



Enhancing Multi-Domain Sentiment Classification Using Adaptive Lexicons and Machine Learning

Saleh Yaakub¹, Mustafa Mat Deris², Ika Safitri Windiarti³

^{1,2}Information Technology, Faculty of Business and Information Technology, Universiti Muhammadiyah Malaysia, Perlis, Malaysia

¹Sistem Informasi, Sains Teknologi dan Kesehatan, Universitas Muhammadiyah Jambi, Jambi, Indonesia.

³Computer Science, Business and Information Technology, Universitas Muhammadiyah Palangkaraya, Kalimantan Tengah, Indonesia

*saleh@umjambi.ac.id

Abstract

Multi-domain sentiment classification remains challenging because lexical polarity and expression patterns vary across application contexts. This study evaluates an enhanced iVADER framework that combines domain-specific lexicon adaptation with text preprocessing and machine-learning-supported feature representations. Five public English-language datasets were analyzed: Amazon Cells Labelled (1,000 records), Yelp Labelled (1,000), Tweets with Sarcasm and Irony (19,000), Apple Twitter Sentiment Texts (4,500), and ChatGPT Tweets (2,500), yielding 28,000 records. A stratified 80:20 split produced 22,400 training records and 5,600 test records. Preprocessing included noise and URL removal, case folding, tokenization, stopword removal, WordNet lemmatization, stemming, and negation handling. The framework expanded domain-specific lexical information and combined it with TF-IDF, n-grams, bag-of-words, and embedding-based representations. Enhanced iVADER achieved pooled accuracy of 92%, precision of 90%, recall of 89%, and an F1-score of 0.91, compared with 85%, 82%, 75%, and 0.80 for conventional VADER. Across the five datasets, accuracy ranged from 87% to 94%, with an unweighted mean of 91.0% and a sample standard deviation of 2.74 percentage points. The results indicate that domain-adaptive lexical information can improve VADER-based sentiment classification across the observed domains while retaining an interpretable scoring process.

Keywords: adaptive lexicon; machine learning; multi-domain sentiment classification; text classification; VADER

How to Cite: [S. . Yaakub, M. M. . Deris, and I. S. . Windiarti, "Enhancing Multi-Domain Sentiment Classification Using Adaptive Lexicons and Machine Learning", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1455–1463, Oct. 2026, doi: [10.29207/resti.v10i5.7862](https://doi.org/10.29207/resti.v10i5.7862).
Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7862>

Received: May 17, 2026

Accepted: September 19, 2026

Available Online: October 14, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.ikatanahliinformatika.org/)

1. Introduction

User-generated reviews and social-media posts provide a large and continuously evolving source of evidence about product experience, service quality, technology adoption, and public opinion. However, a sentiment classifier that performs well on one corpus may lose accuracy on another because domains differ in vocabulary, document length, informality, and the linguistic cues used to express evaluation [1-4].

VADER remains attractive for practical sentiment analysis because its lexicon and linguistic rules are transparent and efficient for short informal text. Its fixed polarity assignments, however, do not fully capture domain-dependent meaning; terms associated with price, delivery, service quality, product design, or

emerging technologies may shift in evaluative force across corpora [5-8].

Recent hybrid approaches address this limitation by combining interpretable lexical resources with corpus-derived features, contextual representations, and domain-specific knowledge. Evidence from adaptive lexicons, TF-IDF, n-grams, embeddings, negation handling, and contextual alignment suggests that classification can improve when lexical evidence is adjusted to the target domain, although figurative language and domain shift remain difficult [9-14].

Table 1 summarizes the studies most directly related to the present framework and clarifies its position within domain-adaptive sentiment classification.

Table 1. Comparison of Related Multi-Domain Sentiment Studies

Study (Year)	Dataset focus	Approach	Classification method	Relevance to present study
Qi and Shabrina (2023) [5]	Twitter/social-media text	Lexicon- and machine-learning comparison	Lexicon and ML classifiers	Shows performance differences between interpretable lexical and data-driven methods
Xing et al. (2019) [6]	Domain-adaptive sentiment data	Context- and semantics-aware lexicon adaptation	Lexicon adaptation	Provides a foundation for context-sensitive polarity adjustment
Sanagar and Gupta (2020) [13]	Multiple genres/domains	Corpus-generated polarity seeds for lexicon learning	Lexicon-based classification	Supports corpus-driven expansion of domain sentiment vocabulary
Barik and Misra (2024) [8]	Customer reviews	Improved VADER with domain dictionary	VADER-based classification	Directly motivates adaptive extension of VADER for review domains
This study (2026)	Five public datasets	Adaptive lexicons with enhanced iVADER and corpus-derived features	Enhanced VADER-based scoring	Reports pooled and domain-level results with explicit reproducibility constraints

Against this background, this study evaluates an enhanced iVADER framework that integrates domain-specific lexical adaptation with machine-learning-supported feature representations for multi-domain sentiment classification. The research question is whether the enhanced framework improves accuracy, precision, recall, and F1-score over conventional VADER on the same observed-domain partitions, and how performance varies between explicit review text and context-dependent social-media text.

