Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 30, Sayı 2, 466-480, 2026

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 30, Issue 2, 466-480, 2026

DOI: 10.19113/sdufenbed.1963184

Imb-LLMClass: A Large Language Model-Based Framework for Imbalanced Text Classification

Fatma Gülşah TAN[*1]

[1]Burdur Mehmet Akif Ersoy University, Bucak Zeliha Tolunay School of Applied Technology and Business Administration, Information Systems and Technologies,15000, Burdur, Türkiye

(Alınış / Received: 03.06.2026, Kabul / Accepted: 12.08.2026, Online Yayınlanma / Published Online: 25.08.2026)

Keywords
Large Language Models, Text Classification, Imbalanced Data, Synthetic Data Generation, Class-Weighted Learning, LLM-Based Classification

Abstract: Class imbalance constitutes a significant challenge in text classification tasks, particularly due to the insufficient representation of minority-class instances. Traditional machine learning methods trained on imbalanced datasets tend to exhibit a bias toward majority classes, thereby limiting their ability to accurately classify minority-class samples. To address this issue, this study proposes a classification framework, termed Imb-LLMClass, to investigate the effectiveness of large language model-based classification approaches for imbalanced text data. Experimental results demonstrate that the proposed Imb-LLMClass framework achieved the highest performance among the compared methods when evaluated based on the average of three experimental runs, attaining a Macro F1-Score of 0.878 and a Balanced Accuracy of 0.870 across three datasets. Furthermore, the increase in the average minority-class recall to 0.834 indicates that the proposed approach not only improves overall classification performance but also enhances the learning of underrepresented classes. These findings indicate that, in imbalanced text classification tasks, evaluating large language models solely based on overall performance metrics may be insufficient and that their ability to effectively represent minority classes should also be taken into consideration.

Imb-LLMClass: Dengesiz Metin Sınıflandırmasına Yönelik Büyük Dil Modeli Tabanlı Bir Sınıflandırma Çerçevesi

Anahtar Kelimeler
Büyük Dil Modelleri, Metin Sınıflandırması, Dengesiz Veri, Sentetik Veri Üretimi, Sınıf Ağırlıklı Öğrenme, Büyük Dil Modeli Tabanlı Sınıflandırma

Öz: Sınıf dengesizliği, özellikle azınlık sınıfına ait örneklerin yetersiz temsil edilmesi nedeniyle metin sınıflandırma görevlerinde önemli bir zorluk oluşturmaktadır. Dengesiz veri kümeleri üzerinde eğitilen geleneksel makine öğrenmesi yöntemleri, çoğunluk sınıflarına yönelik bir yanlılık gösterme eğiliminde olup, bu durum azınlık sınıfı örneklerini doğru şekilde sınıflandırma yeteneklerini sınırlandırmaktadır. Bu sorunu ele almak amacıyla, bu çalışmada dengesiz metin verileri üzerinde büyük dil modeli tabanlı sınıflandırma yaklaşımlarının etkinliğini araştırmak için Imb-LLMClass adı verilen bir sınıflandırma çerçevesi önerilmektedir. Deneysel sonuçlar, önerilen Imb-LLMClass çerçevesinin üç deneysel tekrarın ortalaması üzerinden değerlendirildiğinde karşılaştırılan yöntemler arasında en yüksek performansı elde ettiğini ve üç veri kümesi genelinde 0.878 Makro F1-Skoru ile 0.870 Dengeli Doğruluk değerine ulaştığını ortaya koymuştur. Bunun yanında, azınlık sınıfı duyarlılığının ortalama 0.834 seviyesine yükselmesi, önerilen yaklaşımın yalnızca genel sınıflandırma performansını artırmakla kalmayıp aynı zamanda yetersiz temsil edilen sınıfların öğrenilmesini de geliştirdiğini göstermektedir. Bu bulgular, dengesiz metin sınıflandırma problemlerinde büyük dil modellerinin yalnızca genel performans ölçütleri üzerinden değil, azınlık sınıfları temsil etme kapasitesi açısından da değerlendirilmesinin yararlı olabileceğini göstermektedir.

