Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Sentiment analysis: a comparison of deep learning neural network algorithms with ensemble learning algorithms

Eslam Ashraf Elhadidi[1], Ahmad Salah[2], Marwa abdellah[1], Saad M. Darwish[3],
[1]*Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Egypt*
[2]*College of Computing and Information Sciences, University of Technology and Applied Science, Ibri, Oman*
[3]*Department of Information Technology, Institute of Graduate Studies and Research, Alexandria University, Alexandria, 21526, Egypt*
e.elhadedy020@fci.zu.edu.eg, ahmad.salah@utas.edu.om, MMAboelazm@fci.zu.edu.eg, saad.darwish@alexu.edu.eg

ARTICLE INFO

Received: 25 Dec 2024

Revised: 09 Feb 2025

Accepted: 22 Feb 2025

ABSTRACT

Sentiment analysis presents difficulties, especially in multi-language contexts, when dealing with the meanings based on context, mockery, and complex language. In contemporary NLP applications, the exact classification is necessary for researchers and institutions to obtain valuable visions and increase customer participation and improve decisions. Our method improves the performance of Sentiment analysis by using many advanced deep learning models, such as XLNET, Roberta and Bert. In order to capture a variety of language patterns and increase classification accuracy, we plan to integrate these models using ensemble techniques such as stacking, bagging, and boosting. We applied this approach to the dataset [cardiffnlp/tweetsentimentmultilingual], ensuring a comprehensive evaluation of each model performance in addition to the effectiveness of the group as a whole. The results show that our collection method works better than independent models, where you get F1 degrees and higher accuracy in various sentiment classes. Mixing predictions from various models often uncovers surprising insights. In most cases, ensemble learning really shines when tackling today's NLP challenges and reliably nails sentiment classification even if things sometimes get a bit messy. We took a deep dive into a range of neural network models designed for sentiment analysis, checking out both their solo performance and what happens when they team up. Stacking, bagging, and boosting were thrown into the mix to craft a fresh, hybrid method that bumps up accuracy noticeably. Compared to models running on their own, our approach usually scores higher on the key measures, proving that joining forces is a pretty solid trick for sorting sentiments.

Keywords: Sentiment analysis, XLNet, RoBERTa, BERT, stacking, bagging, boosting.

INTRODUCTION

Sentiment analysis is the process of extracting and classifying words into positive, negative, or neutral categories utilizing text analysis, statistics, and natural language processing [15], [18],[5]. SA is important in social media, customers and market for social media by Detecting and extracting human subjective evaluations of things posted on social media by computational means is known as social media sentiment analysis. Sentiment analysis on social media has encom passed network relationships, interactions, multi-modal texts, temporal dynamics, and sentiment transmission. Additionally, certain sentiment intensity and feelings are identified [12],[2], for Customer is a data processing method that analyzes and deciphers the thoughts, feelings, and attitudes of customers based on their verbal or written comments [1]. Businesses can better understand and address consumer impressions thanks to this analysis, which converts subjective into actionable data [1] , and for markert research strategy and brand reputation management in the digital era wants to provide a workable solution [6].

An informal writing style, sarcasm, irony, and language-specific difficulties are some of the problems that are linked to sentiment analysis and natural language processing. Many words in various languages have meanings and orientations that vary based on the context and field in which they are used. As a result, not many tools and resources are available for every language. Irony and sarcasm are two of the most important issues that researchers have

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

recently focused on [17]. Powerful architectures like transformer-based models with hybrid techniques combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been used in recent developments in deep learning models for sen timent analysis to increase the accuracy of sentiment classification.The ability to detect sentiment has improved, particularly in social media and product eval uations, thanks to large language models (LLMs) and multimodal sentiment analysis algorithms that integrate audio,textual,and visual data. Furthermore, sentiment analysis in healthcare, e-commerce and finance has improved as a result of fine-tuning previously trained models using domain-specific datasets [8]. Although the sentiment classification models that are now in use perform well on their own, they frequently suffer from class imbalance and misclassification in situations that are unclear. Even while transformer-based models produce excellent outcomes, nothing is known about how well they perform in comparison across various datasets and sentiment classes. Furthermore, en semble learning strategies that might capitalize on the advantages of several models for better generalization are rarely the subject of current research. This emphasizes that in order to get more reliable and cutting-edge sentiment catego rization, a thorough comparison of models and an optimized ensemble technique are required. Our method offers a thorough analysis of several transformer-based models, such as XLM-RoBERTa, LaBSE, MPNet, XLM-R-dist, mBERT, and mUSEdist to determine how well they perform in multilingual sentiment classification. We provide a new ensemble approach that maximizes model contributions and improves overall predictive accuracy by utilizing stacking, bagging, and boosting (XGBoost). The influence of these ensemble approaches in enhancing sentiment classification ability is illustrated by our empirical findings on the CardiffNLP Tweet Sentiment Multilingual dataset. By using this method, we surpass individual models and demonstrate the advantages of ensemble learning for reliable multilingual sentiment analysis, resulting in improved F1-scores and accuracy. This is how the remainder of the paper is organized: discusses earlier methods for ensemble learning and multilingual sentiment analysis in this review of relevant work. describes the dataset in detail, including its properties and the preprocessing procedures. explains the training setups and topologies of the several transformer-based models that were employed in our investigation. discusses our ensemble learning strategy, including the XGBoost, bagging, and stacking techniques. describes the experimental setup, including the software, hardware, and measurements for evaluation. compares individual models with ensemble techniques, examines error patterns, then presents and discusses the findings. provides important findings, contributions, and suggestions for further research at the end of the work.

