Review

A Review of Multilingual Sentiment Analysis

Xingjian Xie * and Xuefei Wu

Posted Date: 23 January 2026

doi: 10.20944/preprints202601.1816.v1

Keywords: multilingual sentiment analysis; machine learning; deep learning; ensemble learning; review

Review

A Review of Multilingual Sentiment Analysis

Xingjian Xie [1,*] and Xuefei Wu [2]

[1] School of Computer Science, University of Nottingham Ningbo China, China
[2] School of Computer Science, University of Nottingham, United Kingdom
* Correspondence: scyxx4@nottingham.edu.cn

Abstract

Multilingual sentiment analysis evaluates people's opinions and attitudes in a scenario of multiple languages. In today's digitally globalized world, online platforms serve as spaces where people from different cultural and linguistic backgrounds share their opinions, producing a significant volume of multilingual data every day. Such multilingual data holds substantial value, highlighting the growing importance of multilingual sentiment analysis. This paper reviews the development of multilingual sentiment analysis over the past decade, covering preprocessing methods, classification methods, feature extraction techniques, evaluation metrics, representative datasets, and experimental results. Additionally, this paper discusses the limitations of current research and future prospects of multilingual sentiment analysis.

Keywords: multilingual sentiment analysis; machine learning; deep learning; ensemble learning; review

1. Introduction

Multilingual sentiment analysis (MSA) is a technique that uses natural language processing (NLP) to identify and examine a user's emotional state from text data in multiple languages. In an increasingly globalized digital environment, multilingual sentiment analysis plays a vital role in bridging language barriers and understanding customer feedback, public opinion, and social trends in a comprehensive way. For example, [1] studied the impact of COVID-19 on people's emotions through multilingual sentiment analysis of tweets since the COVID-19 epidemic.

Multilingual sentiment analysis has applications in multiple fields, including cross-cultural market research [2], political sentiment monitoring [3], and so forth. The basic process of sentiment analysis usually begins with text preprocessing, which includes operations such as tokenization, stemming, and the removal of stop words. Feature extraction is then performed using methods such as Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and GloVe to represent the text data as features. Finally, machine learning or deep learning models are applied to identify the polarity of text emotions, classifying them as positive, negative, or neutral. With the continuous development of language processing technology, the complexity and application fields of multilingual sentiment analysis are also changing constantly, making it an innovative and interdisciplinary research field full of dynamics [4].

In contrast to other surveys on multilingual sentiment analysis [4–7], this paper presents a comprehensive overview covering the period from 2015 to 2025, focusing on machine learning, deep learning, and ensemble learning, with an active exploration into data preprocessing, feature extraction, classification models, and datasets being used, along with the experimental results of each work. This paper also summarizes several publicly available datasets for multilingual sentiment analysis and evaluation metrics, and discusses the limitations of current work and future research prospects.

The key contributions of this paper are:

- We provide an overview of research on multilingual sentiment analysis over the past decade, categorizing it based on the classifiers utilized. For each paper reviewed, we provide a brief

 doi:10.20944/preprints202601.1816.v1

introduction covering aspects such as classification model, supported languages, preprocessing techniques, datasets used, feature extraction methods, and experimental results.
- We summarize several representative datasets for multilingual sentiment analysis and provide an overview of the evaluation metrics.
- We discuss the current limitations in the field of multilingual sentiment analysis and provide several possible perspectives for future research.

2. Algorithms

This section reviews current approaches to multilingual sentiment analysis, which usually include several steps: dataset collection or selection, data preprocessing, feature extraction, and sentiment classification.

Multilingual sentiment analysis and conventional monolingual sentiment analysis are similar in terms of data preprocessing, feature extraction, and classification. However, unlike traditional monolingual sentiment analysis, multilingual sentiment analysis relies on datasets that cover two or more languages, which increases the difficulty of the sentiment classification task. A model that is effective in one language may not achieve the same level of performance on the data in another language.

Data preprocessing is essential in multilingual sentiment analysis, as it eliminates irrelevant information and transforms the text data, thereby improving overall task performance. Common preprocessing techniques include stopword removal, normalization, case folding, stemming, lemmatization, and tokenization.

Feature extraction is another crucial step in multilingual sentiment analysis, which can help represent data in a more compact and meaningful way. Some feature extraction methods include TF-IDF, word embeddings such as GloVe and Word2Vec.

In multilingual sentiment analysis, the classifier serves as the key component, as the final decision process relies on it. Machine learning relies on statistical models for multilingual sentiment analysis. Representative machine learning examples include Random Forest (RF), Support Vector Machines (SVM) and Logistic Regression (LR). Deep learning classifiers, on the other hand, use neural networks, where Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are among the most widely adopted models. Ensemble learning methods improve performance by leveraging the strengths of multiple models, for instance by combining several deep learning models through voting strategies.

Considering that classification models are critical to model performance, in this section, we categorize the papers we reviewed in terms of sentiment analysis methods, namely machine learning, deep learning, and ensemble learning.

2.1. Machine Learning

Conventional machine learning methods for multilingual sentiment analysis typically involve several steps: data preprocessing, feature extraction using methods such as TF-IDF, and sentiment classification through machine learning classifiers. Representative classifiers include Support Vector Machine (SVM), Random Forest(RF), K-Nearest Neighbor(KNN), Naive Bayes(NB), and Decision Trees(DT).

Ref. [8] performed sentiment analysis for Arabic and English text data, using three machine learning methods: Decision trees (DT), Naïve Bayes (NB), and support vector machines (SVM). Four English datasets were constructed by collecting from Amazon, regarding movies, books, DVDs, and electronics, while the Arabic dataset was collected from YouTube Arabic pages. The Arabic dataset was preprocessed by removing HTML tags, non-textual content, chunking, normalization, stop words removal, and stemming. The Sebti-Word lexicon for Arabic vocabulary was created, and TF-IDF was utilized for the feature extraction. The experimental results indicated that SVM achieved an accuracy of 0.877 on the movie dataset with feature selection, and NB achieved an accuracy of 0.8934 on the Arabic dataset without feature selection.

 doi:10.20944/preprints202601.1816.v1

Ref. [9] presented an SVM-based classifier supported by a lexicon expansion technique based on a distributional thesaurus and sentence-level co-occurrences for multilingual sentiment analysis. SAIL tweets in Hindi and Bengali were utilized for the experiment. Preprocessing techniques such as conversion of URLs, Twitter handles, multiple white spaces, and tokenization were applied. The results showed that the proposed method achieved an accuracy of 0.432 for Bengali and 0.4968 for Hindi, both in constrained setups.

Ref. [10] tested SVM using Greek and English datasets. They collected the Greek dataset and utilized the data from SemEval 2013 for the English dataset. The preprocessing techniques include removal of noise in the data, including URLs, mentions of other users, irrelevant abbreviations, and stop words, case folding, and accent marks. Bag-of-words representation with unigrams and term presences was employed for the representation. SVM achieved average F1 scores of 0.686 on the Greek dataset and 0.642 on the English dataset.

Ref. [11] proposed an approach based on an SVM classifier for multilingual sentiment analysis. Self-collected data from websites across four languages: Czech, German, English, and Spanish, was utilized for the evaluation. Tokenization was employed as a preprocessing technique. TF-IDF was utilized for the feature extraction. Notably, the proposed method achieved a best accuracy of 0.9531 for English, 0.9323 for Spanish, and 0.8905 for Czech, using certain document frequency and 2-gram optimization.

Ref. [12] tested several machine learning methods, such as Naïve Bayes, DMNBtext, NB Multinomial, and SVM, for multilingual sentiment analysis. Two datasets of Ukrainian and Russian news texts were employed for the experiment. Methods such as TF-IDF, correlation-based feature subset selection, information gain, and bag of words were utilized for the extraction of features. The experimental results indicated that Naive Bayes outperformed the other selected models, with an average F1 score of 0.82 on the Russian and Ukrainian datasets.

Ref. [13] evaluated several machine learning methods, including SVM, Naive Bayes(NB), maximum entropy(ME), decision tree (DT), random forest(RF), and KNeighbors(KNN). The selected datasets contain the English movie reviews dataset, Hindi movie reviews dataset, and HindEnCorp, which is a parallel corpus of English and Hindi. Machine translation was used in the model evaluation. SVM outperforms other models with an accuracy of 0.81 in the native English classification and 0.78 in the native Hindi classification using a certain feature set.

Ref. [14] proposed a method based on an SVM classifier for multilingual sentiment analysis. Several datasets, including TASS-15, SemEval 15-16, and SENTIPOLC-14, and some other datasets were selected across more languages: English, Spanish, Italian, Arabic, German, Portuguese, Russian, and Swedish. Preprocessing techniques include removal of stop words, stemming, negation clue attachment, and tokenization. TF-IDF, word based n-grams, and character q-grams were utilized for the feature extraction. The proposed model achieved accuracies of 0.637 on TASS-15 and 0.578 on SemEval-16 under specific settings.

Ref. [15] evaluated Naïve Bayes(NB) and Decision Tree(DT) for multilingual sentiment analysis. Three datasets in English, modern standard Arabic (MSA), and dialect Arabic (DA) were collected for the experiment. The authors of the paper preprocessed the data by techniques such as correction of misspellings, removal of repeated letters in a word, URLs, punctuation, numerals, and non-English words, and stop word filter. The results showed that Naive Bayes achieved an accuracy of 0.508 on the DA dataset, 0.895 on the MSA dataset, and 0.84 on the English dataset. Decision Tree obtained an accuracy of 0.544 on the DA dataset, 0.97 on the MSA dataset, and 0.84 on the English dataset.

Ref. [16] tested several machine learning methods, including methods such as the support vector classifier with a linear kernel (LSVC), SVC with the Radial Basis Function kernel(RSVC), Multinomial NB(MNB), KNN, decision tree classifier (Tree), linear Logistic Regression(LR), and Random Forest classifier(RF). The dataset includes 963 Gig worker and employer comments across three languages: German, French, and English. Additionally, a Twitter dataset was constructed for the model training. Preprocessing techniques such as case folding, punctuation removal, and tokenization were utilized

 doi:10.20944/preprints202601.1816.v1

for the text data. TF-IDF and N-gram were utilized for the feature extraction. The experiment showed that LR performed best among selected ML methods with an accuracy of 0.869 using 10-fold validation on the dataset of comments.

