Efficient Multi-Class Analysis of Consumer Complaints Using Frozen MiniLM Embeddings and Neural Networks

Sri Vishnu Gopinathan[1], Muhammad Faraz Manzoor[2]*
University of Houston-Victoria, United States[1]
Department of Artificial Intelligence, University of Management and Technology, Lahore, Pakistan[2]

Abstract—**Text classification is a critical task in domains generating large volumes of unstructured text, such as finance, healthcare, and consumer services. However, accurately classifying such data remains challenging due to its noisy, imbalanced, and context-dependent nature. While pre-trained language models have improved general text classification, their direct application often overlooks domain-specific cues and sentiment patterns that are important for nuanced understanding. In this study, we propose a novel framework that extends the MiniLM language model by integrating domain-relevant cues and sentiment features with textual embeddings. This integration allows the model to capture both semantic richness and domain-specific patterns, enhancing reliability and interpretability. Comparative experiments against baselines including TF-IDF + Logistic Regression, Word2Vec + Logistic Regression, TF-IDF + Naïve Bayes, and Word2Vec + Naïve Bayes shows that the proposed approach consistently outperforms traditional methods, achieving an accuracy of 0.8653, precision of 0.8697, recall of 0.8653, F1-score of 0.8668, Cohen's Kappa of 0.7862, and MCC of 0.7870. Ablation studies further demonstrate the critical role of cues and sentiment features in improving performance. These findings indicate that combining pre-trained embeddings with carefully selected domain features offers a more robust and context-aware solution for text classification, establishing a foundation for future work integrating transformer-based models with explainable AI techniques in domain-specific applications.**

Keywords—Consumer complaints; text classification; sentence embeddings; MiniLM; class imbalance; sentiment analysis; domain adaptation; contextual embeddings

I. Introduction

The increased number of consumer complaints with financial institutions, as well as regulating bodies, is a major challenge to the traditional manual operations and analytical procedures [1]. These complaints are largely unstructured free-text stories that encompass an expansive scope of concerns pertinent to credit reporting, debt collection, mortgage performance, and fraudulent banking practices [2]. On the one hand, in contemporary financial ecosystems, the management of these complaints is not only a customer service issue, but it is a regulatory requirement because complaints that are not resolved or improperly assigned can result in compliance failures, reputational damage, and fines. The right and timely classification of such stories is thus critical to immediate solution, regulatory adherence, and successive trend identification in customer complaints [3]. But manual analysis is time-consuming, irregular, and inaccurate, especially when the textual input is noisy, ambiguous, or domain-dependent [4]. Moreover, the constantly growing number of consumer complaints overwhelms human reviewers and forms an operational bottleneck, thus automation is a pressing need.

This issue has been successfully discussed using traditional text-classification algorithms. The majority of the current methods are based on the bag-of-words paradigm, TF-IDF vectorization, or shallow learning models, including logistic regression and support vector machines [5], [6]. Despite the relative ease of interpretation and computational efficiency, these techniques naturally do not represent long-range correlations, semantic details and contextual associations in text [7]. This deficiency is especially prominent in the consumer complaints; words in these lines tend to refer much more specifically to the financial world, to technical aspects of transactions, legal allusions, and emotional terminology. Consequently, the performance of these models tends to be inadequate, reflecting the semantic richness of complaint narratives, so their performance in classification is also suboptimal.

Recent progress in the field of natural language processing (NLP) has seen the emergence of transformer-based language models, which are trained to learn contextualized text representations by taking advantage of self-attention processes [7], [8]. In comparison to the traditional encodings, transformers are able to capture semantic correlations at varying granularity levels, and they are more appropriate in analyzing unstructured data of financial complaints [9]. BERT and RoBERTa models have reached state of the art performance in a broad array of NLP tasks. They are, however, expensive to compute and difficult to train like their better-performing counterparts in that they consume high memory, large training information, and specialized equipment. In systems with large volumes of complaints to handle or where real-time applications are being used, these requirements are unrealistic, especially when a financial institution has to weigh efficiency and accuracy versus regulatory requirements.

In this study, these limitations will be solved by introducing a light but useful framework utilizing sentence embeddings produced by a frozen pre-trained transformer model without the computational cost of fine-tuning. The encoder generates semantic representations that can be used to gain insights about the complaint narratives, whereas a simple feedforward neural

*Corresponding author.

network classifier guarantees effective prediction. The framework classifies complaints into five predefined financial product categories, thus enabling an automated process of triage and resolution. The study uses synthetic oversampling, stratified cross-validation, and hyperparameter optimization as a means to balance the classes and guarantee the model's robustness. Thus, the framework integrates the strengths of transformer-based embeddings with the efficiency and adaptability of lightweight models for scalable deployment.

Following are the contributions of this study:

- A lightweight framework is introduced that integrates frozen transformer-based embeddings with a feedforward neural network to efficiently process and categorize consumer complaint narratives across predefined financial product categories.
- A systematic comparison of the proposed framework against traditional text-classification approaches, such as TF-IDF with logistic regression and support vector machines, demonstrates performance gains in terms of semantic representation and predictive reliability.
- An ablation study is conducted to examine the individual contributions of transformer embeddings and neural network components, providing insights into the effectiveness and interpretability of the model's architecture.

The structure of the study is organized as follows: Section II: Related Work reviews prior studies on text classification methods in the financial domain. Section III: Methodology explains the proposed model, including the integration of transformer-based embeddings with a neural network architecture. Section IV: Results and Discussion describes the dataset, evaluation metrics, and implementation strategies, followed by a detailed presentation of results, including comparisons with baseline models and the ablation study. Section V: Conclusion and Future Work summarizes the key contributions of the study and outlines directions for further research.

II. Related Work

The classification of customer complaints has been examined across multiple domains using both traditional machine learning and modern deep learning approaches. Early works emphasized feature engineering and linguistic cues, while more recent studies have explored transformer-based models and large language models (LLMs) to capture semantic richness.

