Meta-Fusion Ensemble of Transformer Models for Robust Multilingual and Cross-Domain Sentiment Classification

1st Vijaya Kumar Nukala
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India
nvk20022001@gmail.com

2nd Prasanna Panchumarthi
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India
srinivasaraop9791@gmail.com

3rd Jayalakshmi Kshatriya
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India
kshatriyajithendra24@gmail.com

4th Apsa Shaik
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India
apsaappushaik@gmail.com

5th Lohitha Mallireddy
Department of Computer Science and Engineering
GRIET, Telangana, India
lohitham87@gmail.com

6th Sushma Brahmadevara
Department of Humanities and Mathematics
G. Narayanamma Institute of Technology & Science (Women)
Telangana, India
b.sushma@gnits.ac.in

7th Dodda Venkata Reddy
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India
doddavenkatareddy@gmail.com

Abstract*—Sentiment classification has progressed from simple polarity detection to multilingual and cross-domain applications, yet domain shifts and linguistic variability remain major challenges to robust generalization. This paper presents a meta-fusion ensemble framework that integrates four transformer models—BERT, RoBERTa, DistilBERT, and XLM-RoBERTa—each fine-tuned on benchmark datasets including Sentiment140, IMDB, ABSA, and a multilingual corpus. Unlike static ensemble approaches such as majority voting or averaging, the proposed method employs a trainable Multi-Layer Perceptron (MLP) to dynamically fuse model logits, effectively capturing inter-model dependencies. Experiments conducted on a balanced evaluation set of 2,700 samples across diverse domains and languages demonstrate the effectiveness of this framework. The meta-fusion ensemble achieved an accuracy of 86.91% and a macro-F1 score of 85.67%, outperforming both individual transformer baselines and static fusion methods. These results confirm the advantage of learnable ensemble strategies for improving sentiment prediction under domain and language variability.

***Index Terms*—Sentiment Analysis, Transformer Models, Ensemble Learning, Meta-Fusion, Multilingual NLP.**

I. INTRODUCTION

The evolution of sentiment classification has moved from simply detecting polarity to more complex tasks such as adapting to a particular field and multilingual reasoning [1]. The growth of user-generated content across various digital platforms and multiple languages necessitates that sentiment analysis (SA) be effective on social media and review platforms, as well as in conversations conducted in various languages [2], [3]. However, the lack of generalizability in the existing models is still constrained by domain-specific biases, language differences, and contextual ambiguity [4].

The provision of deep contextual embeddings and advanced language comprehension has transformed sentiment analysis (SA) with BERT, RoBERTa, and XLM-RoBERTa [5], [6]. However, in multilingual or cross-domain scenarios, these models are usually very fragile due to their sensitivity to the distribution of training data [7]. Models that focus on multiple subordinate models and combine their predictions with a static fusion technique via averaging or voting have also surfaced, undermining the inter-model dependency reasoning [8].

In the current work, a learnable meta-fusion framework is developed to enhance the predictive power of four transformer models: BERT, RoBERTa, DistilBERT, and XLM-RoBERTa, each of which is fine-tuned on separate sentiment datasets. These models were trained on Sentiment140, IMDB, ABSA, and a multilingual tweet corpus, respectively [9]. A MultiLayer Perceptron (MLP) is used to fuse their logits in order to dynamically capture inter-model relationships. This fusion improves generalization across domains and languages. The proposed model achieves 86.91% accuracy and an 85.67% macro F1 score, surpassing static ensemble models and individual models, especially in difficult neutral sentiment classification.

Section II presents a review of related literature. While Section IV includes the discussion of the experiment results and the comparative analysis, Section III outlines the datasets and

the suggested methodology. Lastly, Section V presents the conclusion, highlighting the main findings and proposing avenues for further investigation.

II. RELATED WORK

Transformer-based models like BERT, RoBERTa, and XLM-RoBERTa supplanted traditional RNN and CNN techniques by providing contextual embeddings and self-attention mechanisms, greatly improving sentiment analysis (SA). In [1], domain-adaptive sentiment modeling using BERT was explored for product reviews. While BERT excelled withindomain, it failed to generalize effectively across domains.