The study makes three contributions. First, it extends VADER with domain-specific term expansion, contextual word alignment, and sentiment-score adjustment while retaining an interpretable lexicon-based decision process. Second, it combines lexical evidence with TF-IDF, n-grams, bag-of-words, and embedding-based representations. Third, it evaluates the framework across five public datasets and reports both pooled and domain-level performance while explicitly delimiting claims that cannot be supported without repeated runs, record-level predictions, and execution logs.

2. Methods

2.1 Research Framework

The experimental framework comprises six core stages: multi-domain dataset acquisition, text preprocessing, adaptive domain-lexicon construction, feature representation, enhanced iVADER classification, and performance evaluation. The design combines the rule-based transparency of VADER with corpus-derived information intended to improve sensitivity to domain vocabulary and contextual variation [6],[8], [9], [13].

Figure 1 summarizes the complete processing sequence. The input consists of five English-language datasets representing e-commerce product reviews, restaurant and service reviews, sarcasm and irony in social media, technology-related tweets, and broader discussion of ChatGPT. The documented corpus contains 28,000 records and uses a stratified random 80:20 split within each dataset, resulting in 22,400 training records and 5,600 test records.

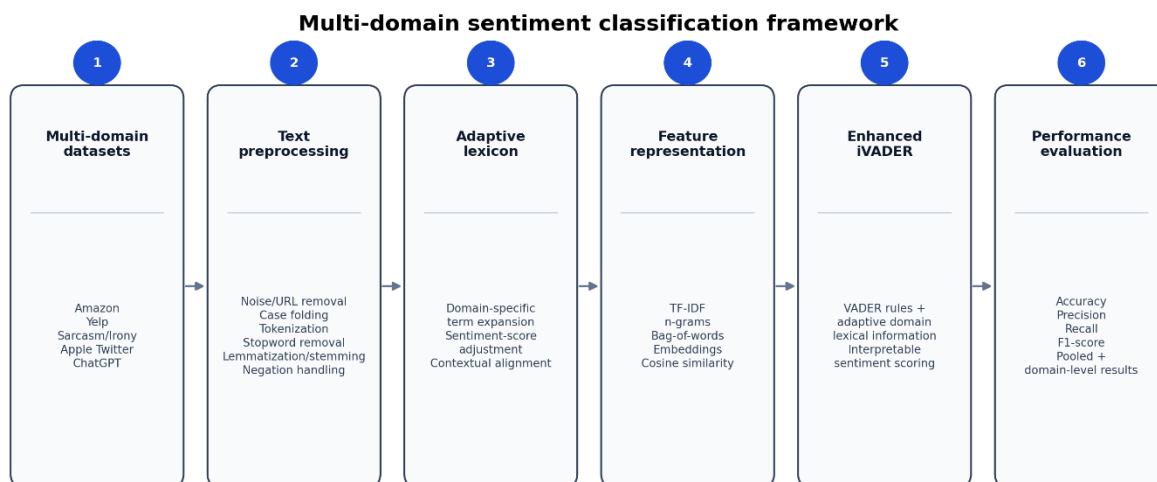


Figure 1. Proposed Multi-Domain Sentiment Classification Framework

Preprocessing standardizes raw text before feature generation. Domain-specific lexical adaptation then expands or adjusts sentiment-bearing vocabulary from the training corpus, while TF-IDF, n-grams, bag-of-words, embeddings, cosine similarity, contextual word alignment, and negation handling provide complementary lexical and semantic evidence [10], [11], [15], [16]. The resulting representation is scored by enhanced iVADER to produce the final sentiment decision.

The framework preserves the interpretable VADER scoring logic while allowing domain vocabulary to influence polarity estimation. This design is intended to improve sensitivity to domain-specific expressions without replacing the lexical decision process with a fully opaque classifier.

Enhanced iVADER is compared with conventional VADER on the same collection of observed domains. Because the documented experiment does not use a leave-one-domain-out protocol, the reported results are interpreted as multi-dataset performance on held-out records from observed domains rather than as unsupervised transfer to a completely unseen target domain.

2.2 Dataset Collection and Experimental Split

The datasets were obtained from the UCI Machine Learning Repository and Kaggle. The documented sample contains 1,000 Amazon reviews, 1,000 Yelp reviews, 19,000 sarcasm-and-irony tweets, 4,500 Apple-related tweets, and 2,500 ChatGPT-related tweets, for a total of 28,000 records. Their principal characteristics are summarized in Table 2.

