SentiFuse: Deep Multi-model Fusion Framework for Robust Sentiment Extraction

Hieu Minh Duong[1]✉, Rupa Ghosh[1], Cong Hoan Nguyen[1], Eugene Levin[2], Todd Gary[2], and Long Nguyen[1]

[1] University of Louisville, Louisville, Kentucky, USA
{hieu.duong,rupa.ghosh,conghoan.nguyen,l.nguyen}@louisville.edu
[2] Meharry Medical College, Nashville, Tennessee, USA
{tpgary,elevin}@mmc.edu

Abstract. Sentiment analysis models exhibit complementary strengths, yet existing approaches lack a unified framework for effective integration. We present SentiFuse, a flexible and model-agnostic framework that integrates heterogeneous sentiment models through a standardization layer and multiple fusion strategies. Our approach supports decision-level fusion, feature-level fusion, and adaptive fusion, enabling systematic combination of diverse models. We conduct experiments on three large-scale social-media datasets: Crowdflower, GoEmotions, and Sentiment140. These experiments show that SentiFuse consistently outperforms individual models and naive ensembles. Feature-level fusion achieves the strongest overall effectiveness, yielding up to 4% absolute improvement in F1 score over the best individual model and simple averaging, while adaptive fusion enhances robustness on challenging cases such as negation, mixed emotions, and complex sentiment expressions. These results demonstrate that systematically leveraging model complementarity yields more accurate and reliable sentiment analysis across diverse datasets and text types.

Keywords: sentiment analysis · model fusion · text classification · natural language processing.

1 Introduction

Large Language Models (LLMs) and deep learning architectures have transformed sentiment analysis by capturing semantic dependencies and contextual nuances beyond the reach of traditional approaches. Rule-based systems such as VADER [15] offer interpretability and efficiency but are brittle in the presence of sarcasm, domain-specific language, or complex modifiers. Statistical models provide lightweight computation but cannot adequately capture sequential dependencies and contextual relationships [12]. In contrast, deep learning models excel in modeling semantics but demand large labeled datasets and risk overfitting [24]. These complementary strengths and weaknesses are especially evident in social media, where sentiment expressions are highly contextual and often

include irony, mixed emotions, or pragmatic cues that challenge single-model methods [22, 25]. For instance, the sentence "Great, another delayed flight – exactly what I needed today!" is likely to be misclassified by a lexicon-based model as positive due to the words great and needed, whereas a context-sensitive model could recognize the sarcastic intent. Such cases highlight both the limitations of individual sentiment analysis approaches and the potential gains from systematically integrating heterogeneous models to leverage their complementary perspectives.

Despite the promise of ensemble learning, existing sentiment analysis fusion methods are typically simplistic and ad hoc. Majority voting, unweighted averaging, and confidence-weighted schemes assume homogeneous base models with compatible outputs and aligned decision boundaries [16, 21]. These naive methods neglect fundamental challenges in heterogeneous model integration, such as inconsistencies in output formats (probabilities, logits, discrete labels), disparities in confidence calibration, and the lack of adaptive mechanisms that adjust to varying linguistic complexities. To address these gaps, we propose SentiFuse, a model-agnostic framework that integrates heterogeneous sentiment models through standardized output processing and multiple fusion strategies. Our work focuses on three Research Questions: ***Does systematic fusion outperform naive combination methods?*** (RQ1), ***How does fusion effectiveness vary across text characteristics?*** (RQ2), ***Does the framework generalize across different model combinations?*** (RQ3). To this end, our contributions are threefold.

- First, we introduce a standardization layer that unifies diverse outputs into probability distributions, enabling integration without architectural modifications.
- Second, we design three complementary fusion strategies—decision-level fusion with learned weights, feature-level fusion using meta-classification, and adaptive fusion guided by automatically extracted text characteristics.
- Finally, we conduct a comprehensive evaluation based on three Research Questions, demonstrating that SentiFuse improves performance over both standalone models and traditional ensembles across diverse and complex sentiment scenarios.