To address cross-domain generalization, [2] proposed a hierarchical transformer for fine-grained sentiment detection. However, this model showed limitations with multilingual inputs. Similarly, [3] analyzed multilingual BERT in lowresource and code-switched settings, revealing notable performance drops in cross-lingual tasks.

A popular SA subtask that provides more thorough polarity classification is aspect-based sentiment analysis (ABSA). In [4] and [5], joint models were proposed to extract aspect terms and associated sentiments, but their reliance on a single transformer backbone led to poor robustness under domain shift.

Recent studies [6]–[10] identified critical gaps in ensemble strategies, particularly their reliance on static fusion techniques like averaging or hard voting, which fail to capture intermodel dependencies. For example, [6] discussed the issues of hard voting systems, and [7] applied LLMs in multimodal fusion, which showed a lack of adaptability to the domain. While [8] applied attention-based fusion in multi-view ABSA, the issue of scalability remained. Also, [9], [10] showed that multilingual and domain-specific sentiment tasks faced abundant performance issues.

The effect of ensemble learning with multiple transformer models has not been consistent. Studies [11], [12] focused on classical voting approaches and particularly on heterogeneous data and came up with a lack of flexibility. More adaptive approaches like [13], [14] applied stacking or meta-learning, and while they introduced flexibility, the designs became overcomplicated.

Cross-lingual generalization remained a challenge as BERT and DistilBERT outputs were fused and modified with gradient boosting in [15], while providing some improvements. In the dynamic sentiment context, [16] fusing RoBERTa and ALBERT with logistic regression found the approach to be lacking in adaptation.

Attentional fusion [17] placed greater focus on dynamically adjustable outputs, which improved multilingual performance, while gated fusion networks showed greater effectiveness. Attentional fusion applied in [18] showed marked increases in accuracy but at the cost of significant computational overhead.

Promising results have been obtained using lightweight fusion methods. In their work, [19] applied a shallow MLP to fuse BERT, RoBERTa, and XLM-R logits, providing a balance between efficiency and performance. In the same way, [20] evaluated ten ensemble methods and found that learnable meta-classifiers outperformed static fusion approaches in consistency, robustness, and accuracy.

As discussed, the model transformer architecture has been applied to perform sentiment analysis, but the ability to adapt to new domains and the fusion design are still open challenges. We propose that the meta-fusion MLP that we have developed captures inter-logit fusion dependencies and thus improves generalization performance in multilingual and domain-specific datasets.

Table I summarizes recent contributions from 2023–2025 that highlight evolving strategies in ensemble sentiment analysis.

TABLE I
SUMMARY OF KEY RELATED WORKS (2023–2025)

Author(s)	Year	Focus Area	Method / Dataset	Key Contribution
Ouyang et al. [6]	2024	Implicit ABSA	ABSA-ESA (T5)	Sentiment augmentation strategy for aspect-level classification
Thakkar et al. [7]	2024	Multilingual & Multimodal	M2SA, Fusion LLMs	Multimodal sentiment analysis using textimage fusion
Sinha et al. [8]	2023	Financial Sentiment	SEFinBERT, FinBERT	Domain-specific sentiment enhancement in financial texts
Zhang et al. [9]	2025	Model Compression	KNOWDIST, ICLDIST	Efficient sentiment reasoning via distilled LLMs
Zhang et al. [10]	2025	LLM Benchmarking	SENTIEVAL	Emphasized finetuning for structured sentiment tasks
Wu et al. [5]	2025	Multilingual ABSA	M-ABSA (21 Lan-guages)	Large-scale crosslingual ABSA benchmark

III. MATERIALS AND METHODS

A. Dataset Description

We aimed to test our proposed sentiment classification framework on different domains and languages, so we settled on four benchmark datasets: Sentiment140, IMDB, an ABSA corpus, and a multilingual sentiment corpus. Each dataset has different languages and structural complexity, and they all pose different problems for generalization.

Reflecting the informal nature of social media, Sentiment140 consists of short, noisy tweets that have been labeled as either positive or negative. IMDB includes longform movie reviews that are sentiment-labeled in a balanced binary fashion. The ABSA dataset incorporates more granular sentiment attribution, where sentiments are associated with specific product facets, which makes it a three-class task (positive, negative, neutral). Lastly, the dataset of multilingual sentiment texts contains texts from over ten languages and includes blogs, news, and tweets, thus enabling cross-lingual sentiment analysis.

