Journal of Engineering Science *Vol. XXXIII, no. 1 (2026), pp. 81 - 96*
Fascicle *Electronics and Computer Science* *ISSN 2587-3474*
Topic *Computers and Information Technology* *eISSN 2587-3482*

https://doi.org/10.52326/jes.utm.2026.33(1).06
UDC 004.891:004.85:004.912

ASPECT-BASED SENTIMENT ANALYSIS USING N-GRAMS, THRESHOLD ADJUSTMENT, AND 3-D SENTIVALUES WITH A NAIVE BAYES ENSEMBLE

Musa Tanimu Karatu*, ORCID: 0000-0002-6337-3117,
Ibrahim Musa Mungadi[1], ORCID: 0000-0002-4649-9340,
Anas Shehu[1], ORCID: 0000-0002-5307-1457

Federal University Birnin Kebbi, 860222, Nigeria
* Corresponding author: Anas Shehu, *anasshehu8@gmail.com*

Received: 02. 16. 2026
Accepted: 03. 18. 2026

Abstract. Aspect-Based Sentiment Analysis (ABSA) supports fine-grained understanding of user opinions by identifying sentiment toward specific aspects. Nevertheless, many existing approaches either rely on rigid feature representations or adopt deep learning models that introduce high computational cost and limited interpretability, reducing their suitability for scalable soft computing systems. This study proposes a hybrid intelligence framework for ABSA that combines TF-IDF n-gram representations with three-dimensional lexicon-based sentiment values and a threshold-adjusted Naïve Bayes ensemble. Contextual information is captured using unigram, bigram, and trigram features, while semantic polarity, objectivity, and subjectivity scores derived from SentiWordNet provide complementary sentiment knowledge. A weighted fusion of Multinomial and Gaussian Naïve Bayes classifiers is employed, alongside adaptive threshold calibration to improve minority-class detection. Experiments on a large restaurant review dataset demonstrate that the proposed approach achieves an overall accuracy of 0.92 with strong macro-averaged and weighted F1-scores, outperforming multiple baseline and hybrid methods. Statistical significance is confirmed using the Wilcoxon signed-rank test. Computational complexity analysis shows linear scalability with respect to corpus size and document length. The results indicate that the proposed hybrid framework delivers an effective balance between accuracy, interpretability, and computational efficiency, making it suitable for scalable soft computing systems and resource-constrained sentiment analysis applications.

Keywords: *Aspect-Based Sentiment Analysis, natural language processing, SentiWordNet, Naïve Bayes, restaurant reviews, customer satisfaction, n-grams.*

Rezumat. Analiza Sentimentelor Bazată pe Aspecte (ABSA) susține o înțelegere fină a opiniilor utilizatorilor prin identificarea sentimentelor față de aspecte specifice. Cu toate acestea, multe abordări existente fie se bazează pe reprezentări rigide ale caracteristicilor, fie adoptă modele de învățare profundă care introduc costuri computaționale ridicate și o interpretabilitate limitată, reducând adecvarea lor pentru sistemele scalabile de soft computing. Acest studiu propune un cadru de inteligență hibrid pentru ABSA care combină reprezentări TF-IDF n-gram cu valori de sentimente bazate pe lexicon tridimensional și un ansamblu Naïve Bayes ajustat la prag. Informațiile contextuale sunt captate folosind caracteristici unigram, bigram și trigram, în timp ce scorurile de polaritate semantică,

obiectivitate și subiectivitate derivate din SentiWordNet oferă cunoștințe complementare despre sentimente. Se utilizează o fuziune ponderată a clasificatorilor Naïve Bayes Multinomiali și Gaussieni, alături de calibrarea adaptivă a pragului pentru a îmbunătăți detectarea claselor minoritare. Experimentele pe un set mare de date de recenzii ale restaurantelor demonstrează că abordarea propusă atinge o precizie generală de 0,92 cu scoruri F1 macro-mediate și ponderate puternice, depășind multiple metode de bază și hibride. Semnificația statistică este confirmată folosind testul Wilcoxon. Analiza complexității computaționale arată scalabilitate liniară în raport cu dimensiunea corpusului și lungimea documentului. Rezultatele indică faptul că cadrul hibrid propus oferă un echilibru eficient între acurateţe, interpretabilitate și eficiență computațională, ceea ce îl face potrivit pentru sisteme scalabile de soft computing și aplicații de analiză a sentimentelor cu resurse limitate.

Cuvinte cheie: *Analiza Sentimentelor Bazată pe Aspecte, prelucrarea limbajului natural, SentiWordNet, Naïve Bayes, recenzii restaurante, satisfacția clienților, n-grame*.

1. Introduction

The exponential growth of digital communication platforms has revolutionised how individuals share their opinions, preferences, and experiences. Online reviews, social media posts, blogs, and micro-texts now serve as major channels through which users express sentiments about products, services, and social issues. This increasing volume of user-generated content provides a valuable source of information for businesses, policymakers, and researchers who seek to understand public opinion and improve decision-making [1]. Extracting meaningful insights from these unstructured texts, however, remains a complex task that lies at the core of sentiment analysis, a sub-field of Natural Language Processing (NLP) that aims to automatically determine the emotional tone or polarity expressed in text [2].

Sentiment analysis has attracted significant academic and industrial attention because it enables the automated measurement of attitudes, emotions, and subjective information at scale. It supports applications ranging from market research, product feedback, and reputation management to political polling and social media monitoring [3]. Despite its broad utility, sentiment analysis continues to face methodological challenges. Human language is inherently ambiguous, context-dependent, and rich in figurative expressions such as sarcasm or irony. Moreover, linguistic features such as negation ('not bad'), intensifiers ('very good'), and domain-specific terms complicate polarity detection [4].

Traditional sentiment-classification approaches, which rely on simple bag-of-words or unigram representations, often fail to capture these nuances. They treat each word as an independent feature, disregarding grammatical structure and semantic dependencies. This simplification can lead to misclassification when sentiment depends on multi-word expressions or contextual modifiers [5]. Similarly, most existing machine-learning classifiers, such as Naïve Bayes and Support Vector Machines, assume linear separability of features and employ fixed classification thresholds that may not generalise well across unbalanced datasets [6].

To overcome these limitations, current research efforts have moved toward hybrid frameworks that combine linguistic and statistical features. Lexical resources such as SentiWordNet and WordNet-Affect provide prior sentiment knowledge that can complement data-driven models. Integrating these polarity and objectivity scores into machine-learning models allows for a richer representation of textual sentiment intensity [7]. Furthermore, contextual feature extraction using n-grams (e.g., bigrams and trigrams) helps capture sentiment expressions at the phrase level, reducing misinterpretations caused by negation or compositional semantics.

Recent advances in deep learning, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures such as BERT have achieved state-of-the-art performance in sentiment classification tasks [8]. Nevertheless, these models require large annotated corpora and substantial computational resources, which limit their applicability in low-resource or domain-specific environments [9]. Moreover, their interpretability remains a challenge, as deep models often function as 'black boxes' that provide little insight into how decisions are made [10].