2 Related Work

Early sentiment analysis relied on statistical methods such as TF-IDF vectorization with traditional classifiers [4, 6, 9, 26], which were efficient but ignored sequential dependencies and semantics. Rule-based systems like VADER [15] and SentiWordNet [1] improved interpretability with linguistic heuristics, yet struggled with context sensitivity and domain adaptation [5, 11, 23, 29]. Deep learning models [2, 27, 31, 33] addressed these issues: recurrent architectures such as BiLSTM captured long-range dependencies [32], while attention mechanisms highlighted sentiment-bearing spans [34]. Transformer-based models, notably BERT [8] and RoBERTa [18], achieved state-of-the-art results with bidirectional

context encoding, though at the cost of data demands and sensitivity to domain shifts.

Beyond single models, ensemble methods have been widely applied. Traditional approaches like bagging [3] and voting [10] improved robustness but failed to fully exploit heterogeneous systems. More recent meta-ensembles [17] and adaptive fusion strategies [13] dynamically weight model outputs or modalities, yielding stronger results but often requiring high computational resources or paired multimodal data. Meanwhile, the emergence of large language models (LLMs) has shifted sentiment analysis research: studies report ChatGPT achieving competitive or superior performance in zero- and few-shot scenarios [30] compared to fine-tuned transformers [19]. Cross-lingual ensembles also demonstrate strong results by combining translation with transformer ensembles [20], though error propagation remains a challenge. Recent surveys emphasize that many fusion approaches still rely on static weighting and lack systematic evaluation against simpler baselines [28], which underscores the need for model-agnostic, adaptive frameworks.

3 Methodology

In this study, we propose an innovative multi-model framework, SentiFuse, designed to integrate multiple heterogeneous sentiment analysis models. Our proposed sentiment analysis framework is composed of four interconnected components: (1) multiple heterogeneous sentiment analysis models, (2) a standardization layer, (3) a fusion strategy selector, and (4) sentiment classification. This structured approach enables flexible integration and coherent combination of sentiment predictions from diverse model types, ranging from simple lexicon-based methods to complex language models.

Multi-model Sentiment Analysis. The first component consists of a diverse set of sentiment models represented as:

$$M = M_1, M_2, ..., M_n. \tag{1}$$

Each model M_i processes input text to produce an output $O_i = M_i(x)$. Our implementation specifically includes:

- **Lexicon-based models:** Estimate sentiment by aggregating word-level polarity from curated lexicons, typically with heuristics for intensity, positional emphasis, and valence shifters (e.g., negation).
- **Pattern-based models:** Infer sentiment from the presence and strength of predefined sentiment-bearing patterns (e.g., idioms, emoji sequences, dependency templates), weighting patterns by frequency and reliability.
- **Machine-learning models:** Learn a mapping from engineered features (e.g., n-grams, TF–IDF, syntactic cues, lexicon features) to sentiment labels using supervised classifiers (e.g., logistic regression, SVM).

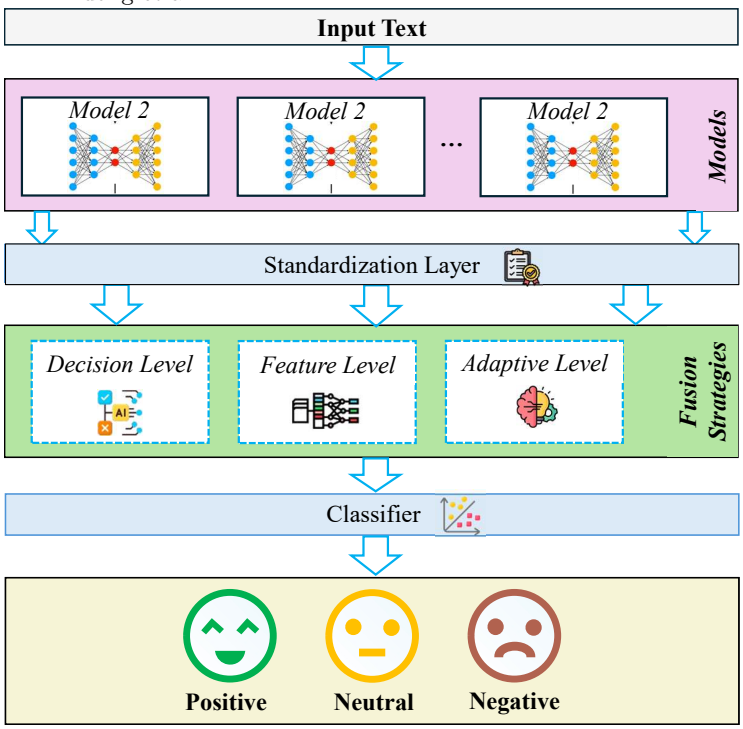

Fig. 1. Overall architecture of the proposed SentiFuse framework.