RELATED WORK

ADVANCED DEEP LEARNING MODELS FOR SENTIMENT ANAL YSIS

The BERT model (Bidirectional Encoder Representations from Transformers) is used to analyze sentiment analysis to address the challenges associated with addressing the complex emotional text. The way to take advantage of the BERT bipped adapter structure and integrate it with deep learning for augmented contextual understanding [19].

The study evaluates the performance of the DistilBERT model against that of more traditional word vector models like FastText, Word2Vec, and GloVe using the SST2 dataset (Stanford Sentiment Treebank). The results indicate that DistilBERT outperforms traditional methods in sentiment classification, with considerable accuracy gains, particularly when modified. However, there are restrictions, such as the need for extensive accurate control and the possibility of overcoming smaller data groups, which confirms the importance of allocating the methods of willingness and persuasion to specific work burdens. The results show the possibilities of BERT models in increasing the accuracy and effectiveness of feelings analysis in a variety of applications [19].

The Robustly Optimized BERT Pretraining Approach (RoBERTa) extends the BERT architecture to get around its drawbacks and improve natural language processing skills. The use of a larger dataset (160 GB compared to BERT's 16 GB), dynamic masking, the removal of the Next Sentence Prediction (NSP) objective, longer tra ining sequences, and improved hyperparameters make RoBERTa more robust and flexible for language compre hension tasks [14].

Solving problems in mental health care, including barriers that prevent personal care, early detection, and treatment methods that can be easily accessible using traditional methods that are hindered by shame stigma, lack of financing, and cost [14]. Using NLP models such as Roberta and BERT, mental health care revolution is done by analyzing linguistic data by providing accurate diagnosis, treatment recommendations and actual time support [14]. Gathering and preparing different types of textual data, optimizing RoBERTa and BERT for mental health applications,

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

developing diagnostic tools, creating algorithms for personalized treatment, and deploying virtual platforms and therapeutic chatbots are experimented in [14]. The study found that RoBERTa has a testing accuracy of 99.88%, outperforming both BERT (99.73%) and LSTM (99.65%) [14].

Significant progress in the treatment of natural language is the modeling of the flipping Permutation Language Modeling (PLM), a new component of XLNET that improves in the traditional models of automatic slope such as GPT [10]. Unlike the traditional models that process the text in a fixed order from left to right or right to the left, XLNET maintains the benefits of automatic training while allowing the mixture of the two -way context by producing the exchange of input serials [10]. Two important problems in NLP: Automatic decline models to accommodate dual direction and prior contrast to training in convincing language models such as BERT [10]. They are looking into techniques including band learning, adapting to the field, distracting knowledge, parameter, and enlarging data that can help XLNET to perform better [10]. Their experimental results validate the superiority of XLNet with state-of-the-art performance on tasks like as text classification, question answering, and summarization. However, the model does have certain shortcomings such as its computational complexity, low robustness on out-of-domain data, and susceptibility to adversarial attacks. Although XLNet has good generalization and job flexibility, its implementation in resource-constrained contexts still must be improved [10].

ENSEMBLE METHODS IN SENTIMENT ANALYSIS

To increase accuracy in classification by combining predictions from many different models, sentiment analysis (SA) has heavily relied on traditional ensemble methods such as bagging (e.g., Random Forest) and boosting (e.g., Ad aBoost, XGBoost, Gradient Boosting) [16]. Nevertheless, there is disagreement regarding how well these techniques perform in comparison [16]. Treating huge amounts of irregular text, enhancing classification performance, addressing issues such as detection of mockery, negative treatment, and field relying on some basic concerns in SA that are addressed in this work [16]. Eight groups models are analyzed using measures such as accuracy, f-score, and operating time across four reference datasets [16]. These models include bagging-based techniques (Random Forest, Extra-Trees, Bagging) and boosting-based techniques (AdaBoost, Gradient Boosting, XGBoost, CatBoost, LightGBM). The results demonstrate that bagging techniques, specifically Random Forest, outperform boosting strategies in terms of accuracy and runtime economy, while boosting strategies hold up well with complex features and unbalanced datasets [16]. The vulnerability of the study, owing to static data sets, absence of real data assessment, limited action for fine-grained emotion recognition, and abandonment of advanced neural techniques such as transformers potentially leading to superior performance or generalization, stands in stark contrast to the advancements taking place [16]. In many domains, such as natural language processing, image classification, and medical imaging, deep learning ensemble techniques are gaining popularity as a way to improve prediction performance [16].