Ref. [17] performed multilingual sentiment analysis with a novel algorithm and several machine learning methods, including Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbour (KNN), and Random Forest (RF). A multilingual dataset was constructed, containing 200 tweets of different languages, including English, Telugu, Hindi, Bengali, Urdu, Arabic, and other undefined languages. Preprocessing techniques include removal of stop words, URLs, numbers, HTML tags, punctuations, special symbols and emoticons, substitution of contraction words, negation words handling, and spelling correction. Bag-of-words vectors were created for the feature extraction after the preprocessing. The results showed that SVM outperformed other selected methods with an accuracy of 0.95 on the multilingual dataset. Additionally, Random Forest achieved an accuracy of 0.93 on the multilingual dataset.

Ref. [18] performed multilingual sentiment analysis, using a machine learning approach with the integration of Hausa features and English features. Several classifiers were tested in the experiment, including Naive Bayes, SVM, and Maximum Entropy. A Twitter multilingual corpus, comprising 12,405 annotated tweets in English and Hausa with three sentiment classes and Hausa WordNet lexical (HWN) resource, was used for the proposed method. Preprocessing procedures, including tokenization, noise removal, stop word removal, and stemming, were applied. Techniques such as TF-IDF and n-gram were used for weighted English features, and term frequency was utilized to generate Hausa feature vectors. The results showed that Naive Bayes achieved best performance on the multilingual test dataset with an accuracy of 0.68 using n-gram English features and Hausa hexical features.

Ref. [19] evaluated several machine learning methods, including Naive Bayes, KNeighbors' Classifier, Decision Tree, and Random Forest. A dataset containing Hindi and Kannada user comments in Twitter was constructed, with two sentiment classes: positive and negative. The data underwent preprocessing such as removal of irrelevant hyperlinks, emojis, and acronyms, and word-level tokenization. The experimental results indicated that the Random Forest classifier achieved accuracies of 0.997 for Hindi and 0.995 for Kannada.

Ref. [20] tested 5 machine learning models, namely, Light Gradient Boosting Machine (LGBM), Random Forest (RF), Support Vector Machine with both linear and RBF kernels (RBF-SVM), and Logistic Regression (LR) for sentiment analysis and depression detection for English and Arabic tweets. The authors of the paper collected and annotated tweets in preparation for the dataset. Preprocessing techniques, such as removal of user mentions, punctuations, URLs, hash tags, short or meaningless words, repeating characters, symbols, English stop words, lemmatization, tokenization and so forth are employed. Tf-IDF and Bag of Words (BOW) were utilized for the feature extraction. Notably, Random Forest achieved an accuracy of 0.85 using TF-IDF and combined sampling with SMOTETomek, and Rbf-SVM achieved an average F1-score of 0.966 across all features on Arabic dataset with TF-IDF.

Ref. [21] evaluated several machine learning methods for multilingual sentiment analysis, including Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM). Two Arabic datasets (ASTD and AJGT) and their English translations were used were utilized in the experiment. Preprocessing techniques include tokenization, stop word removal, punctuation removal, and removal of repeated characters and speech parts, lemmatization, and stemming. Notably, SVM achieved its highest accuracy of 0.71 on the ASTD dataset and 0.86 for the AJGT dataset translated into English.

Ref. [22] proposed Stochastic Gradient Descent (SGD) for sentiment analysis of Arabic and English movie reviews. TF-IDF was utilized for the feature extraction. An Arabic dataset of 100,000 Arabic reviews and an English dataset containing 50,000 movie reviews from IMDB were employed in the experiment. Preprocessing techniques include data cleaning, label encoding, stop words removal,

 doi:10.20944/preprints202601.1816.v1

stemming, emoji removal, and tokenization. The proposed method achieved accuracies of 0.8489 on the Arabic dataset and 0.8744 on the IMDB dataset.

In summary, a variety of machine learning models, such as SVM, NB, DT, KNN, and RF have been explored for multilingual sentiment analysis, many of them achieving high performance. TF-IDF was the most common method for feature extraction along with these machine learning methods. A majority of algorithms are tested on self-collected datasets, indicating the diversity of data in this field. Table 1 provides a summary of the machine learning approaches reviewed in this paper.

Table 1. Summary of Machine Learning Methods.

Literature	Dataset	Language(s)	Features	Classifier	Accuracy
[8]	Self-collected dataset (from Amazon and Youtube)	English, Arabic	TF-IDF	DT, NB, SVM	0.877(English), 0.8934 (Arabic)
[9]	SAIL 2015	Hindi, Bengali	-	SVM	0.432 (Bengali), 0.4968 (Hindi)
[10]	Self-collected Greek dataset, SemEval 2013	Greek and English	-	SVM	0.686 (mF1 Greek), 0.642 (mF1 English)
[11]	Self-collected dataset(from websites)	Czech, German, English, Spanish	TF-IDF, 2-gram	SVM	0.9531 (English), 0.9323 (Spanish), and 0.8905 (Czech)
[12]	Self-collected dataset(from news)	Ukrainian, Russian	TF-IDF, correlation-based feature subset selection, information gain, and bag of words	NB, DMNBtext,NB Multinomial, SVM	0.82 (F1-score)
[13]	HindEnCorps, English movie reviews dataset, Hindi movie reviews dataset	Hindi, English	-	SVM, NB, ME, DT, RF, KNN	0.81 (English), 0.78 (Hindi)
[14]	TASS-15, SemEval 15-16, SENTIPOLC-14, other datasets in different languages	English, Spanish, Italian, Arabic, German, Portuguese, Russian, and Swedish	TF-IDF, word based n-grams, and character q-grams	SVM	0.637 (TASS-15 dataset), 0.456 (SemEval-16)
[15]	Self-collected dataset	MSA, DA, English	NB, DT	-	NB: 0.508 (DA), 0.895 (MSA) and 0.84 (English). DT: 0.544 (DA), 0.97 (MSA), 0.84 (English)
[16]	Self-collected dataset (Gigworker and Company Ratings, Twitter)	German, French, and English	TF-IDF, N-gram	LSVC, RSVC, MNB, KNN, Tree, LR, RF	LR: 0.869
[17]	Self-collected dataset(from Twitter)	English, Telugu, Hindi, Bengali, Urdu, Arabic and other undefined languages	Bag-of-words	MNB, LR, SVM, DT, KNN, RF	SVM: 0.95; RF: 0.93
[18]	HWN lexical resource, Self-collected dataset (from Twitter)	Hausa and English	TF-IDF, term frequency, n-gram	NB, SVM, ME	NB: 0.68
[19]	Self-collected dataset(from Twitter)	Hindi, Kannada	-	NB, KNN, Tree, RF	RF: 0.997 (Hindi) 0.995 (Kannada)
[20]	Self-collected dataset (from Twitter)	Arabic, English	TF-IDF, BOW	LR, RF, NB,SVM	RF: 0.85; Rbf-SVM: 0.966(F1-score)

 doi:10.20944/preprints202601.1816.v1

Literature	Dataset	Language(s)	Features	Classifier	Accuracy
[21]	ASTD, AJGT, English translated datasets	Arabic, English	-	LR, RF, NB,SVM	SVM: 0.71(ASTD), 0.86 (AJGT translated into English)
[22]	Arabic Review Dataset, IMDB	Arabic, English	TF-IDF	SGD	0.8489 (Arabic dataset); 0.8744(IMDB)

2.2. Deep Learning

In recent years, deep learning methods have gained increasing attention in the field of multilingual sentiment analysis. Multilingual data are first preprocessed and subsequently employed with pretrained embeddings, such as GloVe and Word2Vec. Deep learning classification methods were utilized, such as convolutional neural networks (CNN), recurrent neural networks (RNN), gated recurrent units (GRU), bidirectional long short-term memory (Bi-LSTM), and transformers.

Ref. [23] presented an attention-based LSTM for multilingual sentiment analysis. The model was tested on a cross-lingual sentiment classification dataset of NLP&CC 2013, which includes 4,000 reviews in English for training and 4,000 Chinese reviews for testing. The data underwent preprocessing, such as case folding and tokenization, while Word2Vec was utilized to produce word embeddings for feature extraction. LSTM was employed along with a hierarchical attention mechanism. The proposed model achieved an average accuracy of 0.824 using the proposed attention-based model on the multilingual dataset.

Ref. [24] proposed a CNN-based model with character level embedding for multilingual sentiment analysis. A Twitter subset, containing approximately 128k annotated tweets in four languages: English, Spanish, Portuguese, and German. The experimental results showed that the proposed architecture achieved an overall accuracy of 0.695 and an F1-score of 0.722 on multilingual datasets.

Ref. [25] proposed an approach based on CNN for multilingual sentiment analysis. The model was trained on a large Twitter dataset containing tweets across multiple languages, in a three-phase process: word embedding creation with skip-gram model, distant supervised, and supervised with a multi-layer CNN. The preprocessing techniques include substitution of URLs and usernames, case folding, and tokenization using the NLTK tokenizer. The experimental results showed that single-language CNN, which is trained for each single language, achieved an F1-score of 0.6779 for Italian, 0.6509 for German, 0.6479 for French, and 0.6226 for English.

Ref. [26] presented a sentiment classification model based on CNN. Multilingual restaurant review datasets were constructed constructed in English, French, and Greek. GloVe was utilized for feature extraction. The experimental results showed that the proposed model obtained an f1-score of 0.88 for trilingual analysis.

Ref. [27] presented a RNN-based model for multilingual sentiment analysis. Pre-trained global vectors are employed before the RNN-based model with bidirectional GRU layers with dropout. The model was experimented on restaurant reviews in four languages, Spanish, Turkish, Dutch and Russian from multiple public datasets, including Amazon reviews, Yelp restaurant reviews, Competition restaurant reviews, and SemEval-2016 Task 5. The results demonstrated that the presented model trained on English reviews obtained an accuracy of 0.8561 in the Russian dataset, 0.7436 on the Turkish dataset, 0.8177 on the Dutch dataset, and 0.8421 in the Spanish dataset.

Ref. [28] tested CNN and LSTM on unstructured text data. The datasets for the experiment comprise reviews of hotels and restaurants, collected from websites across three languages: French, English, and Greek. Neither preprocessing nor translation was utilized for the data. CNN obtained an accuracy of 0.9125 for sentiment analysis on the dataset in three languages while LSTM achieved an accuracy of 0.9127.

Ref. [29] presented the CNN-based model named NNLS to classify both the sentiment and language of the tweet text data. The model has two variations: NNLS-v1 and NNLS-v2. The model was evaluated on a subset of a twitter dataset across 4 languages: English, Spanish, Portuguese, and

 doi:10.20944/preprints202601.1816.v1

German. The experimental results showed that both variations of NNLS achieved an accuracy of over 0.72 for single-task sentiment analysis.

