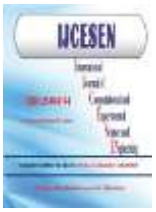
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 4250-4265
http://www.ijcesen.com

ISSN: 2149-9144



Copyright © IJCESEN

Research Article

A Feature-Oriented Textual Ensemble Framework for Robust Emotion Detection Across Diverse Classification Paradigms

Mukesh C Jain[1*], Farha Haneef[2]

[1]Research Scholar, Oriental University, Indore (M.P.) India
* **Corresponding Author Email:** mukesh.ch.jain@gmail.com- **ORCID:** 0009-0008-0254-4294

[2]Professor, Oriental University, Indore (M.P.) India
Email: farhahaneef2014@gmail.com- **ORCID:** 0000-0002-7320-1394

Article Info:

DOI: 10.22399/ijcesen.4934
Received : 01 April 2025
Accepted : 30 April 2025

Keywords

Emotion Detection,
Ensemble Learning,
Text Representation,
Transformer Models

Abstract:

Emotion detection from text plays a crucial role in natural language processing applications such as social media analysis, opinion mining, and conversational systems. Existing approaches typically rely on isolated representation paradigms, including lexical features, semantic embeddings, or contextualized models, which often limits robustness across datasets and task formulations. This study presents a Feature-Oriented Textual Ensemble (FOTE) framework that integrates lexical, semantic, and contextual representations within a unified probability-level ensemble architecture. Unlike conventional ensembles that emphasize model-level diversity, the proposed framework explicitly combines heterogeneous representation paradigms designed to capture complementary linguistic properties. Parallel representation branches are employed to model explicit emotion cues, sentence-level semantics, and context-dependent word meaning, followed by probability-level fusion using ensemble strategies, while maintaining a consistent implementation pipeline across experimental settings. Configuration-level adaptations are applied to support binary, multi-class, and multi-label emotion classification without architectural modification. The proposed framework is evaluated on multiple benchmark and validation datasets spanning diverse domains and emotion classification paradigms. Experimental results indicate that the ensemble consistently improves performance over a strong transformer-based baseline across most datasets, demonstrating enhanced robustness and generalization, while avoiding performance degradation in challenging scenarios. The results highlight the effectiveness of representation-level integration as a generalizable alternative to single-paradigm emotion detection models.

1. Introduction

Emotion detection from text has attracted significant attention in natural language processing due to its relevance in applications such as social media monitoring, opinion mining, customer feedback analysis, and human–computer interaction. Unlike conventional sentiment analysis, which primarily focuses on polarity classification, emotion detection seeks to identify a broader range of affective states expressed in text, making it a more fine-grained and challenging task [1], [2]. Early research predominantly relied on lexical representations and traditional machine learning techniques, including bag-of-words and term frequency–inverse documentfrequency features combined with classifiers such as Naïve Bayes and support vector machines, often supplemented with emotion lexicons and affective resources [12]–[18], [19]–[22]. Although effective in controlled environments, these approaches depend heavily on surface-level word distributions and predefined vocabularies, limiting their ability to capture implicit emotional expressions and contextual nuances. The adoption of deep learning enabled automatic feature learning and reduced reliance on handcrafted representations. Convolutional and recurrent neural architectures improved performance by modeling compositional and sequential patterns in text [26]–[32], while hybrid models combined local feature extraction with long-range dependency modeling [33], [34].However, approaches based on static embeddings or

purely sequential processing remain limited in modeling dynamic word meaning and often exhibit sensitivity to dataset characteristics and task formulations [35], [36]. Transformer-based architectures further advanced emotion detection through context-aware representation learning via self-attention mechanisms. Contextualized embeddings derived from bidirectional transformers have demonstrated strong performance across emotion and sentiment tasks by dynamically modeling word meaning within context [37]–[41], with sentence-level representations improving semantic and affective modeling [42], [43]. Nevertheless, transformer models typically rely on a single representation paradigm and may underutilize explicit lexical cues or sentence-level structural information, resulting in dataset-dependent performance variations [44]–[47]. Ensemble learning has been explored to enhance robustness and generalization by combining multiple classifiers or models. Techniques such as voting, boosting, and stacking have been widely applied to text classification tasks [48]–[50], and ensemble-based emotion detection has shown improved predictive stability. However, most existing approaches operate at the model-output level or combine models trained on similar feature representations, limiting their ability to exploit complementary linguistic information across heterogeneous paradigms. Motivated by these limitations, this study proposes a feature-oriented ensemble framework that integrates lexical, semantic, and contextual representations for emotion detection. The framework employs parallel representation learning and probability-level fusion within a unified implementation pipeline. Extensive evaluation on benchmark studies [55], [56] and additional validation datasets assesses robustness and generalization across binary, multi-class, and multi-label emotion classification tasks.

2. Related Work

2.1 Lexical and Traditional Machine Learning Approaches

Early emotion detection research primarily relied on lexical representations and traditional machine learning techniques. Common approaches employed bag-of-words and term frequency–inverse document frequency features combined with classifiers such as Naïve Bayes, support vector machines, logistic regression, and decision trees [12]–[18]. Lexicon-based methods further incorporated predefined affective resources to capture explicit emotion cues [19]–[22].While computationally efficient and interpretable, these methods depend heavily on surface-level word distributions and domain-specific vocabularies, limiting their ability to model implicit expressions, contextual ambiguity, and cross-domain generalization [23]–[25].

2.2 Deep Learning–Based Emotion Detection

Deep learning models were introduced to reduce reliance on handcrafted features through automatic representation learning. Convolutional and recurrent neural networks, including LSTM and GRU architectures, improved performance by capturing hierarchical and sequential patterns in text [26]–[32]. Hybrid architectures further combined local and long-range dependency modeling for enhanced feature extraction [33], [34]. However, approaches based on static embeddings or purely sequential architectures remain constrained in modeling dynamic contextual meaning and often exhibit dataset-specific sensitivity [35], [36].

2.3 Transformer-Based Contextual Representation Models

Transformer architectures advanced emotion detection through context-aware modeling using self-attention mechanisms. Contextualized embeddings from bidirectional transformer models demonstrated strong performance across emotion and sentiment tasks [37]–[41], while sentence-level representations improved semantic and affective modeling [42], [43]. Benchmark studies, including those by Yousaf et al. [55] and Batbaatar et al. [56], reported competitive results across binary and multi- class settings. Nevertheless, most transformer-based approaches rely on a single representation paradigm and may underutilize complementary lexical or structural cues, leading to dataset-dependent performance variations [44]–[47].

2.4 Ensemble Learning Strategies in Emotion Analysis

Ensemble learning has been explored to enhance robustness by combining multiple classifiers. Techniques such as boosting, voting, and stacking have been widely applied to text classification tasks [48]–[50], with ensemble-based emotion detection improving predictive stability through output aggregation. However, most existing ensemble methods operate at the model-output level or combine models trained on similar feature representations. Explicit integration of heterogeneous lexical, semantic, and contextual paradigms within a unified probability-levelframework remains relatively limited, particularly across binary, multi-class, and multi-label settings [56].Table 1 summarizes representative studies in text- based emotion detection from 2016 to 2024, highlighting the progression of representation paradigms addressed in prior work.

2.5 Summary and Research Gap

The reviewed literature indicates that lexical, deep learning, and transformer-based approaches each offer distinct advantages for emotion detection from text. Nevertheless, when applied in isolation, these paradigms exhibit inherent limitations related to contextual ambiguity, semantic coverage, and cross-dataset generalization. Existing ensemble strategies provide partial solutions but predominantly emphasize model-level diversity rather than representation-level integration. These observations motivate the exploration of a feature- oriented ensemble framework that explicitly integrates complementary textual representations and evaluates performance consistently across diverse emotion classification paradigms and datasets.

3. Feature-Oriented Textual Ensemble Framework

This section presents the proposed Feature- Oriented Textual Ensemble (FOTE) framework for text-based emotion detection. The framework integrates complementary lexical, semantic, and contextual representations within a unified ensemble architecture. It emphasizes representation-level diversity and probability-level fusion while maintaining a consistent implementation pipeline across experimental settings.