- **Encoding models:** Use deep contextual encoders (e.g., BERT, RoBERTa) to obtain sequence representations; a pooled vector (e.g., [CLS] token) is passed to a task-specific classifier after pretraining and fine-tuning.

We distinguish classical classifiers trained on engineered features from neural encoding models that learn contextual representations; for classification, the encoder is paired with a task-specific prediction head and fine-tuned.

Standardization Layer. To facilitate seamless integration of heterogeneous model outputs, we implement a standardization function S that converts different model outputs into unified probability distributions over sentiment classes as:

$$S(O_i) = \begin{cases} \{p_{\text{pos}}, p_{\text{neg}}\}, & \text{if output is probability} \\ \left\{\frac{1+s}{2}, \frac{1-s}{2}\right\}, & \text{if output is score } s \in [-1,1] \\ \{\sigma(v_{\text{pos}}), \sigma(v_{\text{neg}})\}, & \text{if output is logits} \end{cases}. \quad (2)$$

Additionally, we define feature extraction functions $\phi_i(O_i)$ specific to each model type as:

$$\phi_i(O_i) = [f_1(O_i), f_2(O_i), \ldots, f_k(O_i)]^\top, \quad (3)$$

where each f_i represents a feature extraction operation regarding the model type.

Fusion Strategies. We formalize three distinct fusion methods:

- **Decision-level Fusion**: Combines standardized probability outputs from each model using weighted averages defined as follows:

$$F_{\mathrm{d}}\ \left(S\left(O_1\right), \ldots, S\left(O_n\right)\right) = \frac{\sum_{i=1}^{n} w_i \cdot S\left(O_i\right)}{\sum_{i=1}^{n} w_i}, \tag{4}$$

  with model-specific weights $w_i \in [0, 1]$.

- **Feature-level Fusion**: Aggregates extracted features from multiple models into a unified vector for classification represented as follows:

$$F_{\mathrm{f}}\left(S\left(O_1\right), \ldots, S\left(O_n\right)\right) = g\left(\left[\phi_1\left(O_1\right) \oplus \ldots \oplus \phi_n\left(O_n\right)\right]\right), \tag{5}$$

  where $\oplus$ indicates concatenation, ϕ_i is a feature mapping for model i, and g is a trained meta-classifier

- **Adaptive Fusion:** Dynamically re-weights model contributions based on text characteristics defined as:

$$F_{\mathrm{a}}(x, S(O_1), \ldots, S(O_n)) = \frac{\sum_{i=1}^{n} w_i(x) \cdot S(O_i)}{\sum_{i=1}^{n} w_i(x)}, \tag{6}$$

  where $w_i(x)$ are adaptive weights determined from textual features $\psi(x)$ (e.g., negation presence, text length, emotional complexity).

Sentiment Classification. The fused output is mapped to sentiment probabilities by a classification function defined as follows:

$$C\left(F\left(S\left(O_1\right), \ldots, S\left(O_n\right)\right)\right) = \left\{p_{c_1}, p_{c_2}, \ldots, p_{c_k}\right\}, \tag{7}$$

where p_{c_i} is the predicted probability for sentiment class c_i. The final sentiment label $L(x)$ is determined using a confidence threshold δ. For example, if the dataset has three types of label as follows:

$$L(x) = \begin{cases} \text{positive}, & p_{\text{pos}} > p_{\text{neg}} + \delta \\ \text{negative}, & p_{\text{neg}} > p_{\text{pos}} + \delta \\ \text{neutral}, & \text{otherwise} \end{cases}. \tag{8}$$

This generalized formulation allows our framework to seamlessly handle multi-class sentiment tasks and provides greater flexibility in diverse sentiment analysis scenarios.

Training and Adaptation. Fusion weights are trained on labeled data. Decision-level fusion tunes weights on validation sets, while feature-level fusion uses logistic regression with L2 regularization. Adaptive fusion adjusts weights by text type: transformers get more weight with negation or mixed emotions, lexicons with short texts. Models start equal and are modified by simple rules, keeping the system efficient and easy to extend.