Ref. [30] proposed a method of language-agnostic deep neural network based on a modified Bi-LSTM. The proposed method was evaluated on SemEval-2016 task 5, GermEval-2017 task on ABSA, and a Hindi ABSA dataset, from where reviews in English, Hindi, Spanish, French, Dutch, and German are chosen. The authors propose four architectures in their approach, mainly differing in ways features are passed to the model. For example, architecture 1 only utilized word embeddings while architecture 2 uses a combination of word embedding and hand-crafted manual features. The experimental results show that the presented approach achieved the best accuracy of 0.834 in English, 0.669 in Hindi, 0.871 in Spanish, 0.753 in French, 0.819 in Dutch, and 0.872 in German using certain architecture with regard to aspect classification.

Ref. [31] proposed a hybrid model based on CNN and BiLSTM. The model comprises sub-word level representations generated by CNNs, a dual encoding with two BiLSTMs, and a Feature Network, which is built on surface features and monolingual sentence vectors. A Hindi-English code-mixed dataset comprises 3,879 sentences from popular Facebook pages in India. The proposed model achieved an accuracy of 0.8354 and an F1 score of 0.827.

Ref. [32] proposed a CNN-BiLSTM hybrid model for sentiment analysis on the code-switch text data. Subword level embedding was employed to extract features. An English-Kannada code-switch dataset comprising 10.4k comments from YouTube was constructed. Preprocessing techniques include removal of symbols, special characters and digits, case folding, and tokenization. The experimental results indicated that the proposed model achieved an accuracy of 0.776.

Ref. [33] proposed a Bi-LSTM based curriculum learning strategy for multilingual sentiment analysis. The model consists of an embedding layer followed by two bi-LSTM layers. Character trigram technique was emplyed during the feature extraction process. Three Hindi-English code-mixed datasets sourced from social media platforms were selected for the experiment. Preprocessing techniques, such as text normalization, masking of user mentions and URLs with special characters, and tokenization, were applied. The proposed approach achieved an accuracy of 0.7251.

Ref. [34] proposed a language-independent sentiment analysis model using a combination of CNN and LSTM networks. The model aims to predict sentiment from text data without relying on the specific language of the text. The study evaluated the model on two datasets: the English-language Sanders Twitter corpus and the German-language GermEval dataset. The proposed model with 300 dimensional embeddings achieved an accuracy of 0.8094 on the English dataset, 0.7475 on the German dataset, and 0.7508 on the mixed-language dataset.

Ref. [35] proposed a model based on LSTM, Adversarial Auto Encoders, and BiGRU for multilingual sentiment analysis. The model was evaluated on a cross-lingual dataset, which contains Amazon product comments for three product categories in Chinese, English, and German. Word2Vec and LSTM were utilized to create word embeddings. The results showed that the proposed model achieved an average accuracy of 0.78757.

Ref. [36] performed multilingual sentiment analysis using two deep learning methods, Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R). A MultiSenti dataset was built, containing tweets during the general elections of Pakistan in the year 2018, across two languages: English and Roman Urdu. Preprocessing techniques were performed, including case folding and removal of tweets involving the 'single word' text. The dataset was divided into 80% for training and 20% for testing. The results showed that mBERT achieved an accuracy of 0.69 on the targeted dataset, and XLM-R obtained an accuracy of 0.71, both with fine-tuning of hyperparameters.

Ref. [37] proposed an approach based on graph convolutional networks (GCN) with Multi-headed Attention. The selected Dravidian code-mixed datasets comprise 15,744 Tamil-English and 6,739 Malayalam-English comments from YouTube. The preprocessing techniques contain tokenization, removal of hashtags, numbers, URLs, user mentions, and transliteration. Then, the pre-trained fastText model was utilized to create word embeddings. The experimental results indicated that the proposed

 doi:10.20944/preprints202601.1816.v1

model achieved accuracies of 0.73 on the code-mixed Malayalam–English dataset and 0.71 on the code-mixed Tamil–English dataset.

Ref. [38] tested CNN, RNN and combined CNN-RNN models on multilingual sentiment analysis datasets. A balanced experimental dataset was built from three tweet datasets of three different languages: English, Portuguese and Arabic, with each contributing 3,000 positive tweets and 3,000 negative tweets. Preprocessing techniques include removal of URLs, usernames, stopwords, special characters, negation substitution, emojis replacement, case folding, removal of characters repeated in a word at least three times, and stemming. Word2Vec's CBOW model was utilized to create the word embedding. The experimental results demonstrated that CNN outperforms other models, including Bi-LSTM, Bi-GRU, CNN-LSTM, and CNN-GRU on the targeted dataset, with an accuracy of 85.91%.

Ref. [39] utilized XLM-T, a fine-tuning transformer-based language model targeting Twitter dataset, for multilingual sentiment analysis. The authors continued training an XLM-language model on 198 million collected tweets from available checkpoints. Then, the model was evaluated on a balanced multilingual sentiment analysis dataset comprising 24,262 tweets in 8 languages. The results showed that the presented model achieved an average F1 of 0.6935 with all languages at once.

Ref. [40] presented an approach named BPA, based on BiLSTM for multilingual sentiment analysis. The proposed model utilized networks and techniques, including BiLSTMs, Pooling Layers, and BERT embeddings. Datasets in English, Spanish, and Portuguese were selected, encompassing Stanford Sentiment Treebank(SST-2) dataset, SemEval 2017 Subtask A, TASS 2017 Task 1 General Corpus, and two other datasets built by the authors of the paper, namely, CCMD-ES in Spanish and CCMD-PT in Portuguese. The model obtained the best accuracies of 0.901, 0.865, and 0.923 for SST-2, CCMD-ES, and CCMD-PT, respectively.

Ref. [41] presented a BiLSTM-based model with an optimized ReLU layer for multilingual aspect-based sentiment analysis. The dataset for the experiment combines text from multiple sources in multiple languages such as Sentiment Lexicons for 81 Languages and Habeas Corpus Corpora, and so forth. Data underwent preprocessing techniques, such as tokenization, noise elimination, removal of special characters, text normalization, and Parts of Speech (POS) tagging. For the feature extraction, word embedding is identified with multiple approaches applied such as continuous Bag-of-Words (CBOW), Cosine similarity, and word mover's distance. The results showed that the proposed model achieved an accuracy of 0.9725 and a precision of 0.965 in the experiment.

Ref. [42] performed aspect-based sentiment analysis on both Chinese and English datasets using an interactive GCN-based model. The proposed model comprises the BERT-based embedding layer, the GCN layer, the multi-head self-attention–based information interaction layer, and the output layer. Four Chinese datasets(Car, Phone, Notebook, Camera) and six English datasets(Tshirt, Twitter, MAMS, REST14, REST15, REST16) were selected for the experiment. Notably, the proposed model achieves an accuracy of 0.9771 on the camera dataset and 0.9302 on the REST16 dataset.

Ref. [43] presented an approach based on pretrained BERT. An Indonesian-English code-mixed dataset was constructed for the experiment. Preprocessing techniques include emoji conversion, translation, slang word normalization, the removal of offensive content, duplicates, acronyms, hashtags, user mentions, and URLs, as well as case folding, translation, and extraneous word elimination. The results showed that the proposed model achieved an accuracy of 0.7656 and a precision of 0.7607 on the Indonesian-English code-mixed dataset when preprocessing techniques are applied.

Ref. [44] presented a Bi-LSTM based model for multilingual sentiment analysis. The proposed model utilized the shared parameters of a siamese network with a contrastive loss function. The datasets for the experiment include the English Twitter dataset, SemEval 2013, and Hindi-English Code-Mixed (HECM), which contains 3,879 Hindi-English code-mixed sentences. Techniques such as skip-gram, character trigram embedding were applied for the feature representation. The experimental results showed that the proposed model achieved an accuracy of 0.78 on HECM with preprocessing.

Ref. [45] proposed an RNN-BiLSTM model for multilingual sentiment analysis. A publicly available dataset of 1.37 million tweets relevant to Pakistan's government in three formats: Urdu,

 doi:10.20944/preprints202601.1816.v1

English, and transliterated Urdu. The data underwent preprocessing, including removal of non-English text, missing values, non-alphabetical characters, and blank words, integer encoding, sequence padding and tokenization. The experimental results showed that the proposed model achieved an accuracy of 0.9564, surpassing mBERT and ROBERTa on the multilingual dataset.

Ref. [46] proposed a hybrid diagonal gated recurrent neural network (FE-DGRNN) for sentiment analysis and hate speech detection on English and German text data. The model combines the strengths of RNNs and GRUs to boost the overall performance. The HASOC 2019 dataset with three subtasks, containing posts from Twitter and Facebook, was utilized for the experiment. Notably, the proposed model achieved an accuracy of 0.9687 for HASOC-2019 English dataset, and 0.9331 for HASOC-2019 German dataset.

Ref. [47] proposed an approach for multilingual sentiment analysis based on Graph Convolutional Network(GCN). A single heterogeneous text graph is constructed to model the corpus, after which the graph node representation was learned by Slightly Deep Graph Convolutional network. Six Amazon review datasets based on combinations of 4 languages, namely, English, German, French, and Spanish, were constructed, sourcing from MARC. Additionally, a movie reviews dataset was built by merging three movie reviews datasets. Preprocessing techniques include the removal of URLs, special characters and HTML tags, case folding, and tokenization. The results showed that the proposed method achieved an accuracy of approximately 0.88 on the movie review dataset and approximately 0.87 on the FR-DE Amazon review dataset.

Ref. [48] conducted multilingual sentiment analysis using an LSTM-based model and transfer learning in two phases: in two phases: sentiment analysis and transfer learning. Multilingual datasets of Amazon reviews across English, Modern Standard Arabic(MSA), and Bahraini dialects (BD) and an additional Brahrani dialect product review dataset were built for the experiment. Preprocessing techniques, such as text normalization, removal of stop words and non-alphanumeric characters, conversion of emojis, tokenization and so forth, were implemented. Notably, the experimental results indicated that the proposed model achieved accuracies of 0.9704 on the English dataset and 0.971 on the MSA dataset.

Ref. [49] performed multilingual sentiment analysis with DistilBERT model. The model was tested on datasets, including SST2, GLUE, and self-collected multilingual Twitter datasets, which contain tweets from six languages: English, French, German, Arabic, Italian, and Indonesian. Preprocessing techniques include the removal of special characters, usernames and URLs. Machine translation was utilized if the language entered is not English. Notably, the proposed model obtained an accuracy of 0.989 on the SST2 dataset, 0.911 on GLUE, and 0.915 on the Indonesian dataset.