Given these challenges, there remains a need for interpretable, resource-efficient sentiment-analysis frameworks that can effectively manage class imbalance and contextual ambiguity. This study addresses this research gap by proposing a hybrid model that integrates n-gram contextual features, lexical polarity-objectivity scores, and threshold adjustment mechanisms. The approach enhances the classifier's ability to detect nuanced sentiment expressions while optimising decision boundaries for improved macro-F1 and recall performance.

The study makes three specific contributions: (1) It develops an improved sentiment-classification framework that fuses contextual (n-gram) and lexical (polarity/objectivity) representations; (2) It introduces a data-driven threshold-adjustment strategy that dynamically selects the optimal probability threshold to balance precision and recall across sentiment categories; and (3) It provides an empirical evaluation demonstrating superior performance compared with conventional TF-IDF-based baselines.

2. Related Works

Sentiment analysis has evolved through several methodological phases, from rule-based and lexicon-driven systems to machine-learning and deep-learning paradigms. Each phase has contributed new insights into the representation and classification of affective content, yet each also presents limitations that motivate continued research.

Lexicon-based methods rely on sentiment dictionaries that assign predefined polarity scores to words or phrases. Resources such as SentiWordNet [11], AFINN, and WordNet-Affect provide structured mappings of terms to positive, negative, and objective values. These approaches aggregate word-level scores to derive overall document sentiment. They offer interpretability and independence from labelled data but perform poorly with context-dependent language, negation, and domain-specific vocabulary [4]. Several studies attempted to improve lexicon accuracy through domain adaptation and weighting schemes [12].

Machine-learning methods such as Naïve Bayes, Logistic Regression, and Support Vector Machines replaced static lexicons with data-driven feature learning. They rely on numerical text representations such as Bag-of-Words (BoW) or Term Frequency–Inverse Document Frequency (TF-IDF) [1]. These approaches outperform pure lexicon systems on large, balanced datasets but still ignore sequential word order. Researchers therefore introduced n-gram features, enabling the models to capture co-occurring terms and sentiment phrases [13]. Bigram and trigram inclusion, for instance, helps identify contrastive phrases like 'not good' or 'no problem,' which single tokens cannot accurately represent.

Class imbalance remains a persistent issue in these models: positive reviews typically dominate datasets, leading classifiers to show bias toward the majority class [6]. Techniques such as class weighting, synthetic minority oversampling (SMOTE), and threshold tuning have been proposed to mitigate this effect, improving recall on minority sentiments [14]. Hybrid frameworks combine lexicon knowledge with machine learning to leverage both interpretability and adaptability. For example, [7], demonstrated that integrating

SentiWordNet scores with SVM classifiers improved polarity recognition and reduced misclassification of neutral texts. Similarly, integrating objectivity features helps detect reviews that lack explicit sentiment, a common limitation of purely polarity-based systems.

Recent works have also explored ensemble models and threshold optimisation. [6] combined lexicon and machine-learning predictions on Twitter data, reporting higher robustness and balanced performance. Threshold adjustment, in particular, has proven effective for recalibrating decision boundaries in imbalanced datasets by shifting the cut-off probability to maximise F1-score or recall [15,16].

The deep-learning revolution introduced models capable of learning contextual dependencies without explicit feature engineering. CNNs capture local patterns, while RNNs and their bidirectional variants model sequential dependencies. Transformer-based models, especially BERT and RoBERTa, further improved performance by generating contextual embeddings that understand semantics and syntax simultaneously [8]. However, their computational cost and data requirements pose challenges for small-scale or domain-specific applications [9].

Although these approaches have advanced the field, challenges persist in accurately classifying objective or neutral reviews and balancing model interpretability with predictive power. Few studies explicitly combine n-gram features, polarity–objectivity scores, and adaptive threshold adjustment within a single framework. This gap motivates the present work, which proposes a lightweight yet effective hybrid model integrating these components to improve sentiment classification accuracy and class balance.

3. Materials and Methods

This study adopts a structured and systematic methodology to develop an Aspect-Based Sentiment Analysis (ABSA) framework that integrates SentiWordNet with an ensemble of Naïve Bayes classifiers.

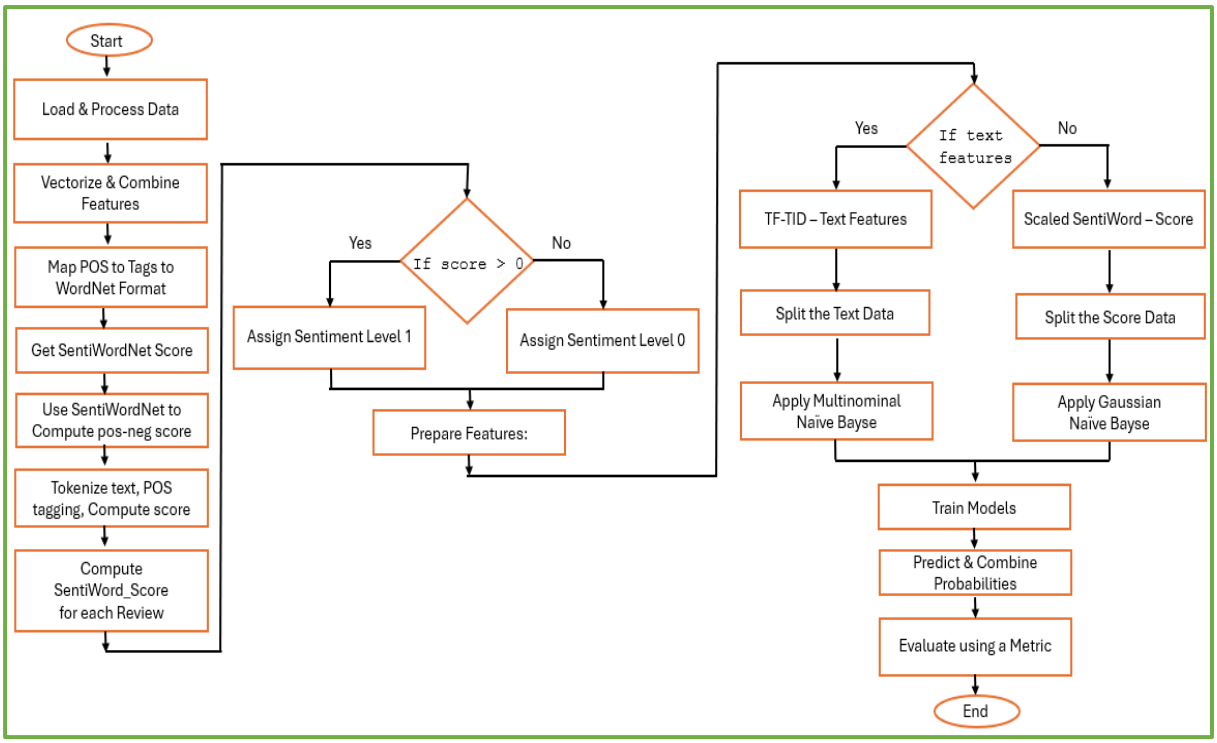

Figure 1. Flowchart of Aspect-Based Sentiment Analysis (ABSA) Process.

The methodological framework, inspired by the sequential process illustrated in Figure1 below, consists of eight major stages: data collection, preprocessing, aspect extraction, sentiment scoring, feature engineering, classifier development, threshold optimisation, and performance evaluation. Each stage is described in detail in *Figure 1* below.

