from positive ones. A higher value of specificity would therefore mean lower false positives and better separation between the classes. The results point out the comparative performance of two translation systems, namely DeepL Translator and Google Translate, along with four sentiment analysis models: OpenAI's GPT-4, Twitter-RoBERTa-Base, BERT-Base-Multilingual-Uncased-Sentiment, ZSC, GPT-3.5 and the proposed Ensemble Model. In the case of the DeepL Translator, the specificity values range from 0.85 to 0.91. The suggested Ensemble Model has maintained the peak specificity in the four languages, reaching 0.91 in French and constantly holding 0.90 for Arabic, Chinese, and Italian. This shows that the ensemble approach effectively minimizes false positive classifications, thus providing more precise sentiment differentiation. GPT-4 OpenAI also demonstrates its strong and balanced performance, with specificity values ranging between 0.87 and 0.89, underlining its stability across data in various languages. In the case of all settings, the Proposed Ensemble Model in Google Translate once again displays the highest specificity score, ranging from 0.90 to 0.92; in contrast, BERT-Base-Multilingual-Uncased-Sentiment and Twitter-RoBERTa-Base give somewhat lower but almost consistent scores between 0.85 and 0.89, reflecting reliable yet comparatively moderate performance. Overall, Table 7 depicts that the Proposed Ensemble Model yields the best and most consistent specificity across all translation systems and languages, thereby further indicating its robustness in minimizing false positive sentiment predictions while securing the appropriate identification of true

negative sentiments, a key determining element to improve multilingual sentiment analysis's precision and reliability. The performance of several multilingual sentiment analysis models is contrasted in Table 8. The proposed ensemble model integrates two translation systems, namely Google Translate and DeepL Translator, and two popular multilingual pre-trained BERT and XLNet models. Among the evaluation metrics are accuracy, precision, recall, F1-score, and specificity. which give a holistic measure of the capability of each model to handle multilingual sentiment classification tasks. The findings show that on every criterion, both iterations of the proposed ensemble model outperformed the baseline multilingual models. With a precision of 0.9624, the DeepL-based ensemble model outperformed the others. The F1-score was 0.9574, demonstrating very good performance in the task of sentiment classification with almost balanced precision and recall. The Google Translate-based ensemble also showed competitive performance with 0.9582 accuracy and 0.9357 specificity, thus proving robust across languages. In contrast, BERT and XLNet achieved slightly lower performances but still very strong ones, where in most metrics, XLNet outperformed BERT, especially regarding recall and F1-score. XLNet's bidirectional context modeling has a slight edge over capturing complex sentiment expressions in multilingual datasets. Overall, the ensemble-based approaches proved to be stronger and more reliable, which again shows their superior generalization and adaptiveness to multilingual sentiment analysis scenarios.

Table 8: Comparative outcomes of the multilingual model experiment.

Model Name	Accuracy	Precision	Recall	F1-score	Specificity
Proposed Ensemble Model - Google Translate	0.9582	0.9543	0.9518	0.953	0.9357
Proposed Ensemble Model - DeepL Translator	0.9624	0.9587	0.9562	0.9574	0.9405
mBERT [33]	0.9368	0.9324	0.9285	0.9304	0.9212
XLNet [19]	0.9427	0.9381	0.9342	0.9361	0.9275

4.2 Threats to Validity

The proposed framework has been tested on standard multilingual benchmark datasets with Arabic (ARA), Chinese (CHI), French (FRA) and Italian (ITA). Although these datasets are popular cannot be generalized for all languages, nor all application domains. Furthermore, the framework uses MT, which can have an impact on translation errors in sentiment classification, and then, in order to guarantee high reliability, two MT systems and multiple pre-trained LMs and the same experimental setup were used.

Justification of Evaluation Criteria

These metrics, such as Accuracy, Precision, Recall, F1-score, and Specificity, are commonly used and the gold standard metrics for sentiment classification evaluation, and to maintain impartiality during comparative analysis, selected state-of-the-art baseline models are recent and commonly used systems to compare with the proposed system. These evaluation measures help us compare the performance and confidence of the classification performed by the proposed architecture and its versatility across different languages.