Ref. [50] conducted a study on sentiment analysis across multiple languages using a subset of Multilingual Amazon Reviews Corpus(MARC), which includes 50,000 samples in French, German, Spanish, Japanese, Chinese. The authors evaluated the performance of transformer models such as BERT and XLM-RoBERTa on both original and machine-translated texts. For sentiment classification, the F1 score was used as the evaluation metric. Before machine translation, the models achieved F1 scores ranging from 0.86715 (Chinese XLM) to 0.90730 (German BERT). After machine translation, the F1 scores generally decreased, with the highest score being 0.89965 (Spanish XLM) and the lowest being 0.82391 (Chinese BERT).

Ref. [51] proposed a transformer-based joint learning framework for multilingual sentiment analysis. Input sentence representation is given to the task-specific BERT encoder to produce contextual sentence features. The features were combined and were then passed to task-specific dense layers, followed by classifiers to output labels for sentences in parallel. Four datasets were employed for multilingual sentiment classification tasks, including an English Dataset, a Hinglish code-mixed dataset (SemEval 2020 Task 9), a Punglish code-mixed dataset, and the HindiMD dataset. The preprocessing techniques included the removal of URLs, user mentions and retweet symbols, and conversion of emojis. The proposed model using BERT achieved 0.7431 on code-mixed Hindi-English data and 0.734 on Punjabi-English code-mixed data.

 doi:10.20944/preprints202601.1816.v1

Ref. [52] performed multilingual sentiment analysis for sarcasm detection using Bi-LSTM and LSTM. The paper experimented with Word embedding techniques, such as Word2Vec, FastText, GloVe, and BERT. Several benchmark datasets across English and Arabic were employed for the experiment, including the ArSarcasm-v2 dataset, iSarcasmEval dataset, IMDB dataset, and the SentiMixArEn dataset. Preprocessing techniques included elimination of noise and insignificant elements, removal of hashtags, repeated character handling, and stop word removal for both languages. Additionally, non-Arabic text removal and text normalization were implemented for the Arabic data. The results showed that BiLSTM with BERT stands out the most for the sentiment analysis with an accuracy of 0.8132 on Arsarcasm-v2 dataset and 0.9061 on IMDB movie review dataset using 10-fold validation.

Ref. [53] introduced a hybrid deep learning model optimized for multilingual sentiment analysis through the application of the Grey Wolf Optimization (GWO) algorithm. The proposed method combines CNNs and LSTMs and utilizes language-specific word embeddings. Additionally, the GWO algorithm was utilized for hyperparameter tweaking. The model was evaluated on a collected dataset consisting of 89,852 movie reviews with three sentiment types: positive, negative, and not assigned. The dataset covers six languages: Spanish, German, French, Polish, Czech, and Slovak. The preprocessing includes techniques such as case conversion, removal of HTML tags, URLs, punctuations, extra spaces, numbers, symbols and stop words, tokenization, lemmatization, and vectorization. The experimental results show that the presented model achieved an accuracy of 0.9598.

Ref. [54] performed multilingual sentiment analysis with a hybrid CNN-GRU model. A multilingual dataset of different languages, including English, Portuguese, Tamil, and French, was built using data collected from Kaggle. Preprocessing methods, such as tokenization, removal of punctuation, numbers, stopwords and HTML tags, and lemmatization, were used in preparation for the sentiment analysis. The results of the experiment showed that the proposed hybrid model achieved an accuracy of 0.98, 0.92, 0.91, and 0.89 for English, Portuguese, Tamil, and French, respectively.

Ref. [55] presents mBERT-based approach for multilingual sentiment analysis. A multilingual code-mixed dataset across Roman Punjabi and Roman Urdu was built, collecting comments from YouTube and messages from WhatsApp. Preprocessing techniques include stop word removal, negation handling, punctuation and emoji removal, lowercase conversion, and tokenization. BERT was utilized to create word embeddings for feature extraction. The results showed that the proposed model achieved an accuracy of 0.7994 for the code-mixed sentences.

On multilingual datasets, deep learning methods often achieve better performance than conventional machine learning methods, one reason being their ability to effectively capture complex semantic patterns. For deep learning approaches, the most popular feature extraction method is pretrained word embedding, such as GloVe and Word2Vec. Unlike conventional feature extraction methods, pretrained word embeddings can encode words in a way that can better capture semantic proximity. In addition to high-resource languages like English, which has been the main focus of sentiment analysis research, to a limited extent, some researchers have also explored low-resource ones, including Tamil. Table 2 provides a summary of the discussed deep learning methods in this subsection.

Table 2. Summary of Deep Learning Methods.

Literature	Classifier	Dataset	Language(s)	Features	Accuracy
[23]	attention-based LSTM	NLP&CC 2013	English, Chinese	Word2Vec	0.824
[24]	CNN	4-language tweets	English, Spanish, Portuguese, and German	-	0.695
[25]	CNN	Self-collected dataset(from websites)	Italian, German, French, English	skip-gram, word embedding	0.6779(F1-score, Italian), 0.6509(F1-score, German), 0.6479(F1-score, French), 0.6226(F1-score, English)
[26]	CNN	Self-collected	English, French and Greek	-	0.88 (F1-score, trilingual)

 doi:10.20944/preprints202601.1816.v1

Literature	Classifier	Dataset	Language(s)	Features	Accuracy
[27]	RNN, BI-GRU	restaurant reviews (from multiple sources)	Spanish, Turkish, Dutch and Russian	Pre-trained global vectors	0.8561(Russian), 0.8421(Spanish), 0.7436 (Turkish), 0.8177 (Dutch)
[28]	CNN, LSTM	Self-collected dataset(reviews from website)	French, English, Greek	-	CNN: 0.9125; LSTM: 0.9127
[29]	CNN	subset of a Twitter corpora	English, Spanish, Portuguese, and German	-	0.72
[30]	Bi-LSTM	SemEval-2016 task 5 and GermEval-2017 task on ABSA	English, Hindi, Spanish, French, Dutch, and German	word embedding, hand-crafted manual features	0.834 (English), 0.669 (Hindi), 0.871 (Spanish), 0.753 (French), 0.819 (Dutch), 0.872 (German)
[31]	CNN, BiLSTM	HECM	Hindi, English	sub-word representation	0.8354
[32]	CNN, BiLSTM	Self-collected	English, Kannada	Sub word level embedding	0.776
[33]	Bi-LSTM	Hindi-English code-mixed datasets	Hindi, English	character trigram	0.7251
[34]	CNN, LSTM	Sanders Twitter corpus, GermEval dataset	English, German	-	0.8094(English), 0.7475(German), 0.7508(mixed)
[35]	LSTM,Adversarial Auto Encoder, Bi-GRU	cross-lingual Amazon product dataset	Chinese, English, and German	word embedding	0.78757
[36]	mBERT, XLM-R	self-collected	English, Roman Urdu	-	mBERT: 0.69; XLM-R:0.71
[37]	GCN	Dravidian code-mixed dataset	Tamil, Malayalam, English	word embedding	0.73 (Malayalam-English), 0.71 (Tamil-English)
[38]	CNN, RNN, CNN-RNN	self-collected	English, Portuguese and Arabic	word embedding	0.8591
[39]	XLM-T	self-collected dataset	8 languages (for evaluation)	-	0.6953(average F1)
[40]	BiLSTM	SST-2, SemEval 2017 Subtask A, TASS 2017 Task 1 General Corpus, two self-collected datasets	English, Spanish, and Portuguese	-	0.901(SST-2), 0.865(CCMD-ES), 0.923(CCMD-PT)
[41]	BiLSTM,	self-constructed datasets(from multiple sources)	multiple languages	BERT	0.9725
[42]	GCN	4 Chinese datasets(Car, Phone, Notebook, Camera), 6 English datasets(Tshirt, Twitter, MAMS, REST14, REST15, REST16)	Chinese, English	BERT	0.9771(camera); 0.9302(REST16)
[43]	pretrained BERT	self-collected	Indonesian, English	-	0.7656
[44]	Bi-LSTM	English Twitter dataset, SemEval 2013, and Hindi-English Code-Mixed (HECM)	Hindi, English	Skip-gram, character trigram embedding	0.78(HECM)
[45]	RNN-BiLSTM	Pakistan Government Twitter Dataset	Urdu, English	-	0.9564
[46]	hybrid diagonal gated RNN	HASOC 2019	English, German	-	0.9687(English), 0.9331(German)
[47]	GCN	self-constructed (MARC, 3 movie review datasets)	English, German, French, Spanish	-	0.87(FR-DE), 0.88(movie review)
[48]	LSTM	self-constructed	English, MSA, BD, BD dialect	-	0.9704(English), 0.971(MSA)

 doi:10.20944/preprints202601.1816.v1

Literature	Classifier	Dataset	Language(s)	Features	Accuracy
[49]	DistilBERT	SST2, GLUE, self-collected multilingual Twitter datasets	English, French, German, Arabic, Italian, Indonesian	-	0.989 (SST2), 0.911(GLUE) and 0.915 (Indonesian dataset)
[50]	BERT, XLM-RoBERTa	subset of MARC	French, German, Spanish, Japanese, Chinese	-	No MT:0.86715 (Chinese XLM), 0.90730 (German BERT) MT: 0.89965 (Spanish XLM), 0.82391 (Chinese BERT)
[51]	task-specific BERT	English Dataset, Hinglish code-mixed dataset(SemEval2020 Task 9 dataset), Punglish code-mixed dataset, HindiMD dataset	English, Hindi, Punjabi	-	0.7431 (Hindi-English), 0.734 (Punjabi-English)
[52]	BiLSTM, LSTM	ArSarcasm-v2, iSarcasmEval, IMDB, SentiMixArEn	English, Arabic	word embedding	0.8132 (Arsarcasm, BiLSTM with BERT), 0.9061 (IMDB, BiLSTM with BERT)
[53]	CNN, LSTM	self-collected	Czech, German, Spanish, French, Polish, and Slovak	word embedding	0.9598
[54]	CNN-GRU	Self-constructed(from Kaggle)	English, Portuguese, Tamil, and French	-	0.98(English), 0.92(Portuguese), 0.91(Tamil), 0.89(French)
[55]	mBERT	self-collected (from Youtube and Whatsapp)	Roman Punjabi, Roman Urd	BERT	0.7994

2.3. Ensemble Learning

In recent years, ensemble learning has also become a relatively popular approach in multilingual sentiment analysis. By combining multiple models, ensemble learning achieved better model performance than each of the individual models it combines. Ensemble strategies such as bagging, stacking,voting, and boosting are employed to combine multiple models, to leverage the strengths of individual models.