3.1 Overview of the FOTE Framework

FOTE adopts a parallel representation learning strategy in which heterogeneous textual representations are learned independently and integrated through ensemble fusion. As illustrated in Figure 1, the architecture comprises three parallel branches, each designed to capture a distinct linguistic property of the input text. Each branch produces class-wise probability outputs that are subsequently combined at the ensemble stage. By explicitly modeling surface-level lexical cues, sentence-level semantic meaning, and context-dependent word interpretation, the framework leverages complementary linguistic information beyond single-paradigm approaches. Probability-level fusion preserves uncertainty information from individual branches and enables effective integration across representations. The architecture remains unchanged across binary, multi-class, and multi-label emotion classification tasks, requiring only configuration-level adaptations.

3.2 Branch-wise Representation Design

Each branch of the FOTE framework is designed with a specific representation objective, ensuring that distinct linguistic information is captured during learning. The branch-wise design is summarized in Table 2.

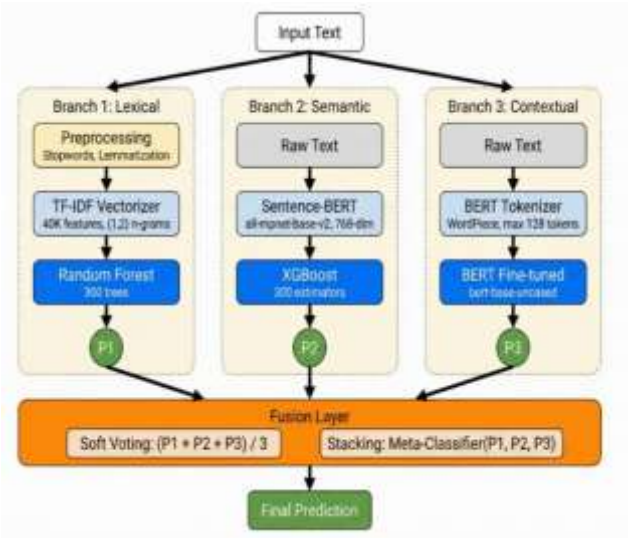


Figure 1. *Overall architecture of the Feature-Oriented Textual Ensemble (FOTE) framework*

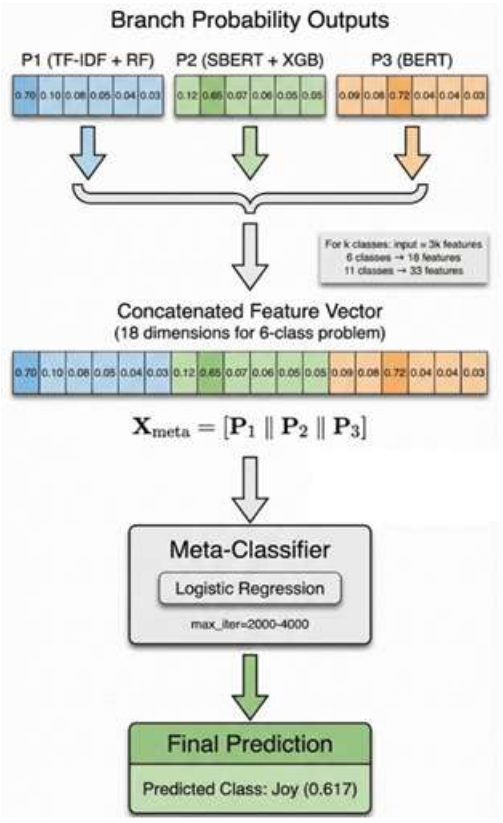


Figure 2. *Probability-level fusion strategy employed in*

3.3 Probability-Level Ensemble Fusion Strategy

Two ensemble strategies are employed within the framework. In soft voting, probability distributions from individual branches are combined using weighted or unweighted averaging to obtain final predictions. In stacking, branch-level probability outputs are concatenated and provided as input to a

Table 1. *Representative Studies in Text-Based Emotion Detection (2016–2024)*

Author(s) / Year	Findings	Research Gap
Perikos et al., 2016 [6]	Sentence-level ensemble (NB + MaxEnt + linguistic features) shows consistent gains over single classifiers on ISEAR and Affective Text.	Limited to emotion presence and polarity using BoW and handcrafted features; lacks fine-grained intensity modeling and cross-dataset validation.
Dandan Jiang et al., 2017 [1]	Introduces a Word Emotion Association Network for event-level emotion analysis on 48,396 social media posts, demonstrating multi-emotion trend consistency with emoticon-based benchmarks.	Emotion inference is lexicon- and emoticon-dependent at the word level, lacking sentence- level context modeling and cross-domain validation.
Bandhaka vi et al., 2017 [8]	Proposes a probabilistic unigram mixture model for domain-specific emotion lexicon generation, showing improved ranking and classification performance across news, blogs, and tweets.	Remains lexicon-centric without integration into full sentence-level classification or ensemble frameworks, and lacks systematic cross-dataset robustness evaluation.
Poria et al., 2017 [9]	Comprehensive survey of affective computing, reviewing representations, models, fusion strategies, and key challenges in emotion analysis.	Highlights challenges in context and generalization but lacks standardized evaluation protocols and validated ensemble frameworks for text-based emotion detection.
Agrawal et al., 2018 [11]	Introduces Emotion Word Embeddings learned via distant supervision and LSTM training, demonstrating improved classification and similarity performance on ISEAR and narrative datasets.	Focuses on word-level emotion features with single-model classification, lacking multi- representation integration and comprehensive robustness evaluation.
Shirsat et al. 2018 [12]	Examines sentence-level sentiment detection in news using Naïve Bayes and SVM with dictionary features, evaluated on the BBC dataset with standard metrics.	Limited to polarity-based sentiment with shallow lexical features; lacks fine-grained emotion modeling, contextual semantics, and cross-domain validation.
Hasan, et al. 2019 [14]	Explores supervised emotion detection in streaming Twitter data using word embeddings, highlighting challenges of temporal drift and evolving language.	Evaluates single classifiers with standard metrics, lacking ensemble integration, class- wise robustness analysis, and cross-dataset generalization.
Tadesse et al., 2019 [15]	Uses TF-IDF, LIWC, and LDA features for Reddit-based depression detection, with MLP achieving the best accuracy and F1 through feature combination.	Models emotion indirectly using binary, platform-specific handcrafted features, lacking fine-grained emotion analysis and cross-dataset generalization.
AlSheikh et al. 2019 [17]	Analyzes trust-related emotional cues in Airbnb seller descriptions, linking textual emotion indicators to buyer feedback patterns in marketplace contexts.	Application-specific and limited to a single affective construct, with indirect emotion inference and no standardized classification or multi-metric comparative evaluation.
Kim et al. 2019 [18]	Investigates CNN depth and pooling strategies for sentiment classification on MR, CR, and SST datasets, showing improved accuracy and F1 with deeper architectures.	Focuses on sentiment polarity instead of fine-grained emotions, with evaluation limited to standard metrics and no analysis of class-wise behavior or ensemble extension to emotion tasks.
Abas et al., 2020 [19]	Proposes a BERT-based model with attention, Bi-GRU, and CNN layers for aspect-level sentiment analysis, achieving improved	Focuses on aspect-level polarity within a single-model framework, lacking emotion-

	accuracy and F1 on SemEval-2014 and Twitter datasets.	category modeling, ensemble fusion, and cross-dataset generalization analysis.