Deep learning extends these techniques, which include bagging, boosting, and stacking utilizing a variety of topologies, hyperparameter tuning, and advanced fusion methods like meta-learning and voting [11]. The paper cited the limitations of current ensemble techniques like being overly reliant on simple majority voting, uniformity of baseline models, and lack of generalization. It offers a taxonomy for dividing ensemble methods according to training paradigms (sequential or parallel), data sampling methods, and combination algorithms and sheds light on how to properly combine various deep learning models [11]. The result shows that, in spite of a couple of drawbacks in the form of processing cost time-consuming hyperparameter optimization, and the danger of overfitting on short datasets, ensemble deep learning usually outperforms individual models in terms of accuracy, particularly when combined with sophisticated tactics such as stacking and boosting [11]. They emphasize the necessity for more innovative and computationally efficient methods for truly tapping into the possibilities offered by ensemble deep learning [11].

In NLP, hybrid ensemble approaches mix different deep learning architectures (such as CNNs, RNNs, and Transformers) or deep learning with conventional machine learning models to enhance performance [7]. Such core issues like overfitting, bad generalization, taking care of very long-range dependencies, computational inefficiencies, and sparsity of data are solved with such methods [7]. Suggested methodologies include machine translation Transformer-based ensembles, CNN-LSTM hybrids for sentiment analysis, and BiLSTM-CRF combined with BERT for named entity recognition [7]. Even big ensembles such as FrugalGPT and LLM-Blender help achieve efficiency

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

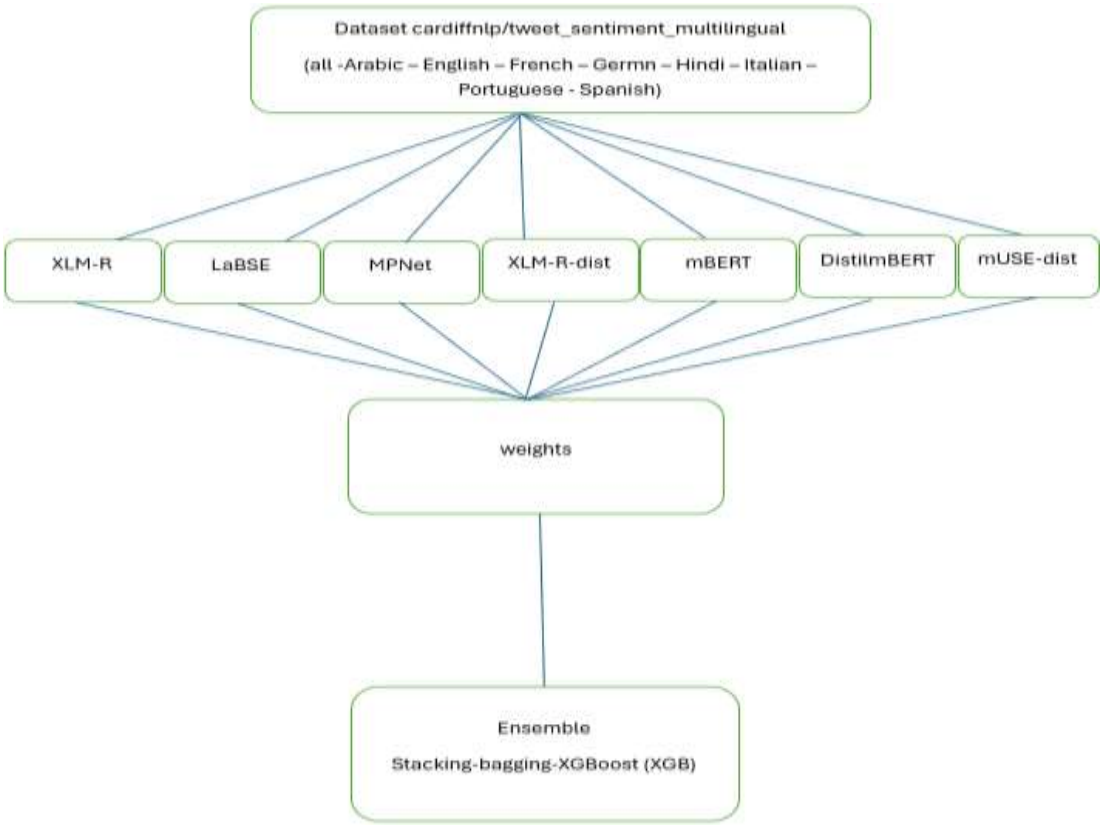
Research Article

and cost [7]. The outcome is more accurate and stable results compared to standalone models for a range of NLP tasks. High computational costs, complexity of deployment, decreased interpretability, risks of overfitting, and latency issues are some of the drawbacks. Regardless of these shortcomings, hybrid ensemble techniques provide great advances in NLP by strategically combining models to sacrifice efficiency for effectiveness [7].