4 Experiments and Results

4.1 Experiment Setup

Datasets. We evaluate SentiFuse on three sentiment datasets:

- **Crowdflower US Airline Twitter** (14.6k tweets): A benchmark dataset with balanced sentiment labels (positive, negative, neutral) focusing on airline customer experiences.
- **GoEmotions** (211k Reddit posts) [7]: A comprehensive emotion dataset from Google, containing 28 fine-grained emotions categorized into positive, negative, and neutral sentiment classes.
- **Sentiment140** (1.6M tweets) [14]: A large-scale tweet dataset for sentiment classification, labeled as negative, neutral, positive.

All datasets undergo consistent preprocessing with text normalization and sentiment label standardization. We apply 80-10-10 stratified splits for training, validation, and testing.

Baseline models. To evaluate our framework, we deliberately employ a diverse set of sentiment analysis models that represent different methodological paradigms:

- **Classical machine learning models**. We incorporate TF–IDF vectorization combined with linear classifiers such as SVM and XGBoost. These models rely on bag-of-words style representations, which capture lexical patterns effectively but ignore deep contextual semantics.
- **Deep neural models.** To represent state-of-the-art contextual embeddings, we include BERT, RoBERTa, and DistilBERT. The versions used in the experiments are uncased from Hugging Face Transformers, pretrained on English Wikipedia and BookCorpus. No additional pretraining was performed; we fine-tuned these models directly on the sentiment datasets. These transformers encode rich semantic and syntactic information, leading to strong performance across benchmarks but at higher computational cost.

For fair comparison, we also evaluate several standard ensemble rules: simple averaging, confidence-weighted averaging, majority voting, median averaging, and max-confidence selection.

4.2 Research Question 1

The first research question, "***Does systematic fusion outperform naive combination methods?***", evaluates whether structured fusion strategies provide measurable benefits over naive ensemble rules and individual models. The fixed model pool for this analysis consists of **VADER**, **DistilBERT**, and a **TF–IDF** classifier, chosen to represent lexicon-based, neural, and statistical paradigms. We compare three categories of methods: (i) the best-performing individual model (typically DistilBERT), (ii) naive ensembles including simple averaging, confidence-weighted averaging, majority voting, median averaging, and

max-confidence selection, and (iii) structured fusion methods, namely decision-level, feature-level, and adaptive fusion. Table 1 reports Accuracy, Precision, Recall, F1-Score of strategies across datasets.

Table 1. Performance (percentage) of fusion strategies across datasets.

Strategy	Crowdflower				GoEmotions				Sentiment140			
	Recall	Prec	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec	Acc	F1
Best Individual	77.80	58.79	87.60	66.97	53.59	**77.29**	77.32	63.29	77.28	75.60	76.11	76.43
Simple Average	**80.34**	61.29	88.63	69.53	**71.61**	64.12	75.01	67.66	69.40	76.99	74.27	73.00
Confidence Weighted	78.44	61.73	88.66	69.09	68.93	63.51	74.21	66.11	65.51	77.78	73.33	71.12
Majority Vote	70.82	69.07	90.16	**69.94**	59.86	74.43	77.85	66.36	53.92	**79.95**	70.13	64.40
Median Average	79.49	60.65	88.35	68.80	71.26	67.24	76.84	**69.19**	71.94	75.80	74.42	73.82
Max Confidence	77.59	61.68	88.59	68.73	67.07	62.91	73.55	64.92	62.40	77.67	72.16	69.20
Decision Fusion (Ours)	**80.34**	61.29	88.63	69.53	**71.61**	64.12	75.01	67.66	69.40	76.99	74.27	73.00
Feature Fusion (Ours)	65.54	**73.99**	**90.71**	69.51	60.46	75.70	**78.49**	67.23	**77.85**	77.93	**77.85**	**77.89**
Adaptive Fusion (Ours)	**80.34**	60.22	88.25	68.84	70.95	63.43	74.47	66.98	68.27	77.25	74.02	72.48