A. Traditional Machine Learning and Hybrid Approaches

A number of studies have relied on traditional machine learning techniques combined with feature engineering to classify customer complaints (see Table I). Khedkar and Shinde [10] evaluated hotel reviews using linguistic features such as nouns, verbs, and intensifiers, comparing machine learning classifiers, ensemble methods, and deep learning models. Their ensemble approach showed improved accuracy, but the reliance on handcrafted features limited generalizability. Similarly, Bozyiğit et al. [11] focused on Turkish food product complaints, demonstrating that TF-IDF with Chi-Square feature selection outperformed word2vec embeddings when paired with classifiers such as XGBoost and SVM.

Beyond consumer products, similar methods have been applied in healthcare and telecommunications. Li et al. developed a complaint classification system for hospital records using Logistic Regression, Naive Bayes, and SVM, highlighting the role of SMOTE in handling class imbalance. Elmessiry et al. [12] further explored automated triage of patient complaints, reporting over 80% accuracy using standard classifiers. In the telecom sector, Mahmoud et al. [13] employed ensemble methods on large-scale service logs, showing that AdaBoost with Decision Trees effectively predicted complaint escalation. These works collectively highlight the utility of ML models in structured or moderately sized datasets, but they often depend on domain-specific preprocessing and lack adaptability to new contexts.

B. Deep Learning and Transformer-Based Approaches

With the rise of deep learning, researchers have increasingly adopted neural architectures and pre-trained models for complaint classification. Vinayak and Jyotsna [14] tested CNN, LSTM, Bi-LSTM, and GRU with embeddings such as DistilBERT on CFPB complaints, achieving strong performance with CNN-embedding combinations. Alamsyah et al. [15] extended this to large-scale banking data, where a neural network trained on one million complaints demonstrated the feasibility of text mining for real-world deployment. Idrissi-Yaghir et al. [16] evaluated transformer-based models on German feedback data, showing that domain adaptation improved results compared to off-the-shelf pre-trained models.

Recent advances in LLMs have introduced new opportunities for zero-shot and few-shot complaint classification. Roumeliotis et al. [17] benchmarked leading LLMs and reasoning models, including GPT-4 and Claude, for financial complaint classification on CFPB data. Their findings highlighted the promise of zero-shot approaches for emerging categories but also pointed out challenges in handling overlapping complaint classes. Other work by Basha and Rajput [18] focused on aspect-level sentiment analysis, using probabilistic latent models to capture semantic relations in hotel reviews, demonstrating how advanced NLP can extract actionable insights. These studies confirm the strength of deep learning and LLMs in capturing semantic complexity, though issues such as interpretability, efficiency, and imbalance remain open challenges.

While prior studies have explored various approaches to complaint and feedback classification across domains such as healthcare, hospitality, food, banking, and telecommunications, they often exhibit limitations that constrain their applicability. Earlier works relied heavily on handcrafted linguistic features, traditional machine learning classifiers, or domain-specific adaptations of transformer models, which restricted their generalizability, interpretability, or performance in low-resource settings. In contrast, the proposed framework leverages a pre-trained MiniLM model enhanced with domain-relevant cues and sentiment features, enabling it to capture both semantic richness and contextual nuances across consumer complaints.

TABLE I. COMPARISON OF RELATED STUDIES

Ref.	Year	Focus	Dataset	Technique	Limitation
[12]	2017	Patient Triage	PARS Complaint Dataset	ML Classifiers	Binary physician vs. non-physician
[18]	2019	Aspect Mining	Hotel Reviews (Wang et al.)	LARR, NLP, MLRC	More sentiment-focused than classification
[10]	2020	Praise/Complaint	Hotel Reviews (Kaggle)	Hybrid features + ML, Ensemble, CNN	Relies on handcrafted linguistic features
[11]	2022	Food Complaints	Turkish Food Complaints	TF-IDF, word2vec, XGBoost, SVM	Limited to non-English context
[15]	2022	Bank Complaints	BRI Dataset (1M records)	Neural Network + TF-IDF	Limited interpretability
[16]	2023	German Feedback	GermEval 2017	Transformers (BERT, domain-adapted)	Requires domain-specific adaptation
[14]	2023	Consumer Complaints	CFPB Dataset	CNN, LSTM, DistilBERT	High-resource training required
[13]	2024	Telecom Complaints	Egyptian Telecom Logs	AdaBoost + DT, RFC, SVM	Focused on structured data logs
[19]	2025	Patient Complaints	Hospital Records (1465 + test set)	Logistic Regression, SVM, Naive Bayes	Moderate dataset, domain-specific
[17]	2025	Financial Complaints	CFPB Complaints (Kaggle)	Zero-shot LLMs & Reasoning Models	Overlap between financial categories

III. METHODOLOGY

The methodology of this study is centered on leveraging semantic representations from a frozen pre-trained transformer language model, which serves as the encoder for capturing contextual meaning in consumer complaint narratives. These embeddings are then fed into a lightweight feedforward neural network that performs classification across multiple product categories. To demonstrate the effectiveness of the proposed framework, its performance is systematically compared with traditional baseline models, including TF-IDF [20] and Word2Vec [21] embeddings combined with logistic regression and Naïve Bayes classifiers.

A. Dataset Collection

The dataset comprises 162,421 consumer complaint records, each containing a narrative text and an associated product category label. These product classifications fall into five distinct categories (see Fig. 1). The dataset exhibits a markedly imbalanced class distribution, with some categories represented by significantly more narratives than others [22].