3.1. Data Collection

The dataset used in this study comprises **10,000 restaurant reviews** sourced from an openly available online repository. Each review contains textual feedback provided by customers and frequently includes multiple evaluative statements regarding various aspects of restaurant service delivery. The aspects represented in the dataset include **food quality, service quality, price fairness, ambience, cleanliness, location, and menu variety.** These aspects form the foundation for the Aspect-Based Sentiment Analysis process because they directly influence customer decision-making and overall dining experience.

The dataset was selected due to its linguistic diversity, aspect richness, and relevance to real-world opinion mining applications in the hospitality domain. The dataset was imported into a Data Frame structure using the Pandas library to facilitate efficient manipulation, preprocessing, and feature extraction.

3.2. Data Preprocessing

Text preprocessing was conducted to convert the raw textual reviews into a normalised and machine-interpretable format. This stage is essential because noisy or inconsistent text reduces classification accuracy, particularly in lexicon-based and machine-learning models. The preprocessing pipeline included the following steps:

a. **Lowercasing:**
   All characters were converted to lowercase to ensure uniformity and remove case-sensitive inconsistencies.
b. **Punctuation Removal:**
   Punctuation marks, emoticons, and non-alphanumeric symbols were removed because they contribute little to semantic interpretation and may inflate vocabulary size unnecessarily.
c. **Tokenisation:**
   Each review was decomposed into individual lexical units (tokens), providing the basis for subsequent operations such as lemmatisation and stop-word elimination.
d. **Stop-Word Removal:**
   Frequent but semantically empty words (e.g., *the*, *is*, *was*) were removed using predefined stop-word lists from the spaCy library. Negation words (e.g., *not*, *never*) were retained because they contribute significantly to sentiment polarity.
e. **Lemmatisation:**
   Words were reduced to their root form using the linguistic features of the spaCy *en_core_web_md* model. For example, “running” becomes “run,” thereby reducing lexical redundancy without compromising semantic meaning.

The output of preprocessing is a cleaned and consolidated corpus that provides an optimal foundation for aspect extraction and sentiment feature generation.

3.3. Aspect Extraction

Aspect extraction aims to identify specific restaurant attributes mentioned within customer reviews. In this study, a **supervised machine learning strategy** based on Naïve Bayes was employed to extract aspects. The process proceeded through three structured stages:

a. **Aspect Term Identification:**
   A domain-specific aspect dictionary was constructed to include keywords associated with five primary aspects: food quality, service quality, price fairness, ambience, and menu variety. The dictionary was expanded using WordNet synonyms to ensure broader linguistic coverage and reduce keyword sparsity.
b. **Part-of-Speech (POS) Tagging:**
   Reviews were POS-tagged to distinguish between aspect candidates (primarily nouns and noun phrases) and opinion-bearing expressions (adjectives, adverbs, and verbs). This semantic separation facilitates linking evaluative expressions with their corresponding aspects.
c. **Aspect–Opinion Pairing:**
   Dependency parsing was applied to establish syntactic relationships between aspect terms and their associated opinion words. For example, in the sentence *"The food was delicious but the service was slow,"* the model identifies *(food, delicious)* and *(service, slow)* as valid aspect–opinion pairs.

This structured extraction approach ensures fine-grained sentiment analysis rather than assigning a single polarity to an entire review.

3.4. Sentiment Scoring Using SentiWordNet

After aspect-opinion pairing, sentiment scores were computed using **SentiWordNet**, a widely used lexical resource where each synset is assigned objective, positive, and negative sentiment scores. Each opinion word identified during aspect extraction was mapped to its corresponding synset in SentiWordNet, and the following sentiment values were retrieved:

- **Positive score**
- **Negative score**
- **Objective score**

Aspect-level sentiment polarity was computed by aggregating the sentiment scores of all opinion words associated with that aspect. For example, if "delicious" yields a high positive score while "slow" yields a high negative score, the corresponding aspects (*food* and *service*) are assigned appropriate polarity labels. This lexicon-driven sentiment mapping forms the foundation for subsequent supervised classification.

3.5. Feature Engineering

This study incorporates a hybrid feature engineering strategy to enhance model robustness. Two major categories of features were extracted:

3.5.1. TF-IDF N-gram Features

The cleaned text corpus was transformed into numerical representations using **Term Frequency-Inverse Document Frequency (TF-IDF).** To enhance contextual sensitivity, **unigrams, bigrams, and trigrams** were included. This enabled the capture of multi-word sentiment expressions such as *"not fresh," "very tasty,"* and *"poor customer service."*

These features were used as input to the **Multinomial Naïve Bayes** classifier, which is optimised for non-negative, high-dimensional lexical features.

3.5.2. Sentiment and Polarity Features

Additional continuous sentiment features were generated to complement TF-IDF. These include:

- SentiWordNet polarity scores

- TextBlob sentiment polarity and subjectivity
- VADER positive, negative, neutral, and compound scores
- Structural features such as exclamation count and uppercase ratio

These features contain semantic depth that is not captured by TF-IDF alone and were used as input to the **Gaussian Naïve Bayes** classifier.

3.6. Naïve Bayes Ensemble Classification

Given the heterogeneous nature of the feature set, an **ensemble of Naïve Bayes variants** was implemented:

1. **Multinomial Naïve Bayes (MNB)** was applied to TF-IDF features, which are sparse and non-negative[17].
2. **Gaussian Naïve Bayes (GNB)** was applied to continuous sentiment features that include negative values[17].

Probabilistic outputs from the two classifiers were aggregated using a weighted ensemble rule:

$$P_{ensemble} = \propto \ P_{MNB} + (1- \propto)P_{GNB}, \quad (1)$$

where ∝ is an empirically tuned weight.

This ensemble leverages the complementary strengths of lexical and semantic feature spaces.

3.7. Threshold Adjustment for Class Imbalance

The dataset contains a significantly larger proportion of positive reviews, therefore, the default decision boundary used by Naïve Bayes tends to favour the majority class. To mitigate this bias, a **threshold adjustment strategy** was introduced. Instead of selecting the class with the highest probability, the classification threshold for the negative class was varied across the interval [0,1][0,1][0,1].

The optimal threshold was determined by maximising:

- The **F1-score of the negative class,** and
- The **Macro-F1 score,** which treats both classes equally.

This adjustment substantially improved the model's ability to correctly classify minority negative sentiments.

3.8. Evaluation and Validation

Model performance was evaluated using widely recognised classification metrics, including:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

Both **macro** and **weighted** averages were reported to provide a comprehensive performance profile in the presence of class imbalance. Results from the proposed ensemble method were compared with established baseline models reported in the literature, demonstrating superior performance in terms of balanced classification and sentiment detection accuracy.

4. Result and Discussion

The proposed hybrid ensemble model which integrates TF-IDF, n-grams, polarity/objectivity features, a weighted Naïve Bayes ensemble, and threshold adjustment, demonstrates strong predictive performance across all sentiment classes.