METHODOLOGY

MODEL ARCHITECTURE

Figure 1: The architecture of the ensemble learning model



The foundation model of the research is a transformer-based architecture specifically built for sequence classification tasks, sentiment analysis. It uses pre-trained models from the Hugging Face Transformers package, XLM-R, LaBSE, MPNet, XLM-R-dist, mBERT, DistilmBERT and mUSE-dist, optimized on multilingual datasets. To represent intricate context relationships between the words in a phrase, regardless of their word order, they use a deep learning model which utilizes self-attention mechanisms. These transformer-based models possess the primary advantage of being able to process sequences of input in parallel without maintaining context in longer pieces of text and thus being effective for sentiment classification tasks as well as other natural language tasks. Three categories of sentiment negative, neutral, and positive are represented by the classification head of every base model, and that too is accompanied by an embedding layer as well as multiple self-attention layers. The self-attention layers process the dense vector representations created by the embedding layer to capture semantic dependencies. Class probabilities are obtained by passing the output from these layers through a fully connected layer with a softmax activation function. The model is able to generalize well to the subtleties of the sentiment analysis task in other languages by fine-tuning its internal parameters using the task-specific data. With the learning rate set to 5e-5, Adam optimizer and Sparse Categorical Crossentropy loss are utilized to fine-tune the base models for retaining model stability and convergence speed. To ensure stable input sizes across batches, the input sequences are either padded or chopped when the token size is limited to 128. On the training phase, early stopping and dropout layers are both integrated to ensure generalization and also avoid overfitting. All base models learn general language

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

patterns as well as particular sentiment markers by fine-tuning the pre-trained models on the multilingual sentiment data. These are then integrated in the ensemble learning process to improve overall classification performance.

In real terms, raw data text is transformed into something interpretable, organized, and ready for modeling in data preparation. The most commonly deployed preprocessing pipeline components include normalizing text (lowercasing, stripping special characters, and resolving contractions), stripping stopwords, and tokenization. Preprocessing in sentiment analysis includes handling imbalanced classes by oversampling or undersampling and consistency across the models involved. Models are therefore ensured to train efficiently and generalize to new input well. As a result, models are guaranteed to learn efficiently and generalize well to new data. Transformer-based models such as XLM-R, LaBSE, MPNet, DistilBERT, mBERT, and XtremeDistil implement tokenization, which is the process of dividing the text into smaller pieces (tokens) for machine learning models to understand, using subword tokenization techniques such as WordPiece and Byte Pair Encoding (BPE). Input text tokenization utilizing the Hugging Face transformers library's Auto Tokenizer preserves model consistency. For efficient usage of memory as well as preservation of contextual information, tokenization is followed by truncation and padding to a fixed length (128 tokens for this example). Then tokenized text is converted into tensors to enable effective batch processing while training deep learning models.

Training is an improvement process of transformer-based multilingual sentiment analysis models. The models are initialized with a learning rate of 5e-5, optimized by the Adam optimizer, and the Sparse Categorical Cross entropy loss function was used for training. 128 tokens are set as the maximum length for the sequence to ensure that it's processed efficiently, thus keeping a very important context. Batch size is chosen as 16 for training and 64 for testing, and every model is trained for five epochs using shuffled data to maximize generalization. Early stopping isn't specifically employed but monitoring the performance on the validation set to prevent overfitting of models is done. The models' weights are also stored after training to make reloading easier for ensemble learning. In order to provide a balanced distribution of sentiment, 80% of the dataset is used for training and 20% for testing. Before combining their outputs in an ensemble approach, the models (XLM-R, LaBSE, MPNet, DistilBERT, and mBERT) are all individually tweaked. Both macro and micro F1 ratings are used to evaluate the models' performance, taking into consideration class imbalances. The models forecast sentiment in three different classes: neutral, positive, and negative.

ENSEMBLE STRATEGY

An ensemble technique called stacking ensemble teaches a new model how to combine predictions from several different models as effectively as feasible. It integrates forecasts from several distinct training models [13]. Bootstrap aggregation, another name for bagging, is a well-liked ensemble technique that helps to lower excessive variance. Some observations in the resulting data sets can be replicated since the bagging technique adopts a sampling with replacement strategy in an attempt to construct several data sets from the initial training data set [9]. XGB is regarded as a distributed gradient boosting library that is designed to be extremely effective. This cross-purpose library is dedicated to solving problems of natural language processing (NLP), like those relating to classification, regression, and ranking. There are some typical reasons why XGB is used[3]. This approach tunes different models, like XLM-RoBERTa, LaBSE, mBERT, and MPNet, separately to capture various patterns and subtleties in the data. In comparison to any individual base model, such a layered model allows the ensemble to average various models' bias and variances, improving the generalization ability and overall performance.

The predictions from several transformer-based models are combined to maximize the ensemble weights. Probability distributions across the sentiment classes (positive, neutral, and negative) are extracted for every model. To make sure that each model contributes proportionately to its predictive performance, these forecasts are combined, and the weights are tuned using a softmax-based averaging technique. Using the macro F1-score as the main measure to address possible class imbalances, the optimization procedure iteratively modifies the weights in order to minimize the validation loss. The ensemble optimizes over all accuracy and generalization across multilingual sentiment analysis tasks by giving better-performing models higher weights and preserving stability with random seeds.