4.3 Discussion

The proposed framework builds on, rather than incrementally refining, a specific AI paradigm such as neural machine translation, transformer-based language models, large language models, or ensemble learning, to a unified multilingual classification framework, offering significant knowledge beyond individual models, a limitation often shared in previous research works that only enhance a specific model. Furthermore, by introducing a robust, systematic method of multi-translator validation, weighted ensemble decision, and contextual knowledge fusion, this work designs a framework with high generalization and reliable, scalable performance, leading to effective cross-lingual sentiment classification. The proposed framework therefore not only sheds light on new methodology but also establishes a guideline for building effective, scalable, and reliable cross-lingual NLP systems.

5. CONCLUSION AND FUTURE WORK

The multilingual sentiment classification challenge resulting from linguistic variability, insufficient annotated resources and cross-lingual contextual ambiguity for current translation-based

and single-model techniques was tackled. A comprehensive framework leveraging the synergy between the capabilities of Google Translator, DeepL Translator, Twitter-RoBERTa, BERT-Multilingual, ZSC, GPT-3.5 and GPT-4, based on a weighted ensemble, has been implemented and assessed on multinational standard datasets. From the conducted experiments, the proposed multilingual model invariably outperformed individual pre-trained models according to metrics of Accuracy, Precision, Recall, F1-score and Specificity, verifying the efficacy of its application to enhancing cross-lingual sentiment classification. The investigation is a progressive exploration in contrast to the contemporary approaches that generally concentrated on one translation service or individual transformer models, providing a scalable and potent multilingual sentiment mining solution via a process of translation-aided combined model that provides deepened contextual interpretation and excellent cross-language performance. Therefore, this work presents an advancement in the research field with regard to the combined utilization of a set of neural machine translation services and a diverse collection of pre-trained language models in multilingual applications and lays down a precedent for more work with a broader range of low-resource languages, adaptive ensemble techniques, and specific-domain multilingual cases. Even though this framework proves very effective, it presents some limitations, such as the evaluation is only run on four languages, which is not representative enough of the total amount of communication exchanges in the world. Future work will then adapt this framework to various languages to improve the high-quality resource languages.

REFERENCES:

[1] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, and M. T. Martin-Valdivia, "Automatic medical protocol classification using machine learning approaches," *Comput. Methods Programs Biomed.*, vol. 200, p. 105939, 2021.

[2] Barbieri, F., Anke, L.E. and Camacho-Collados, J., 2022, June. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In Proceedings of the thirteenth language resources and evaluation conference (pp. 258-266).

[3] Han, Wenjuan, Bo Pang, and Ying Nian Wu. "Robust transfer learning with pretrained language models through adapters." In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers), pp. 854-861. 2021.

[4] Kassner, N., Dufter, P. and Schütze, H., 2021, April. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 3250-3258).

[5] Doddapaneni, Sumanth, Gowtham Ramesh, Mitesh Khapra, Anoop Kunchukuttan, and Pratyush Kumar. "A primer on pretrained multilingual language models." ACM Computing Surveys 57, no. 9 (2025): 1-39

[6] J. Mao and W. Liu, "Factuality classification using the pre-trained language representation model BERT," in *Proc. IberLEF@SEPLN*, 2019, pp. 126–131.

[7] Dhananjaya, V., Demotte, P., Ranathunga, S. and Jayasena, S., 2022, June. Bertifying sinhala-a comprehensive analysis of pre-trained language models for sinhala text classification. In Proceedings of the thirteenth language resources and evaluation conference (pp. 7377-7385).

[8] S. S. Badawi, "Using multilingual bidirectional encoder representations from transformers on medical corpus for Kurdish text classification," *ARO Sci. J. Koya Univ.*, vol. 11, pp. 10–15, 2023

[9] H. Rathnayake, J. Sumanapala, R. Rukshani, and S. Ranathunga, "Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification," *Knowl. Inf. Syst.*, vol. 64, pp. 1937–1966, 2022.

[10] S. Liu, F. Le, S. Chakraborty, and T. Abdelzaher, "On exploring attention-based explanation for transformer models in text classification," in *Proc. IEEE Int. Conf. Big Data*, 2021, pp. 1193–1203.

[11] E. Lee, C. Lee, and S. Ahn, "Comparative study of multiclass text classification in research proposals using pretrained language models," *Appl. Sci.*, vol. 12, p. 4522, 2022.