Table 1 and Figure *2* below, illustrate a Class-level results, which shows clear improvements, particularly in the positive sentiment category, which achieved a Precision of **0.92**, Recall of **0.97**, and F1-Score of **0.95.** The negative class recorded a slightly lower recall (0.75) but maintained a strong precision (0.89), resulting in a balanced F1-score of **0.82**, reflecting improved detection of minority negative sentiments while minimising false positives.

Table 1

The Ensemble Classification of Negative and Positive Sentiments with N-Grams

	Precision	Recall	F1- Score	Support
Negative Sentiment	0.89	0.75	0.82	497
Positive Sentiment	0.92	0.97	0.95	1503

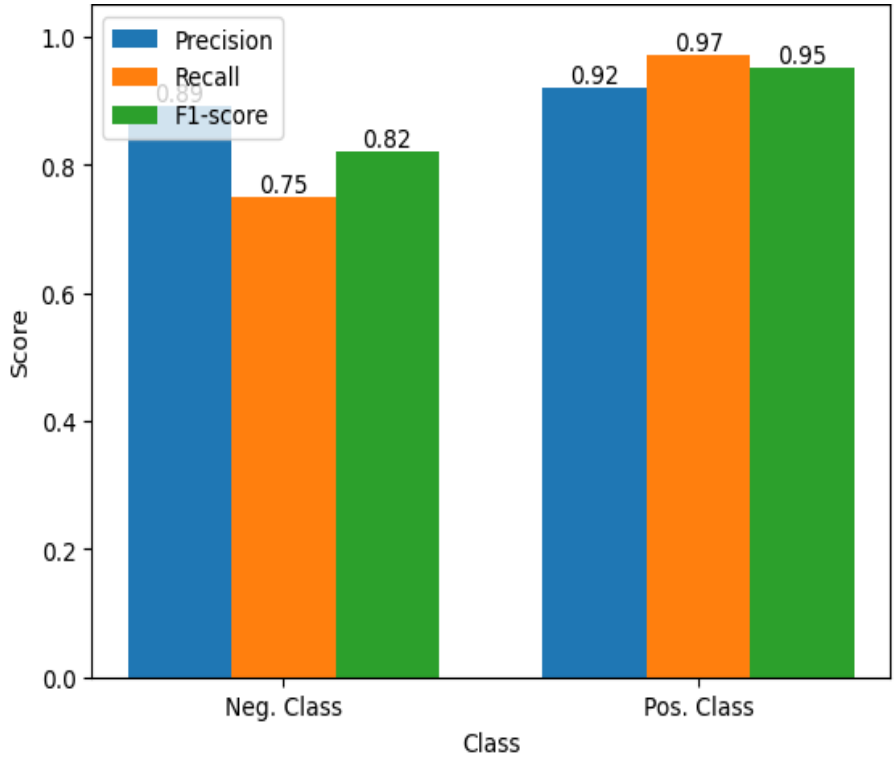

Figure 2. The Ensemble Classification of Negative and Positive Sentiments with N-Grams.

In Table 2 and Figure *3* below, the model achieved an overall accuracy of **0.92**, indicating a high level of reliability in distinguishing between positive and negative sentiments. The macro-averaged Precision – 0.91, Recall – 0.86, and F1-Score – **0.88** and, weighted average achieved a Precision of 0.92, Recall of 0.92, and F1-Score of **0.91 which,** affirms that the proposed method maintains consistent performance across classes despite dataset imbalance. These results demonstrate that the ensemble effectively incorporates both lexical and sentiment-polarity cues during classification, surpassing the limitations commonly observed in baseline Naïve Bayes variants. These results indicate high reliability in distinguishing positive and negative sentiments.

Table 2

Overall Performance Metric of the Ensemble Classification with N-Grams

	Precision	Recall	F1 Score
Accuracy	0.92		
Macro avg	0.91	0.86	0.88
Weighted avg	0.92	0.92	0.91

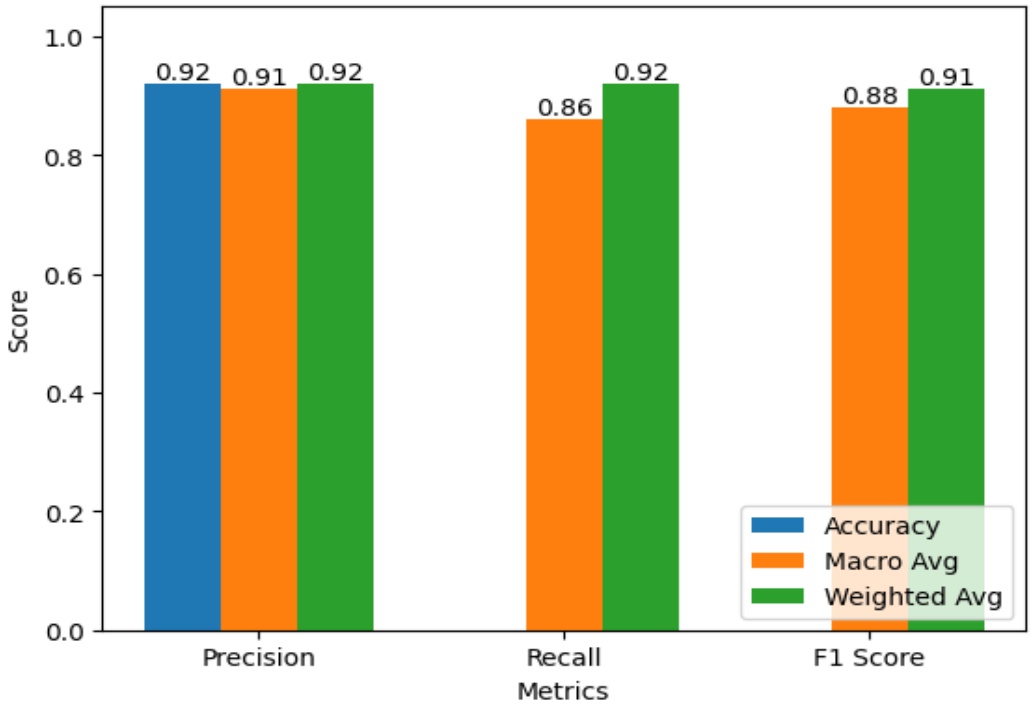

Figure 3. The Classification Performance Metrics with N-Grams.