Acheampong et al., 2020 [21]	Proposes a two-stage BERT + Bi-LSTM model for sentence-level emotion recognition, achieving 0.73 weighted F1 on ISEAR and outperforming traditional baselines.	Employs a single feature pipeline and classifier, without exploring alternative representations, ensemble strategies, or cross-dataset validation for robustness analysis.
Dobša, et al. 2021 [28]	Compares text-based models with qualitative self-report features on ISEAR, showing improved accuracy when combining textual and non-textual variables.	Relies on non-textual features, limiting portability; ensemble use is limited to feature fusion, with no evaluation of text-only sentence-level generalization across datasets.
Li et al., 2022 [29]	Reviews common machine learning algorithms and feature extraction methods (e.g., TF-IDF, n-grams), outlining their impact on text classification performance.	Emphasizes shallow features and task-specific evaluation, without addressing fine-grained emotion modeling, contextual semantics, or standardized multi-metric assessment.
Tu et al., 2022 [31]	Proposes Sentic-GAT, combining dialogue transformers with sentiment-aware graph attention and external knowledge, achieving consistent micro-/weighted-F1 gains on DailyDialog, MELD, and EmoryNLP.	Tightly dependent on external knowledge and dialogue contexts, with no evaluation on sentence- or document-level emotion tasks and limited robustness analysis beyond F1 metrics.
Tan et al., 2022 [32]	Integrates RoBERTa embeddings with LSTM for contextual–sequential sentiment modeling, achieving high F1-scores on IMDb, Twitter US Airline, and Sentiment140.	Focuses on sentiment polarity within a single hybrid pipeline, without exploring ensemble fusion across diverse representations or robustness on emotion datasets.
Yohanes et al., 2023 [37]	Compares LSTM, BiLSTM, and GRU with Word2Vec features on ISEAR, reporting GRU as the best-performing RNN baseline.	The study relies on static embeddings and single-model learning. Contextual representations, ensemble modeling, and cross-dataset validation are not explored.
Mozhdehi & Eftekhari Moghadam, 2023 [39]	Fine-tunes a BERT model for emotion detection, outperforming RNN baselines with higher F1-scores on Twitter and MELD datasets.	The approach centers on a single transfer-learning pipeline. Ensemble strategies, hybrid feature engineering, and cross-dataset generalisation beyond the tested corpora are not investigated.
He & Abisado, 2024 [46]	Proposes a hybrid BERT–CNN–BiLSTM–Attention model for short-text sentiment analysis, achieving improved accuracy and F1 on a large Douban review dataset.	The study addresses binary sentiment polarity only. Fine-grained emotion modeling, ensemble decision fusion, and robustness across datasets are not explored.
Sharma, et al. 2024 [47]	Compares transformer-based LLMs for sentiment analysis of YouTube EV comments, showing improved weighted F1 after fine-tuning, with RoBERTa performing consistently best.	The analysis is limited to sentiment polarity and single-model comparisons. Emotion-level modeling, ensemble fusion, and robustness metrics beyond standard scores are not examined.
Al Maruf et al., 2024 [48]	Presents a comprehensive survey of lexicon-based, ML, DL, and transformer approaches for text emotion detection, synthesizing datasets, challenges, and emerging trends.	Identifies challenges like contextual ambiguity and inconsistent evaluation, but does not empirically validate unified ensemble frameworks or standardized multi-metric protocols.
Esfahani & Adda, 2024 [49]	Compares classical ML and large language models for emotion recognition on ISEAR, showing higher accuracy with fine-tuned Mistral models across multiple feature settings.	Evaluates models independently without heterogeneous ensemble fusion, and lacks analysis of sentence-level robustness, class-wise behavior, and unified ensemble optimization.
Petrescu et al., 2024 [50]	Proposes an ensemble combining classical ML, LSTM, and transformer models for event-level sentiment analysis, achieving consistent gains on Sentiment140 over individual models.	The framework focuses on event-level sentiment polarity rather than fine-grained emotion categories. Ensemble fusion is limited to majority voting, without adaptive or emotion-aware fusion strategies.

Table 2. *Branch-wise representation design of the FOTE framework*

Branch	Representation Paradigm	Captured Linguistic Property	Motivation
Branch 1	Lexical (TF–IDF + ML classifier)	Word frequency patterns and explicit emotion cues	Captures discriminative lexical signals
Branch 2	Semantic (Sentence- level embeddings)	Sentence-level meaning and semantic similarity	Handles paraphrasing and semantic variation
Branch 3	Contextual (Transformer-based embeddings)	Context-dependent word meaning	Resolves contextual ambiguity

Table 3. *Overview of datasets used in the experimental evaluation*

#	Dataset	Source	Samples	Classes	Type	Domain
1	Twitter Sentiment	Yousaf et al. [55]	99,989	2	Binary	Social Media
2	Women's Clothing	Yousaf et al. [55]	23,486	5	Multi-class	E-commerce
3	Twitter Hatred	Yousaf et al. [55]	29,530	2	Binary	Social Media
4	DailyDialog	Batbaatar et al. [56]	14,160	6	Multi-class	Conversations
5	CrowdFlower	Batbaatar et al. [56]	26,693	6	Multi-class	Twitter
6	TEC	Batbaatar et al. [56]	19,678	6	Multi-class	Twitter
7	Tales-Emotion	Batbaatar et al. [56]	5,129	6	Multi-class	Narratives
8	ISEAR (5-class)	Batbaatar et al. [56]	5,477	5	Multi-class	Survey
9	EmoInt	Batbaatar et al. [56]	6,755	4	Multi-class	Twitter
10	Emotion-Cause	Batbaatar et al. [56]	2,224	6	Multi-class	News
11	IMDB [51]	Maas et al. [51]	50,000	2	Binary	Reviews
12	SemEval 2018 E-c [53]	Mohammad et al. [53]	10,983	11	Multi-label	Twitter
13	Citation Sentiment [13]	Raza et al. [13]	8,736	3	Multi-class	Academic
14	ISEAR (7-class) [52]	Scherer et al. [52]	7,666	7	Multi-class	Survey

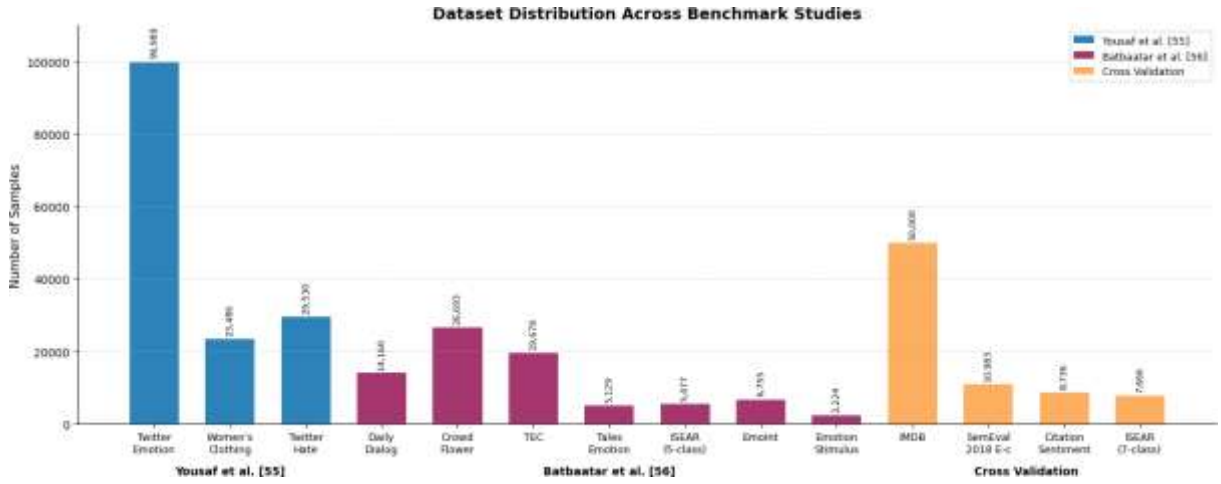


Figure 3. *Distribution of datasets across emotion classification paradigms and domains*

4. Experimental Setup

This section describes the datasets, implementation protocol, and evaluation metrics used to assess the proposed Feature-Oriented Textual Ensemble (FOTE) framework. The experimental design follows a unified

implementation pipeline across all settings, with configuration-level adaptations applied to support different emotion classification paradigms.

Datasets

An overview of all datasets used in the experiments, including their task type, domain, and class distribution, is provided in Table 3. The diversity of datasets across paradigms and domains is further illustrated in Figure 3, highlighting the breadth of evaluation settings considered in this study.

4.1 Implementation Details

The FOTE framework is implemented using a fixed execution pipeline across all experimental settings, consisting of text preprocessing, parallel representation learning, independent branch training, probability output generation, ensemble fusion, and final prediction. This structure remains consistent across datasets and task formulations to ensure fair and comparable evaluation. The overall implementation consistency is illustrated in Figure 4. To ensure reproducibility, all experiments were executed in a controlled environment with fixed random initialization and identical preprocessing pipelines across datasets. The framework wasimplemented in Python 3.8 using GPU acceleration on Google Colaboratory (NVIDIA Tesla T4, 16GB VRAM). A global seed value of 42 was applied to all stochastic operations, including data shuffling, weight initialization, and sampling procedures.

4.1.1 Data Splitting Protocol

Stratified partitioning was used to preserve class distribution across training, validation, and testing subsets. Feature extractors were fitted exclusively on training data to avoid information leakage. The splitting strategy was adapted based on dataset characteristics as summarized in Table 4.