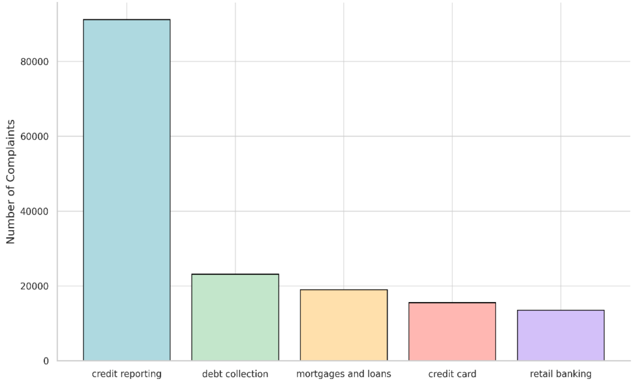

Fig. 1. Target class distribution.

The distribution of complaint narrative length shows a highly skewed pattern, where the majority of complaints contain fewer than 200 words, with a sharp peak below 100 words (see Fig. 2). A long tail extends beyond 500 words, indicating that while most complaints are concise, a small number are significantly longer, exceeding even 2,000 words.

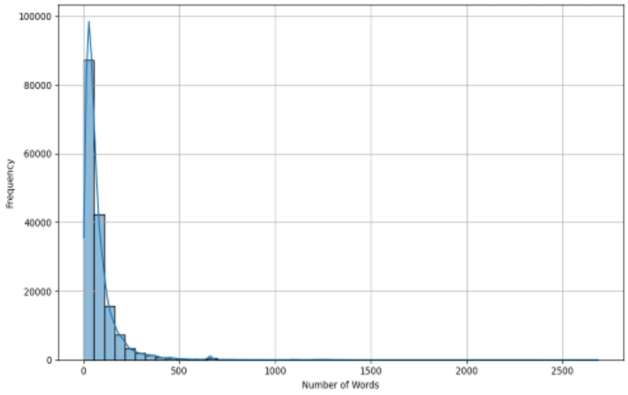

Fig. 2. Distribution of complaints narrative length.

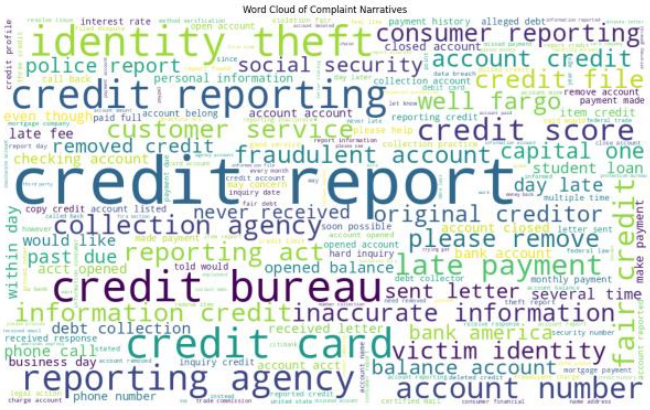

Fig. 3. Word cloud representation of consumer complaint narratives.

The word cloud shown in Fig. 3 provides a visual summary of the most frequently occurring terms in the consumer complaint narratives. Dominant terms such as credit report, identity theft, late payment, collection agency, and fraudulent highlight the common themes and concerns raised by consumers. These include issues related to credit reporting inaccuracies, debt collection practices, and unauthorized transactions. The prevalence of entity names like Capital One, Wells Fargo, and Bank of America also suggests repeated mentions of specific financial institutions in the complaints.

B. *Data Cleaning and Preprocessing*

There were numerous inconsistencies in the consumer complaint narratives, such as special characters, email addresses, and inconsistent word casing. A structured preprocessing pipeline was created in order to prepare the data to be used with machine learning models. This pipeline consisted of text normalization, feature enrichment, data balancing strategies, and partitioning.

1) Text cleaning and normalization: Initial standardization of the narratives was achieved by converting all the text to lower case and removing any non-desired patterns like email addresses and non characters using regular expressions. Removal of fronting and preceding spaces increased uniformity further. This normalization step guaranteed the sanity of the text of a complaint, which is vital to both consistent embedding generation and feature extraction.

2) Label encoding: The product-classes that were the target variables were classified as complaint datasets. These product labels underwent label encoding as they were converted into an integer representation so as to facilitate learning. This supported categorical output processing of the models without a loss of mapping between original class names and coded values.

3) Structured cue features: In addition to the purged textual representations, domain-specific characteristics were developed to pinpoint the important complaint cues. Keywords used to develop binary indicators included the words: refund, dispute, chargeback and cancellation. These structured hints had explicit patterns of complaints, complementing with the full meaning of the textual embeddings and easy interpretancy features.

4) Sentiment extraction: Sentiment polarity scores computed with the TextBlob library were used to add emotion to the tone of each story. These scores measured the consumer sentiment on a negative to positive scale. The combination of sentiment, and structured/unstructured features added to the dataset due to the reflection of the content but also the emotional context of the complaints.

5) Data balancing with SMOTE: Class imbalance in the dataset was also present with some of the product categories having more representations than others. The training set was mitigated with the Synthetic Minority Oversampling Technique (SMOTE) [23]. SMOTE creates artificial examples of underrepresented classes thereby balancing the classes and allowing the model to generalize better across the entire categories.

6) Dataset splitting and cross-validation: Upon balancing, the data were divided into training and testing sets using a stratified method to maintain the distributions of classes. Eighty per cent of the data was to be used in training and twenty percent in testing. In order to make the model evaluation robust, 5-fold cross-validation was used in the course of training. This method gave a more consistent measure of performance by averaging over the uses of multiple folds and made it less likely to overfit to one partition.

C. *Proposed Model*

The proposed model combines a frozen pre-trained transformer language model to generate sentence embeddings that capture the semantic and contextual details of consumer complaint narratives (Fig. 4). These embeddings are then passed into a feedforward neural network, which delivers accurate and efficient classification across different product categories. The approach relies on the representational strength of transformer-based embeddings while leveraging the simplicity of a lightweight neural network, striking a balance between predictive accuracy and computational cost, making it well-suited for large-scale complaint management systems.