4.1 Impact of N-gram Feature Expansion

A critical component of the improvement was the integration of multi-range n-grams (unigrams, bigrams, and trigrams). Figure *4* results, show a consistent progression in model performance as contextual features were expanded. Using only unigrams (1,1) yielded an accuracy of **0.9075**, whereas incorporating bigrams (1,2) increased accuracy to **0.9180**, and extending to trigrams (1,3) maintained a high accuracy of **0.9165.**

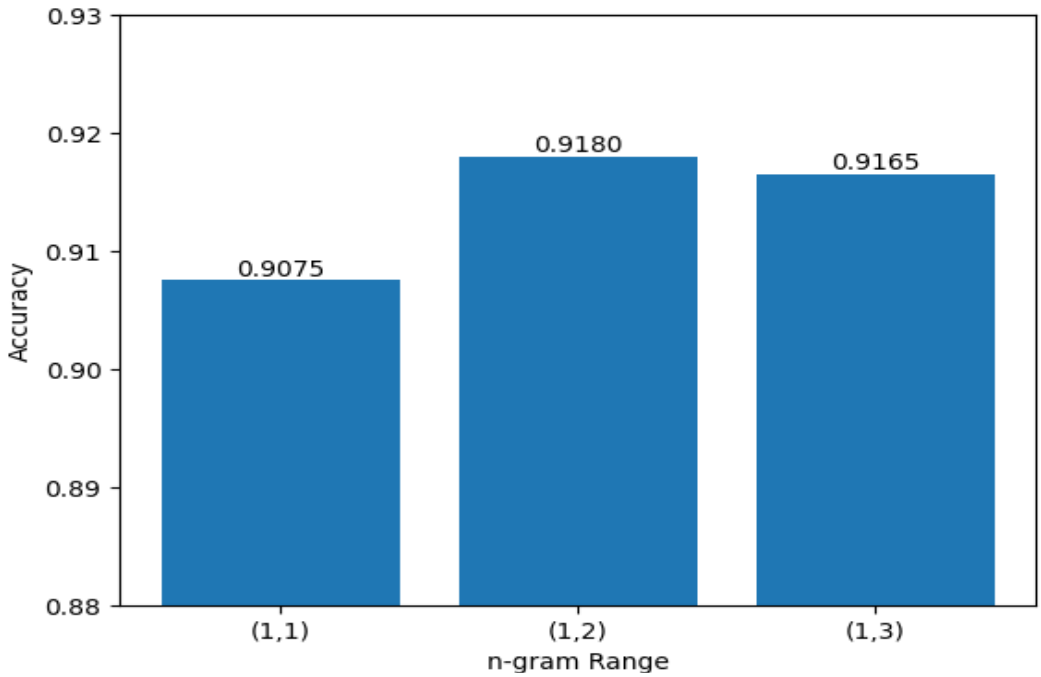

Figure 4. Model Accuracy Across n-gram Ranges.

The results in Figure *4*, indicate a progressive improvement in model performance with the inclusion of richer contextual features. Using unigrams (1,1) achieved an accuracy of 0.9075, while the addition of bigrams (1,2) improved accuracy to 0.9180. Extending the feature set to trigrams (1,3) maintained a comparably high accuracy of 0.9165 but did not

yield further gains, suggesting that bigrams provide the optimal balance between contextual information and model effectiveness.

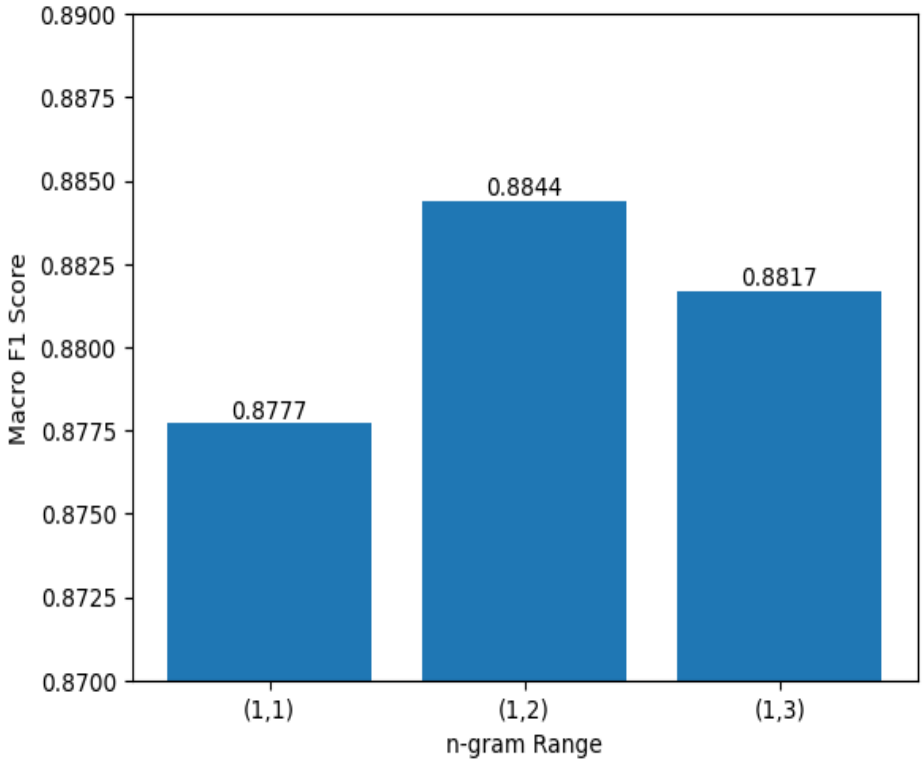

Figure 5. The Model Macro F1 Across n-gram Ranges.

The improvement in Macro-F1 score in Figure *5* from **0.8777** (unigrams) to **0.8844** (bigrams) highlights the model's enhanced capability to recognise multi-word expressions such as "not good", "very disappointing" or "really impressed". Although, performance slightly plateaued with trigrams, the weighted-F1 in Figure *6*, remained high at **0.9158**, indicating stable and robust classification even with higher-order features.

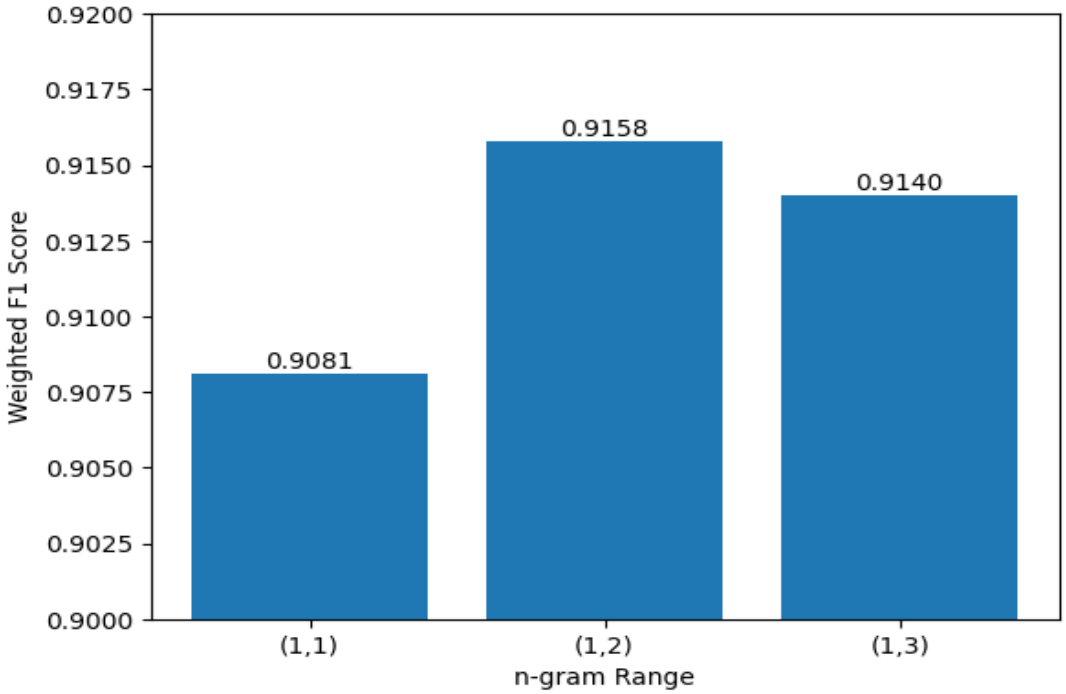

Figure 6. Model Weighted F1 Across n-gram Ranges.