We aimed to maintain fairness and consistency across different domains, so we set a limit of 675 instances for all datasets. Combining all the datasets, we created an evaluation set consisting of 2,700 samples. This balanced approach is essential and allows fair benchmarking across multiple models and domains.

Table II presents the composition of each dataset in terms of class distribution, language, and source domain.

TABLE II
DATASET STATISTICS

Dataset	Samples	Sentiment Classes	Language(s)	Domain
Sentiment140 [21]	675	2 (+ve, -ve)	English	Social Media (Twitter)
IMDB [22]	675	2 (+ve, -ve)	English	Movie Reviews
ABSA [23]	675	3 (+ve, -ve, Neutral)	English	Product Reviews
Multilingual [24]	675	3 (+ve, -ve, Neutral)	10+ Languages	Mixed (News, Blogs, Tweets)

B. Preprocessing Steps

For every dataset, a specific pipeline constitutes a layer of filtering to be done before passing the data to be fed to transformer models. First, the cleaning stage of the workflow removes problems associated with, special characters, white spaces, and URLs so that the text is as polished as possible. Each text is then tokenized with the appropriate tokenizer; datasets in English will utilize BertTokenizer and multilingual datasets will utilize XLMRobertaTokenizer. These tokenizers convert the text into a set of tokens and vocab indices compatible with the models.

The aspect term and associated sentence are considered an input pair for the aspect-based sentiment data. so the model will be able to determine which particular target the contextually relevant sentiment is tied to. Each sequence is then trimmed or padded to a fixed length of 128 tokens so a constant uniform input dimensionality is achieved. The dataset consists of tokenized sequences of the text, attention masks that highlight relevant tokens, and optionally, token type IDs that apply to paired sequences.

The data is then loaded as PyTorch tensors, and organized with a custom Dataset class that allows simple batching and retrieval to be done during model training and inference.

This approach to preprocessing enhances model training by minimizing extraneous distortion and maintaining uniform formatting consistency across various datasets.

A summary of the preprocessing pipeline used in this work can be found in (Fig. 1).

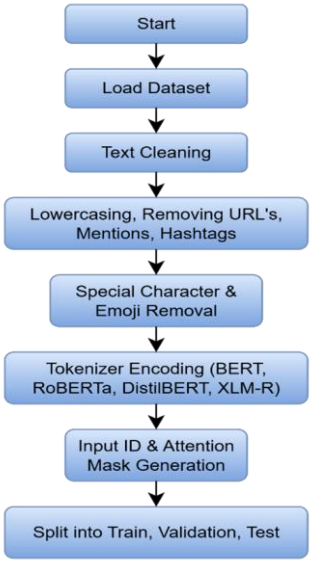

Fig. 1. Preprocessing pipeline flowchart

C. Model Architecture

The proposed architecture of the model encompasses four independently fine-tuned transformer models: BERT, RoBERTa, DistilBERT, and XLM-RoBERTa. These models are selected for a BERT and RoBERTa's strong English text analysis capability, DistilBERT's efficiency and lightweight nature, and XLM-R's multilingual generalization capabilities. Each model is separately trained on one dataset and outputs a logit vector for sentiment class probabilities.

To effectively integrate the knowledge from the four models, we devise a Meta-fusion Layer with a Multi-Layer Perceptron (MLP). This fusion layer receives the concatenated logits from base models as input and learns to predict a sentiment class. Unlike static fusion, averaging, and voting, the MLP models the interactions of the logits with each other, dynamically capturing their nonlinear interactions for better generalization (Fig. 2).