1) Embedding representation using pre-trained language models: To convert raw complaint narrative into fixed-size numerical vectors to be used in classification, we utilized sentence-level embeddings of a pre-trained transformer encoder. To make computations efficient and to maintain the semantic representations that had been learnt during pre-training, we selected a lightweight frozen encoder to avoid fine-tuning.

Let $x_i \in R^T$ denote the i-th tokenized complaint narrative of length T. The sentence encoder maps this input to a dense vector $v_i \in R^d$, where d is the embedding dimension:

$$v_i = f_{encoder}(x_i) \tag{1}$$

The resulting vectors $\{v_1, v_2, , , , , v_n\}$ were generated in inference-only mode and subsequently served as the primary textual features for classification.

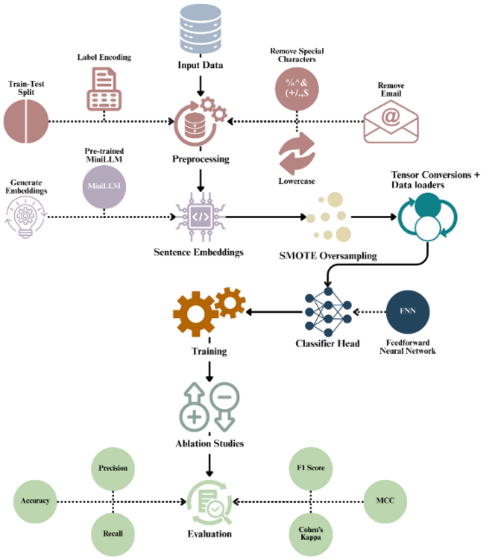

Fig. 4. Architecture of the proposed model integrating frozen transformer embeddings with a feedforward neural network.

2) Structured feature augmentation: In addition to embeddings, domain-specific cues were incorporated to enrich the representation of each narrative. Four binary features were extracted to indicate the presence of keywords related to refund, dispute, chargeback, and cancellation. These were obtained by rule-based pattern matching within each narrative.

Furthermore, a sentiment polarity score was computed for every complaint narrative. This score, ranging from −1-1−1 (negative) to +1+1+1 (positive), was appended as a continuous feature. The final feature vector for each instance was represented as:

$$z_i = [v_i \parallel s_i \parallel k_i] \tag{2}$$

where, v_i denotes the embedding representation, s_i is the sentiment score, k_i is the structured keyword feature vector, and ∥ indicates concatenation.

3) Neural classifier head architecture: The classification model was implemented as a multi-layer feedforward neural network. The input layer dimension matched the size of the augmented feature vector ziz_izi. This was followed by two fully connected hidden layers with ReLU activation functions. Dropout regularization was applied at each hidden layer to mitigate overfitting.

Let the classifier be represented by a function $g_\theta(\cdot)$, parameterized by θ, which maps the embedding v_i to a probability vector $p_i \in R^C$, where C is the number of complaint categories:

$$p_i = softmax(g_\theta(v_i)) \tag{3}$$

The model was trained using the cross-entropy loss function, defined as:

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{C=1}^{C} y_{ic}\log(p_{ic}) \tag{4}$$

where, y_{ic} is the one-hot encoded true label for instance i and class c, and p_{ic} is the predicted probability. Optimization was performed using the Adam optimizer with a fixed learning rate of 10^{-3}, and learning rate schedulers were optionally considered for stabilization.

TABLE II. HYPERPARAMETER CONFIGURATION OF THE PROPOSED MODEL

Hyperparameter	Value
Hidden layer sizes	[256, 128]
Activation function	ReLU
Dropout rate	0.3
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	32
Number of epochs	20
Loss function	Categorical cross-entropy

4) Hyperparameter optimization: Hyperparameters for the classifier were optimized using randomized search. This approach provided a computationally efficient alternative to exhaustive grid search by exploring a broad search space with fewer evaluations. A representative configuration of the proposed model is summarized in Table II. Also, the algorithm of proposed model is shown in Algorithm 1.

Formally, the objective of hyperparameter tuning was to identify parameters θ that minimized the validation loss L_{val}averaged across all folds:

$$\theta^* = \arg\min_{\theta} \frac{1}{K}\sum_{k=1}^{K} L_{val}^{(k)}(\theta) \tag{5}$$

where, K = 5 is the number of folds in cross-validation.

Algorithm 1: Proposed Frozen MiniLM-Based Model

```
Input:
Complaint text dataset $D = \{x_1, x_2, \ldots .x_n\}$with corresponding labels Y
Pre-trained Frozen MiniLM encoder E(·)
Feedforward Neural Network Classifier C(·)
Output:
Predicted labels $\hat{Y}$
    1. Initialization
        o Load frozen MiniLM encoder E
        o Initialize classifier parameters θ
    2. Encoding
        o For each complaint text $X \in$ D:
              $h_i = E(x_i)$
where $h_i \in R^d$ is the sentence embedding
    3. Classification
        o Pass embeddings through the classifier:
              $\hat{Y_i} = C(h_i, \theta)$
    4. Optimization
        o Minimize loss function $L(\hat{Y_i}, Y)$ using training set
    5. Evaluation
        o Compare predictions $\hat{Y}$ with ground truth Y on test set
        o Compute evaluation metrics: Accuracy, Precision, Recall, F1, Kappa, MCC
    6. Output
        o Final predictions $\hat{Y}$ for unseen complaint narratives
```

D. Baseline Framework

A baseline framework was developed using conventional feature extraction procedure and classification to give a reliable comparison to the proposed model (see Fig. 5). These baselines give a benchmark of reference to assess whether improved architecture causes significant differences in performance: TF-IDF and Word2VE were used in combination with Logistic Regression (LR) [24], Multinomial Naive Bayes (NB) [25], and Gaussian Naive Bayes (GNB) as classical classifiers [26].