Ref. [56] proposed an ensemble of an LSTM model based on character-trigrams and a Multinomial Naive Bayes based on word-ngrams. Hindi-English code-mixed dataset (HECM), containing 3,879 Facebook user comments from two Indian public figures was utilized for the experiment. The data was divided into 70% for training, 10% for testing, and 20% for validation. The results showed that the proposed model achieved an accuracy of 0.708 and an F1-score of 0.661 in the experiment.

Ref. [57] presented an ensemble model with sampling and emotion lexicon. The proposed model combines different classifiers, including Fasttext, RCNN, and CNN, using a simple voting technique. Data from the NLPCC2018 task1, which contains in total around 8,000 posts in English and Chinese, was utilized for the experiment. The results of the experiment showed that the proposed model achieved an F1-score of 0.734 for happiness and 0.616 for Sadness, both on the test set.

Ref. [58] proposed an ensemble model for Bengali and Hindi sentiment analysis, integrating two Multinomial Naive Bayes models and an SVM-based model. Bengali and Hindi SAIL 2015 datasets of tweets were utilized for the experiment, after which the data were preprocessed with irrelevant character removal. Techniques such as word n-gram, character n-gram, and unigram were used for feature extraction. The results showed that the proposed model using majority voting achieved the highest accuracy of 0.5753 for Bengali and the proposed model using average of probabilities achieved the highest accuracy of 0.6263 for Hindi.

Ref. [59] proposed a stacking ensemble of seven deep learning models, including a Bi-LSTM-Bi-GRU model, CNN-Bi-GRU model, CNN-Bi-LSTM, Bi-LSTM-LSTM model, Bi-GRU model, BI-LSTM model, and Deep CNN model. Datasets, including Emirati Sentiment Analysis Annotated Dataset

 doi:10.20944/preprints202601.1816.v1

(ESAAD), Arabic company reviews dataset (ACRD), preprocessed sentiment analysis dataset (PSAD), and a Twitter-based benchmark Arabic sentiment analysis dataset (ASAD) were used for training and evaluating the model, containing comments from Instagram, tweets, Arabic Company reviews, and Amazon product reviews, across three languages: Emirati, Arabic, and English. Preprocessing procedures, including tokenization, stop word removal, and normalization, were applied in the experiment. Training, testing, and validating constitute 80%, 10%, 10% of the utilized datasets, respectively. AraBERT and MARBERT were utilized for the feature extraction. The results showed that the proposed ensemble with a meta classifier Random Forest achieved an accuracy of 0.9462 for the undersampled ESAAD dataset, 0.946 on the ACRD dataset, 0.983 for the balanced oversampled ASAD dataset, and 0.9779 on the English dataset PSAD.

Ref. [60] proposed an ensemble model combining Cardiffnlp Base, Cardiffnlp XLM, and distilUSE-base-multilingual-case model (distiluse) for multilingual sentiment analysis. In the stacking process, Multinomial Naive Bayes (MultinomialNB) and Support Vector Classifier (SVC) were selected as the estimators for Ternary Classificationw and Binary classification, respectively. Spanish-English and Hindi-English data from the SemEval2020 Task 9 dataset, Tamil-English and Malayalam-English data derived from Dravidian-CodeMix challenge, and Code-Mixed Telugu-English Tex (CMTET) dataset, were utilized for the experiment. Google Translation API was applied to translate the targeted dataset into English, after which preprocessing techniques were employed, including removal of emojis, punctuations, and web links. The experimental results showed that the proposed ensemble model achieved an accuracy of 0.87 for the binary classification and 0.74 for the ternary classification on the original dataset. Regarding sentence-level translation with a known language source, the ensemble model obtained an accuracy of 0.84 for binary classification and 0.7 for ternary classification.

Ref. [61] performed multilingual sentiment analysis using an ensemble learning approach for classification. The proposed stacking ensemble combines LSTM, GRU, and BiLSTM with a meta classifier SVM. BERT was utilized to generate word embeddings for the feature extraction. Two English datasets(Sentiment140 and Twitter US Airline Sentiment), two Arabic datasets (Arabic-Reviews of Hotels Dataset and Arabic Twitter dataset) and two Moroccan dialect datasets (MAC dataset and Dialect dataset) were utilized for the experiment. Under-sampling and oversampling were used for the data augmentation. The results showed that the proposed ensemble achieved the best accuracy of 0.94 on the Sentiment140 dataset, 0.94 on the Arabic Twitter Dataset, and 0.943 on the MAC dataset, all using 10-fold validation and SMOTE, an over-sampling technique.

Ref. [62] proposed a voting ensemble model, combining two machine learning methods, the Random Forest and the Logistic Regression. A Hindi-English dataset containing comments on the videos from social media or YouTube with different themes and sentiment was built. Irrelevant conjunctions and articles are excluded from the experimental data and tokenization was employed during the preprocessing. Methods such as TF-IDF and word embeddings(Word2Vec or GloVe) were utilized for feature extraction. The results of the experiment showed that the ensemble model achieved an accuracy of 0.92 on the bilingual dataset.

Ref. [63] proposed two ensemble models for multilingual sentiment analysis. Both ensembles combine an AraBERT or RoBERTa Model and a multilingual BERT model with a feed-forward network. One of them employed a multi-head attention after the pretrained models and before the feed-forward network. The data used in the experiment were from SemEval-17 and Arabic Sentiment Tweet dataset(ASTD) in English and Arabic. Preprocessing techniques include removal of symbols, URLs and invisible characters, and tokenization using Byte-pair encoding tokenizers. The ensemble model with multi-head attention Feed Forward achieved an accuracy of 0.674 in English and 0.663 in Arabic. The ensemble model with Feed Forward achieved an accuracy of 0.6891 in English using English training data and 0.6767 in Arabic using Arabic training data.

Ref. [64] performed multilingual sentiment analysis using an ensemble of Support Vector Machine, Random Forest, Logistic Regression, and Naive Bayes. UCL Roman Urdu dataset, comprising 20,228 sentences, and IMDB dataset were utilized for the experiment. Preprocessing techniques include case

 doi:10.20944/preprints202601.1816.v1

folding, removal of digits, special characters and punctuation, ASCLL control characters, HTML tags, stop words, and sentences completely in English. Bag-of-words and TF-IDF were explored for the feature extraction. The experimental results indicated that the proposed model employing a stacking technique, TF-IDF, and combined unigram–bigram features achieved accuracies of 0.803 on the UCL dataset and 0.9092 on the IMDB dataset.

Ref. [65] proposed an ensemble model of transformers and LLM for multilingual sentiment analysis. Machine translation methods, LibreTranslate and Google Translate, were implemented before the ensemble of pretrained models, including Twitter-Roberta-Base-Sentiment-Latest, bert-base-multilingual-uncased-sentiment, and GPT-3. The proposed model adopts majority voting as their ensemble strategy. The model was tested on a dataset in Arabic, Chinese, French and Italian, collected from multiple sources. The results demonstrated that the proposed model achieved an accuracy of 0.8671 using Google translate.

Ref. [66] proposed an ensemble-driven multilingual sentiment analysis method, which combines Logistic Regression, Random Forest and SVM with Soft Voting Classifier. Multilingual embeddings, such as mBERT and XLM-RoBERTa, were incorporated in the presented method to improve the sentiment classification performance. The CodeSwitched-YTSentiment dataset of YouTube comments was built through the YouTube Data API v3, containing 50 Tamil, 250 English, and 300 Tanglish (TamilEnglish hybrid) comments. Preprocessing techniques, including text cleaning, transliteration, stop words removal, normalization, and noise handling, such as removal of hashtags, URLs, and numerical data. TF-IDF and pretrained word embeddings, including FastText and Word2Vec, were employed for the feature extraction. The results showed that the proposed model achieved an accuracy of 0.97, outperforming traditional methods, including Naive Bayes and standalone SVM.

In conclusion, ensemble learning approaches have demonstrated high performance on various multilingual sentiment analysis datasets. By applying appropriate ensemble strategies that combine various models, including deep learning methods such as CNNs, GRUs, and BiLSTMs, and machine learning methods such as SVM and Random Forest, the strengths of different models can be effectively utilized. The discussed ensemble models were employed on various datasets in multiple languages with different feature extraction methods applied, including BERT, TF-IDF, BOW, and so forth. Table 3 presents a summary of ensemble models discussed in this subsection.

Table 3. Summary of Ensemble Learning Methods

Literature	Classifier	Dataset	Language(s)	Features	Accuracy
[56]	LSTMs, Multinomial Naive Bayes	HECM	Hindi, English	character-trigram, word-ngram	0.708
[57]	Fasttext, RCNN, CNN	NLPCC2018 Task1	English and Chinese.	-	0.734(happiness); 0.616(sadness)
[58]	Multinomial NB, SVM	SAIL 2015	Bengali, Hindi	word n-gram, character n-gram, unigram	0.5753 (Bengali), 0.6263 (Hindi)
[59]	Bi-LSTM-Bi-GRU model, CNN-Bi-GRU model, CNN-Bi-LSTM, Bi-LSTM-LSTM model, Bi-GRU model, BI-LSTM model, Deep CNN model	ESAAD, ASAD, ACRD, PSAD	Emirati, Arabic, English	AraBERT, MARBERT	0.9462 (under-sampled ESAAD), 0.946 (ACRD), 0.983 (balanced oversampled ACRD), 0.9779 (PSAD)
[60]	Cardiffnlp Base, Cardiffnlp XLM, distiluse	SemEval2020 Task 9 dataset, Tamil-English and Malayalam-English datasets, CMTET	Hindi, Spanish, Tamil, Malayalam, Telugu, English	-	Original dataset: 0.84(binary classification), 0.74(ternary classification)

 doi:10.20944/preprints202601.1816.v1

[61]	LSTM, GRU, BiLSTM	Sentiment140, Twitter US Airline Sentiment, Arabic-Reviews of Hotels Dataset, Arabic Twitter dataset, MAC dataset, Dialect dataset	English, Arabic, Moroccan	BERT	0.94 (Sentiment140 dataset), 0.94 (Arabic Twitter Dataset), 0.943(MAC dataset)
[62]	RF, LR	self-collected (from Youtube)	Hindi, English	TF-IDF, word embeddings	0.92
[63]	AraBERT or RoBERTa Model, multilingual BERT	SemEval-2017 task4, ASTD	English, Arabic	-	0.6891 (English), 0.6767 (Arabic)
[64]	SVM, RF, LR, NB	UCL Roman Urdu dataset, IMDB dataset	Roman Urdu, English	BOW, TF-IDF, unigram and bigram	0.803 (UCL dataset), 0.9092 (IMDB)
[65]	transformers and LLM	self-constructed	Arabic, Chinese, French, Italian	-	0.8671(google translate)
[66]	LR, RF, SVM	self-collected (from Youtube)	Tamil, English, Tanglish	mBERT and XLM-RoBERTa	0.97

3. Datasets

Datasets play a crucial role in multilingual sentiment analysis. During the process of our review, we found that many studies in this field employ self-collected datasets, which vary in language coverage and domains, reflecting the diversity of data in this field. Nevertheless, some works also make use of publicly available datasets, which can help reduce the cost of time and manpower. Several representative publicly available datasets for multilingual sentiment analysis are shown below.