Table 2. Characteristics of the Datasets Used in This Study

Dataset	Domain	Records	Label structure	Linguistic characteristic
Amazon Cells Labelled	E-commerce product reviews	1,000	Binary positive/negative	Explicit product-evaluation vocabulary
Yelp Labelled	Restaurant and service reviews	1,000	Binary positive/negative	Service- and food-related evaluation
Tweets with Sarcasm and Irony	Social media	19,000	Four figurative-language categories	Short, context-dependent figurative expressions
Apple Twitter Sentiment Texts	Technology/social media	4,500	Positive, negative, neutral	Brand- and product-specific vocabulary
ChatGPT Tweets	Technology/social media	2,500	Sentiment, topic, emotion annotations	Broad technology opinions and emerging terminology

Each dataset was divided by stratified random sampling, with 80% assigned to training and 20% to testing. This corresponds to 800/200 records for Amazon, 800/200 for Yelp, 15,200/3,800 for the sarcasm-and-irony corpus, 3,600/900 for Apple Twitter Sentiment Texts, and 2,000/500 for ChatGPT Tweets. Table 3 reports the resulting partitions.

Table 3. Datasets and Experimental Split

Dataset	Records	Train	Test
Amazon Cells Labelled	1,000	800	200
Yelp Labelled	1,000	800	200
Tweets with Sarcasm and Irony	19,000	15,200	3,800
Apple Twitter Sentiment Texts	4,500	3,600	900
ChatGPT Tweets	2,500	2,000	500
Total	28,000	22,400	5,600

Note: training and test counts are calculated directly from the documented 80:20 split.

The source datasets do not share a uniform annotation scheme. Amazon and Yelp are described as binary positive/negative datasets; Apple Twitter Sentiment Texts contains positive, negative, and neutral labels; the sarcasm-and-irony corpus contains four figurative-language categories; and the ChatGPT dataset includes sentiment, topic, and emotion annotations. Because an explicit cross-dataset label-mapping script is not preserved, the manuscript reports these source label structures transparently rather than introducing an undocumented harmonization rule. Amazon and Yelp are reported as balanced, with approximately 500 positive and 500 negative records each. The sarcasm-

and-irony dataset is described as approximately balanced, whereas exact class counts are not available for the remaining datasets. Consequently, results are reported per dataset and through the documented pooled metrics without introducing unverified class-level distributions.

The datasets provide complementary linguistic conditions. Amazon and Yelp contain relatively explicit evaluative statements; Apple tweets contain brand- and product-specific vocabulary; ChatGPT tweets represent broader technology opinions; and the sarcasm-and-irony corpus contains short, context-dependent expressions. This variation supports a domain-level assessment of lexical adaptability [8], [12], [14], [17].

All records were processed in English. Because the source label spaces are heterogeneous, the analysis does not assume that every dataset was converted into one common binary label space. This distinction is important when interpreting pooled results across datasets with different annotation structures. The experiment therefore evaluates performance within the five observed datasets and does not establish generalization to a completely unseen domain. The conclusions are limited to this experimental scope [3], [4], [14].

2.3 Text Preprocessing

The preprocessing sequence includes noise removal, URL removal, case folding, handling of missing records, tokenization, stopword removal,

lemmatization, stemming, and negation handling. Noise removal eliminates symbols, numbers, and HTML fragments considered irrelevant to sentiment classification, while case folding reduces orthographic variation [10], [18-20].

Tokenization separates each text into lexical units. WordNet Lemmatizer is used to reduce inflected forms, while stemming provides an additional normalization stage. Stopword removal reduces non-informative dimensions, and negation handling is retained because constructions such as 'not good' can reverse the polarity assigned by a static lexicon [18], [20].

The implementation record does not specify a fixed negation-window length, a _NEG suffix transformation, emoji-conversion rules, or package-specific stopwords versions. These details are therefore not asserted here; the method description is restricted to preprocessing operations that are explicitly documented.

The tokenizer name, dependency versions, and complete software environment are also not preserved. This omission is treated as a reproducibility limitation. A fully reproducible release should archive the preprocessing script, environment specification, and fixed random seed [21].

2.4 Adaptive Lexicon and Feature Representation

The enhanced framework begins with VADER's general sentiment vocabulary and extends it with domain-specific terms identified from the training corpus. Two operations are central to the adaptation process: domain-specific term expansion and sentiment-score adjustment. Their purpose is to represent expressions whose evaluative meaning depends on application context, including vocabulary related to delivery, battery life, restaurant service, and emerging technologies [6-8], [13].

Let t_i represent a sentiment-bearing term within a domain-specific corpus. The adaptive sentiment weight of t_i is calculated using Equation 1.

$$W_{adaptive}(t_i) = \alpha \times TF_{norm}(t_i) + \beta \times Sim_{context}(t_i) \quad (1)$$

$TF_{norm}(t_i)$ denotes the normalized term frequency of sentiment-bearing terms, $Sim_{context}(t_i)$ represents contextual semantic similarity, while α and β are weighting parameters used to control the contribution balance between statistical lexical frequency and contextual semantic similarity.

The contextual semantic similarity was measured using cosine similarity between domain-specific lexical vectors, as expressed in Equation 2.

$$Sim_{context}(t_i, t_j) = \frac{V_i \cdot V_j}{\|V_i\| \times \|V_j\|} \quad (2)$$

V_i and V_j represent lexical embedding vectors generated from textual representations within heterogeneous domains. This mechanism enables the adaptive lexicon

to recognize contextual polarity shifts and domain-dependent sentiment variations more effectively.