1) TF-IDF with Logistic Regression: Term Frequency-Inverse Document Frequency (TF-IDF) representation was used to construct the first baseline. Both text documents were mapped to a sparse 5000-sized vector that included the most informative terms. These features were used to train a Logistic Regression, the decision boundary of which is modeled as:

$$P\,(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}} \tag{6}$$

where, x is the TF-IDF feature vector, w represents learned weights, and b is the bias term. The optimization was performed using maximum likelihood estimation with a maximum of 200 iterations.

2) Word2Vec embedding with Logistic Regression: Word2Vec was trained on the tokenized training corpus using the skip-gram architecture to measure the effectiveness of dense word embeddings. The model trained 200-dimensional embeddings using a context window of 5 and eliminating words that occur less frequently than 2 times. Document vectors d were then calculated as the average of their token embeddings d:

$$d = \frac{1}{N}\sum_{i=1}^{N} v_{w_i} \tag{7}$$

where, N is the number of tokens and vv_{w_i} is the embedding of token w_i. Logistic Regression was then applied to the document vectors with 500 maximum iterations to ensure convergence.

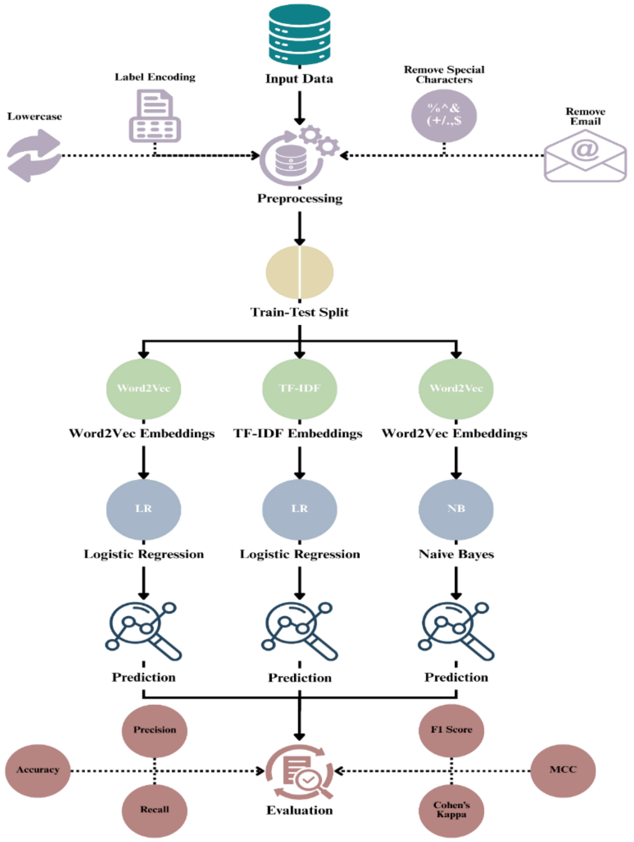

Fig. 5. Baseline framework for comparison.

3) TF-IDF with Multinomial Naïve Bayes: A second baseline on sparse features utilized Multinomial Naïve Bayes, a classical probabilistic classifier suitable for count-based features. The posterior probability for class ccc is defined as:

$$P(c \mid x) = \frac{P(c)\prod_{i=1}^{M} P(x_i|c)}{P(x)}) \tag{8}$$

where, M represents the number of features and x_i the term frequency values.

4) Word2Vec with Gaussian Naïve Bayes: Gaussian Naïve Bayes was applied on the dense Word2Vec document vectors. Unlike the multinomial variant, GNB assumes a Gaussian likelihood for each feature dimension. The class-conditional probability is defined as:

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}\right) \tag{9}$$

where, μ_c denote the mean and variance of feature x_i for class c.

5) Hyperparameter tuning: All baseline models were optimized through Grid Search with 5-fold Cross-Validation to ensure fair comparison. The selected hyperparameters are summarized in Table III.

TABLE III. HYPERPARAMETER SETTINGS FOR BASELINE MODELS (OPTIMIZED USING GRID SEARCH CV)

Model	Hyperparameters
TF-IDF + Logistic Regression	max_features = 5000, max_iter = 200
Word2Vec + Logistic Regression	vector_size = 200, window = 5, min_count = 2, sg = 1, max_iter = 500
TF-IDF + Multinomial NB	alpha = optimized via grid search
Word2Vec + Gaussian NB	assumes Gaussian distribution, no tunable hyperparameters beyond prior smoothing

E. Evaluation Strategy

Evaluation was conducted both during cross-validation and on a separate holdout set. Metrics included accuracy [27], precision [28], recall [29], and F1-score [17] in both macro and weighted forms to account for class imbalance. To assess agreement beyond chance, Cohen's Kappa coefficient [30] was computed. The Matthews Correlation Coefficient (MCC) was also used due to its ability to summarize the confusion matrix into a single metric, even in the presence of imbalanced labels.

Formally, the metrics are defined as follows:

- Accuracy: the proportion of correctly classified instances.

$$\text{Accuracy (Acc)}: \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{10}$$

- Precision: the fraction of correctly predicted positive cases among all predicted positives.

$$\text{Precision (Prec)}: \frac{(TP)}{(TP+FP)} \tag{11}$$

- Recall: the fraction of correctly predicted positive cases among all actual positives.

$$\text{Recall}: \frac{(TP)}{(TP+FN)} \tag{12}$$

- F1-Score: harmonic mean of precision and recall, balancing both metrics.

$$\text{F1 Score: } 2 \times \frac{Precision \times Recall}{(Precision + recall)} \tag{13}$$

- Cohen's Kappa (κ): measures inter-rater agreement while adjusting for chance.

$$k = \frac{p_o - p_e}{1 - p_e} \tag{14}$$

where, p_o is the observed agreement and p_e is the expected agreement by chance.