4.1.2 Branch-wise Model Configuration

Each branch was configured with hyperparametersselected through preliminary validation on held-out subsets. The configurations are detailed in Table 5

4.1.3 Input Preprocessing Strategy

A heterogeneous preprocessing strategy was adopted to leverage the strengths of each representation paradigm, as shown in Table 6.This heterogeneous preprocessing design preserves discriminative lexical cues for traditional models while allowing pretrained language models to exploit natural linguistic structure without information loss

4.1.4 Ensemble Fusion Configuration

Two probability-level fusion strategies were employed for comparative analysis, as shown in Table 7.

The threshold optimization ensures balanced precision-recall tradeoffs for each emotion category independently. Grid search with step=0.01 provides fine-grained threshold selection across 81 candidate values per label.

4.2 Evaluation Metrics

The evaluation of the proposed FOTE framework employs task-appropriate performance metrics for binary, multi-class, and multi-label emotion classification to ensure statistically meaningful assessment. For binary and multi-class tasks, standard metrics including accuracy, precision,recall, and F1-score are used to measure predictive performance. ROC–AUC is additionally reported to assess discriminative capability across varying decision thresholds. To evaluate reliability beyond accuracy-based measures, agreement-based metrics are incorporated: Cohen's Kappa (κ) for multi-class datasets and Matthews Correlation Coefficient (MCC) for binary and imbalanced multi-class settings.For multi-label classification, evaluation includes Jaccard Accuracy (Jaccard Index) along with micro- and macro-averaged precision, recall, and F1-score, reflecting overlap between predicted and ground-truth emotion labels. Statistical significance is assessed using McNemar's test tocompare FOTE with the transformer baseline in applicable binary and multi-class settings, excluding cases with ceiling effects or non-informative distributions. The complete set of evaluation metrics used for each task type is summarized in Table 9.

5. Results and Analysis

This section presents a comprehensive evaluation of the proposed Feature-Oriented Textual Ensemble (FOTE) framework across multiple experimental settings. The analysis spans fourteen datasets covering binary, multi-class, and multi-label emotion classification tasks. The results are organized according to benchmark studies and additional validation datasets, followed by cross-paradigm and statistical analyses to assess robustness, generalization, and significance.

5.1 Results on Benchmark Datasets from Yousaf et al. [55]

The first set of experiments evaluates the performance of the proposed framework on benchmark datasets reported by Yousaf et al. [55], including Twitter Emotion, Women's Clothing Reviews, and Twitter Hatred. These datasets represent diverse textual characteristics, rangingfrom informal social media content to structured review-based expressions. Table 10 presents the consolidated Macro

Table 4. *Data Splitting Strategy by Dataset Category*

Dataset Category	Split Strategy	Rationale
Yousaf et al. [55] datasets	Train–Test (70-30) stratified split	Standard binary/multi-class protocol
Batbaatar et al. [56] datasets	Train–Validation–Test (80-10-10) stratified split	Validation used for early stopping
Large binary datasets (IMDB)	Train–Test (80-20) stratified split	Sufficient samples for direct evaluation
Multi-label dataset (SemEval)	Train–Validation–Test with threshold tuning on validation	Label-wise threshold optimization

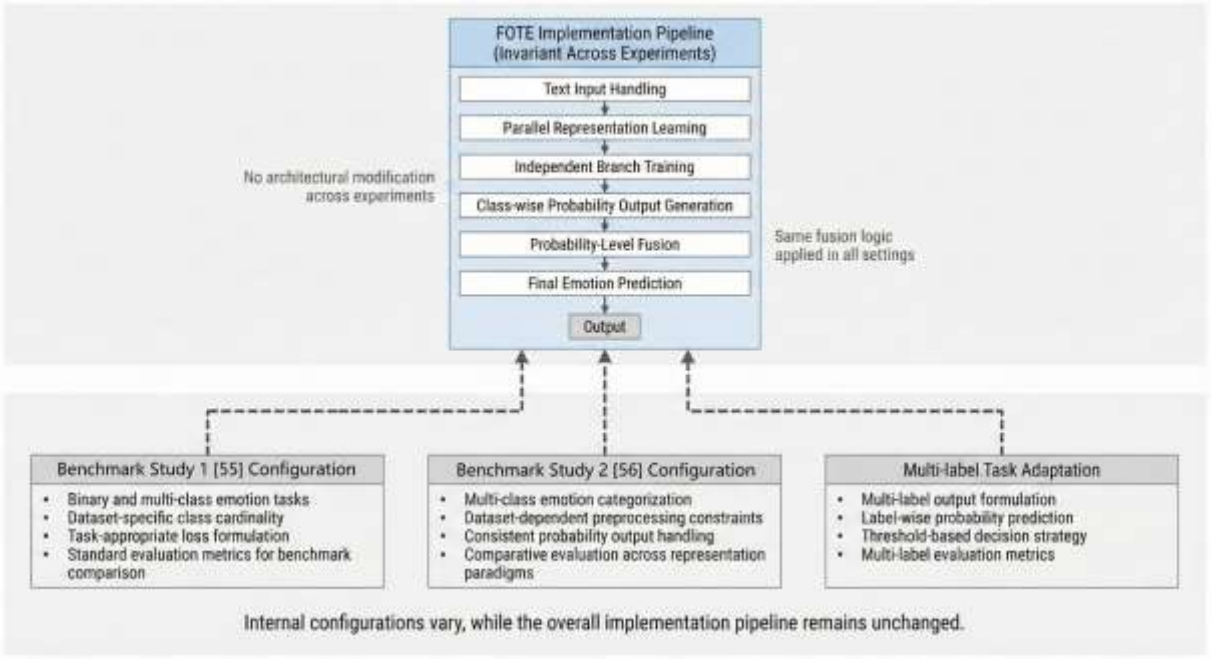


Figure 4. *Implementation consistency of the FOTE framework across experimental settings, illustrating a fixed pipeline with configuration-level adaptations for benchmark and multi-label tasks*

Table 5. *Branch-wise Implementation Configuration*

Branch	Representation	Model	Key Parameters
Branch 1 (Lexical)	TF–IDF (1–2 grams, max_features=40000)	Random Forest	n_estimators=300, max_depth=None, class_weight='balanced'
Branch 2 (Semantic)	Sentence embeddings (all- mpnet-base-v2, 768-dim)	XGBoost	max_depth=6, n_estimators=300, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8
Branch 3 (Contextual)	Transformer (BERT-base- uncased, 12 layers, 768 hidden dim)	BERT fine-tuned	learning_rate=2e-5, batch_size=16, dropout=0.3, max_length=128, early stopping (patience=2, monitor=val_loss)

Table 6. *Branch-specific Preprocessing*

Branch	Input Form	Preprocessing Operations
Lexical	Cleaned text	Lowercase, lemmatization, stopword removal, non-alphabetic character removal
Semantic	Raw text	No preprocessing (preserves sentence structure)
Contextual	Raw text	Tokenization via BERT tokenizer, truncation to 128 tokens

Table 7. *Ensemble Fusion Strategies*

Method	Description	Configuration
Soft Voting	Mean probability aggregation	Equal weight (1/3) per branch

Stacking	Meta-classifier fusion	Logistic Regression (C=1.0, max_iter=2000-4000, solver='lbfgs')

In stacking, the meta-classifier receives concatenated class probability vectors from all three branches as input features. For a K-class problem, the input dimensionality is 3K.Multi-Label Task Adaptation

For multi-label emotion classification (SemEval 2018 Task 1 E-c), the following adaptations were applied without modifying the core architecture, as shown in Table 8.

Table 8. *Multi-Label Configuration*

Component	Adaptation	Implementation Details
Random Forest	Multi-output classifier	Independent trees per label (MultiOutputClassifier wrapper)
XGBoost	Multi-output classifier	11 separate binary models (one per emotion) with scale_pos_weight for class imbalance
BERT	Multi-label classification head	BCEWithLogitsLoss, sigmoid activation
Decision strategy	Label-wise threshold optimization	Grid search (0.10–0.90, step=0.01) on validation set
Stacking	Multi-label meta-classifier	OneVsRestClassifier with Logistic Regression
Evaluation	Multi-label metrics	Jaccard accuracy, micro/macro F1, label-wise precision/recall

Table 9. *Evaluation metrics used across different emotion classification paradigms*

Evaluation Category	Metrics / Tests Used	Applicable Datasets
Basic Classification Metrics	Accuracy, Precision, Recall, F1-score	All binary and multi-class datasets
Discriminative Performance	ROC–AUC	All binary and multi-class datasets
Agreement-Based Metrics	Cohen's Kappa (κ)	Multi-class emotion datasets
	Matthews Correlation Coefficient (MCC)	Binary and multi-class datasets with class imbalance
Multi-Label Metrics	Jaccard Accuracy (Jaccard Index)	SemEval 2018 Task 1 E-c
	Micro / Macro Precision, Recall, F1-score	SemEval 2018 Task 1 E-c
Statistical Significance Testing	McNemar's Test	Binary and multi-class datasets (where applicable)

-F1 performance of lexical, semantic, contextual, and ensemble-based models across these datasets. The results indicate that FOTE consistently achieves superior performance compared to individual representation paradigms, including transformer-based contextual models. Notably, the improvements are more pronounced in datasets characterized by noisy language and sparse emotional cues, where explicit lexical features complement contextual embeddings. Figure 5 further illustrates the comparative performance trends, highlighting the progressive gains obtained when moving from single-representation models to the integrated FOTE framework. These results demonstrate that representation-level fusion effectively enhances emotion discrimination beyond isolated modeling approaches.