Feature relevance is supported by frequency-based representations and semantic proximity. TF-IDF emphasizes terms that are important within a document but less frequent across the corpus, while cosine similarity and embedding-based representations support the identification of semantically related domain terms [9], [11], [15].

Contextual word alignment and negation handling are used to reduce errors caused by polarity reversal and domain-specific phrase structure. The adapted lexicon is then incorporated into iVADER sentiment-intensity scoring while retaining VADER's rules for punctuation, intensification, and informal language [5], [8], [18].

The documented method references Word2Vec, FastText, BERT, n-grams, bag-of-words, TF-IDF, cosine similarity, and domain-specific lexical expansion, but it does not preserve embedding dimensionality, semantic-neighbor counts, similarity thresholds, training epochs, or final coefficient values. Accordingly, the mechanism is described at the level supported by the experimental record rather than with reconstructed hyperparameters.

The technical contribution is the integration of adaptive domain vocabulary with a transparent VADER-based scoring process and corpus-derived feature representations. The intended effect is greater contextual sensitivity while preserving interpretability at the lexical decision layer.

Feature representation includes bag-of-words and n-grams for lexical occurrence patterns, TF-IDF for term importance, and Word2Vec, FastText, or BERT representations for semantic relationships. The documented domain analyses specifically associate performance with TF-IDF, n-grams, word embeddings, cosine similarity, contextual word alignment, and negation handling [9], [11], [15].

The experimental record does not contain an executed search grid, final feature count, embedding dimensions, similarity threshold, or a complete set of selected model parameters. The manuscript therefore does not claim that the reported configuration is a globally optimized hyperparameter solution.

The reproducible evaluation unit is the enhanced iVADER configuration for which aggregate and domain-level metrics are documented. A stronger reproducibility package should provide the feature vocabulary, lexicon-update rules, model configuration, and fold-level validation outputs before claims about statistical optimality are made [21].

No cross-validation scores or next-best configurations are available. The reported 80:20 experiment is therefore interpreted descriptively rather than as evidence that an exhaustive model-selection procedure identified a unique optimum.

The same documented data partition is used for the comparison between enhanced iVADER and conventional VADER, which keeps the reported baseline comparison internally consistent.

This reporting boundary separates the implemented comparison from candidate classifiers discussed in the broader methodology and avoids attributing unrecorded experimental results to SVM, Random Forest, or other models.

2.5 Classification and Evaluation

The primary evaluated method is enhanced iVADER, which combines the VADER lexicon and rule system with domain-specific lexical adaptation and the feature-processing stages described above. Conventional VADER is used as the controlled baseline. Other classifiers are discussed as methodological alternatives, but complete executed metrics are not available for them and they are therefore not treated as experimental baselines in this manuscript.

Evaluation uses accuracy, precision, recall, and F1-score as the reported classification metrics [22]. Aggregate differences between enhanced iVADER and VADER are expressed as absolute percentage-point changes for accuracy, precision, and recall, and as an absolute change for F1-score.

The documented results include one pooled metric set for enhanced iVADER and one for conventional VADER, together with domain-level accuracy, precision, and recall for enhanced iVADER. Domain F1 values in Table 5 are derived from the reported precision and recall using $F1 = 2PR/(P + R)$.

Variability across domains is summarized descriptively using the unweighted mean and sample standard deviation of the five reported domain scores. These statistics describe between-domain dispersion; they are not confidence intervals and do not estimate repeated-run uncertainty.

2.6 Reproducibility and Analysis Constraints

The experimental record does not archive fold-level predictions, bootstrap samples, p-values, effect sizes, confusion-matrix counts, software versions, hardware specifications, runtime measurements, or peak-memory logs. Consequently, the study reports only the numerical results that can be traced to the documented dataset summary and performance tables.

Without record-level predictions, a normalized confusion matrix, paired significance test, and auditable list of misclassified instances cannot be reconstructed reliably. Likewise, runtime and memory comparisons would require re-execution of the original implementation in a documented software and hardware environment.

These constraints do not invalidate the reported descriptive comparison, but they limit the strength of inferential and efficiency claims. The analysis therefore

distinguishes observed performance differences from conclusions that would require repeated runs, ablation experiments, or formal statistical testing.

Aggregate performance is compared using absolute differences between enhanced iVADER and VADER. Domain-specific results are then interpreted in relation to text structure, vocabulary specificity, and figurative-language complexity, without assigning causal effects to individual components in the absence of ablation experiments.

Qualitative analysis uses representative challenging sentences documented with the study to illustrate sarcasm and mixed-polarity failure modes. Because record identifiers and saved predictions are not available for these sentences, they are treated as diagnostic examples rather than as auditable test-set misclassifications.

All findings are therefore bounded by the evidence available from one documented 80:20 split, the reported aggregate metrics, the domain-level metrics, and qualitative descriptions of difficult language patterns.