- Matthews Correlation Coefficient (MCC): a balanced measure of classification quality derived from all four entries of the confusion matrix.

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{15}$$

In order to establish how resistant the model is, assessments on perturbed versions of the test narratives were also done. These distortions covered character level noises like misspellings and deletion of random words. In addition, the domain drift was designed by dividing the dataset into older and newer complaint narratives across a period of time to assess the generalization under the change of language tendencies.

IV. Results and Discussion

The results section offers the evaluation of the proposed model as compared with the baseline frameworks, drawing attention to its feasibility in processing the candid complaint narratives that are not structured. It has quantitative and comparative analysis to determine reliability and robustness of the approach. An ablation study is also described to investigate the impact of each of these components, including cues and sentiment features, and sheds more light on the design and production of the model.

A. Proposed Model

The results of the evaluation indicate that the proposed model performed well in several metrics with accuracy (0.8653) and precision (0.8697), recall (0.8653), and F1-score (0.8668) showing a steady balance between accurate classifications and minimization of errors (see Fig. 6). The high degree of agreement between precision and recall indicates that the model is unlikely to be biased towards a particular class, even when data is skewed. The relatively large F1-score also proves that the model is quite successful in keeping this trade-off, indicating its strengths in real-life situations, where picking a false positive or a false negative may be expensive.

In addition to these standard indicators, the indicators that are based on the agreement are also informative. The value of Kappa (0.7917) and Matthews Correlation Coefficient (0.7924) both indicate that the classifier exhibits a considerable improvement over the chance knowledge, and the MCC again substantiates the reliability of the chosen algorithm despite the imbalance of the classes. These values support the stability and the generalization of the model.

The ROC curve of the proposed model shows consistently high discriminative capability across all classes, with Area Under the Curve (AUC) values ranging from 0.929 to 0.989 (see Fig. 7). Notably, Class 4 achieves the highest AUC (0.989), reflecting near-perfect separability, while Class 2, with an AUC of 0.929, represents the relatively more challenging case but still indicates strong performance. The micro-average AUC of 0.972 demonstrates that the overall classification performance is highly reliable when aggregating across all classes. The steep rise of the curves towards the top-left corner further confirms that the model maintains high true positive rates while keeping false positives low, highlighting its robustness and suitability for handling multi-class classification in imbalanced datasets.

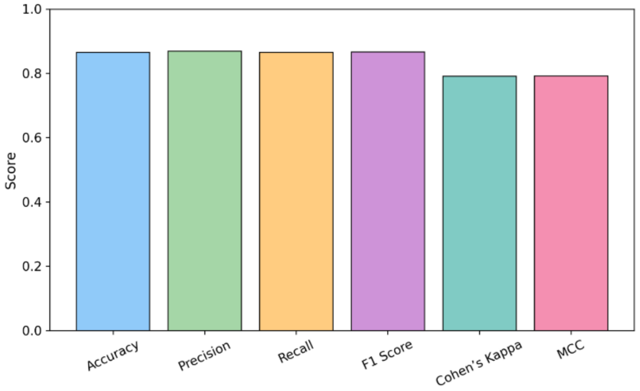

Fig. 6. Evaluation of the proposed model.

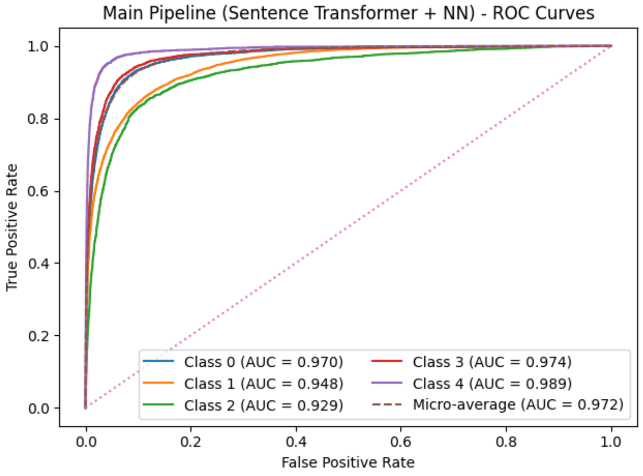

Fig. 7. ROC Curves illustrating the performance of the proposed model.

B. Ablation Study

An ablation study was conducted to measure the contribution of various components to the proposed pipeline, and in doing so, augmentation strategies, cue features, and sentiment features were gradually eliminated (see Table IV). The findings indicate the relative significance of every aspect in determining the overall performance.

The main pipeline achieved balanced performance across all metrics, with an accuracy of 0.8653, F1-score of 0.8668, and strong agreement measures (Kappa = 0.7862, MCC = 0.7870). When data augmentation was removed, accuracy slightly increased to 0.8690, while F1-score stabilized at 0.8677. This suggests that augmentation did not substantially improve core classification ability, although it may have contributed to better generalization, as evidenced by more stable performance across perturbed test sets.

TABLE IV. RESULTS OF THE ABLATION STUDY SHOWING THE IMPACT OF REMOVING DIFFERENT COMPONENTS

Setting	Accuracy	Precision	Recall	F1	Kappa	MCC
Main Pipeline	0.8653	0.8697	0.8653	0.8668	0.7862	0.7870
No Augmentation	0.8690	0.8673	0.8690	0.8677	0.7914	0.7917
No Cues	0.8541	0.8649	0.8541	0.8571	0.7782	0.7808
No Sentiment	0.8577	0.8663	0.8577	0.8602	0.7826	0.7846
No Cues + Sentiment	0.8469	0.8605	0.8469	0.8505	0.7687	0.7720