On **Crowdflower**, attains the best Accuracy (90.71), with high Precision (73.99). Majority vote yields the best F1 (69.94), slightly above feature (69.51) and decision/simple (69.53). On **GoEmotions**, feature fusion gives the best Accuracy (78.49). Interestingly, median averaging yields the best F1 (69.19), exceeding decision/simple (67.66) and feature (67.23). On the large-scale **Sentiment140**, feature fusion is best on both Accuracy (77.85) and F1 (77.89), outperforming the best individual (76.11/76.43). Across datasets, feature-level fusion is the most reliable overall winner (best Acc on all three; best F1 on Sentiment140). We also present ROC and PR curves in Figure 2 specifically for the Sentiment140 dataset due to its large scale (1.6M samples) and its balanced class distribution. Feature-level fusion achieves the strongest performance with ROC-

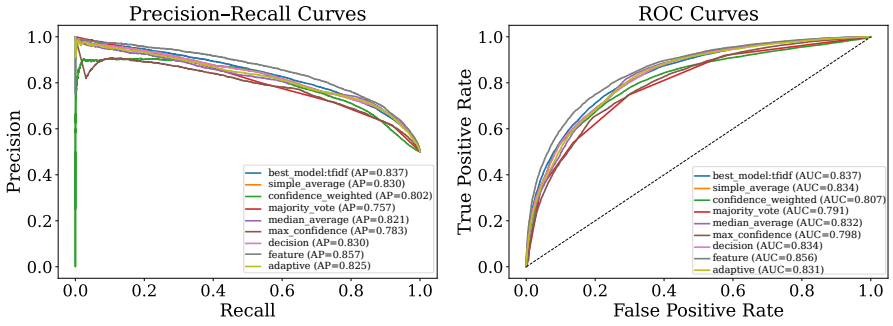

Fig. 2. PR curves and ROC curves on the Sentiment140 dataset.

AUC of 0.856 and PR-AUC of 0.857, outperforming the best individual model (AUC 0.837, PR-AUC 0.837). Decision fusion and simple averaging follow closely (0.834 – 0.830), while majority vote and max-confidence lag behind (AUC below 0.800, PR-AUC as low as 0.757). These results confirm that structured fusion yields superior discriminative ability and more reliable precision–recall trade-offs than naive ensemble rules.

4.3 Research Question 2

The second research question, "How does fusion effectiveness vary across text characteristics?" examines whether different fusion strategies perform better on specific types of texts. Since models have complementary strengths, such as lexicon approaches excel on short texts while transformers capture complex semantics, we hypothesize that structured fusions can adapt to text conditions more effectively than either naive rules or single models. We then compare the best individual model, decision fusion, adaptive fusion, and feature fusion within each category. Figure 3 reports accuracy across the three datasets.

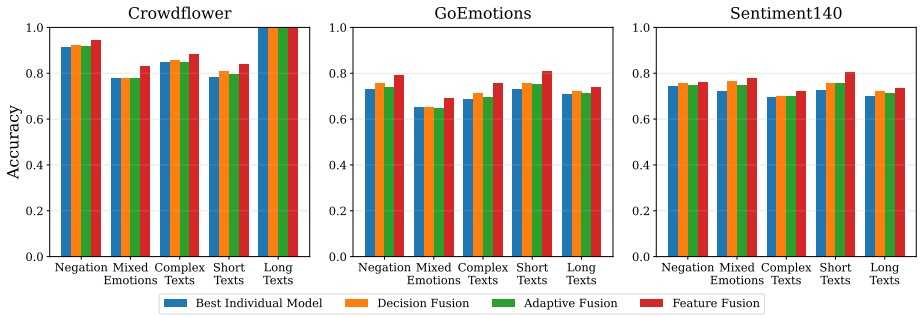

Fig. 3. Accuracy of proposed framework and best individual model across datasets.