No inferential statistical test is reported because repeated-run or fold-level observations are unavailable. Descriptive differences are presented without claiming statistical significance.

The same constraint applies to computational efficiency. Because wall-clock time, throughput, CPU/GPU specifications, and peak-memory measurements are unavailable, the manuscript makes no quantitative efficiency claim.

Future reproducibility work should preserve predictions, execution logs, software versions, random seeds, and profiling results so that statistical uncertainty, error structure, and computational cost can be evaluated directly.

3. Results and Discussions

Table 4 presents the documented pooled comparison. Enhanced iVADER improved every reported metric relative to conventional VADER: accuracy increased from 85% to 92%, precision from 82% to 90%, recall from 75% to 89%, and F1-score from 0.80 to 0.91. The largest gain occurred in recall, indicating that the adapted lexical information recovered sentiment cues that conventional VADER did not identify as consistently.

Table 4. Aggregate Performance of Enhanced iVADER and VADER

Metric	Enhanced iVADER	VADER
Accuracy	92%	85%
Precision	90%	82%
Recall	89%	75%
F1-score	0.91	0.80

The pooled accuracy of 92% should not be interpreted as the arithmetic mean of the five domain accuracies. The unweighted domain mean is 91.0%, whereas the

pooled value reflects the experimental dataset composition. This distinction is important because the sarcasm-and-irony corpus is substantially larger than the other datasets.

3.1 Aggregate Improvement over VADER

The 14-point recall gain is the most pronounced difference between the two methods. Domain-specific vocabulary and contextual processing appear to increase sensitivity to relevant sentiment cues, whereas the smaller precision gain indicates that some ambiguous expressions remain difficult to classify. Similar limitations of static lexicons under domain variation have been reported in multi-domain and hybrid sentiment studies [6], [12-14].

Figure 2 visualizes the same aggregate metrics reported in Table 4. The chart makes the largest relative improvement in recall visible while preserving the same enhanced iVADER and conventional VADER baseline definitions used in the table.

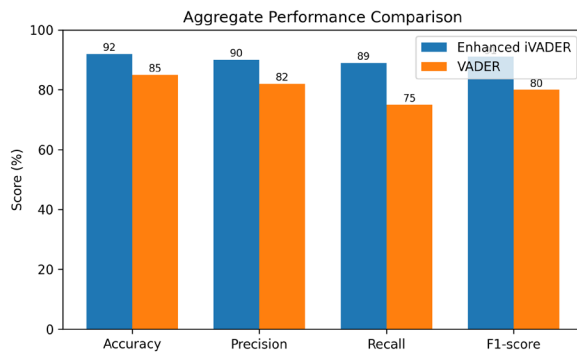


Figure 2. Aggregate performance comparison between enhanced iVADER and conventional VADER.

Across the five reported datasets, enhanced iVADER achieved an unweighted mean accuracy of 91.0% (SD = 2.74), mean precision of 89.4% (SD = 2.70), mean recall of 85.0% (SD = 3.67), and a mean derived F1-score of 87.1% (SD = 3.16). These standard deviations describe heterogeneity across domains rather than repeated experimental uncertainty.

The difference between pooled and unweighted scores also demonstrates why dataset-level reporting is necessary. A single aggregate metric can conceal lower performance on figurative-language data and higher performance on datasets in which sentiment is expressed more explicitly.

Enhanced iVADER's improvement over VADER is therefore interpreted as descriptive evidence of greater sensitivity to domain-specific sentiment cues, not as a statistically verified superiority claim. Repeated-run or fold-level outputs would be required for inferential testing.

3.2 Domain-Level Performance and Explanation

Table 5 shows that enhanced iVADER achieved its highest accuracy on Amazon Cells Labelled (94%) and

its lowest on Tweets with Sarcasm and Irony (87%). The seven-point gap is consistent with substantial differences in linguistic explicitness and annotation structure between the datasets.

Table 5. Domain-Specific Performance of Enhanced iVADER

Dataset	Acc.	Prec.	Recall	F1*
Amazon	94%	92%	88%	90.0%
Yelp	90%	89%	83%	85.9%
Sarcasm/Irony	87%	85%	80%	82.4%
Apple	91%	90%	85%	87.4%
Twitter				
ChatGPT	93%	91%	89%	90.0%
Mean $\pm$ SD	91.0 $\pm$ 2.74	89.4 $\pm$ 2.70	85.0 $\pm$ 3.67	87.1 $\pm$ 3.16

*F1 is calculated from the precision and recall values reported in the experimental record.

Amazon reviews contain direct evaluative vocabulary related to product quality, price, delivery, and customer experience. Their binary and approximately balanced labels also provide a comparatively clear decision boundary. Domain-specific expressions such as 'on-time delivery', 'poor quality', and 'affordable price' therefore align well with lexical adaptation and TF-IDF weighting. ChatGPT Tweets achieved 93% accuracy, followed by Apple Twitter Sentiment Texts at 91% and Yelp Labelled at 90%. Apple tweets contain brand- and product-specific vocabulary such as 'battery life', 'operating system', and 'product design', while Yelp reviews benefit from service- and food-related terms but remain vulnerable to contrastive aspect statements. The sarcasm-and-irony dataset produced the lowest accuracy, precision, recall, and derived F1-score. Short context, pragmatic reversal, irony, and competing positive and negative cues weaken document-level lexical scoring even when negation handling and contextual word alignment are used. This pattern is consistent with studies showing that figurative sentiment often requires information beyond isolated lexical polarity [12], [14], [23], [24].