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

EXPERIMENTAL SETUP

DATASET

The code makes use of the CardiffNLP Tweet Sentiment Multilingual Dataset, which is intended for sentiment analysis (SA) in many languages. This dataset can be used for cross-lingual sentiment classification tasks because it is a component of a vast collection of tweets that have been classified for sentiment in different languages. The dataset has text the tweet in it and three sentiment labels 0: negative, 1: neutral and 2: Positive [4]. The dataset's attributes include Twitter, social media that is the domain of data, many language, Three categories negative, neutral, and positive are applied to tweets distribution of sentiment, thousands of tweets per language, the size varies depending on the language subset and More neutral or positive tweets than negative ones may have unbalanced sentiment labels which is imbalance [4]. Steps in the Pre processing Pipeline makes sure that the unprocessed twitter data is organized and prepared for model training. The steps are Using Hugging Face's datasets package, the dataset is loaded and transformed into Pandas DataFrames for simple processing, Train-Test Split maintains class distribution through stratified sampling,Text Tokenization that Tweets are tokenized into subword tokens using the Hugging Face AutoTokenizer, which provides padding and truncation (max length = 128) , Data formatting Using tf.data the tokenized outputs are transformed into TensorFlow datasets and shuffled batch dataset for training and Label Encoding that sparse categorical cross-entropy loss function is compatible with sentiment labels, which are encoded as integers (0, 1, 2)[4]. Using (train-test-split) from sklearn, the dataset is first split into 80 for training and 20 for testing, with a fixed random seed (SEED = 42) for reproducibility. During model training, the training set is further divided into training and validation datasets, with 10 of the training data automatically allocated for validation through the *model.fit()* function, resulting in roughly 72 of the total data for training, 8 for validation, and 20 for testing. This is done to ensure reliable model evaluation and avoid overfitting. During training, a validation set is in use to track model performance and modify weights, while the test set as sesses how well the model generalizes to new data. Throughout the training, validation, and testing stages, all splits preserve the initial class distribution (stratified sampling) to provide balanced sentiment representation.

IMPLEMENTATION DETAILS

GPU-accelerated environment was used for the training procedure in order to take advantage of TensorFlow's and transformer-based models' computational performance. The hardware configuration featured a multi-core CPU for data preprocessing duties and an NVIDIA GPU (such as the Tesla T4) for quicker model training. TensorFlow 2.x, Scikit Learn, Pandas, Hugging Face Transformers, and Python 3.10 constituted the tech stack. Training has been performed with the Adam optimizer learning rate as 5e-5 with batch size for training as 16 and evaluation as 64. To avoid overfitting, each model was trained for five epochs before being stopped early depending on validation results. A random seed (42) was assigned in libraries in favor of cloning. Definition of measures: The accuracy of the model is calculated as the number of predictions divided into the total number of predictions. The accuracy of unbalanced data can be deceptive even when they are intuitively attractive. Precision: Measures the proportion of all the positive predictions that are actually positive. In scenarios in which false positives are very costly, higher precision means lower false positives, which is very crucial. Recall (sensitivity): Measures the degree to which the model will identify all the cases that are pertinent by the proportion of true positives to the number of actual positives. Because missing positive examples is very expensive, high recall is crucial. F1-Score: Precision and recall are compromised by finding the harmonic mean of the two metrics. It is especially useful when working with unbalanced datasets. In contrast to micro F1-score, which compromises the metric by class frequency, macro F1-score averages it uniformly across classes.

RESULTS AND DISCUSSION

INDIVIDUAL MODEL PERFORMANCE

language	XLM-R(%)	LaBSE(%)	MPNet(%)	XLM-R-dist(%)	mBERT(%)	DistilmBERT(%)	mUSE-dist(%)
All	33	62	33	54	34	56	52
Arabic	36	61	31	37	36	56	42

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

English	36	64	62	67	56	58	58
French	36	78	31	61	65	62	64
Germn	33	73	31	65	65	65	62
Hindi	33	33	31	37	36	31	34
Italian	57	63	31	52	55	57	54
Portuguese	33	64	31	54	53	55	55
Spanish	36	59	31	55	36	50	52

Table 1: Train Accurecy

language	XLM-R(%)	LaBSE(%)	MPNet(%)	XLM-R-dist(%)	mBERT(%)	DistilmBERT(%)	mUSE-dist(%)
All	33	34	33	53	34	57	50
Arabic	31	63	31	44	47	48	57
English	35	68	36	70	60	59	53
French	52	73	36	62	71	70	65
Germn	49	68	36	63	67	66	65
Hindi	33	33	36	36	33	31	34
Italian	31	67	36	51	61	56	55
Portuguese	33	63	36	55	53	55	55
Spanish	33	64	36	55	43	56	46