3.1. SemEval-2016 Task 5

SemEval 2016 [67] shared task on Aspect-Based Sentiment Analysis (ABSA) consists of 70,790 annotated ABSA tuples, including 47,654 sentence-level annotations (SB1) in 8 languages for 7 domains, and 23,136 text-level annotations (SB2) in 6 languages for 3 domains. The languages include English, Chinese, French, Arabic, Russian, Spanish, Dutch and Turkish. The domains of the data include restaurants, laptops, mobile phones, digital cameras, hotels, museums, as well as telecommunication. The SE-ABSA16 task comprises three subtasks: Sentence-level ABSA, Text-level ABSA, and Out-of-domain ABSA.

3.2. Multilingual Amazon Reviews Corpus

The Multilingual Amazon Reviews Corpus (MARC) [68] is a large dataset of Amazon reviews for multilingual sentiment analysis. It includes reviews in 6 languages: English, Japanese, German, French, Spanish, and Chinese, collected from 2015 to 2019.

The dataset is evenly distributed for five-star ratings. For each language, 200,000 reviews are used for training, with 5,000 reviews allocated to both the development and test sets. Techniques such as sampling, filtering, and text processing are utilized to reduce noise in the dataset.

3.3. SemEval-2017 Task 4

SemEval-2017 Task 4 dataset [69] is a comprehensive multilingual dataset for sentiment analysis, comprising over 70,000 tweets of different topics in two languages: English and Arabic. Duplicate tweets are detected and those whose bag-of-words cosine similarity surpasses 0.6 were removed. CrowdFlower was employed to annotate the new training and testing datasets.

The SemEval-2017 Task 4 dataset consists of five distinct subtasks that each explore different facets of sentiment analysis, containing overall sentiment classification, topic-specific classification, and tweet quantification. The labels were set on 2-point, 3-point, and 5-point scales.

 doi:10.20944/preprints202601.1816.v1

3.4. HECM

Hi-En code-mixed dataset (HECM) [70] is a bilingual code-mixed dataset containing comments from popular Facebook pages in India, including Salman Khan, a popular Indian actor, and Narendra Modi, the Prime Minister of India. The dataset is composed of 15% negative comments, 50% neutral comments, and 35% positive comments, with 3,879 annotated sentences in total.

3.5. SemEval-2020 task 9

Semeval-2020 task 9 [71] provides Hinglish (Hindi-English) and Spanglish (Spanish-English) code-mixed datasets, comprising 20k and 19k annotated tweets, labeled as positive, neutral, or negative. Each tweet has word-level language marking and can be used for other tasks besides sentiment analysis.

3.6. Sentiment Analysis in Indian Languages

Sentiment Analysis in Indian Languages (SAIL) [72] provides tweets for researchers in three Indian languages: Bengali, Hindi and Tamil, which are divided into three sentiment polarity classes: positive, negative, and neutral. The numbers of tweets are 1,498, 1,688, and 1,663 for Bengali, Hindi, and Tamil, respectively. The tweets in the dataset contain many spelling variations, abbreviations, and emoticons, contributing to the difficulty of sentiment analysis.

Table 4 shows the summary of discussed multilingual sentiment analysis datasets.

Table 4. Overview of Representative Multilingual Sentiment Datasets

Dataset	Languages	Sentiment Classes	Scale
SemEval-2016 Task 5 [67]	English, Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish	3,4	70,790 tuples
MARC [68]	English, Japanese, German, French, Spanish, and Chinese	5	1.26 M reviews
SemEval-2017 Task 4 [69]	English, Arabic	2,3,5	70 k+ tweets
HECM [70]	Hindi, English	3	3,879 sentences
SemEval-2020 Task 9 [71]	Hindi, English, Spanish	3	20k(Hindi-English), 19k(Spanish-English)
SAIL [72]	Bengali, Hindi, Tamil	3	1,498 tweets (Bengali), 1,688 tweets (Hindi), 1,663 tweets (Tamil)

4. Evaluation Metrics

The evaluation metrics are essential for the multilingual sentiment analysis since they measure the performance of the models. The most common evaluation metrics are Accuracy, Precision, F1-score, and Recall. These evaluation metrics can be calculated based on the confusion matrix in Table 5. [73]

Table 5. Confusion Matrix

	Predicted positive	Predicted negative
Actually Positive	True Positives (TP)	False Negatives (FN)
Actually Negative	False Positives (FP)	True Negatives (TN)

 doi:10.20944/preprints202601.1816.v1

4.1. Accuracy

Accuracy demonstrates the proportion of correct predictions among all the predictions made by models.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.2. Precision

Precision indicates the proportion of true positive predictions among all predicted positive instances.

$$Precision = \frac{TP}{TP + FP}$$

4.3. Recall

Recall reflects the ratio of true positive predictions over all actual positive predictions made by the classifiers.

$$Recall = \frac{TP}{TP + FN}$$

4.4. F1-score

F1-score equals the harmonic average of recall value and precision value.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5. Limitations

Although multilingual sentiment analysis has achieved notable progress, its accuracy and applicability constraints continue to exist. These limitations include the handling of informal and poorly structured data, underexplored low-resource languages, limited research on real-time multilingual sentiment analysis and predominance of low-granularity multilingual sentiment classification.

Informal and poorly structured data: The existing methods for multilingual sentiment analysis are mainly trained and tested on structured data. The performance of models on the informal and ill-structured text data falls far short of expectations, since elements like slang, grammatically incorrect expressions, and expressions with ironic undertones, which require a comprehensive contextual understanding of the text, pose challenges for the classification of the sentiment classes.

Underexplored low-resource languages sentiment analysis: Multilingual sentiment analysis for low-resource languages faces the difficulty caused by the scarcity of data and has received limited attention. Although multilingual sentiment analysis for several major languages such as English, Chinese, and French, has gained significant progress in recent years, multilingual sentiment analysis for low-resource languages lacks sufficient data for model training and remains underexplored.

Real-time multilingual sentiment analysis: Current multilingual sentiment analysis research is largely limited to static data processing, with little exploration of the real-time complex scenarios. The combination of massive data volumes and the high data generation velocity increases the difficulty of performing real-time analysis in the dynamic multilingual real-world settings.

low-granularity multilingual sentiment classification: Current research on multilingual sentiment analysis is largely towards low-granularity sentiment analysis, which commonly involves three sentiment categories: positive, negative, and neutral. The coarse-grained approach fails to effectively capture the diverse range of human emotions and attitudes, such as sarcasm and anger, and is generally less challenging than the fine-grained sentiment analysis.

6. Future Research Directions

To mitigate the challenges discussed in Section 5, future research can focus on broadening the coverage of sentiment analysis and the dataset used. This may include improved text data processing,

 doi:10.20944/preprints202601.1816.v1

development for low-resource languages, handling real-time data and advances in fine-grained multilingual sentiment analysis.

Processing of informal and poorly organized data: Multilingual sentiment analysis models struggle with informal or badly structured text data. Researchers can try approaches of adversarial learning or reinforcement learning and train models over a large-scale, diverse corpus with informal or poorly structured text data to develop models more robust to poorly structured data.

Low-resource sentiment analysis: Currently, the sentiment analysis on multiple languages is mainly focused on some popular languages and there is little research on low-resource languages. Researchers can develop comprehensive datasets with diverse and complex contextual data, using web crawling techniques to collect data from social media and other platforms, which can facilitate further exploration of multilingual sentiment analysis for low-resource sentiment analysis.

Handling real-time data: Multilingual sentiment analysis faces challenges in handling real-time dynamic data. Researchers can explore the use of lightweight multilingual models to enable cost-effective deployment for handling real-time data. With lower sizes and computational costs, lightweight multilingual models can reduce the inference time and enable faster real-time multilingual sentiment analysis.

Fine-grained multilingual sentiment analysis: Most existing studies of multilingual sentiment analysis focus on coarse-grained polarity sentiment classification, which struggles to represent the full spectrum of human emotions and attitudes accurately. Fine-grained multilingual sentiment analysis can overcome the limitation by expanding emotion categories such as anger, fear, surprise and disgust and including emotion intensity. Researchers can explore various kinds of deep learning and ensemble learning strategies for fine-grained multilingual sentiment analysis.

7. Conclusions

Owing to its wide-ranging interdisciplinary applications, multilingual sentiment analysis has become an important topic in natural language processing (NLP). Early multilingual sentiment analysis mainly relied on traditional machine learning techniques. Text data undergoes preprocessing steps such as tokenization, stem extraction, deletion of stop words, and normalization. Then, features are extracted using methods such as TF-IDF, after which the processed data are input into machine learning classifiers, including Naive Bayes, SVM, and DT. However, with the rapid development of NLP, the focus has shifted towards deep learning techniques. In general, the text is first mapped to a dense vector space through pre-trained embeddings such as Word2Vec or contextualized encoders such as mBERT, XLM-R, and DistilBERT. These embeddings capture semantic and syntactic patterns in the text, which are then processed by deep learning methods such as CNN, RNN and Transformers. Recent studies have also explored ensemble learning methods that combine predictions from multiple models to enhance performance. While many researchers used self-collected datasets for multilingual sentiment analysis, some publicly available datasets are helpful in advancing research in this field, such as SemEval-2016 Task 5 (ABSA), the multilingual Amazon review corpus, and the SemEval-2017 Task 4 dataset.