TABLE V. PERFORMANCE COMPARISON OF BASELINE FRAMEWORK AND THE PROPOSED MODEL

Model	Accuracy	Precision	Recall	F1	Kappa	MCC
TF-IDF + LR	0.8548	0.8531	0.8548	0.8539	0.7702	0.7706
Word2Vec + LR	0.8510	0.8486	0.8510	0.8488	0.7619	0.7626
TF-IDF + NB	0.8377	0.8370	0.8377	0.8344	0.7403	0.7417
Word2Vec + NB	0.6942	0.7512	0.6942	0.7073	0.5606	0.5728
Proposed Model	0.8653	0.8697	0.8653	0.8668	0.7862	0.7870

In contrast, removing cue features resulted in a noticeable decline, with accuracy dropping to 0.8541 and F1-score to 0.8571. Similarly, removing sentiment features led to slightly higher values than the no-cue setting but still below the full pipeline (accuracy = 0.8577, F1 = 0.8602). This indicates that both cues and sentiment individually enhance the model's discriminatory power, though their effects are modest. Importantly, removing both cues and sentiment together caused the largest performance decline (accuracy = 0.8469, F1 = 0.8505, Kappa = 0.7687), underscoring their complementary value.

The ablation study demonstrates that while augmentation plays a secondary role, linguistic cues and sentiment information contribute significantly and synergistically to model robustness, particularly in distinguishing subtle variations in consumer complaint narratives. This validates the design choice of incorporating multiple feature perspectives in addition to the sentence transformer representations.

C. Comparison with Baseline Framework

To evaluate the effectiveness of the proposed framework, its performance was compared against multiple baseline models including TF-IDF with Logistic Regression (LR), Word2Vec with LR, TF-IDF with Naïve Bayes (NB), and Word2Vec with NB. As shown in Table V, the proposed model consistently outperforms the baseline frameworks across all metrics. Specifically, the proposed model achieved the highest accuracy (0.8653), precision (0.8697), recall (0.8653), and F1-score (0.8668), demonstrating its superior ability to capture both the semantic and structural features of the data. Furthermore, the agreement-based measures, Cohen's Kappa (0.7862) and MCC (0.7870), indicate stronger reliability and balanced performance compared to the baseline methods.

Among the baseline models, TF-IDF combined with LR performed comparatively well, reaching an accuracy of 0.8548 and F1-score of 0.8539, which is close to the proposed model, but still inferior across all measures. Similarly, Word2Vec with LR achieved slightly lower values (accuracy 0.8510, F1-score 0.8488), highlighting the limitations of traditional embedding methods in capturing contextual information. TF-IDF with NB provided moderate performance (accuracy 0.8377), while Word2Vec with NB yielded the weakest results, with accuracy dropping to 0.6942, reflecting its inability to generalize effectively for the given task. The comparison emphasizes that the integration of sentence-level representations with neural architectures, as in the proposed model, yields a more robust and reliable framework than traditional text representations and shallow learning methods.

V. CONCLUSION AND FUTURE DIRECTIONS

This study demonstrates that integrating contextual embeddings with cues and sentiment features provides a robust framework for automated text classification. Beyond achieving high accuracy and reliable agreement measures, the results highlight the model's ability to effectively capture semantic and contextual nuances that traditional shallow approaches often miss. The ablation study underscores the importance of combining multiple sources of information, showing that cues and sentiment features meaningfully enhance classification performance. Compared to conventional methods such as TF-IDF or Word2Vec with logistic regression, the proposed framework offers a more nuanced understanding of textual data, suggesting its potential for broader applications in domains where reliable interpretation of unstructured text is critical. Overall, the findings illustrate that leveraging advanced neural architectures alongside carefully selected features can yield both accurate and interpretable models, providing a practical pathway for improving automated text analysis in real-world settings.

While the results are promising, there remain several avenues for future research. First, the integration of transformer-based architectures such as BERT or RoBERTa could be explored to enhance contextual embeddings and improve classification robustness. Moreover, incorporating domain adaptation strategies would allow the framework to be applied across different datasets and problem settings with minimal retraining. Finally, future studies could investigate explainability methods to provide better interpretability of the classification decisions, thereby enhancing trust and usability in real-world applications.

REFERENCES

[1] S. Nuthakki, S. Kumar, C. S. Kulkarni, and Y. Nuthakki, “Role of AI Enabled Smart Meters to Enhance Customer Satisfaction,” Int. J. Comput. Sci. Mob. Comput., vol. 11, no. 12, pp. 99–107, 2022, doi: 10.47760/ijcsmc.2022.v11i12.010.

[2] M. Chowdhury, K. M. T. Rahman, and H. A. Maruf, “Whistle Blower: An Insurance Awareness Mobile Application with Insurance Policy Selection, Fraud Detection, Critical Help, Complaint Features,” 8th IEEE Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2024, pp. 1–6, 2024, doi: 10.1109/CSITSS64042.2024.10817002.

[3] E. Fedorova, S. Druchok, and P. Drogovoz, “Impact of news sentiment and topics on IPO underpricing: US evidence,” Int. J. Account. Inf. Manag., vol. 30, no. 1, pp. 73–94, 2022, doi: 10.1108/IJAIM-06-2021-0117.

[4] W. Zheng et al., “MIDLG: Mutual Information based Dual Level GNN for Transaction Fraud Complaint Verification,” Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 5685–5694, 2023, doi: 10.1145/3580305.3599865.

[5] A. Palanivinayagam, C. Z. El-Bayeh, and R. Damaševičius, “Twenty Years of Machine-Learning-Based Text Classification: A Systematic Review,” Algorithms, vol. 16, no. 5, pp. 1–28, 2023, doi: 10.3390/a16050236.

[6] V. Karpukhin et al., “Dense passage retrieval for open-domain question answering,” EMNLP 2020 - 2020 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf., pp. 6769–6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.

[7] S. Nuthakki, “CONVERSATIONAL AI AND LLM ’ S CURRENT AND FUTURE IMPACTS IN IMPROVING AND SCALING HEALTH SERVICES,” Int. J. Comput. Eng. Technol., vol. 14, no. December, pp. 149–155, 2023, doi: 10.17605/OSF.IO/7GHJY.