5.2 Results on Benchmark Datasets from Batbaatar et al. [56]

To assess robustness across fine-grained emotion categorization tasks, the proposed framework is evaluated on multiple benchmark datasets derived from Batbaatar et al. [56]. These datasets span conversational dialogues, narrative texts, survey responses, and social media content, with varying class cardinalities and domain characteristics.Table 11 reports the consolidated performance comparison between transformer-based baselines and the proposed FOTE framework across these datasets. In addition to performance metrics,statistical significance is assessed using McNemar's test where applicable. The results show that FOTE consistently outperforms the contextual baseline across most datasets, with statistically significant improvements observed in several cases. Figure 6 depicts the relative performance improvement of FOTE-Stacking over the contextual baseline across all datasets in this benchmark. The observed improvements indicate that integrating lexical and semantic representations with contextual modeling leads to more stable and discriminative emotion

prediction, particularly in datasets with overlapping emotional expressions and class imbalance.

Table 10. *Consolidated performance on Yousaf et al. [55] benchmark datasets*

Dataset	Lexical F1	SBERT F1	BERT F1	FOTE F1	BERT → FOTE (% Gain)
Twitter Sentiment	0.7241	0.7690	0.8118	0.8226	+1.33%
Women's Clothing	0.2140	0.3250	0.5070	0.5209	+2.74%
Twitter Hatred	0.7956	0.8074	0.8650	0.8801	+1.75%

Table 11. *Performance comparison between BERT and FOTE across Batbaatar et al. [56] datasets*

Dataset	Classes	Domain	BERT F1	FOTE F1	Improvement	McNemar p
DailyDialog	6	Conversational	0.6232	0.6629	+6.37%	0.0086*
CrowdFlower	6	Social Media	0.3552	0.3633	+2.28%	0.0288*
ISEAR	5	Survey	0.8098	0.8142	+0.54%	0.8450
TEC	6	Social Media	0.5493	0.5812	+5.81%	<0.0001**
Tales-Emotion	6	Literary	0.5819	0.6094	+4.73%	0.4731
EmoInt	4	Social Media	0.8779	0.9060	+3.20%	<0.0001**
Emotion-Cause	6	News	0.9811	0.9811	+0.00%	1.0

** indicates p < 0.05 and ** indicates p < 0.0001 statistical significance over the BERT baseline*

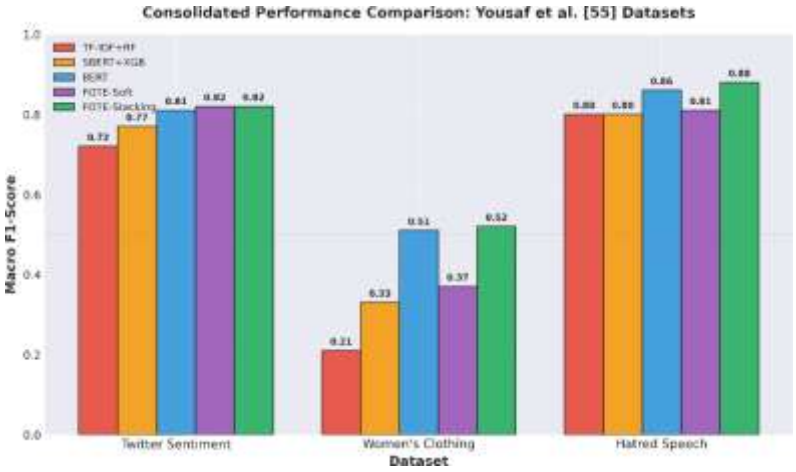


Figure 5. *Comparative performance of lexical, semantic, contextual, and FOTE models on Yousaf et al. [55] datasets*

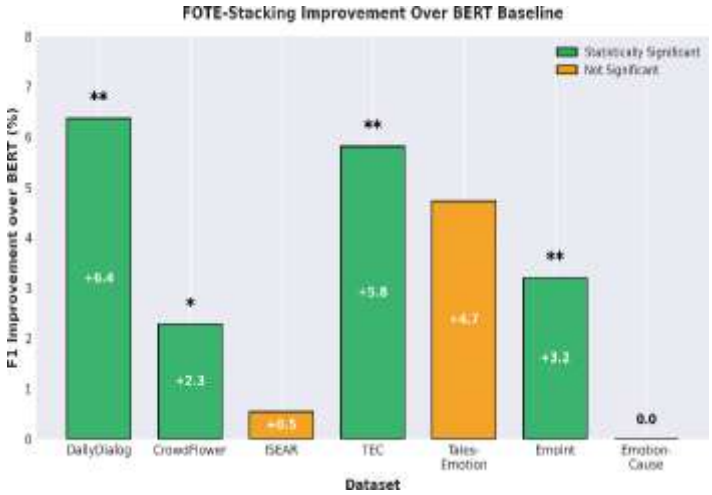


Figure 6. *Relative improvement of FOTE over BERT across Batbaatar et al. [56] datasets*
** denotes p < 0.05 and ** denotes p < 0.01 statistical significance over the BERT baseline*

5.3 Results on Additional Validation Datasets

Beyond benchmark evaluations, the proposed framework is validated on additional datasets to examine its generalization capability across domains and task formulations. These datasets include movie reviews, academic citation sentiment, extended survey-based emotion corpora, and a multi-label emotion classification benchmark. Table 12 summarizes the

performance of FOTE on these additional validation datasets. The results demonstrate that the framework maintains competitive and stable performance across binary, multi-class, and multi-label settings.For the multi-label emotion classification task, evaluation is conducted using Jaccard Accuracy along with micro- and macro-averaged metrics, reflecting the framework's ability to capture overlapping emotional states without architectural modification.For the multi-label SemEval 2018 dataset, Macro F1-score is reported as the primary evaluation metric, while Jaccard Accuracy is additionally computed for complementary analysis. These findings confirm that the proposed framework generalizes effectively beyond benchmark datasets and adapts seamlessly to diverse emotion detection scenarios.

Table 12. *Performance of FOTE on additional validation datasets*

Dataset	Type	Classes	BERT F1	FOTE F1	Improvement	McNemar p
IMDB [51]	Binary	2	0.8925	0.9294	+4.13%	9.2E-24**
Citation Sentiment [13]	Multi-class	3	0.5503	0.6002	+9.07%	---
ISEAR 7-class [52]	Multi-class	7	0.7031	0.7498	+6.64%	3.3E-05**
SemEval 2018 [53]	Multi-label	11	---	0.5833	---	---

** *indicates p < 0.001*

5.4 Cross-Dataset Performance Analysis

To provide an aggregated view of model behavior across all experimental settings, Table 13 presents a master performance summary encompassing all fourteen datasets. This table consolidates results across different task types and domains, offering a holistic comparison of representation paradigms.

Additionally, Figure 7 provides a comprehensive visual comparison of model performance across all datasets. The figure highlights the consistent advantage of the proposed framework over individual lexical, semantic, and contextual models, reinforcing the robustness of representation-level integration.

5.5 Statistical Significance and Branch Contribution Analysis

To validate whether the observed performance gains are statistically meaningful, McNemar's test is employed in selected comparative settings. The statistical analysis confirms that the improvements achieved by FOTE over the contextual baseline are significant for most applicable binary and multi- class datasets, supporting the reliability of the reported results.Furthermore, a branch contribution analysis is conducted to examine the impact of individual representation components within the FOTE framework. The analysis shows that removing any representation branch leads to a noticeable reduction in performance, confirming that lexical, semantic, and contextual representations contribute complementary information. This observation substantiates the design motivation of the proposed framework and justifies the inclusion of multiple representation paradigms within a unified ensemble structure.