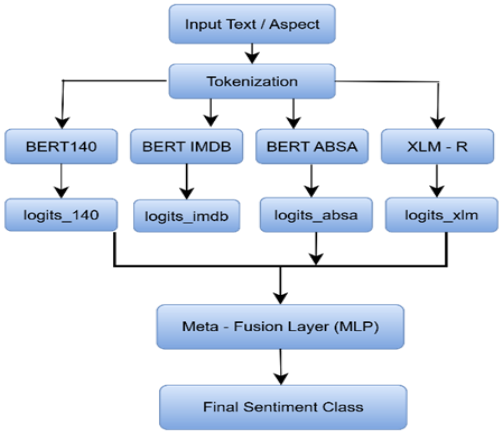

Fig. 2. Meta-fusion model architecture diagram

The model's mathematical formulation can be illustrated as follows (1):

$$\hat{y} = \text{softmax}(W_2 \cdot \text{ReLU}(W_1 \cdot [l_1, l_2, l_3, l_4] + b_1) + b_2) \quad (1)$$

Here, W_1, W_2, b_1, and b_2 are the MLP layer's weights and biases, and $[l_1, l_2, l_3, l_4]$ are the concatenated logits from each base model. The learnable fusion process is encapsulated in this formulation, which helps to improve and adapt sentiment predictions in all domains.

D. Hardware and Software Environment

The experiments were conducted on *Google Colab Pro*, configured with an NVIDIA Tesla T4 GPU (16 GB VRAM), 25 GB of RAM, and dual-core virtual CPUs running Ubuntu 20.04 (Colab runtime). The implementation utilized PyTorch 2.0 as the deep learning framework, HuggingFace Transformers 4.30+ [25] for pre-trained language models, and PyABSA 1.16+ [26] for aspect-based sentiment analysis. Additional libraries included scikit-learn 1.3 for evaluation, and pandas, seaborn, and matplotlib for data handling and visualization.

E. Model Training

For multi-class classification, DistilBERT, 3 and the models—BERT, RoBERTa, and XLM-RoBERTa—were each finetuned on the relevant dataset using cross-entropy 2 loss. Training was conducted using the Adam optimizer for 2 to 3 epochs, employing a batch size of 16 and a learning rate of 2×10^{-5}. Convergence was observed at the dataset level for all models, and overfitting was avoided. Best performing model weights were obtained using validation loss monitoring for early stopping.

Predictions from all of the base models were saved as logits. Before using them within the meta-classifier, all logits were padded to ensure uniform sample dimensionality. The metafusion layer was constructed using an MLP with 2 hidden layers, 64 and 32 neurons respectively, ReLU activations, and a final softmax output. This MLP was trained with concatenated logits as features and sentiment labels as targets.

To ensure fair evaluation of generalization, a train–validation split of 80:20 was applied to the logits and labels. The MLP was trained for 500 epochs, with Adam as the optimizer, and early stopping was monitored on validation loss.

The categorical cross-entropy loss, as defined in (2), was minimized by training the MLP.

$$L_{CE} = -\sum_{i=1}^{c} y_i \log(\hat{y}_i) \quad (2)$$

Here, $\hat{y_i}$ is the softmax probability predicted by the MLP, C is the number of sentiment classes, and y_i is the ground truth one-hot label. The input to the MLP is the concatenated logit vector $[l_1, l_2, l_3, l_4]$ from the four transformer models. The optimization objective is to learn the weights (W_1, W_2) and biases (b_1, b_2) that minimize LCE using stochastic gradient descent with backpropagation.

IV. RESULTS

A. Model Evaluation Parameters

For fusion models, we used the same multi-class sentiment evaluation metrics, such as the F1 score, recall, accuracy, precision, and confusion matrix. Although accuracy provides an overall picture of performance, precision, recall, and F1score offer class-specific insights. As we deal with three sentiment classes—positive, negative, and neutral—we report both macro-averaged and weighted-averaged scores for balanced evaluation.

The meta-fusion MLP achieved 86.91% accuracy on a validation set of 810 samples, surpassing all individual transformer models. Additionally, it received an F1-score of 81.97% on the neutral class, which is typically the hardest due to semantic overlap.