The available domain-specific VADER baselines indicate accuracy gains of 9 points for Amazon (94% versus 85%), 10 points for Yelp (90% versus 80%), 14 points for sarcasm and irony (87% versus 73%), and 16 points for Apple tweets (91% versus 75%). A domain-specific VADER baseline is not documented for ChatGPT Tweets, so no value is inferred.

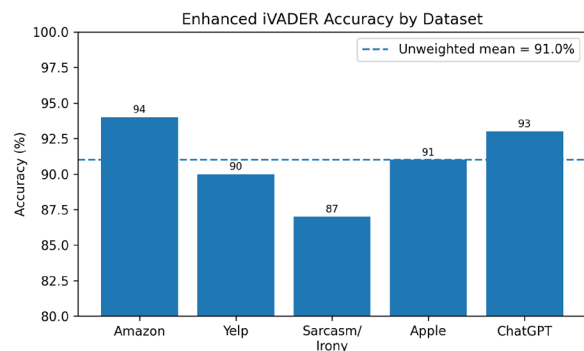


Figure 3. Enhanced iVADER accuracy across the five datasets.

Figure 3 displays enhanced iVADER accuracy across the five datasets together with the unweighted domain mean. The dispersion reinforces the observation that performance is domain-dependent rather than uniform across the collection.

3.3 Limitations and Future Work

Several limitations qualify the findings. The five source datasets use heterogeneous annotation schemes, and an explicit uniform label-mapping procedure is not preserved. In addition, only one 80:20 split is documented, so repeated-run uncertainty and statistical significance cannot be evaluated, and record-level predictions are unavailable for reconstructing a normalized confusion matrix or auditable error cases.

The English-language focus also limits multilingual generalization. Informal spelling, emerging terminology, irony, implicit evaluation, and mixed-aspect statements remain difficult, particularly in social-media data. These limitations are consistent with broader challenges reported for multi-domain and implicit sentiment analysis [3], [12], [14], [24].

Computational evidence is also incomplete because wall-clock time, throughput, memory use, hardware specifications, and software-version metadata were not archived. The reported experiment therefore supports a predictive-performance comparison but does not establish an accuracy-efficiency trade-off.

Future work should release the complete preprocessing and lexicon-update code, fixed random seeds, exact class mappings, dependency versions, fold-level predictions, and profiling logs. Leave-one-domain-out evaluation would provide a stronger test of whether the framework generalizes to an unseen domain rather than only to held-out records from observed datasets.

A further extension should combine adaptive lexical scores with contextual transformer representations and conduct ablation experiments on negation handling, contextual word alignment, TF-IDF, embeddings, and lexicon expansion. Such experiments would identify the contribution of individual components more precisely [11], [12], [25], [26].

3.4 Qualitative Analysis of Difficult Expressions

Representative difficult expressions illustrate two persistent weaknesses. In the sarcastic construction 'Amazing! How terrible this product is!', the positive cue 'amazing' is immediately contradicted by 'terrible', and the intended evaluation depends on pragmatic incongruity rather than a simple sum of lexical scores. A document-level lexical model may therefore assign excessive weight to the initial positive intensifier before resolving the negative proposition.

A second challenge is mixed-aspect evaluation. In 'This laptop is great, but its battery life is terrible', the words 'great' and 'terrible' refer to different product aspects, while a document-level classifier must compress them

into one sentiment label. A similar conflict occurs in 'The service was bad, but the food was excellent.' These cases show why domain adaptation can improve recognition of explicit domain vocabulary without fully resolving aspect-level conflict or discourse structure.

These sentences are treated as diagnostic examples rather than as verified row-level errors because document identifiers and gold/predicted label pairs are not available. The distinction prevents illustrative language analysis from being presented as unsupported empirical error counts.

4. Conclusions

This study evaluated an enhanced iVADER framework for multi-domain sentiment classification by integrating domain-specific lexical adaptation with systematic preprocessing and machine-learning-supported feature representations. The documented corpus comprised 28,000 English-language records drawn from Amazon Cells Labelled, Yelp Labelled, Tweets with Sarcasm and Irony, Apple Twitter Sentiment Texts, and ChatGPT Tweets, using a stratified 80:20 split that produced 22,400 training records and 5,600 test records. Across the pooled experiment, enhanced iVADER achieved 92% accuracy, 90% precision, 89% recall, and an F1-score of 0.91, compared with 85%, 82%, 75%, and 0.80 for conventional VADER.