Despite significant progress, existing methods still face challenges in handling informal, poorly-structured and real-time data, highlighting the need for more context-aware and cost-effective language models. In addition, most current research focuses on coarse-grained sentiment classification, such as positive, negative, and neutral, while research on low-resource languages sentiment analysis is still limited. Future research should explore more fine-grained sentiment analysis models to capture a wider range of emotions and different emotional intensities, while focus more on informal, poorly-structured and low-resource data handling.

Data Availability Statement: No data was used for the research described in the article.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202601.1816.v1

References

1. Kruspe, A.; Häberle, M.; Kuhn, I.; Zhu, X.X. Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic. *arXiv preprint arXiv:2008.12172* **2020**.
2. Oliveira, A.S.; Renda, A.I.; Correia, M.B.; Antonio, N. Hotel customer segmentation and sentiment analysis through online reviews: An analysis of selected European markets. *Tourism & Management Studies* **2022**, *18*, 29–40.
3. Vicente, I.S.; Saralegi, X.; Agerri, R. Real Time Monitoring of Social Media and Digital Press. *arXiv preprint arXiv:1810.00647* **2018**.
4. Abdullah, N.A.S.; Rusli, N.I.A. Multilingual Sentiment Analysis: A Systematic Literature Review. *Pertanika Journal of Science & Technology* **2021**, *29*, 445–470. https://doi.org/10.47836/pjst.29.1.25.
5. Agüero-Torales, M.M.; Abreu Salas, J.I.; López-Herrera, A.G. Deep learning and multilingual sentiment analysis on social media data: An overview. *Applied Soft Computing* **2021**, *107*, 107373. https://doi.org/https://doi.org/10.1016/j.asoc.2021.107373.
6. Mabokela, K.R.; Celik, T.; Raborife, M. Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access* **2023**, *11*, 15996–16020. https://doi.org/10.1109/ACCESS.2022.3224136.
7. Aliyu, Y.; Sarlan, A.; Usman Danyaro, K.; Rahman, A.S.B.A.; Abdullahi, M. Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources. *IEEE Access* **2024**, *12*, 66883–66909. https://doi.org/10.1109/ACCESS.2024.3398635.
8. Elawady, R.; Barakat, S.; Elrashidy, N. Sentimentanalysis for Arabic and English datasets. *International Journal of Intelligent Computing and Information Sciences* **2015**, *15*, 55–70.
9. Kumar, A.; Kohail, S.; Ekbal, A.; Biemann, C. IIT-TUDA: System for Sentiment Analysis in Indian Languages Using Lexical Acquisition. In Proceedings of the Mining Intelligence and Knowledge Exploration; Prasath, R.; Vuppala, A.K.; Kathirvalavakumar, T., Eds., Cham, 2015; pp. 684–693.
10. Makrynioti, N.; Vassalos, V. Sentiment Extraction from Tweets: Multilingual Challenges. In Proceedings of the Big Data Analytics and Knowledge Discovery; Madria, S.; Hara, T., Eds., Cham, 2015; pp. 136–148.
11. Povoda, L.; Burget, R.; Dutta, M.K. Sentiment analysis based on Support Vector Machine and Big Data. In Proceedings of the 2016 39th International Conference on Telecommunications and Signal Processing (TSP), 2016, pp. 543–545. https://doi.org/10.1109/TSP.2016.7760939.
12. Bobichev, V.; Kanishcheva, O.; Cherednichenko, O. Sentiment analysis in the Ukrainian and Russian news. In Proceedings of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017, pp. 1050–1055. https://doi.org/10.1109/UKRCON.2017.8100410.
13. Patel, S.; Nolan, B.; Hofmann, M.; Owende, P.; Patel, K. Sentiment analysis: Comparative analysis of multilingual sentiment and opinion classification techniques. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering* **2017**, *11*, 565–571.
14. Tellez, E.S.; Miranda-Jiménez, S.; Graff, M.; Moctezuma, D.; Suárez, R.R.; Siordia, O.S. A simple approach to multilingual polarity classification in Twitter. *Pattern Recognition Letters* **2017**, *94*, 68–74. https://doi.org/https://doi.org/10.1016/j.patrec.2017.05.024.
15. Abo, M.E.M.; Shah, N.A.K.; Balakrishnan, V.; Abdelaziz, A. Sentiment analysis algorithms: evaluation performance of the Arabic and English language. In Proceedings of the 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018, pp. 1–5. https://doi.org/10.1109/ICCCEEE.2018.8515844.
16. Pustulka-Hunt, E.; Hanne, T.; Blumer, E.; Frieder, M. Multilingual sentiment analysis for a swiss gig. In Proceedings of the 2018 6th International Symposium on Computational and Business Intelligence (ISCBI). IEEE, 2018, pp. 94–98.
17. Arun, K.; Srinagesh, A. Multi-lingual Twitter sentiment analysis using machine learning. *International Journal of Electrical & Computer Engineering* **2020**, *10*, 5992–6000. Accessed: 29 July 2025, https://doi.org/10.11591/ijece.v10i6.pp5992-6000.
18. Abubakar, A.I.; Roko, A.; Bui, A.M.; Saidu, I. An Enhanced Feature Acquisition for Sentiment Analysis of English and Hausa Tweets. *International Journal of Advanced Computer Science and Applications* **2021**, *12*.
19. Sumana.; Kanchan, P. Hindi and Kannada Twitter Sentiment Analysis Using Machine Learning Algorithm. In Proceedings of the 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2022, pp. 370–377. https://doi.org/10.1109/ICAC3N56670.2022.10074259.

 doi:10.20944/preprints202601.1816.v1

20. Helmy, A.; Nassar, R.; Ramdan, N. Depression detection for twitter users using sentiment analysis in English and Arabic tweets. *Artificial Intelligence in Medicine* **2024**, *147*, 102716. https://doi.org/https://doi.org/10.1016/j.artmed.2023.102716.
21. Maree, M.; Eleyat, M.; Mesqali, E. Optimizing Machine Learning-based Sentiment Analysis Accuracy in Bilingual Sentences via Preprocessing Techniques. *The International Arab Journal of Information Technology (IAJIT)* **2024**, *21*, 257–270. https://doi.org/10.34028/iajit/21/2/8.
22. Alharbi, Y.; Khan, S.S. Classifying Multi-Lingual Reviews Sentiment Analysis in Arabic and English Languages Using the Stochastic Gradient Descent Model. *Computers, Materials and Continua* **2025**, *83*, 1275–1290. https://doi.org/https://doi.org/10.32604/cmc.2025.061490.
23. Zhou, X.; Wan, X.; Xiao, J. Attention-based LSTM network for cross-lingual sentiment classification. In Proceedings of the Proceedings of the 2016 conference on empirical methods in natural language processing, 2016, pp. 247–256.
24. Becker, W.; Wehrmann, J.; Cagnini, H.E.L.; Barros, R.C. An efficient deep neural architecture for multilingual sentiment analysis in twitter. *Proceedings of the 30th FLAIRS, 2017, Brasil*. **2017**.
25. Deriu, J.; Lucchi, A.; De Luca, V.; Severyn, A.; Müller, S.; Cieliebak, M.; Hofmann, T.; Jaggi, M. Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. In Proceedings of the Proceedings of the 26th International Conference on World Wide Web, Republic and Canton of Geneva, CHE, 2017; WWW '17, p. 1045–1052. https://doi.org/10.1145/3038912.3052611.
26. Medrouk, L.; Pappa, A. Deep Learning Model for Sentiment Analysis in Multi-lingual Corpus. In Proceedings of the Neural Information Processing; Liu, D.; Xie, S.; Li, Y.; Zhao, D.; El-Alfy, E.S.M., Eds., Cham, 2017; pp. 205–212.
27. Can, E.F.; Ezen-Can, A.; Can, F. Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data, 2018, [arXiv:cs.CL/1806.04511].
28. Medrouk, L.; Pappa, A. Do Deep Networks Really Need Complex Modules for Multilingual Sentiment Polarity Detection and Domain Classification? In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–6. https://doi.org/10.1109/IJCNN.2018.8489613.
29. Wehrmann, J.; Becker, W.E.; Barros, R.C. A multi-task neural network for multilingual sentiment classification and language detection on Twitter. In Proceedings of the Proceedings of the 33rd Annual ACM Symposium on Applied Computing, New York, NY, USA, 2018; SAC '18, p. 1805–1812. https://doi.org/10.1145/3167132.3167325.
30. Akhtar, M.S.; Kumar, A.; Ekbal, A.; Biemann, C.; Bhattacharyya, P. Language-agnostic model for aspect-based sentiment analysis. In Proceedings of the Proceedings of the 13th international conference on computational semantics-long papers, 2019, pp. 154–164.
31. Lal, Y.K.; Kumar, V.; Dhar, M.; Shrivastava, M.; Koehn, P. De-Mixing Sentiment from Code-Mixed Text. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop; Alva-Manchego, F.; Choi, E.; Khashabi, D., Eds., Florence, Italy, 2019; pp. 371–377. https://doi.org/10.18653/v1/P19-2052.
32. Chundi, R.; Hulipalled, V.R.; Simha, J. SAEKCS: Sentiment Analysis for English – Kannada Code SwitchText Using Deep Learning Techniques. In Proceedings of the 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 327–331. https://doi.org/10.1109/ICSTCEE49637.2020.9277030.
33. Dahiya, A.; Battan, N.; Shrivastava, M.; Sharma, D.M. Curriculum Learning Strategies for Hindi-English Code-Mixed Sentiment Analysis. In Proceedings of the Artificial Intelligence. IJCAI 2019 International Workshops; El Fallah Seghrouchni, A.; Sarne, D., Eds., Cham, 2020; pp. 177–189.
34. Shakeel, M.H.; Faizullah, S.; Alghamidi, T.; Khan, I. Language Independent Sentiment Analysis. In Proceedings of the 2019 International Conference on Advances in the Emerging Computing Technologies (AECT), 2020, pp. 1–5. https://doi.org/10.1109/AECT47998.2020.9194186.
35. Shen, J.; Liao, X.; Lei, S. Cross-lingual Sentiment Analysis via AAE and BiGRU. In Proceedings of the 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2020, pp. 237–241. https://doi.org/10.1109/IPEC49694.2020.9115134.
36. Younas, A.; Nasim, R.; Ali, S.; Wang, G.; Qi, F. Sentiment Analysis of Code-Mixed Roman Urdu-English Social Media Text using Deep Learning Approaches. In Proceedings of the 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE), 2020, pp. 66–71. https://doi.org/10.1109/CSE50738.2020.00017.