5.6 Summary of Results

Overall, the experimental results demonstrate that the proposed FOTE framework consistently outperforms individual representation-based models
across diverse datasets and task formulations. The integration of lexical, semantic, and contextual representations at the probability level enhances robustness, generalization, and statistical reliability in text-based emotion detection.

Table 13. *Master performance summary across all fourteen datasets*

Dataset	Type	Classes	Domain	BERT F1	FOTE F1	Improvement
Twitter Sentiment	Binary	2	Social Media	0.8118	0.8226	+1.33%
Women's Clothing	Multi-class	5	E-commerce	0.5070	0.5209	+2.74%
Twitter Hatred	Binary	2	Social Media	0.8650	0.8801	+1.75%
DailyDialog	Multi-class	6	Conversational	0.6232	0.6629	+6.37%
CrowdFlower	Multi-class	6	Social Media	0.3552	0.3633	+2.28%
ISEAR-5	Multi-class	5	Survey	0.8098	0.8142	+0.54%
TEC	Multi-class	6	Social Media	0.5493	0.5812	+5.81%
Tales-Emotion	Multi-class	6	Literary	0.5819	0.6094	+4.73%
EmoInt	Multi-class	4	Social Media	0.8779	0.9060	+3.20%
Emotion-Cause	Multi-class	6	News	0.9811	0.9811	0.00%
IMDB [51]	Binary	2	Movie Reviews	0.8925	0.9294	+4.13%
Citation Sentiment [13]	Multi-class	3	Academic	0.5503	0.6002	+9.07%
ISEAR-7 [52]	Multi-class	7	Survey	0.7031	0.7498	+6.64%
SemEval 2018 E-c [53]	Multi-label	11	Social Media	---	0.5833	---

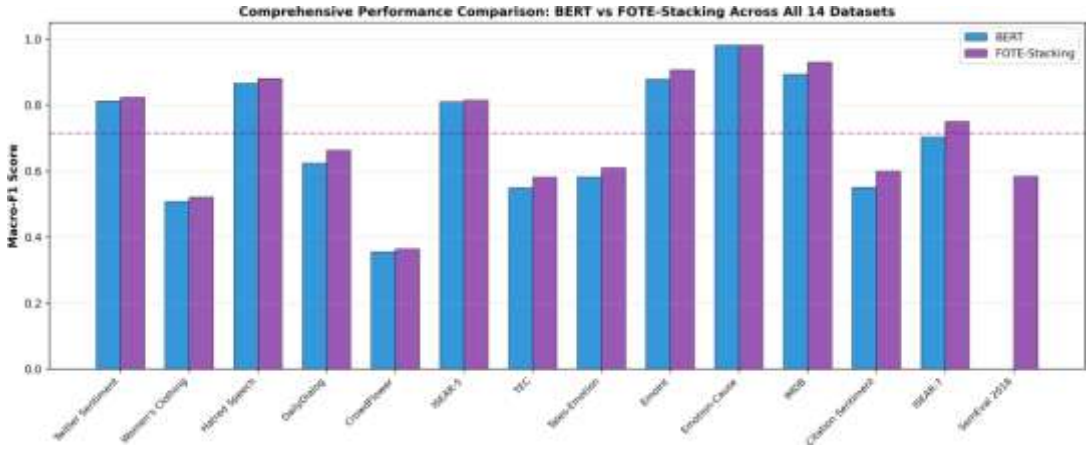


Figure 7. *Overall performance comparison of FOTE across all evaluated data sets*

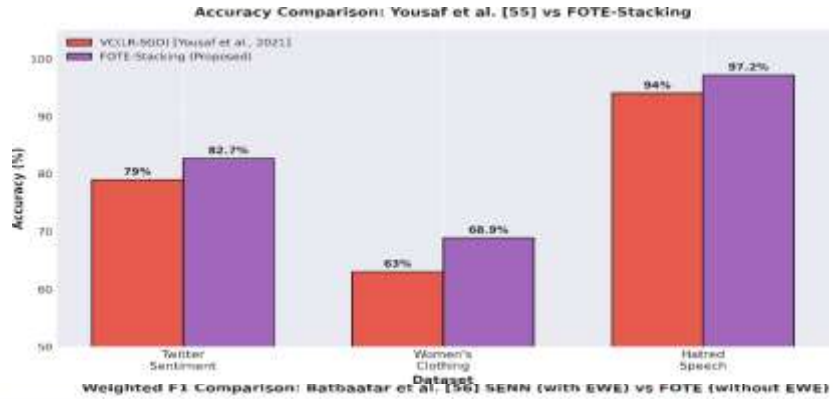
Table 14. *Accuracy Comparison with Yousaf et al. [55]*

Dataset	Yousaf et al. (2020) [55] VC(LR-SGD)	Proposed Approach (FOTE-Stack)	Improvement
Twitter Sentiment	79.00%	82.65%	+3.65%
Women's Clothing	63.00%	68.88%	+5.88%
Hatred Speech	94.00%	97.15%	+3.15%

Table 15. *Weighted F1 Comparison with Batbaatar et al. [56]*

Dataset	Batbaatar et al. (2019) [56] SENN + EWE	Proposed Approach (FOTE-Stack)	Difference
DailyDialog	84.8%	87.1%	+2.3%
CrowdFlower	51.1%	54.2%	+3.1%
TEC	61.3%	65.6%	+4.3%
Tales-Emotion	62.3%	64.1%	+1.8%
ISEAR-5	74.6%	81.4%	+6.8%
EmoInt	91.0%	90.5%	-0.5%
Emotion-Cause	98.8%	98.7%	-0.1%

Figure 8. *Yousaf et al. [55] Accuracy Comparison: VC(LR-SGD) vs FOTE-Stacking*



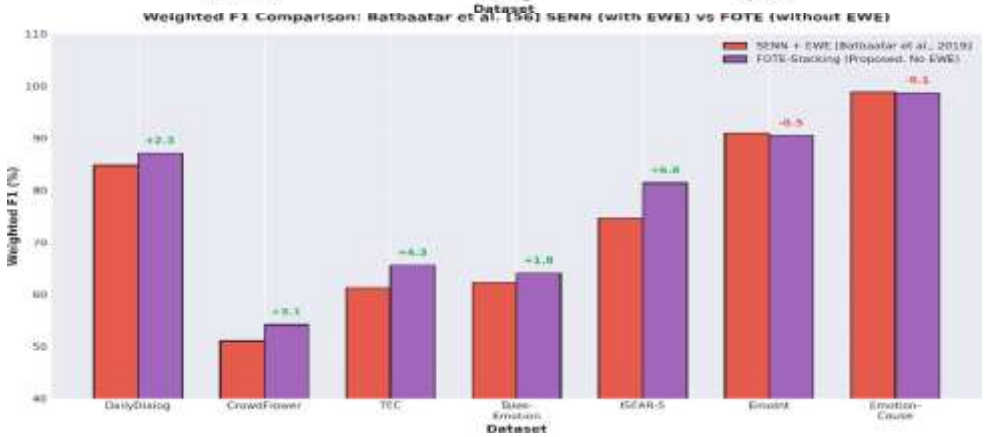


Figure 9. *Batbaatar et al. [56] Comparison: SENN (with EWE) vs FOTE (without EWE)*

Table 16. *Cross-domain performance comparison between reference studies and FOTE-Stacking*

Dataset	Metric	Best Reference Study	Best Reported Value	Proposed Approach (FOTE-Stack)
IMDB	Accuracy	Dang et al. [23]	86.69%	**92.94%**
ISEAR-7	Accuracy	Acheampong et al. [25]	74.31%	**74.97%**
SemEval-2018 E-c	Jaccard Accuracy	NTUA-SLP [54]	58.80%	**58.85%**
Citation Sentiment	Macro-F1	Raza et al. [13]	49.00%	**60.00%**

6. Discussion

This section interprets the experimental results reported in Section 5 and discusses their implications in the context of existing emotion detection research. Across fourteen datasets, the proposed Feature-Oriented Textual Ensemble (FOTE) demonstrates consistent improvements over individual lexical, semantic, and contextual models. The framework achieves superior performance on 13 out of 14 datasets, confirming that heterogeneous representation integration is more effective than single-paradigm modeling. Performance gains range from 0.54% to 9.07%, with the largest improvement observed on the Citation Sentiment dataset (+9.07%) and substantial gains on DailyDialog (+6.37%) and ISEAR-7 (+6.64%), indicating the effectiveness of learned stacking in conversational emotion analysis. Stacking consistently outperforms soft voting, as it learns adaptive weights for each representation instead of assuming equal importance. This explains the substantial gains on complex datasets, where contextual and semantic cues dominate over surface lexical patterns.