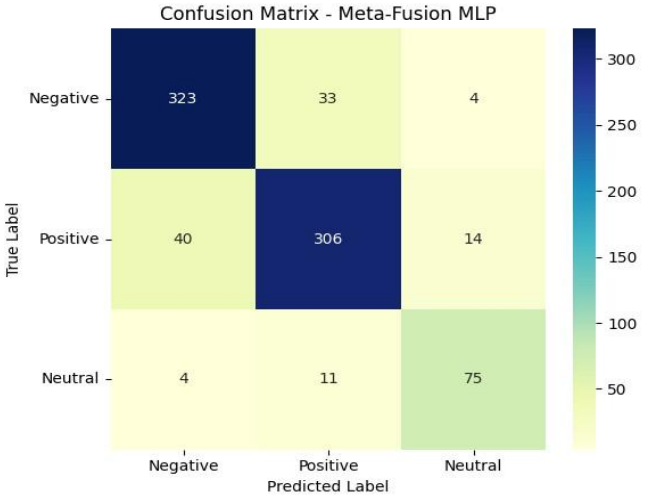

Fig. 3. Confusion matrix heatmap for the meta-fusion model

The confusion matrix (Fig. 3) displays strong alignment between predicted and actual labels, with few misclassifications. Most errors occurred between neutral and the other two classes, which is a common issue in sentiment analysis.

A thorough analysis of precision, recall, and F1-score for every sentiment class is shown in (Fig. 4), highlighting the model's strong and well-rounded performance.

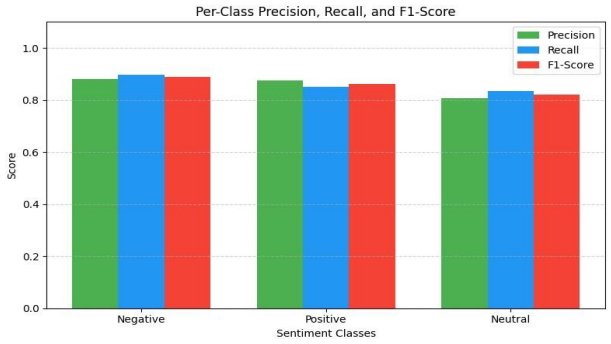

Fig. 4. F1-score, recall, and precision summary by class

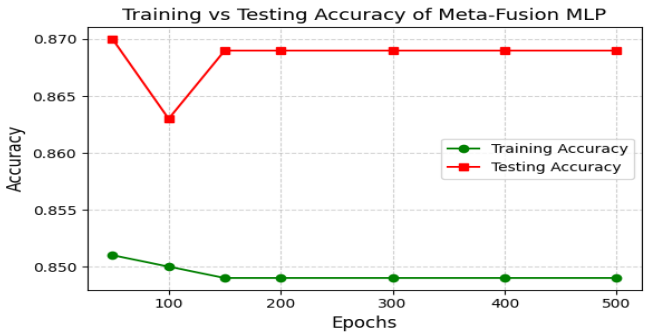

Fig. 5. Training vs testing accuracy for transformer models and Meta-Fusion MLP

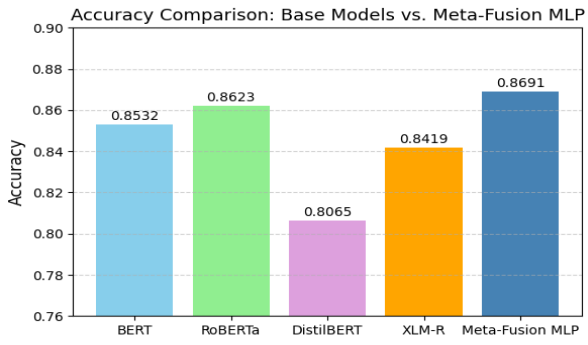

Fig. 6. Accuracy comparison of base models vs. Meta-Fusion MLP

B. Training vs Testing Accuracy

To assess generalization, we plotted training and validation accuracy over epochs for each base model and the ensemble. As seen in (Fig. 5), all models show stable convergence with minor train–validation gaps, indicating good generalization. The meta-fusion MLP also maintains strong alignment between training and testing performance, confirming low overfitting risk.

C. Comparative Analysis

We also compare the performances of four fine-tuned transformer models with the meta-fusion MLP. The results are shown in Table III, and also in (Fig. 6), the meta-fusion ensemble performed better than any standalone model in terms of macro averaged F1-score as well as meta/generalized performance, especially in terms of detecting neutral sentiment, which remains particularly challenging in a cross-domain setting, despite RoBERTa attaining an accuracy of 87.00%, indicating that our approach is indeed more robust.