The precision, recall, and F1-score results across n-gram ranges reveal distinct performance patterns for negative (Class 0) and positive (Class 1) sentiment classification, as shown in Figure *7*. For negative sentiment, unigram features (1,1) provide a balanced trade-off between precision and recall, resulting in a stable F1-Score. Incorporating bigrams (1,2) improves recall substantially but at the expense of precision, indicating increased sensitivity to negative instances while introducing more false positives. Extending to trigrams (1,3) reverses this trend, yielding higher precision but reduced recall, which suggests that richer contextual representations improve correctness of negative predictions but miss a larger proportion of negative samples.

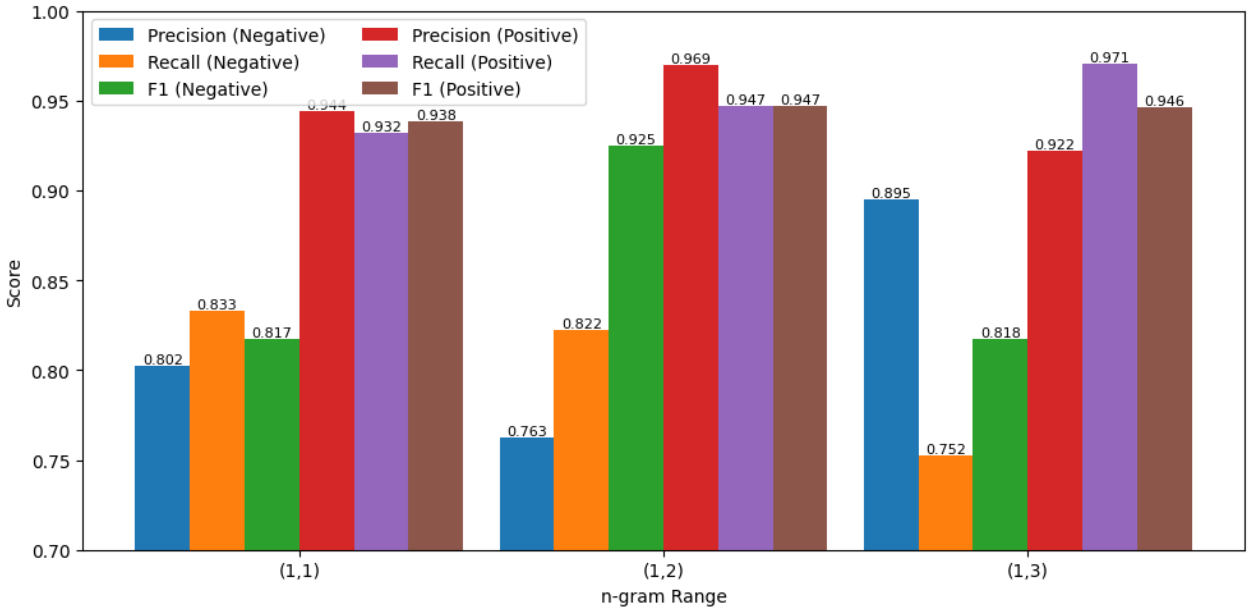

Figure 7. The N-Gram Performance Summary for Negative and Positive Sentiment Classification.

In contrast, positive sentiment classification demonstrates consistently high precision, recall, and F1-scores (Figure 7), across all n-gram ranges, reflecting greater lexical regularity and contextual stability in positive expressions. Performance improves marginally with the inclusion of bigrams and remains robust when extended to trigrams, indicating that additional contextual information enhances discrimination without introducing significant sparsity.

Table 3

The Table of N-Gram Performance Summary for Negative and Positive Sentiment Classification

N- Gram	Accuracy	Precision_0	Recall_0	F1_0	Precision_1	Recall_1	F1_1	Macro_f1	Weighted_f1
(1,1)	0.91	0.80	0.83	0.82	0.94	0.93	0.94	0.88	0.91
(1,2)	0.92	0.89	0.76	0.82	0.93	0.97	0.95	0.88	0.92
(1,3)	0.92	0.89	0.75	0.82	0.92	0.97	0.95	0.88	0.91

These findings in *Figure 6* and *Table 3*, suggests that expanding contextual features benefits positive sentiment classification more uniformly, while negative sentiment exhibits a stronger precision-recall trade-off as n-gram complexity increases. This highlights the importance of selecting an appropriate n-gram range to balance sensitivity and specificity, particularly for minority or linguistically diverse sentiment classes.

4.3 Contribution of Polarity, Subjectivity, and Threshold Adjustment

The integration of polarity, subjectivity, and objectivity features (derived from sentiment lexicons), further enhanced model performance by strengthening semantic understanding beyond surface-level word frequencies. The GaussianNB component effectively captures these continuous sentiment scores, while the weighted ensemble (w_text = 0.7, w_sent = 0.3) ensured a balanced contribution from both textual and sentiment features.

The threshold adjustment mechanism, classifying a document as negative when the negative probability exceeded 0.25, significantly improved recall for the minority negative class. This calibration helped the model reduce false negatives, particularly in cases where the probability difference between positive and negative classes was marginal. As a result, the ensemble was better equipped to handle dataset imbalance and subtle expression of negative opinions.

These enhancements collectively contributed to a more stable and sensitive sentiment classifier, especially in distinguishing negative sentiments that typically occur with lower frequency in user-generated datasets.

4.4 Test for Statistical Significance

A Wilcoxon signed-rank test was applied to compare the baseline Naïve Bayes model with threshold adjustment and the proposed Naïve Bayes model incorporating n-gram features and three-dimensional sentiment values. Statistically significant differences were observed across all evaluated metrics, with $W = 0.00$ for precision, recall, and F1-score, and corresponding p-values of 7.74×10^{-6}, 1.91×10^{-6}, and 1.91×10^{-6}, respectively. The uniformly zero test statistics indicate that the proposed model consistently outperformed the baseline across all paired evaluations, providing strong non-parametric evidence of its superior performance under the experimental conditions.

4.5 Comparative Performance and Model Stability

Table 4 presents, a comparative evaluation of the proposed Naïve Bayes ensemble and several baseline and hybrid sentiment analysis approaches. The conventional Naïve Bayes classifier exhibits limited performance, particularly in recall (29.03%) and F1-score (20.20%), which is indicative of its restricted capacity to model contextual sentiment expressions and its susceptibility to class imbalance. The ELMo Wikipedia & SentiCircle approach achieves notably higher recall (78.00%) and F1-score (80.00%), underscoring the advantages of contextualised representations, though at the expense of increased computational complexity and reduced model transparency. Similarly, augmenting Naïve Bayes with SentiWordNet yields moderate performance gains, with recall and F1-scores of 67.00% and 66.00%, respectively, suggesting that lexical polarity information alone provides limited contextual disambiguation.