The largest improvement was observed in recall, which increased by 14 percentage points, indicating that domain-adaptive lexical information helped recover sentiment cues missed by the static baseline. Domain-level analysis further showed that performance was not uniform: accuracy reached 94% on Amazon reviews and fell to 87% on sarcasm-and-irony tweets, with an unweighted mean of 91.0% and a sample standard deviation of 2.74 percentage points. These results support the use of adaptive lexicons when sentiment is expressed through explicit domain vocabulary, while also showing that irony, pragmatic reversal, and mixed-aspect statements remain challenging for document-level lexical scoring.

The main contribution is therefore an interpretable extension of VADER that incorporates domain-sensitive lexical and semantic evidence without replacing the transparent scoring process with a fully opaque model. The conclusions are limited to the documented observed-domain split because repeated-run outputs, uniform label mappings, record-level predictions, software versions, and runtime profiles were not preserved. Future work should release complete executable configurations, perform leave-one-domain-out and ablation experiments, and evaluate contextual transformer integration so that generalization, component-level contributions, statistical uncertainty, and computational cost can be assessed more rigorously.

Declarations

AI Usage Statement

OpenAI ChatGPT was used during manuscript revision to assist with English language editing. The authors are responsible for reviewing and approving the final text, as well as for the accuracy, integrity, interpretation, and conclusions of the research.

Author Contributions

Saleh Yaakub: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing-original draft, and visualization. Mustafa Mat Deris: conceptualization, methodology, supervision, project administration, and writing-review and editing. Ika Safitri Windiarti: validation, formal analysis, resources, visualization, and writing-review and editing. All authors reviewed and approved the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare no conflict of interest regarding the publication of this study.

Ethical Clearance

This study analyzed previously available, publicly accessible textual datasets and did not involve direct recruitment of human participants or interventions.

Data and Code Availability

This study analyzes five English-language datasets described in Section 2.2. The Amazon Cells Labelled and Yelp Labelled datasets are available in the UCI Machine Learning Repository (Sentiment Labelled Sentences, Dataset 331; <https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences>; DOI: 10.24432/C57604) under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Details regarding the sources of the other mentioned datasets can be found at <https://www.kaggle.com/datasets/nikhiljohnk/tweets-with-sarcasm-and-irony> and <https://www.kaggle.com/datasets/seriousran/appletwitter-sentiment-texts>. The collection of ChatGPT-related tweets is available at <https://www.kaggle.com/datasets/manishabhattach22/tweets-on-chatgpt-chatgpt>. The original VADER baseline code is publicly available at <https://github.com/cjhutto/vaderSentiment> under the MIT license.

References

- [1] C. Puschmann and A. Powell, "Turning Words Into Consumer Preferences : How Sentiment Analysis Is Framed in Research and the News Media," 2018, DOI: <https://doi.org/10.1177/2056305118797724>.
- [2] L. Theodorakopoulos and A. Theodoropoulou, "Review Article Leveraging Big Data Analytics for Understanding Consumer Behavior in Digital Marketing : A Systematic Review," vol. 2024, 2024, DOI: <https://doi.org/10.1155/2024/3641502>.
- [3] K. Alahmadi, S. Alharbi, J. Chen, and X. Wang, "Generalizing sentiment analysis : a review of progress , challenges , and emerging directions," Soc. Netw. Anal. Min., vol. 15, no. 1, pp. 1–28, 2025, DOI: <https://doi.org/10.1007/s13278-025-01461-8>.
- [4] P. Singhal, R. Walambe, S. Ramanna, and K. Kotecha, "Domain Adaptation: Challenges, Methods, Datasets, and Applications," IEEE Access, vol. 11, no. December 2022, pp. 6973–7020, 2023, DOI: <https://doi.org/10.1109/ACCESS.2023.3237025>.
- [5] Y. Qi and Z. Shabrina, "Sentiment analysis using Twitter data : a comparative application of lexicon - and machine - learning - based approach," Soc. Netw. Anal. Min., vol. 13, no. 1, pp. 1–14, 2023, DOI: <https://doi.org/10.1007/s13278-023-01030-x>.
- [6] F. Z. Xing, F. Pallucchini, and E. Cambria, "Cognitive-inspired domain adaptation of sentiment lexicons," Inf. Process. Manag., vol. 56, no. 3, pp. 554–564, 2019, DOI: <https://doi.org/10.1016/j.ipm.2018.11.002>.
- [7] A. Hotho, J. Würzburg, R. Jäschke, and K. Lerman, "Sentiment lexicon adaptation with context and semantics for the social web," vol. 8, pp. 643–665, 2017, DOI: <https://doi.org/10.3233/SW-170265>.
- [8] K. Barik and S. Misra, "Analysis of customer reviews with an improved VADER lexicon classifier," J. Big Data, vol. 11, no. 1, 2024, DOI: <https://doi.org/10.1186/s40537-023-00861-x>.
- [9] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," IEEE Access, vol. 9, no. M1, pp. 52177–52192, 2021, DOI: <https://doi.org/10.1109/ACCESS.2021.3069001>.
- [10] Y. Matrane, F. Benabbou, and Z. Ellaky, "Enhancing Moroccan Dialect Sentiment Analysis through Optimized Preprocessing and transfer learning Techniques," IEEE Access, vol. 12, no. December, pp. 187756–187777, 2024, DOI: <https://doi.org/10.1109/ACCESS.2024.3514934>.
- [11] D. Suresh, A. Naresh, K. Nagwani, and P. Singh, Impact of word embedding models on text analytics in deep learning environment : a review, no. 0123456789. Springer Netherlands, 2023. DOI: <https://doi.org/10.1007/s10462-023-10419-1>.
- [12] E. Zuo, A. Aysa, M. Muhammad, Y. Zhao, and K. Ubul, "Context aware semantic adaptation network for cross domain implicit sentiment classification," Sci. Rep., pp. 1–14, 2021, DOI: <https://doi.org/10.1038/s41598-021-01385-1>.
- [13] S. Sanagar and D. Gupta, "Unsupervised Genre-Based Multidomain Sentiment Lexicon Learning Using Corpus-Generated Polarity Seed Words," vol. 8, 2020, DOI: <https://doi.org/10.1109/ACCESS.2020.3005242>.
- [14] N. A. Abdullah, A. Feizollah, A. Sulaiman, and N. B. Anuar, "Challenges and Recommended Solutions in Multi-Source and Multi-Domain Sentiment Analysis," IEEE Access, vol. 7, pp. 144957–144971, 2019, DOI: <https://doi.org/10.1109/ACCESS.2019.2945340>.
- [15] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist, "ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding," IEEE Trans. Vis. Comput. Graph., vol. 24, no. 1, pp. 361–370, 2018, DOI: <https://doi.org/10.1109/TVCG.2017.2744478>.
- [16] M. Alfreihat, O. S. Almousa, Y. Tashtoush, A. Alsobeh, K. Mansour, and H. Migdady, "Emo-SL Framework: Emoji Sentiment Lexicon Using Text-Based Features and Machine Learning for Sentiment Analysis," IEEE Access, vol. 12, no. February, pp. 81793–81812, 2024, DOI: <https://doi.org/10.1109/ACCESS.2024.3382836>.
- [17] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "International Journal of Research in Marketing More than a Feeling : Accuracy and Application of Sentiment Analysis,"