Table 2: Test Accurecy

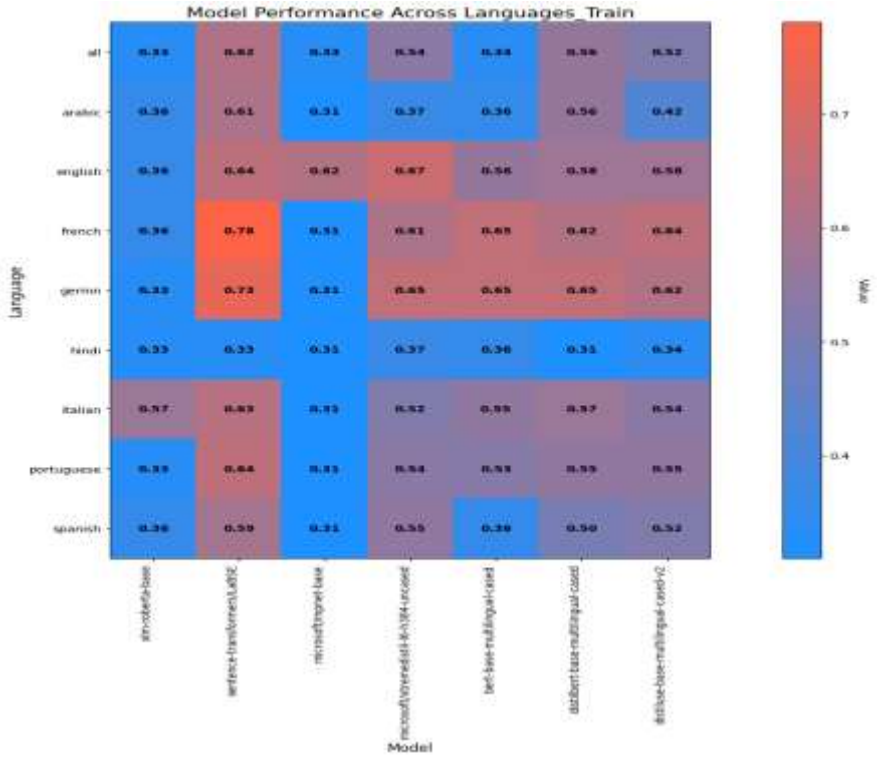


Figure 2: Confusion matrix for models' accuracy on training data across different languages.

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

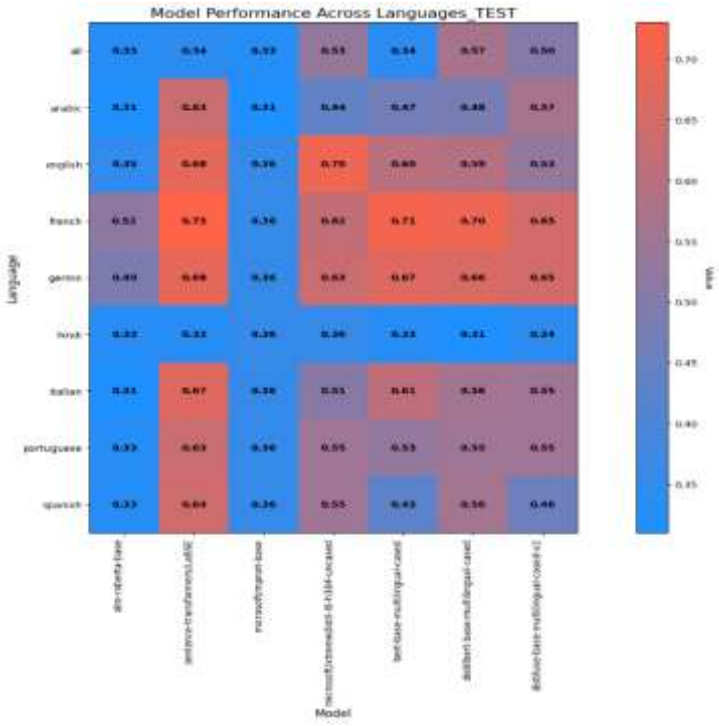


Figure 3: Confusion matrix for models' accuracy on test data across different languages.

Comparison evaluation of performance between different models shows its advantages and limitations towards performing the task of multilingual sentiment analysis. Various measures like F1-score, recall, accuracy, and precision were used in evaluation for several languages. Not repeatedly were XLM-RoBERTa (xlm-roberta-base) and MPNet having a consistently high Accurecy. The model remained exposed for generalizability in other languages. Particularly for neutral sentiment analysis, LaBSE obtained competitive F1-scores due to its superior performance in semantic similarity capture. Better-performing models such as XtremeDistil, mBERT, and mUSE dist were effective for ensemble methods Models introduced efficiency and diversity. By studying the error patterns of individual models, some persistent misclassifications and challenges were identified. The XLM-RoBERTa often confused neutral information with sentiment; it struggled to capture implicit sentiment, especially when sarcasm or undercurrents were involved in tweets. LaBSE performed well semantically; however, it faltered on mixed-sentence tweets, confusing the subjective and neutral sentiments. MPNet produced spurious positives, especially in the identification of negative sentiment, due to its sensitivity towards complex sentence structure and negations. While the light XtremeDistil model was helpful, it often lacked strong contextual cues, resulting in improved misclassification against negative and neutral tweets. The capacity of mBERT in properly identifying sentiment in localized contexts was impaired by its inability to understand idioms, slang, and culturally relevant mentions. Just like mUSE-dist had been challenged by colloquial language, emojis, and code-switching, which resulted in misunderstandings, DistilmBERT was challenged by short, but succinct tweets and tended to under-estimate positive sentiment.