6.1 Comparison with Yousaf et al. [55]

Compared to the classical machine-learning ensemble reported by Yousaf et al. [55], FOTE- Stacking achieves accuracy improvements of 3.65%, 5.88%, and 3.15% on Twitter Sentiment, Women's Clothing (5-class), and Hatred Speech datasets, respectively (Table 14, Figure 8).While the reference study primarily relied on accuracy, the present work additionally reports Cohen's Kappa, MCC, ROC-AUC, and McNemar's test, providing a more reliable assessment of agreement, correlation, and statistical significance. This highlights that the observed gains are not only numerical but also statistically meaningful and balanced across classes.Comparison with Batbaatar et al. [56].Batbaatar et al. [56] introduced SENN with Emotion Word Embeddings (EWE), a domain-specific resource. FOTE, despite not using EWE, outperforms SENN on five of seven datasets and achieves near-equivalent performance on EmoInt and Emotion-Cause datasets (Table 15, Figure 9).The marginal differences on these two datasets (−0.5% and −0.1%) indicate performance saturation rather than model limitation, as both datasets already exhibit near-ceiling scores. This result demonstrates that principled ensemble learning can match or exceed domain-specific feature engineering while offering greater portability. Reference study primarily relied on Weighted F1.

6.2 Cross-Domain Generalization

To evaluate generalization beyond the benchmark datasets, the proposed framework was further assessed on four heterogeneous datasets representing distinct application settings: IMDB (reviews), ISEAR-7 (psychological survey data), SemEval-2018 Task 1 E-c (multi-label emotion classification), and Citation Sentiment (academic discourse). These datasets differ substantially in domain characteristics, class granularity, and label distributions, thereby providing a rigorous test of cross-domain robustness.Table 16 summarizes the best reported results from existing reference studies for each dataset and compares them with the performance of FOTE-Stacking. On the IMDB dataset, FOTE-Stacking achieves an accuracy of 92.94%, exceeding prior best-performing models such as the RNN-based approach of Dang et al. [23], while also reporting strong agreement and correlation measures (Kappa = 0.859, MCC = 0.859) along with a statistically significant McNemar p- value (9.20×10^{-24}). For the ISEAR-7 dataset, FOTE-Stacking attains an accuracy of 74.97% and Macro-F1 of 74.98%, marginally surpassing the RoBERTa-based model reported by Acheampong etal. [25]. Importantly, the proposed framework additionally reports reliability measures that were not included in earlier studies, indicating stronger agreement beyond chance in a challenging seven- class emotion classification task. In the SemEval- 2018 multi-label task, FOTE-Stacking achieves a Jaccard Accuracy of 58.85% and competitive Micro- F1, slightly exceeding the shared-task winning system NTUA-SLP [54]. This result is noteworthy as the winning system employed task-specific affective features, whereas FOTE relies on a general-purpose ensemble architecture without dataset-specific modifications. On the Citation Sentiment Corpus, although accuracy improvements over classical methods are modest, Macro-F1 improves from 49% to 60%, corresponding to a 22.4% relative improvement in minority sentiment detection. This improvement is further supported by increased agreement and correlation values, highlighting the effectiveness of balanced representation integration in highly imbalanced academic text data.Overall, the results in Table 16

demonstrate that the proposed framework maintains competitive or superior performance across diverse domains and task formulations, confirming that the observed improvements are not limited to a specific dataset or emotion taxonomy.

6.3 Research Contributions

The main contributions of this work are summarized below:

1. **Integrated ensemble framework:**
This work presents a Feature-Oriented Textual Ensemble (FOTE) framework that combines lexical, semantic, and contextual representations at the probability level, allowing different types of emotional cues to be captured within a single model.
2. **Consistent fusion strategy across task types:** Soft voting and stacking-based fusion strategies are used to integrate outputs from heterogeneous representations, and the same framework is applied consistently to binary, multi-class, and multi-label emotion classification tasks without changing the model structure.**Domain-independent applicability:** Experimental results across diverse datasets show that the proposed framework is not tied to any specific emotion domain or taxonomy, and can generalize across different forms of emotion expression without relying on domain-specific feature engineering.
3. **Analysis of representation-level contributions:** A detailed comparison of lexical features, sentence-level semantic embeddings, and contextual representations is carried out, highlighting how each representation contributes differently and how their combination improves overall performance.
4. **Extensive validation across datasets:**
The framework is evaluated on fourteen publicly available datasets covering social media text, reviews, conversational data, surveys, and academic writing, demonstrating consistent improvements over individual representation- based models and strong transformer-based baselines.
5. **Improved evaluation methodology:**
In addition to standard accuracy and F1-score, the evaluation includes agreement measures, correlation metrics, and statistical significance testing, providing a more reliable and comprehensive assessment of emotion detection performance.

6. Conclusion and Future Scope

This work investigated text-based emotion detection through heterogeneous ensemble learning and presented a Feature-Oriented Textual Ensemble (FOTE) framework that integrates lexical, semantic, and contextual representations using probability-level fusion. The framework addresses limitations of single-representation and domain-specific approaches by enabling complementary emotion cues to be captured within a unified architecture. Extensive evaluation on fourteen publicly available datasets, covering binary, multi-class, and multi- label emotion classification across diverse domains, demonstrates the effectiveness and generalizability of the proposed approach. The results show that FOTE achieves consistent performance improvements over strong transformer-based baselines on 13 out of 14 datasets, with stacking- based fusion yielding the most stable gains. The analysis further confirms that lexical, semantic, and contextual representations contribute distinct and complementary information, leading to robust emotion detection across varied textual settings. Despite these strengths, the study is limited to English-language datasets and incurs higher computational cost due to multi-model training. In addition, temporal and conversational dynamics are not explicitly modelled. Future work will focus on extending the framework to multilingual settings, improving efficiency through lightweight or distilled models, and incorporating interpretability mechanisms. Further directions include adapting the ensemble principle to multimodal emotion analysis and dynamic learning scenarios, enabling broader applicability in real- world, human-centric AI systems.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions

References

[1] Jiang, D., Luo, X., Xuan, J., & Xu, Z. 2017).Sentiment computing for the news event based on thesocial media big data. *IEEE Access*, *5*, 2373–2382. https://doi.org/10.1109/ACCESS.2016.2607218

[2] Sreeja, P. S., & Mahalakshmi, G. S. (2016). Comparison of probabilistic corpus based method and vector space model for emotion recognition from poems. Asian Journal of Information Technology, 15(5), 908–915.https://doi.org/10.36478/ajit.2016.908.915

[3] Povoda, L., Burget, R., Masek, J., Uher, V., & Dutta,M. K. (2016, September). Optimization methods in emotion recognition system. *Radioengineering*, *25*(3), 565–572.

https://doi.org/10.13164/re.2016.0565

[4] Douiji, Y., Mousannif, H., & Al Moatassime, H. (2016). Using YouTube comments for text-based emotion recognition. *Procedia Computer Science*, *83*, 292–299. https://doi.org/10.1016/j.procs.2016.04.128

[5] Zhou, D., Zhang, X., Zhou, Y., Zhao, Q., & Geng, X. (2016). Emotion distribution learning from texts. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 638–647). https://doi.org/10.18653/v1/D16-1061

[6] Perikos, I., & Hatzilygeroudis, I. (2016, May). Recognizing emotions in text using ensemble of classifiers. *Engineering Applications of Artificial Intelligence*, *51*, 191–201. https://doi.org/10.1016/j.engappai.2016.01.012

[7] Bravo-Marquez, F., Frank, E., Mohammad, S. M., & Pfahringer, B. (2016, October). Determining word-emotion associations from tweets by multi-label classification. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 536–539). https://doi.org/10.1109/WI.2016.0091

[8] Bandhakavi, A. S., Wiratunga, N., Massie, S., & Padmanabhan, D. (2017, January–February). Lexicon generation for emotion detection from text. *IEEE Intelligent Systems*, *32*(1), 102–108. https://doi.org/10.1109/MIS.2017.22

[9] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

[10] Čuš, N., Robnik-Šikonja, M., & Demšar, J. (2018). Emotion recognition on Twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2018.2807817

[11] Agrawal, A., An, A., & Papagelis, M. (2018). Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)* (pp. 950– 961).