[12] Chakravarthi, B.R., Priyadharshini, R., Muralidaran, V., Jose, N., Suryawanshi, S., Sherly, E. and McCrae, J.P., 2022. Dravidiancodemix: Sentiment analysis and

offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, *56*(3), pp.765-806.

[13] X. Meng, "Research and implementation of multilingual text classification system based on deep learning," Ph.D. dissertation, Yanbian Univ., 2018.

[14] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," *Engineering*, 2022.

[15] J. Duan, H. Zhao, Q. Zhou, M. Qiu, and M. Liu, "A study of pre-trained language models in natural language processing," in *Proc. IEEE Int. Conf. SmartCloud*, 2020, pp. 116–121.

[16] E. Çelik and T. Dalyan, "Unified benchmark for zero-shot Turkish text classification," *Inf. Process. Manag.*, vol. 60, p. 103298, 2023.

[17] F. El-Alami, S. O. El Alaoui, and N. En Nahnahi, "A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 6048–6056, 2022.

[18] Y. Kit and M. M. Mokji, "Sentiment analysis using pre-trained language model with no fine-tuning and less resource," *IEEE Access*, vol. 10, pp. 107056–107065, 2022.

[19] A. Alduailej and A. Alothaim, "AraXLNet: Pre-trained language model for sentiment analysis of Arabic," *J. Big Data*, vol. 9, pp. 1–21, 2022.

[20] S. S. Badawi, "Using multilingual bidirectional encoder representations from transformers on medical corpus for Kurdish text classification," *ARO Sci. J. Koya Univ.*, vol. 11, pp. 10–15, 2023.

[21] Y. Kim *et al.*, "A pre-trained BERT for Korean medical natural language processing," *Sci. Rep.*, vol. 12, p. 13847, 2022.

[22] F. El-Alami, S. O. El Alaoui, and N. En Nahnahi, "Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, pp. 8422–8428, 2022.

[23] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Inf. Process. Manag.*, vol. 57, p. 102121, 2020.

[24] S. Romeo, D. Ienco, and A. Tagarelli, "Knowledge-based representation for transductive multilingual document classification," in *Proc. Eur. Conf. Inf. Retrieval*, 2015, pp. 92–103.

[25] X. Chen and J. Ren, "Multi-label text classification with latent word-wise label information," *Appl. Intell.*, vol. 51, pp. 966–979, 2021.

[26] S. Rosenthal, N. Farra, and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," in *Proc. SemEval-2017*, 2017, pp. 502–518.

[27] Bhargava, Maddineni, Karthika Vijayan, Oshin Anand, and Gaurav Raina. "Exploration of transfer learning capability of multilingual models for text classification." In Proceedings of the 2023 5th International Conference on Pattern Recognition and Intelligent Systems, pp. 45-50. 2023.

[28] R. Vinayakumar *et al.*, "DEFT 2017–Texts Search@TALN/RECITAL 2017: Deep analysis of opinion and figurative language on tweets in French," in *Proc. TALN*, 2017, p. 99.

[29] N. Novielli *et al.*, "SENTIPOLC 2016 dataset," Dataset, 2021. [Online]. Available: https://doi.org/10.57771/N279-Q780

[30] Van Nguyen, M., Lai, V.D., Veyseh, A.P.B. and Nguyen, T.H., 2021, April. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations (pp. 80-90).

[31] Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. "TimeLMs: Diachronic language models from Twitter." In *Proceedings of the 60th annual meeting of the association for computational linguistics: System demonstrations*, pp. 251-260. 2022.

[32] Voloshchuk, Yurii, and Oleksandr Mitsa. "DETERMINING THE EFFECTIVENESS OF GPT-4.1-MINI FOR MULTICLASS TEXT CATEGORIZATION IN ENGLISH AND UKRAINIAN." Eastern-European Journal of Enterprise Technologies 5 (2025).

[33] L. Khan, A. Amjad, N. Ashraf, and H. T. Chang, "Multi-class sentiment analysis of Urdu text using multilingual BERT," *Sci. Rep.*, vol. 12, p. 5436, 2022.

[34] Turunen, J., 2025. Text classification using language models: leveraging pre-trained language models for hierarchical multi-class

text classification under complex information management plans.

[35] Heseltine, M., 2025. Comparing Large Language Models for Text Classification: Model Selection Across Tasks, Texts, and Languages.