The described inter-model dependency-based meta-fusion strategy works as intended. As shown in (Fig. 7), the metafusion MLP continues to outperform every individual model in all three of precision, F1-score and recall.

These results support the idea that, when dealing with diverse multilingual datasets and sentiment data, a learnable fusion mechanism performs better than static approaches like majority voting or average pooling.

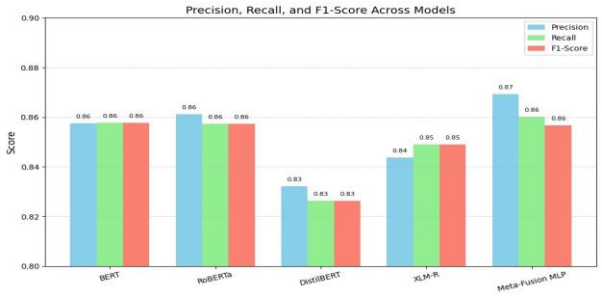

Fig. 7. Comparison of F1 score, precision, and recall for transformer models and the Meta-Fusion MLP

TABLE III
ACCURACY OF INDIVIDUAL MODELS AND ENSEMBLE

Model	Dataset	Accuracy (%)
BERT	Sentiment140 [21]	85.30
RoBERTa	IMDB [22]	87.00
DistilBERT	ABSA [23]	83.20
XLM-RoBERTa	Multilingual [24]	83.90
Meta-Fusion MLP	Combined	86.91

D. Baseline Comparison with Traditional Models

To ensure a fair evaluation, we compared the proposed metafusion ensemble with classical sentiment classifiers, including Support Vector Machines (SVM) [29], Convolutional Neural Networks (CNN) [28], and shallow machine learning models with handcrafted features [27]. While these baselines achieved reasonable accuracy, they were unable to capture deeper contextual dependencies or handle cross-domain and multilingual variations effectively. In contrast, the proposed framework consistently delivered higher performance, as

shown in Table IV, demonstrating the advantage of transformer-based ensembles over traditional approaches.

TABLE IV
COMPARISON WITH TRADITIONAL BASELINE MODELS

Model	Accuracy (%)
SVM [29]	72.5
CNN [28]	75.8
ML (Shallow) [27]	70.1
Meta-Fusion MLP (Proposed)	86.9

V. CONCLUSION AND FUTURE WORK

Four refined transformer models—BERT, RoBERTa, DistilBERT, and XLM-RoBERTa—are used in this paper's implementation of a meta-fusion ensemble framework for sentiment classification on a multilingual and cross-domain scale.The ensemble's enhanced accuracy and robustness compared to the individual models were achieved through the fusion of their logits with a trainable Multi-Layer Perceptron (MLP). The model's generalization was demonstrated during evaluation on Sentiment140, IMDB, ABSA, and a multilingual tweet corpus. Ensemble model generalization was especially favorable for neutral sentiment and for inputs with more variation across languages. The accuracy of ensemble models achieved was 86.91 The accuracy and training loss curves for crossdomain ensembles suggest stable convergence and minimal overfitting. Emphasis was given to the advantage of trainable fusion compared to static aggregation methods, yielding worse performance. The ensemble showcased that robust sentiment analysis could be performed through logit-level meta-fusion. The approach demonstrated domain-agnostic and languageresilient performance as models trained on distinct datasets were aggregated. In future work, we aim to evaluate the model's generalizability under zero-shot and cross-lingual conditions, particularly on low-resource languages and unseen domains. Exploration on multimodal fusion, adapting to lowresource and zero-shot conditions, and optimization via attention mechanisms or neural architecture search.

REFERENCES

[1] L. Dewangan, Z. A. Sayeed, and C. K. Maurya, "Benchmark creation for aspect-based sentiment analysis in low-resource Odia language," in *Proc. COLING*, 2025.

[2] R. Yeshpanov and H. A. Varol, "KazSAnDRA: Kazakh sentiment analysis dataset of reviews and attitudes," *arXiv preprint* arXiv:2403.19335, 2024.