On **Crowdflower**, feature fusion consistently outperforms the other approaches across nearly all categories. It reaches about 0.95 accuracy on negation and nearly 1.0 on long texts, demonstrating that combining models captures subtle polarity shifts and benefits from richer context. Even on short texts, where lexicon-based methods usually excel, feature fusion attains 0.82, exceeding decision (0.75), adaptive (0.76), and the best individual model (0.75). This shows that feature-level integration can compensate for sparse signals by pooling complementary features. For **GoEmotions**, the advantage of feature fusion is again visible, particularly in challenging categories such as negation (0.77) and complex texts (0.73). Short texts also favor feature fusion (0.79 vs. 0.77 for decision and 0.73 for the best individual). Interestingly, on mixed emotions, the gap between strategies narrows, suggesting that all models find this phenomenon inherently difficult, and fusion only partially mitigates the challenge. On long texts, all methods converge near 1.0, reflecting that with sufficient context, both individual models and ensembles perform robustly. On the large-scale **Sentiment140**, feature fusion maintains a consistent edge across categories, but the margins are smaller than in the other datasets. It scores 0.74 on short texts versus 0.73 for the best individual, and 0.69 on complex texts versus 0.68 for decision and adaptive fusion. This suggests that on very large datasets, strong individual models already capture much of the available signal, and fusion strategies yield incremental but reliable improvements.

Overall, RQ2 confirms that the effectiveness of fusion varies with text characteristics. Feature fusion is most effective on negation, complex sentiment, and

short texts, where single models struggle. Decision and adaptive fusion provide stable performance, often close to feature fusion, but without consistently surpassing it. On long texts, all approaches converge, highlighting that fusion adds the most value when input is limited or ambiguous.

4.4 Research Question 3

The third research question, "***Does the framework generalize across different model pools?***", examines whether the proposed framework remains effective when applied to different sets of heterogeneous models. Across the three datasets, our results show that systematic fusion consistently matches or outperforms naive ensembles, which is illustrated in Table 2 and Table 3.

Table 2. Performance (percentage) of fusion strategies with combination of TextBlob + RoBERTa + SVM across datasets.

Strategy	Crowdflower				GoEmotions				Sentiment140			
	Recall	Prec	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec	Acc	F1
Simple Average	58.33	73.25	78.11	64.68	62.92	71.06	75.90	**66.70**	52.02	75.58	64.17	56.33
Confidence Weighted	59.33	72.69	78.11	64.92	62.05	71.53	75.94	66.39	49.61	75.61	63.65	54.98
Majority Vote	**62.83**	72.65	**80.43**	**66.91**	59.84	**72.35**	75.97	65.53	49.34	76.06	63.60	54.95
Median Average	60.67	73.09	79.18	66.30	61.52	71.71	75.98	66.25	**52.14**	75.55	**64.37**	**56.40**
Max Confidence	60.17	72.53	78.93	65.68	59.73	71.95	75.60	65.09	45.86	**76.24**	62.61	52.07
Decision Fusion (Ours)	57.00	74.18	77.89	64.27	**62.65**	71.37	75.75	66.26	51.63	75.11	63.56	56.12
Feature Fusion (Ours)	61.67	73.67	79.80	66.61	61.92	71.56	**76.09**	66.18	51.49	75.68	63.63	55.93
Adaptive Fusion (Ours)	55.50	**75.83**	77.39	64.13	62.06	71.99	75.96	66.16	50.39	75.46	63.58	55.79

Table 3. Performance (percentage) of fusion strategies with combination of AFINN + BERT + XGBoost across datasets.

Strategy	Crowdflower				GoEmotions				Sentiment140			
	Recall	Prec	Acc	F1	Recall	Prec	Acc	F1	Recall	Prec	Acc	F1
Simple Average	75.66	61.94	88.16	68.20	71.08	64.37	75.20	67.29	66.78	76.07	73.63	71.59
Confidence Weighted	75.16	62.19	88.14	68.20	70.09	64.99	75.18	67.23	63.23	76.40	72.46	70.07
Majority Vote	71.00	67.62	89.91	69.02	62.59	72.98	77.83	66.59	56.29	**78.17**	70.22	65.34
Median Average	74.00	62.38	87.89	67.80	70.83	66.18	76.36	67.27	69.85	75.71	74.71	72.70
Max Confidence	73.66	62.47	88.20	67.90	69.05	64.91	75.14	66.60	60.55	76.89	72.06	68.38
Decision Fusion (Ours)	75.66	61.94	88.16	68.20	71.07	64.37	75.20	67.29	66.78	76.07	73.63	71.59
Feature Fusion (Ours)	63.83	**72.69**	**90.39**	**69.08**	61.10	**75.97**	**78.59**	67.39	**77.07**	76.89	**76.89**	**76.90**
Adaptive Fusion (Ours)	**79.16**	60.62	88.27	68.51	**72.52**	63.67	75.18	**67.82**	66.44	76.89	73.49	71.36