ENSEMBLE PERFORMANCE

language	LaBSE(%)	XLM-R-dist(%)	mBERT(%)	DistilmBERT(%)	mUSE-dist(%)	Stack(%)	Bagg(%)	XGB(%)
Arabic	61	37	36	56	42	62	54	64
English	64	67	56	58	58	67	52	67
French	78	61	65	62	64	76	60	74
Germn	73	65	65	65	62	70	57	69
Italian	63	52	55	57	54	61	59	68
Portuguese	64	54	53	55	55	61	56	61

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Spanish	59	55	36	50	52	60	49	61

Table 3: train ensemble

language	LaBSE(%)	XLM-R-dist(%)	mBERT(%)	DistilmBERT(%)	mUSE-dist(%)	Stack(%)	Bagg(%)	XGB(%)
Arabic	63	44	47	48	57	63	54	64
English	68	70	60	59	53	69	52	68
French	73	62	71	70	65	73	60	76
Germn	68	63	67	66	65	75	57	73
Italian	67	51	61	56	55	68	59	71
Portuguese	63	55	53	55	55	68	56	64
Spanish	64	55	43	56	46	55	49	61

Table 4: test ensemble

Ensemble approach outcompeted individual models. While a single model like LaBSE or XLM-R-dist were exceptionally good in picking up deep semantic relationships, at times these individual models lost grip over certain linguistic subtleties like sarcasm or idiom. The variations were bridged by ensemble approach, as this combined prediction via weighted averaging and provided more stable and reliable sentiment classification. Specifically, in the case of underrepresented classes like negative emotions, for which individual models tended to perform worse, the ensemble yielded improved F1-scores and recall. Aside from generalization across heterogeneous tweet structures and sentiment nuances, combining the views of multiple models alleviated model-specific bias. All else being equal, the ensemble performed better than single models, yielding more accurate and predictable results, especially when addressing uncertain or contextually rich tweets. Performance optimization and error reduction were affected differently by the varying combinations of XGBoost (XGB), stacking, and bagging. Stacking optimized accuracy and F1-scores by combining predictions of several base models into a meta learner and thereby leveraging their strengths. Diverse decision patterns from models such as LaBSE, XLM-R-dist , mBERT , DistilmBERT and mUSE-dist were success fully captured by this method. Model stability was not enhanced by bagging (Bootstrap Aggregating), training base models on subsets of data and averaging their predictions. Consistently strong performance was achieved by XGBoost (XGB), a meta-classifier applied to the base model probability outputs, particularly in the presence of complex data patterns and class imbalances. Its inherent regularization features prevented overfitting and tended to yield superior F1-scores to bagging.

COMPARATIVE ANALYSIS

language	[4]	Our Work	Proposed Model
Arabic	45.98%	64%	XGB
English	50.58%	70%	microsoft/xtremedistil-l6-h384-uncased
French	54.82%	76%	XGB
Germn	59.56%	75%	Stacking
Hindi	37.08%	36%	mpnet-base
Italian	54.65%	71%	XGB
Portuguese	55.05%	68%	Stacking
Spanish	50.06%	64%	sentence-transformers/LaBSE

Table 5: Proposed methods comparison against the state-of-the-art (reference [4]) on the F1-score.

Table 5 displays the results of state-of-the-art and our work in 8 languages: Arabic, English, French, German, Hindi, Italian, Portuguese, and Spanish. In Arabic, our paper is better achieving 64% vs 45.98% with the proposed model XGB. In English, our paper is better achieving 70% vs 50.85% with the proposed model microsoft/xtremedistil-l6-h384-uncased. In French, our paper is better achieving 76% vs 54.82% with the proposed model XGB. In German, our paper is better achieving 75% vs 59.56% with the proposed model Stacking. In Hindi, our paper is not better achieving 36% vs 37.08% with the proposed model mpnet-base. In Italian, our paper is better achieving 71% vs

54.65% with the proposed model XGB. In Portuguese, our paper is better achieving 68% vs 55.5% with the proposed model Stacking. In Spanish, our paper is better achieving 64% vs 50.06% with the proposed model sentence-transformers/LaBSE. We offer eight languages, train each model, and build an ensemble for the top three models that achieve an F1 score of 56% or higher in each language and model. We discovered that the Indian language is less than 56% in all models in both test and train. The results are displayed in Tables 1 and 3.