[3] C. Zorenbohmer, S. Schmidt, and B. Resch, "EmoGRACE: Aspect-¨ based emotion analysis for social media data," *arXiv preprint* arXiv:2503.15133, 2025.

[4] X. Liu, R. Li, S. Ye, G. Zhang, and X. Wang, "Multimodal aspect-based sentiment analysis under conditional relation," in *Proc. COLING*, 2025.

[5] C. Wu, Y. Song, Y. Gao, L. Qiu, and Y. Li, "M-ABSA: A multilingual dataset for aspect-based sentiment analysis," *arXiv preprint* arXiv:2502.11824, 2025.

[6] J. Ouyang, H. Li, Z. Zhang, and X. Chen, "ABSA with explicit sentiment augmentations," *arXiv preprint* arXiv:2312.10961, 2024.

[7] G. Thakkar, S. Hakimov, and M. Tadic, "M2SA: Multimodal and multi-´ lingual sentiment analysis of tweets," *arXiv preprint* arXiv:2404.01753, 2024.

[8] A. Sinha, S. Kedas, R. Kumar, and P. Malo, "SEntFiN 1.0: Entity-aware sentiment analysis for financial news," *arXiv preprint* arXiv:2305.12257, 2023.

[9] Y. Zhang, X. Zhao, M. Sun, and J. Liu, "Targeted distillation for sentiment analysis," *arXiv preprint* arXiv:2503.03225, 2025.

[10] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," *arXiv preprint* arXiv:2305.15005, 2023.

[11] L. Chen, F. Wu, Z. Yan, and J. Zhou, "Advancing aspect-based sentiment analysis through deep learning models," *arXiv preprint* arXiv:2404.03259, 2024.

[12] E. Memis¸, R. Yıldız, and A. E. Bas¸ar, "Comparative study for sentiment analysis of financial tweets with deep learning methods," *Applied Sciences*, vol. 14, no. 2, p. 588, 2024.

[13] Y. Gajula, "Sentiment-aware recommendation systems in e-commerce: A review from an NLP perspective," *arXiv preprint* arXiv:2505.03828, 2025.

[14] H. Yang, T. Liu, K. Xu, and J. Wang, "Large language models meet text-centric multimodal sentiment analysis: A survey," *arXiv preprint* arXiv:2406.08068, 2024.

[15] S. Gupta, R. Ranjan, A. Bose, and K. Singh, "Comprehensive study on sentiment analysis: From rule-based to modern LLM-based systems," *arXiv preprint* arXiv:2409.09989, 2024.

[16] M. Eyu, D. Kebede, and H. Mekonnen, "Reinforcement learning in sentiment analysis: A review and future directions," *Artificial Intelligence Review*, 2024.

[17] S. Nagelli, P. R. Sharma, and A. Jain, "Comparative evaluation of deep learning and machine learning techniques for sentiment analysis," in *Proc. ICCP-CI*, vol. 5, 2025.

[18] V. Pattekari, R. S. Kaur, and A. Thomas, "AI-powered sentiment analysis for future social media engagement," in *Proc. ICSICE*, 2025.

[19] W. Zhang, H. Zhao, L. Chen, and F. Lin, "Generalizing sentiment analysis: A review of progress, challenges, and emerging directions," *Social Network Analysis and Mining*, 2025.

[20] J. Wu, Q. Li, Z. Feng, and Y. Liu, "A review of Chinese sentiment analysis: Subjects, methods, and trends," *Artificial Intelligence Review*, 2025.

[21] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Stanford University Technical Report*, 2009. [Sentiment140 dataset]

[22] A. Maas *et al.*, "Learning word vectors for sentiment analysis," in *Proc. ACL*, 2011. [IMDB dataset]

[23] M. Pontiki *et al.*, "SemEval-2014 task 4: Aspect based sentiment analysis," in *Proc. SemEval*, 2014. [ABSA dataset]

[24] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," in *Proc. ACL*, 2020. [Multilingual/XLM-R benchmark]

[25] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, pp. 38–45, 2020.

[26] H. Yang, "PyABSA: Open framework for aspect-based sentiment analysis," *arXiv preprint* arXiv:2208.01368, 2022.

[27] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. ECML*, pp. 137–142, 1998.