[8] C. Wang, Y. Yang, C. Gao, Y. Peng, H. Zhang, and M. R. Lyu, “No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence,” ESEC/FSE 2022 - Proc. 30th ACM Jt. Meet. Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng., pp. 382–394, 2022, doi: 10.1145/3540250.3549113.

[9] Q. Ren et al., “A survey on fairness of large language models in e-commerce: progress, application, and challenge,” May 2024.

[10] S. Khedkar and S. Shinde, “Deep Learning and Ensemble Approach for Praise or Complaint Classification,” Procedia Comput. Sci., vol. 167, no. 2019, pp. 449–458, 2020, doi: 10.1016/j.procs.2020.03.254.

[11] F. BOZYİĞİT and D. K. Onur DOĞAN, “Categorization of Customer Complaints in Food Industry Using Machine Learning Approaches,” J. Intell. Syst. Theory Appl., vol. 5, no. June 2021, pp. 85–91, 2022, doi: 10.38016/jista.954098.

[12] A. Elmessiry, W. O. Cooper, T. F. Catron, J. Karrass, Z. Zhang, and M. P. Singh, “Triaging Patient Complaints : Monte Carlo Cross-Validation of Six Machine Learning Classifiers Corresponding Author :,” vol. 5, pp. 1–10, 2017, doi: 10.2196/medinform.7140.

[13] M. M. Gewaly, Machine Learning Approach for Complaint Prediction in the Telecom Industry. IEEE, 2024. doi: 10.1109/IMSA61967.2024.10652677.

[14] V. Vinayak and C. Jyotsna, “Consumer Complaints Classification using Deep Learning & Word Embedding Models,” 2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023, pp. 1–5, 2023, doi: 10.1109/ICCCNT56998.2023.10307286.

[15] D. P. Alamsyah, T. Arifin, Y. Ramdhani, F. A. Hidayat, and L. Susanti, “Classification of Customer Complaints: TF-IDF Approaches,” 2022 2nd Int. Conf. Intell. Technol. CONIT 2022, pp. 1–5, 2022, doi: 10.1109/CONIT55038.2022.9848056.

[16] A. Idrissi-Yaghir, H. Schäfer, N. Bauer, and C. M. Friedrich, “Domain Adaptation of Transformer-Based Models Using Unlabeled Data for Relevance and Polarity Classification of German Customer Feedback,” SN Comput. Sci., vol. 4, no. 2, pp. 1–13, 2023, doi: 10.1007/s42979-022-01563-6.

[17] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, “Think Before You Classify: The Rise of Reasoning Large Language Models for Consumer Complaint Detection and Classification,” Electron., vol. 14, no. 6, 2025, doi: 10.3390/electronics14061070.

[18] S. M. Basha and D. S. Rajput, “An innovative topic-based customer complaints sentiment classification system,” Int. J. Bus. Innov. Res., vol. 20, no. 3, pp. 375–391, 2019, doi: 10.1504/IJBIR.2019.102718.

[19] X. Li, Q. Shu, C. Kong, J. Wang, and G. Li, “An Intelligent System for Classifying Patient Complaints Using Machine Learning and Natural Language Processing : Development and Validation Study Corresponding Author :,” vol. 27, 2025, doi: 10.2196/55721.

[20] J. A. Nasir, A. Karim, G. Tsatsaronis, and I. Varlamis, “A knowledge-based semantic kernel for text classification,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7024 LNCS, no. 3, pp. 261–266, 2011, doi: 10.1007/978-3-642-24583-1_25.

[21] N. Alami, M. Meknassi, and N. En-nahnahi, “Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning,” Expert Syst. Appl., vol. 123, pp. 195–211, 2019, doi: 10.1016/j.eswa.2019.01.037.

[22] V. Nayyar, “Kaggle,” Kaggle. Accessed: Apr. 16, 2025. [Online]. Available: https://www.kaggle.com/code/vikram92/multiclass-complaints-classification-using-bi-lstm/notebook

[23] E. F. Swana, W. Doorsamy, and P. Bokoro, “Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset,” Sensors, vol. 22, no. 9, 2022, doi: 10.3390/s22093246.

[24] A. Shanbhag, S. Jadhav, A. Thakurdesai, R. Sinare, and R. Joshi, “BERT or FastText? A Comparative Analysis of Contextual as well as Non-Contextual Embeddings,” 2024, [Online]. Available: http://arxiv.org/abs/2411.17661

[25] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, “Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification,” 2019 Int. Conf. Autom. Comput. Technol. Manag. ICACTM 2019, pp. 593–596, 2019, doi: 10.1109/ICACTM.2019.8776800.

[26] W. Dai, G.-R. Xue, Q. Yang, and Y. Yu, “Transferring Naive Bayes Classifiers for Text Classification,” in AAAI, 2007, pp. 540–545.

[27] A. Abdi, S. Hasan, S. M. Shamsuddin, N. Idris, and J. Piran, “A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion,” Knowledge-Based Syst., vol. 213, p. 106658, 2021, doi: 10.1016/j.knosys.2020.106658.

[28] M. E. Mowlaei, M. Saniee Abadeh, and H. Keshavarz, “Aspect-based sentiment analysis using adaptive aspect-based lexicons,” Expert Syst. Appl., vol. 148, p. 113234, 2020, doi: 10.1016/j.eswa.2020.113234.

[29] M. Z. Asghar et al., “Performance evaluation of supervised machine learning techniques for efficient detection of emotions from online content,” Comput. Mater. Contin., vol. 63, no. 3, pp. 1093–1118, 2020, doi: 10.32604/CMC.2020.07709.

[30] M. J. Warrens, “Five ways to look at Cohen’s kappa,” J. Psychol. Psychother., vol. 05, no. 04, 2015, doi: 10.4172/2161-0487.1000197.