[12] Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2018). Sentence level sentiment identification and calculation from news articles using machine learning techniques. In *Computing, Communication and Signal Processing: Proceedings of ICCASP 2018* (pp. 371–376). Singapore: Springer.

[13] Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications*, *10*(12), 157–165.

[14] Hasan, M., Rundensteiner, E. A., & Agu, E. (2019, March). Automatic emotion detection in text streams by analyzing Twitter data. *International Journal of Data Science and Analytics*, *7*(1), 35–51. https://doi.org/10.1007/s41060-018-0096-z

[15] Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019).Detection of depression-related posts in Reddit social media forum. *IEEE Access*, *7*, 44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180

[16] Shrivastava, K., Kumar, S., & Jain, D. K. (2019). An effective approach for emotion detection in multimedia text data using sequence-based convolutional neural network. *Multimedia Tools and Applications*, *78*(20), 29607–29639. https://doi.org/10.1007/s11042-019-07813-9

[17] Alsheikh, S. S., Shaalan, K., & Meziane, F. (2019). Exploring the effects of consumers' trust: A predictive model for satisfying buyers' expectations based on sellers' behavior in the marketplace. *IEEE Access*, *7*, 73357–73372.

[18] Kim, H., & Jeong, Y.-S. (2019). Sentiment classification using convolutional neural networks. *Applied Sciences*, *9*(11), Article 2347. https://doi.org/10.3390/app9112347

[19] Abas, A. R., El-Henawy, I., Mohamed, H., & Abdellatif, A. (2020). Deep learning model for fine-grained aspect-based opinion mining. *IEEE Access*, *8*, 128845–128855.https://doi.org/10.1109/ACCESS.2020.3009487

[20] Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok,K. (2020, October). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)* (pp. 105–114). https://doi.org/10.1145/3340531.3412003

[21] Adoma Acheampong, F. A., Nunoo-Mensah, H., Chen, W., & Andre, N. R. (2020, December). Recognizing emotions from texts using a BERT-based approach. In *Proceedings of the IEEE 17th International Computing Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 62–66).

[22] Acheampong, F. A., Chen, W., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), Article e12189. https://doi.org/10.1002/eng2.12189

[23] Dang, N. C., Moreno-García, M. N., & De la Prieta,F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, *9*(3), Article 483. https://doi.org/10.3390/electronics9030483

[24] Asghar, M. Z., Subhan, F., Imran, M., Kundi, F. M., Khan, A., Shamshirband, S., Mosavi, A., Csiba, P., & Várkonyi-Kóczy, A. R. (2020). Performance evaluation of supervised machine learning techniques for efficient detection of emotions from online content. *Computers, Materials & Continua*, *63*(3), 1093–1118.https://doi.org/10.32604/cmc.2020.07709

[25] Adoma Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2020, December). Comparative analyses of BERT, RoBERTa, DistilBERT, and XLNet for text-based emotion recognition. In *Proceedings of the IEEE 17th International Computing Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 117– 121).

[26] Singh, G., Brahma, D., Rai, P., & Modi, A. (2021). Fine-grained emotion prediction by modeling emotion definitions. In *Proceedings of the IEEE 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1–8).

[27] Luo, J., Bouazizi, M., & Ohtsuki, T. (2021). Data augmentation for sentiment analysis using sentence compression-based SeqGAN with data screening. *IEEE*

Access, 9, 99922–99931.https://doi.org/10.1109/ACCESS.2021.3094023

[28] Dobša, J., Šebalj, D., & Bužić, D. (2021). Classification of emotions based on text and qualitative variables. In *Proceedings of the IEEE 44th International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 383–388).

[29] Li, R., Liu, M., Xu, D., Gao, J., Wu, F., & Zhu, L.(2022). A review of machine learning algorithms for text classification. In *Cyber Security* (Communications in Computer and Information Science, Vol. 1506, pp. 181–194). Springer. https://doi.org/10.1007/978-981-16-9229-1_14

[30] Bharti, S. K., Varadhaganapatby, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022). Text-based emotion recognition using deep learning approach. *Computational Intelligence and Neuroscience, 2022*, Article 2645381. https://doi.org/10.1155/2022/2645381

[31] Tu, G., Wen, J., Liu, C., Jiang, D., & Cambria, E. (2022). Context- and sentiment-aware networks for emotion recognition in conversation. *IEEE Transactions on Artificial Intelligence, 3*(5), 699–708. https://doi.org/10.1109/TAI.2022.3190243

[32] Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A hybridmodel for sentiment analysis with transformer and recurrent neural network. *IEEE Access, 10*, 21517–21525.https://doi.org/10.1109/ACCESS.2022.3152828

[33] Habib, M. A., Akhand, M. A. H., & Kamal, M. A. S. (2022, December). Emotion recognition from microblog managing emoticon with text and classifying using 1D CNN. *Journal of Computer Science, 18*(12), 1170–1178.https://doi.org/10.3844/jcssp.2022.1170.1178

[34] Adak, A., Pradhan, B., & Shukla, N. (2022). Sentiment analysis of customer reviews of food delivery services using deep learning and explainable artificial intelligence: Systematic review. *Foods, 11*(10), Article 1500. https://doi.org/10.3390/foods11101500

[35] Frye, R. H., & Wilson, D. C. (2022). Comparative analysis of transformers to support fine-grainedemotion detection in short-text data. In *Proceedings of the International FLAIRS Conference* (Vol. 35).

[36] Tesfagergish, S. G., Kapočiūtė-Dzikienė, J., & Damaševičius, R. (2022). Zero-shot emotion detection for semi-supervised sentiment analysis using sentence transformers and ensemble learning. *Applied Sciences, 12*(17), Article 8662. https://doi.org/10.3390/app12178662

[37] Yohanes, D., Putra, J. S., Filbert, K., Suryaningrum,K. M., & Saputri, H. A. (2023). Emotion detection in textual data using deep learning. *Procedia Computer Science, 227*, 464–473. https://doi.org/10.1016/j.procs.2023.10.547

[38] Machová, K., Szabóová, M., Paralič, J., & Mičko, J. (2023). Detection of emotion by text analysis using machine learning. *Frontiers in Psychology, 14*, Article 1190326. https://doi.org/10.3389/fpsyg.2023.1190326

[39] Hadikhah Mozhdehi, M., & Moghadam, A. E. (2023). Textual emotion detection utilizing a transfer learning approach. *The Journal of Supercomputing, 79*(12), 13075–13089. https://doi.org/10.1007/s11227-023-05168-5

[40] de León Languré, A., & Zareei, M. (2023). Breaking barriers in sentiment analysis and text emotion detection: Toward a unified assessment framework. *IEEE Access, 11*, 125698–125715.https://doi.org/10.1109/ACCESS.2023.3331323

[41] Sinha, A., & Chakma, K. (2023). A comparative analysis of sentiment analysis and emotion detection problems on texts. In Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (pp. 465–475). Singapore: Springer Nature Singapore.

[42] Saffar, A. H., Mann, T. K., & Ofoghi, B. (2023). Textual emotion detection in health: Advances and applications. Journal of Biomedical Informatics, 137, Article 104258. https://doi.org/10.1016/j.jbi.2022.104258

[43] Yao, L., & Zheng, N. (2024). Sentiment analysis based on improved transformer model and conditional random fields. IEEE Access, 12, 90145–90157.https://doi.org/10.1109/ACCESS.2024.3418847

[44] de León Languré, A., & Zareei, M. (2024). Evaluating the effect of emotion models on the generalizability of text emotion detection systems. IEEE Access, 12, 70489–70500.https://doi.org/10.1109/ACCESS.2024.3401203

[45] de León Languré, A., & Zareei, M. (2024). Improving text emotion detection through comprehensive dataset quality analysis. IEEE Access, 12, 166512–166536.https://doi.org/10.1109/ACCESS.2024.3491856

[46] He, A., & Abisado, M. (2024). Text sentiment analysis of Douban film short comments based on BERT-CNN-BiLSTM-Att model. IEEE Access, 12, 45229–45237.https://doi.org/10.1109/ACCESS.2024.3381515

[47] Sharma, H., Ud Din, F., & Ogunleye, B. (2024). Electric vehicle sentiment analysis using large language models. Analytics, 3(4), 425–438. https://doi.org/10.3390/analytics3040023