On **TextBlob + RoBERTa + SVM (Combo 2)**, naive ensembles produce mixed outcomes depending on the dataset. Majority voting achieves the highest accuracy on Crowdflower (80.4%) and F1 (66.9%), while median averaging performs best on GoEmotions (F1 66.3%). These outcomes highlight that naive ensembles are sensitive to dataset conditions and fail to generalize consistently. In contrast, the SentiFuse strategies yield more stable performance. Feature fusion achieves competitive results across datasets (79.8% accuracy and 66.6% F1 on Crowdflower), while adaptive fusion reaches the highest precision on Crowdflower (75.8%), indicating robustness in skewed distributions. Decision fusion

remains reliable, closely tracking simple averaging but with lower variance. On **AFINN + BERT + XGBoost (Combo 3)**, structured fusion demonstrates clearer advantages. Feature fusion delivers the strongest overall balance, including 90.4% accuracy on Crowdflower and 75.97% precision with 78.59% accuracy on GoEmotions, outperforming both naive and individual models. Adaptive fusion emphasizes recall, reaching 79.2% on Crowdflower, while decision fusion again shows stable behavior aligned with simple averaging. On Sentiment140, feature fusion achieves the highest overall performance (76.9% accuracy and F1), demonstrating that structured integration maintains benefits even in large-scale settings.

Overall, RQ3 shows that SentiFuse generalizes effectively across heterogeneous model fusions. While naive ensembles occasionally perform well in isolated cases, their outcomes vary substantially across datasets. Structured methods, by contrast, consistently enhance reliability: feature fusion excels in fine-grained scenarios, adaptive fusion strengthens recall and robustness, and decision fusion provides stable performance. These results highlight that the framework is not tied to a particular model set but provides a general architecture for integrating diverse sentiment classifiers.

5 Conclusion

In this paper, we introduced SentiFuse, a flexible and model-agnostic framework for sentiment analysis that integrates diverse models through a unified standardization and fusion pipeline. By supporting decision-level, feature-level, and adaptive fusion strategies, the framework improves performance across multiple datasets and challenging text phenomena such as negation, mixed emotions, and short or complex expressions. Our experiments demonstrate that no single strategy dominates in all cases: feature-level fusion provides strong overall gains, while adaptive and decision-level methods offer robustness in heterogeneous contexts. Feature-level fusion is particularly strong on categories such as negation and mixed emotions, where complementary cues from lexicon and neural embeddings align. Adaptive fusion tends to generalize better across varied categories by weighting models dynamically. These findings underline the methodological contribution of offering a general fusion architecture that can be applied across model families, the analytical contribution of revealing when and why certain fusion strategies succeed on specific text types, and the practical contribution of showing that such strategies generalize effectively across different model pools. Together, these contributions provide a more reliable foundation for sentiment classification and point toward broader applications where robustness and complementarity are essential. While SentiFuse introduces some additional inference cost from running multiple models, it remains lightweight: feature fusion adds only linear concatenation overhead, and adaptive fusion requires a small meta-classifier. Although our accuracies fluctuate around 80%, this is consistent with prior state-of-the-art sentiment work on noisy social media text. Tweets and Reddit posts often include sarcasm, slang, and mixed emotions, and annotator

agreement itself is rarely above 85–90%. We expect that stronger models such as LLaMA or Gemma, if included in the pool, would further improve individual baselines. However, our framework is model-agnostic: fusion still helps in cases where even large models misclassify ambiguous or sarcastic inputs. Thus, SentiFuse can complement LLMs rather than compete with them. Future work will explore more context-aware adaptive mechanisms and extend the framework to multilingual and domain-specific settings.

Funding: This work was supported by NSF - USA CNS-2219614.