- Int. J. Res. Mark., vol. 40, no. 1, pp. 75–87, 2023, DOI: <https://doi.org/10.1016/j.ijresmar.2022.05.005>.
- [18] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, “Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis,” *Inf. Process. Manag.*, vol. 52, no. 1, pp. 36–45, 2016, DOI: <https://doi.org/10.1016/j.ipm.2015.04.003>.
- [19] N. Alnazzawi, P. Thompson, and S. Ananiadou, “Mapping Phenotypic Information in Heterogeneous Textual Sources to a Domain-Specific Terminological Resource,” pp. 1–27, 2016, DOI: <https://doi.org/10.1371/journal.pone.0162287>.
- [20] M. Xiang, J. Grove, and A. Giannakidou, “Semantic and pragmatic processes in the comprehension of negation: An event related potential study of negative polarity sensitivity,” *J. Neurolinguistics*, vol. 38, pp. 71–88, 2016, DOI: <https://doi.org/10.1016/j.jneuroling.2015.11.001>.
- [21] A. Demircioğlu, “Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics,” *Insights Imaging*, pp. 1–10, 2021, DOI: <https://doi.org/10.1186/s13244-021-01115-1>.
- [22] M. A. Talukder et al., “A hybrid deep learning model for sentiment analysis of COVID-19 tweets with class balancing,” *Sci. Rep.*, vol. 15, no. 1, pp. 1–19, 2025, DOI: <https://doi.org/10.1038/s41598-025-97778-7>.
- [23] P. Yadav, I. Kashyap, and B. S. Bhati, “Contextual Ambiguity Framework for Enhanced Sentiment Analysis,” *Teh. Glas.*, vol. 18, no. 3, pp. 385–393, 2024, DOI: <https://doi.org/10.31803/tg-20231227064230>.
- [24] E. Zuo, H. U. I. Zhao, B. O. Chen, and Q. Chen, “Context-Specific Heterogeneous Graph Convolutional Network for Implicit Sentiment Analysis,” *IEEE Access*, vol. 8, pp. 37967–37975, 2020, DOI: <https://doi.org/10.1109/ACCESS.2020.2975244>.
- [25] A. Onan, “Improving Turkish Text Sentiment Classification Through Task-Specific and Universal Transformations: An Ensemble Data Augmentation Approach,” *IEEE Access*, vol. 12, no. December 2023, pp. 4413–4458, 2024, DOI: <https://doi.org/10.1109/ACCESS.2024.3349971>.
- [26] L. Kelebercová, “Cross-domain sentiment classification in Slovak: a comparative analysis of transformer-based models,” 2026, DOI: <https://doi.org/10.7717/peerj-cs.3641>.