CONCLUSION AND FUTURE WORK

On a sentiment analysis dataset, our method trains many state-of-the-art deep learning models, such as BERT, RoBERTa, and XLNet, and tests each model separately. In order to obtain the best possible classification accuracy, we then use an ensemble method that uses the bagging, stacking, and boosting techniques. The ensemble model transcends the limitations of the individual models by combining the strengths of many architectures. In terms of accuracy, F1-score, and general robustness, our results indicate that the ensemble approach consistently surpasses that of single models. We reveal overarching misclassification patterns through thorough error analysis and demonstrate how ensemble learning mitigates these errors. We provide empirical justification for enhanced sentiment classification performance, an optimized ensemble framework, and comprehensive comparison analysis of state-of-the-art NLP models. These findings are evidence of the success of ensemble learning on contemporary NLP tasks such as multilingual sentiment analysis. There are still problems, such as augmented arithmetic complexity and long training times due to the combination of the plural, although the group's approach is better. The mysterious or contextual emotion continues to cause a wrong classification, which implicitly means continuous requirements for improvement in the treatment of advanced language. Future studies should investigate more effective ensemble methods to maximize performance while lowering computational costs, including neural architecture search or adaptive weighting procedures. Further broadening the dataset to incorporate a wider range of language and cultural variances may also improve the model's generalizability.

REFRENCES

[1] Maryam Abdullah AL-Barrak and Adel Ismail Al-Alawi. Sentiment analy sis on customer feedback for improved decision making: a literature review. In 2024 ASU International Conference in Emerging Technologies for Sus tainability and Intelligent Systems (ICETSIS), pages 207–212. IEEE, 2024.

[2] Ohud Alsemaree, Atm S Alam, Sukhpal Singh Gill, and Steve Uhlig. Sen timent analysis of arabic social media texts: A machine learning approach to deciphering customer perceptions. Heliyon, 10(9), 2024.

[3] Omar AlZoubi, Saja Khaled Tawalbeh, and Mohammad AL-Smadi. Affect detection from arabic tweets using ensemble and deep learning techniques. Journal of King Saud University- Computer and Information Sciences, 34(6, Part A):2529–2539, 2022.

[4] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 258–266, Marseille, France, June 2022. Eu ropean Language Resources Association.

[5] Erik Cambria. An introduction to concept-level sentiment analysis. In Ad vances in Soft Computing and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II 12, pages 478–483. Springer, 2013.

[6] Akhilesh Ingole, Prathamesh Khude, Sanket Kittad, Vishakha Parmar, and Archana Ghotkar. Competitive sentiment analysis for brand reputation monitoring. In 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), pages 1–7. IEEE, 2024.

[7] Jianguo Jia, Wen Liang, and Youzhi Liang. A review of hybrid and ensem ble in deep learning for natural language processing, 2024.

[8] Jamin Rahman Jim, Md Apon Riaz Talukder, Partha Malakar, Md Mohsin Kabir, Kamruddin Nur, and Mohammed Firoz Mridha. Recent advance ments and challenges of nlp-based sentiment analysis: A state-of-the-art review. Natural Language Processing Journal, page 100059, 2024.

[9] Chandra Mouli Madhav Kotteti, Xishuang Dong, and Lijun Qian. Ensem ble deep learning on time-series representation of tweets for rumor detection in social media. Applied Sciences, 10(21), 2020.

Journal of Information Systems Engineering and Management
2025, 10(34s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[10] Pankaj Malik, Vidhi Gupta, Vanshika Vyas, Rahul Baid, and Parth Kala. Understanding and enhancing xlnet: comprehensive eling. exploration of permutation language A mod https://www.researchgate.net/profile/Pankaj-Malik 4/publication/380753147UnderstandingandEnhancingXLNetAComprehensiveExplorationofPermutation and–Enhancing –XLNet–A–Comprehensive–Exploration–of – Permutation –Language–Modeling.pdf,2024.
[11] Ammar Mohammed and Rania Kora. A comprehensive review on ensem ble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences, 35(2):757–774, 2023.
[12] Joyce Y. M. Nip and Benoit Berthelier. Social media sentiment analysis. Encyclopedia, 4(4):1590–1598, 2024.
[13] Omer Sagi and Lior Rokach. Ensemble learning: A survey. Wiley interdis ciplinary reviews: data mining and knowledge discovery, 8(4):e1249, 2018.
[14] Arun Singla. Roberta and bert: Revolutionizing mental healthcare through natural language. Shodh Sagar Journal of Artificial Intelligence and Ma chine Learning, 1(1):10–27, Mar. 2024.
[15] Hoong-Cheng Soong, Norazira Binti A Jalil, Ramesh Kumar Ayyasamy, and Rehan Akbar. The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and tech niques. In 2019 IEEE 9th symposium on computer applications & industrial electronics (ISCAIE), pages 272–277. IEEE, 2019.
[16] Dimple Tiwari, Bharti Nagpal, Bhoopesh Singh Bhati, Ashutosh Mishra, and Manoj Kumar. A systematic review of social network sentiment anal ysis with comparative study of ensemble-based techniques. Artificial Intel ligence Review, 56(11):13407–13461, 2023.
[17] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and chal lenges. Artificial Intelligence Review, 55(7):5731–5780, 2022.
[18] Adam Westerski. Sentiment analysis: Introduction and the state of the art overview. Universidad Politecnica de Madrid, Spain, pages 211–218, 2007.
[19] Yichao Wu, Zhengyu Jin, Chenxi Shi, Penghao Liang, and Tong Zhan. Research on the application of deep learning-based bert model in sentiment analysis. Applied and Computational Engineering, 71:14–20, 05 2024