Disclosure of Interests: The authors declare no conflict of interest.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). pp. 2200–2204. European Language Resources Association (ELRA) (2010)
2. Behera, R.K., Jena, M., Rath, S.K., Misra, S.: Co-lstm: Convolutional lstm model for sentiment analysis in social big data. Information Processing & Management **58**(1), 102435 (2021)
3. Breiman, L.: Bagging predictors. Machine Learning **24**(2), 123–140 (1996)
4. Chikersal, P., Poria, S., Cambria, E.: Sentu: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In: Nakov, P., Zesch, T., Cer, D., Jurgens, D. (eds.) Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 647–651 (2015)
5. Chiny, M., Chihab, M., Bencharef, O., Chihab, Y.: Lstm, vader and tf-idf based hybrid sentiment analysis model. International Journal of Advanced Computer Science and Applications **12**(7) (2021)
6. Das, B., Chakraborty, S.: An improved text sentiment classification model using tf-idf and next word negation (2018)
7. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: Goemotions: A dataset of fine-grained emotions. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4040–4054. Association for Computational Linguistics, Online (2020)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186 (2019)
9. Dey, R., Das, A.: Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis. Multimedia Tools and Applications **82**(21), 32967–32990 (2023)
10. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. pp. 1–15. Springer (2000)
11. Elbagir, S., Yang, J.: Twitter sentiment analysis using natural language toolkit and vader sentiment. In: Proceedings of the International Multiconference of Engineers and Computer Scientists. vol. 122 (2019)

12. Feldman, R.: Techniques and applications for sentiment analysis. Communications of the ACM **56**(4), 82–89 (2013). https://doi.org/10.1145/2436256.2436274
13. Gan, C., Fu, X., Feng, Q., Zhu, Q., Cao, Y., Zhu, Y.: A multimodal fusion network with attention mechanisms for visual–textual sentiment analysis. Expert Systems with Applications **242**, 122731 (2024)
14. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Tech. rep., Stanford University (2009)
15. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 8, pp. 216–225 (2014)
16. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research **50**, 723–762 (2014)
17. Kora, R., Mohammed, A.: An enhanced approach for sentiment analysis based on meta-ensemble deep learning. Social Network Analysis and Mining **13**, 38 (2023)
18. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
19. Lossio-Ventura, J.A., Weger, R., Lee, A.Y., et al.: A comparison of chatgpt and fine-tuned open pre-trained transformers (opt) against widely used sentiment analysis tools: Sentiment analysis of covid-19 survey data. JMIR Mental Health **11**, e50150 (2024)
20. Miah, M., et al.: Ensemble model combining transformers and llm for multilingual sentiment analysis. In: International Conference on Natural Language Processing (2024)
21. Mohammad, S.M., Kiritchenko, S., Zhu, X.: Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises (SemEval-2013). pp. 321–327 (2013)
22. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems **89**, 14–46 (2015)
23. Ray, P., Chakrabarti, A.: A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Applied Computing and Informatics **18**(1/2), 163–178 (2022)
24. Reusens, M., Stevens, A., Tonglet, J., De Smedt, J., Verbeke, W., vanden Broucke, S., Baesens, B.: Evaluating text classification: A benchmark study. Expert Systems with Applications **254**, 124302 (2024)
25. Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., Cuenca-Jiménez, P.M.: A review on sentiment analysis from social media platforms. Expert Systems with Applications **223**, 119862 (2023)
26. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management **24**(5), 513–523 (1988)
27. Singh, C., Imam, T., Wibowo, S., Grandhi, S.: A deep learning approach for sentiment analysis of covid-19 reviews. Applied Sciences **12**(8), 3709 (2022)
28. Singh, U., Abhishek, K., Azad, H.: A survey of cutting-edge multimodal sentiment analysis. ACM Computing Surveys **56**(9), 1–38 (2024)
29. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational Linguistics **37**(2), 267–307 (2011)
30. Wang, L., et al.: Is chatgpt a good sentiment analyzer? a preliminary study (2024)
31. Wang, X., Gan, Y.: Multi-level adversarial training for stock sentiment prediction models. Social Network Analysis and Mining (2023)
32. Xu, G., Meng, Y., Qiu, X., Yu, Z., Wu, X.: Sentiment analysis of comment texts based on bilstm. IEEE Access **7**, 51522–51532 (2019)

33. Yadav, A., Vishwakarma, D.K.: Sentiment analysis using deep learning architectures: A review. Artificial Intelligence Review **53**(6), 4335–4385 (2020). https://doi.org/10.1007/s10462-019-09794-5
34. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 207–212 (2016)