 doi:10.20944/preprints202601.1816.v1

37. Dowlagar, S.; Mamidi, R. Graph Convolutional Networks with Multi-headed Attention for Code-Mixed Sentiment Analysis. In Proceedings of the Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages; Chakravarthi, B.R.; Priyadharshini, R.; Kumar M, A.; Krishnamurthy, P.; Sherly, E., Eds., Kyiv, 2021; pp. 65–72.
38. Mboutayeb, S.; Majda, A.; Nikolov, N.S. Multilingual Sentiment Analysis: A Deep Learning Approach. In Proceedings of the Conference: International Conference On Big Data, Modelling And Machine Learning (BML'21), 2021, pp. 1–7.
39. Barbieri, F.; Anke, L.E.; Camacho-Collados, J. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond, 2022, [arXiv:cs.CL/2104.12250].
40. Chaves, I.C.; Martins, A.D.F.; Praciano, F.D.; Brito, F.T.; Monteiro, J.M.; Machado, J.C. BPA: A Multilingual Sentiment Analysis Approach based on BiLSTM. In Proceedings of the ICEIS (1), 2022, pp. 553–560.
41. Suresh Kumar, K.; Helen Sulochana, C. Local search five-element cycle optimized reLU-BiLSTM for multilingual aspect-based text classification. *Concurrency and Computation: Practice and Experience* **2022**, *34*, e7374.
42. Yang, Q.; Kadeer, Z.; Gu, W.; Sun, W.; Wumaier, A. Affective Knowledge Augmented Interactive Graph Convolutional Network for Chinese-Oriented Aspect-Based Sentiment Analysis. *IEEE Access* **2022**, *10*, 130686–130698. https://doi.org/10.1109/ACCESS.2022.3228299.
43. Astuti, L.W.; Sari, Y.; Suprapto. Code-mixed sentiment analysis using transformer for twitter social media data. *International Journal of Advanced Computer Science and Applications* **2023**, *14*.
44. Choudhary, N.; Singh, R.; Bindlish, I.; Shrivastava, M. Sentiment Analysis of Code-Mixed Languages Leveraging Resource Rich Languages. In Proceedings of the Computational Linguistics and Intelligent Text Processing; Gelbukh, A., Ed., Cham, 2023; pp. 104–114.
45. Deole, Y.; Jagadish, T.; Shabrez, M.; M, P.; G, A. Multi-Lingual Sentiment Analysis of Urdu and English Tweets Using RNN with Bidirectional LSTM. In Proceedings of the 2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC), 2023, pp. 813–817. https://doi.org/10.1109/ICRTAC59277.2023.10480788.
46. Kar, P.; Debbarma, S. Sentimental analysis & Hate speech detection on English and German text collected from social media platforms using optimal feature extraction and hybrid diagonal gated recurrent neural network. *Engineering Applications of Artificial Intelligence* **2023**, *126*, 107143. https://doi.org/https://doi.org/10.1016/j.engappai.2023.107143.
47. Mercha, E.M.; Benbrahim, H.; Erradi, M. SlideGCN: Slightly Deep Graph Convolutional Network for Multilingual Sentiment Analysis. In Proceedings of the Advances in Computational Intelligence; Rojas, I.; Joya, G.; Catala, A., Eds., Cham, 2023; pp. 91–103.
48. Omran, T.M.; Sharef, B.T.; Grosan, C.; Li, Y. Transfer learning and sentiment analysis of Bahraini dialects sequential text data using multilingual deep learning approach. *Data & Knowledge Engineering* **2023**, *143*, 102106. https://doi.org/https://doi.org/10.1016/j.datak.2022.102106.
49. Gupta, N.; Bhattarai, M.; Chavan, A. Multilingual Sentiment Analysis using DistilBert. In Proceedings of the 2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA), 2024, pp. 1–6. https://doi.org/10.1109/ICAIQSA64000.2024.10882364.
50. Kathunia, A.; Kaif, M.; Arora, N.; Narotam, N. Sentiment Analysis Across Languages: Evaluation Before and After Machine Translation to English, 2024, [arXiv:cs.CL/2405.02887].
51. Mamta.; Ekbal, A. Transformer based multilingual joint learning framework for code-mixed and english sentiment analysis. *Journal of Intelligent Information Systems* **2024**, *62*, 231–253.
52. Yacoub, A.D.; Aboutabl, A.E.; Slim, S.O. Multilingual Sarcasm Detection for Enhancing Sentiment Analysis using Deep Learning Algorithms. *Journal of Communications Software and Systems* **2024**, *20*, 278–289. https://doi.org/10.24138/jcomss-2024-0071.
53. Jain, V.; Malviya, L.; . S, A. Optimized hybrid deep learning for cross-linguistic sentiment analysis: a novel approach. *Journal of Cloud Computing* **2025**, *14*, 30.
54. Jain, V.; Mohanani, G.; Gaur, A.; Patheja, P.S. Enhanced Multilingual Sentiment Analysis Using Hybrid CNN-GRU Deep Learning Architecture. In Proceedings of the Advanced Network Technologies and Intelligent Computing; Verma, A.; Verma, P.; Pattanaik, K.K.; Buyya, R.; Dasgupta, D., Eds., Cham, 2025; pp. 357–370.
55. Nazir, M.K.; Faisal, C.N.; Habib, M.A.; Ahmad, H. Leveraging Multilingual Transformer for Multiclass Sentiment Analysis in Code-Mixed Data of Low-Resource Languages. *IEEE Access* **2025**, *13*, 7538–7554. https://doi.org/10.1109/ACCESS.2025.3527710.

 doi:10.20944/preprints202601.1816.v1

56. Jhanwar, M.G.; Das, A. An Ensemble Model for Sentiment Analysis of Hindi-English Code-Mixed Data, 2018, [arXiv:cs.CL/1806.04450].
57. Zhang, X.; Zhang, C.; Shi, H. Ensemble of Binary Classification for the Emotion Detection in Code-Switching Text. In Proceedings of the Natural Language Processing and Chinese Computing; Zhang, M.; Ng, V.; Zhao, D.; Li, S.; Zan, H., Eds., Cham, 2018; pp. 178–189.
58. Sarkar, K. Heterogeneous classifier ensemble for sentiment analysis of Bengali and Hindi tweets. *Sādhanā* **2020**, *45*, 196.
59. Al Shamsi, A.A.; Abdallah, S. Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects. *Journal of King Saud University - Computer and Information Sciences* **2023**, *35*, 101691. https://doi.org/https://doi.org/10.1016/j.jksuci.2023.101691.
60. Aryal, S.K.; Prioleau, H.; Washington, G.; Burge, L. Evaluating Ensembled Transformers for Multilingual Code-Switched Sentiment Analysis. In Proceedings of the 2023 International Conference on Computational Science and Computational Intelligence (CSCI), 2023, pp. 165–173. https://doi.org/10.1109/CSCI62032.2023.00032.
61. Habbat, N.; Nouri, H.; Anoun, H.; Hassouni, L. Sentiment analysis of imbalanced datasets using BERT and ensemble stacking for deep learning. *Engineering Applications of Artificial Intelligence* **2023**, *126*, 106999. https://doi.org/https://doi.org/10.1016/j.engappai.2023.106999.
62. Choudhary, C.; Thakur, J.; Singh, M.; Anurag. Sentiment Analysis in Code-Mixed Video Comments Using Machine Learning on a Hinglish Dataset. In Proceedings of the 2024 First International Conference on Software, Systems and Information Technology (SSITCON), 2024, pp. 1–6. https://doi.org/10.1109/SSITCON62437.2024.10796129.
63. Hasan, M.A. Ensemble Language Models for Multilingual Sentiment Analysis, 2024, [arXiv:cs.CL/2403.06060].
64. Hassan, M.E.; Maab, I.; Hussain, M.; Habib, U.; Matsuo, Y. Polarity Classification of Low Resource Roman Urdu and Movie Reviews Sentiments Using Machine Learning-Based Ensemble Approaches. *IEEE Open Journal of the Computer Society* **2024**, *5*, 599–611. https://doi.org/10.1109/OJCS.2024.3476378.
65. Miah, M.S.U.; Kabir, M.M.; Sarwar, T.B.; Safran, M.; Alfarhood, S.; Mridha, M. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports* **2024**, *14*, 9603.
66. Dharini, N.; Madhuvanthi, M.; Aswini, C.S.; Lakshya, R.; Triumbika, M.; Saranya, N. Ensemble-Driven Multilingual Sentiment Analysis Framework for YouTube Comments with Dashboard. In Proceedings of the 2025 International Conference on Visual Analytics and Data Visualization (ICVADV), 2025, pp. 1524–1529. https://doi.org/10.1109/ICVADV63329.2025.10961107.
67. Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In Proceedings of the Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); Bethard, S.; Carpuat, M.; Cer, D.; Jurgens, D.; Nakov, P.; Zesch, T., Eds., San Diego, California, 2016; pp. 19–30. https://doi.org/10.18653/v1/S16-1002.
68. Keung, P.; Lu, Y.; Szarvas, G.; Smith, N.A. The Multilingual Amazon Reviews Corpus, 2020, [arXiv:cs.CL/2010.02573].
69. Rosenthal, S.; Farra, N.; Nakov, P. SemEval-2017 Task 4: Sentiment Analysis in Twitter, 2019, [arXiv:cs.CL/1912.00741].
70. Joshi, A.; Prabhu, A.; Shrivastava, M.; Varma, V. Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text. In Proceedings of the Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers; Matsumoto, Y.; Prasad, R., Eds., Osaka, Japan, 2016; pp. 2482–2491.
71. Patwa, P.; Aguilar, G.; Kar, S.; Pandey, S.; PYKL, S.; Gambäck, B.; Chakraborty, T.; Solorio, T.; Das, A. SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets, 2020, [arXiv:cs.CL/2008.04277].
72. Patra, B.G.; Das, D.; Das, A.; Prasath, R. Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview. In Proceedings of the Mining Intelligence and Knowledge Exploration; Prasath, R.; Vuppala, A.K.; Kathirvalavakumar, T., Eds., Cham, 2015; pp. 650–655.
73. Padmaja, S.; Fatima, S.S. Opinion mining and sentiment analysis-an assessment of peoples' belief: A survey. *International Journal of Ad hoc, Sensor & Ubiquitous Computing* **2013**, *4*, 21.

 doi:10.20944/preprints202601.1816.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.