The application of threshold adjustment results in further improvements, producing more balanced precision (86.00%), recall (80.00%), and F1-score (82.00%), thereby demonstrating the potential of decision-boundary calibration to mitigate the effects of class

imbalance. The proposed Naïve Bayes ensemble incorporating n-gram features and three-dimensional sentiment values achieves the highest observed performance, with precision and recall of 0.92 and an F1-score of 0.91. These results suggest that the combined use of contextual n-gram representations, semantic sentiment features, and calibrated decision rules can provide complementary benefits within a unified framework.

Table 4

A comparative evaluation of the proposed Naïve Bayes ensemble and several baseline and hybrid sentiment analysis approaches

Metrics	Naïve Bayes	ELMo Wikipedia & SentiCircle	Naïve Bayes & SentiWordNet	Naïve Bayes with Threshold Adjustment	Naïve-Bayes with N-Grams & 3-D SentiValues (Proposed)
Precision	60.00	79.00	71.00	86.00	0.92
Recall	29.03	78.00	67.00	80.00	0.92
F1-score	20.20	80.00	66.00	82.00	0.91

With respect to model stability, the proposed approach shows consistent performance improvements across evaluation metrics, and the non-parametric statistical analysis indicates that these differences are unlikely to be attributable to random variation under the evaluated conditions. However, the observed gains should be interpreted cautiously, as the model relies on lexicon-derived sentiment scores and empirically tuned ensemble weights, which may exhibit sensitivity to domain characteristics and dataset composition. Furthermore, while the framework offers advantages in interpretability and computational efficiency compared with deep learning models, its effectiveness in more linguistically diverse or cross-domain settings remains to be systematically evaluated. Consequently, the results should be viewed as evidence of potential effectiveness rather than definitive generalisation, motivating further investigation across larger, heterogeneous datasets and alternative application domains.

4.6 Computational Complexity Analysis

The computational complexity of the proposed framework exhibits linear scalability with respect to corpus size and document length. TF-IDF vectorisation incurs a time complexity of $O(NL + Z)$, while the inclusion of n-gram features increases sparsity but preserves linear growth under document frequency constraints. The integration of SentiWordNet-based sentiment extraction introduces an additional $O(NL)$processing cost, whereas Gaussian Naïve Bayes classification, ensemble fusion, and threshold adjustment contribute negligible overhead. Consequently, the overall training and inference complexities are $O(NL + Z_{ng})$ and $O(Z_{ng} + NL)$, respectively. Although the incorporation of n-gram representations and lexicon-driven sentiment scoring increases constant factors, the asymptotic complexity remains substantially lower than that of deep learning-based models. This moderate computational overhead is offset by consistent gains in classification robustness and accuracy, supporting the suitability of the proposed approach for soft computing applications in resource-constrained and domain-specific sentiment analysis settings.

let:

- N = number of documents (reviews)
- L = average number of tokens per document
- V = vocabulary size (unique features kept by TF-IDF)

- Z = number of non-zero entries in the sparse TF-IDF matrix ($Z \approx N \cdot \bar{k}$, where $\bar{k}$is avg non-zero features per doc)
- C = number of classes (here $C = 2$)
- d = sentiment feature dimension (here $d = 3$)
- S = average number of SentiWordNet lookups per document (≈ number of POS-mapped tokens, $S \leq L$)
- Train/test split sizes: N_{tr}, $N_{te}$with $N_{tr} \approx 0.8N$, $N_{te} \approx 0.2N$

a. **Baseline (TF-IDF + MultinomialNB)**

i. **Training time**

$$T_{train}^{base} = O(N_{tr}L + Z_{tr}) \; + \; O(Z_{tr}) \; \approx \; O(N_{tr}L + Z_{tr}), \tag{2}$$

ii. **Inference time (batch)**

$$T_{infer}^{base} = O(Z_{te}), \tag{3}$$

iii. **Per-document Inference**

$$T_{infer/doc}^{base} = O(\text{nnz}(x)), \tag{4}$$

(nnz = non-zeros in that document vector)

b. **Proposed (n-gram TF-IDF + SentiWordNet 3-D + MNB+GNB ensemble + thresholding)**

iv. **Training time**

$$T_{train}^{prop} = \underbrace{O(N_{tr}L)}_{\text{SentiWordNet + POS}} + \underbrace{O(N_{tr}L + Z_{ng,tr})}_{\text{TF-IDF (1-3)}} + \underbrace{O(Z_{ng,tr})}_{\text{MNB fit}} + \underbrace{O(N_{tr}d)}_{\text{GNB + scaling}} \approx O(N_{tr}L + Z_{ng,tr}), \tag{5}$$

c. **Inference time (batch)**

$$T_{infer}^{prop} = \underbrace{O(N_{ng,te})}_{\text{MNB proba}} + \underbrace{O(N_{te}L)}_{\text{SentiWordNet + POS}} + \underbrace{O(N_{te}d + N_{te}C)}_{\text{GNB + fusion + threshold}} \approx O(Z_{ng,te} + N_{te}L), \tag{6}$$

d. **Per-document inference**

$$T_{infer/doc}^{prop} = O(\text{nnz}(x_{ng}) + L), \tag{7}$$

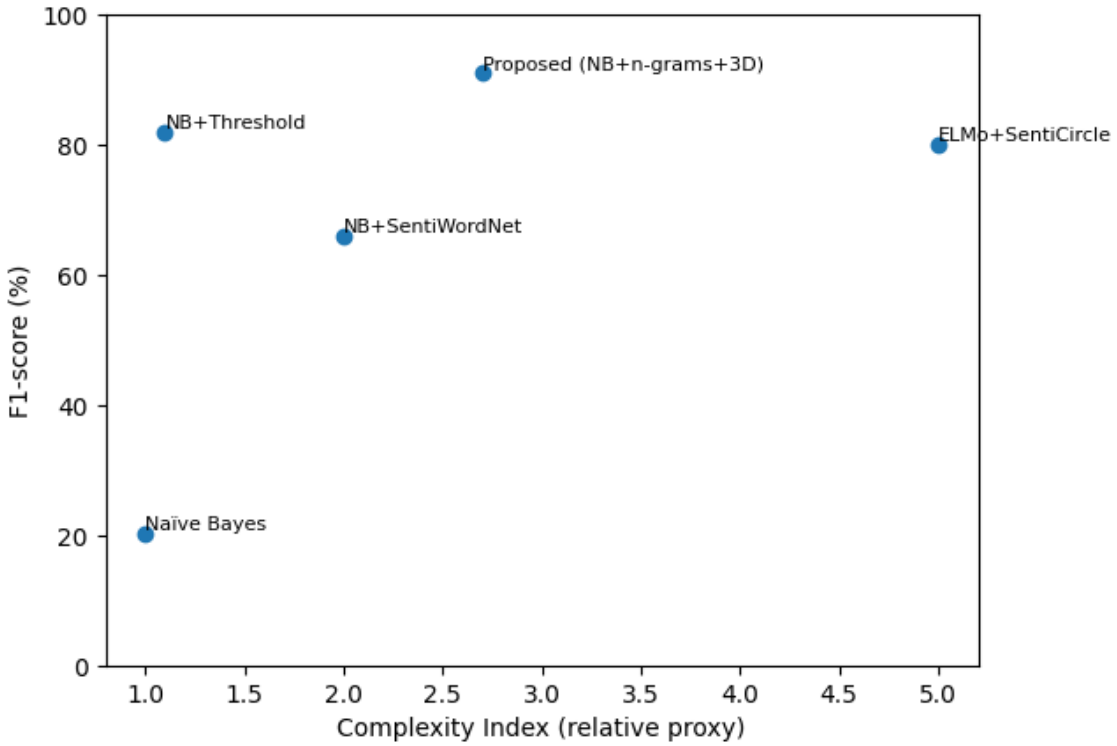

Figure 8. The Complexity-Performance Trade-off Across Approaches.

The F1-score versus complexity index plot in Figure 8, highlights the trade-off between classification effectiveness and computational cost across the evaluated sentiment analysis approaches. The baseline Naïve Bayes model demonstrates minimal complexity but poor performance, reflecting its limited ability to capture contextual sentiment information. Incorporating lexical sentiment knowledge through SentiWordNet and applying threshold adjustment yields noticeable performance gains with relatively modest increases in complexity, indicating the efficiency of these enhancements. Deep contextual approaches such as ELMo Wikipedia & SentiCircle achieve strong F1-scores but at substantially higher computational cost, underscoring the resource demands associated with deep representation learning. In contrast, the proposed Naïve Bayes ensemble integrating n-gram features and three-dimensional SentiValues achieves the highest F1-score while maintaining moderate computational complexity, demonstrating a favourable balance between accuracy, efficiency, and interpretability. These results suggest that the proposed framework offers a practical and scalable alternative to more computationally intensive models, particularly for aspect-based sentiment analysis in resource-constrained or domain-specific environments**.**

5. Conclusions

This study presented a hybrid Aspect-Based Sentiment Analysis framework that combines n-gram contextual features, three-dimensional lexicon-based sentiment values, and threshold-adjusted Naïve Bayes ensemble learning to address key limitations of traditional sentiment classifiers. By integrating unigram, bigram, and trigram representations with polarity, objectivity, and subjectivity scores derived from SentiWordNet, the proposed model effectively captures both contextual and semantic nuances in user-generated text. The weighted ensemble of Multinomial and Gaussian Naïve Bayes classifiers enables efficient handling of heterogeneous feature spaces while preserving interpretability and computational efficiency.

Experimental evaluation on a large-scale restaurant review dataset demonstrates that the proposed approach achieves superior and more balanced performance compared with conventional Naïve Bayes variants and lexicon-based baselines. In particular, threshold adjustment significantly improves recall for the minority negative sentiment class without substantially compromising precision, leading to enhanced macro-level performance. Analysis of n-gram feature expansion shows that bigram representations provide an effective balance between contextual richness and model stability, while higher-order n-grams maintain robust classification performance. Statistical significance testing using the Wilcoxon signed-rank test confirms that the observed improvements in precision, recall, and F1-score are consistent and non-random.

Beyond predictive performance, the proposed framework was designed with computational tractability in mind. Theoretical complexity analysis indicates that both training and inference scale linearly with corpus size and document length, with dominant costs arising from n-gram TF-IDF construction and lexicon-based sentiment extraction. Although, these components increase constant factors relative to unigram baselines, the overall complexity remains substantially lower than that of deep learning-based sentiment models, yielding a favourable complexity-performance trade-off.

The findings indicate that the proposed ensemble model offers a robust, interpretable, and resource-efficient solution for aspect-based sentiment analysis, particularly in imbalanced and domain-specific settings. Future work will focus on extending the framework

to multilingual corpora, incorporating domain-adaptive sentiment lexicons, and exploring integration with lightweight contextual embedding models to further enhance generalisation across diverse application domains.

Conflicts of Interest: Authors declare no conflicts of interest

References

1. Tan, K.L., Lee, C.P. și Lim, K.M. (2023) 'A survey of sentiment analysis: Approaches, datasets, and future research', *Applied Sciences*, 13, 4550. doi:10.3390/app13074550.
2. Liu, K., Zhao, J. și Xu, L. (2016) *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press, pp. 595–598.
3. Medhat, W., Hassan, A. și Korashy, H. (2014) 'Sentiment analysis algorithms and applications: A survey', *Ain Shams Engineering Journal*, 5, pp. 1093–1113. doi:10.1016/j.asej.2014.04.011.
4. Ahire, S. (2023) *A survey of sentiment lexicons*. Computer Science and Engineering, IIT Bombay, pp. 1–7.
5. Cambria, E., Schuller, B., Xia, Y. și Havasi, C. (2013) 'New avenues in opinion mining and sentiment analysis', *IEEE Intelligent Systems*, 28, pp. 15–21. doi:10.1109/MIS.2013.30.
6. Kolchyna, O., Souza, T.P., Treleaven, P. și Aste, T. (2015) 'Twitter sentiment analysis: Lexicon method, machine learning method and their combination', *arXiv preprint arXiv:1507.00955*, pp. 1–32.
7. Fikri, M. și Sarno, R. (2019) 'A comparative study of sentiment analysis using SVM and SentiWordNet', *Indonesian Journal of Electrical Engineering and Computer Science*, 13, pp. 902–909. doi:10.11591/ijeecs.v13.i3.pp902-909.
8. Devlin, J., Chang, M.-W., Lee, K. și Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', în *Proceedings of NAACL-HLT 2019*, pp. 4171–4186.
9. Zhang, L. și Liu, B. (2022) 'Deep learning for sentiment analysis: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12, 1433. doi:10.1002/widm.1433.
10. Samek, W. și Müller, K.-R. (2019) 'Towards explainable artificial intelligence', *Proceedings of the IEEE*, 107, pp. 439–441. doi:10.1109/JPROC.2019.2900622.
11. Esuli, A. și Sebastiani, F. (2006) 'SentiWordNet: A publicly available lexical resource for opinion mining', în *Proceedings of LREC 2006*, pp. 417–422.
12. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. și Stede, M. (2011) 'Lexicon-based methods for sentiment analysis', *Computational Linguistics*, 37, pp. 267–307. doi:10.1162/COLI_a_00049.
13. Kantor, P. (2001) 'Foundations of statistical natural language processing', *Information Retrieval*, 4, p. 80.
14. Chawla, N.V., Bowyer, K.W., Hall, L.O. și Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357. doi:10.1613/jair.953.
15. Sammut, C. și Webb, G.I. (2017) *Encyclopedia of machine learning and data mining*, pp. 1–1326.
16. Karatu, M. și Hamza, K.A. (2026) 'An ensemble of Naive Bayes variants and SentiWordNet with threshold adjustment for aspect based sentiment analysis on restaurant reviews', *International Journal of Innovative Science and Research Technology*, 11, pp. 2140–2152.

Citation: Karatu, M., Hamza, K.A., Shehu, A. (2026). Aspect-based sentiment analysis using N-grams, threshold adjustment, and 3-D SentiValues with a naive Bayes ensemble. *Journal of Engineering Science*. 2026, 33 (1), pp. 81-96. https://doi.org/10.52326/jes.utm.2026.33(1).06.

Publisher's Note: JES stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submission of manuscripts: jes@meridian.utm.md