1. Introduction

Text classification is widely employed in numerous natural language processing (NLP) applications, including sentiment analysis, spam detection, hate speech identification and fake news detection. One of the primary challenges encountered in these tasks is class imbalance. Class imbalance refers to a situation

*Corresponding author: ftan@mehmetakif.edu.tr

in which certain classes are represented by substantially fewer instances than others within a dataset, leading classification models to become biased toward majority classes [1]. In particular, the limited availability of samples belonging to critical categories, such as hate speech, cyberbullying and fraudulent messages, makes it difficult for models to effectively learn minority-class patterns. Consequently, classification performance in domains such as healthcare, security and social media analytics may be significantly affected.

To mitigate the class imbalance problem, various approaches based on oversampling, undersampling, cost-sensitive learning and ensemble learning have been proposed in the literature [2]. Among these, Synthetic Minority Over-sampling Technique (SMOTE) and related synthetic data generation methods aim to improve learning performance by increasing the number of minority-class instances [3]. However, most existing studies have primarily focused on traditional text representation techniques, such as Term Frequency–Inverse Document Frequency (TF-IDF), Bag-of-Words (BoW) and Word2Vec. In contrast, research investigating the impact of class imbalance on large language model (LLM)-based classification approaches remains relatively limited.

In recent years, transformer-based LLMs have achieved substantial performance improvements across a wide range of NLP tasks. Through few-shot learning capabilities, these models can perform effective classification using only a limited number of labeled examples. Nevertheless, current studies have not yet sufficiently demonstrated the extent to which LLMs can effectively learn and represent minority classes in imbalanced datasets. Furthermore, the role of LLM-driven synthetic text generation in the data balancing process has received comparatively limited attention.

This study proposes Imb-LLMClass, an integrated framework combining balanced prompting, synthetic data generation and class-weighted learning for imbalanced text classification. Its effectiveness is evaluated on three benchmark datasets using Macro F1, Balanced Accuracy and minority-class recall. The remainder of this paper is organized as follows. First, the related literature is reviewed. Next, the proposed Imb-LLMClass framework is described, followed by the experimental setup and evaluation metrics. Finally, the experimental results are presented and discussed.

2. Related Works

The benchmark tasks considered in imbalanced text classification include question classification, emotion recognition and offensive-language identification. In their hierarchical question classification framework developed for open-domain question answering systems, Li and Roth [4] demonstrated that semantic features significantly improve classification performance. Similarly, Saravia et al. [5] showed that contextual emotion representations employed in the CARER model can effectively capture implicit emotional information in short social media texts. In the context of Turkish NLP, Çöltekin [6] introduced a social-media corpus for Turkish offensive-language identification, providing an important benchmark for studying offensive content in a relatively under-resourced language.

One of the pioneering approaches for addressing the class imbalance problem is the SMOTE proposed by Chawla et al. [3]. This method aims to reduce model bias toward majority classes by generating synthetic instances for minority classes. Susan and Kumar [1] conducted a comprehensive review of oversampling, undersampling and hybrid sampling strategies and reported that intelligent sampling techniques can substantially enhance class discrimination capability. Likewise, Indrawati et al. [7] demonstrated that SMOTE-based approaches improve overall performance in imbalanced text classification tasks, although they may exhibit limitations when applied to high-dimensional textual data. Mahmoudi et al. [8] proposed a GPT-based synthetic data generation approach to address class imbalance in text classification. Their results demonstrated that LLM-generated minority-class samples significantly improved multi-class classification performance and outperformed conventional balancing techniques.

With the widespread adoption of transformer-based LLMs, LLM-based approaches have increasingly been explored for imbalanced text classification problems. The Bidirectional Encoder Representations from Transformers (BERT) model introduced by Devlin et al. [9] achieved substantial performance improvements across a variety of NLP tasks through its bidirectional contextual representation capability. Mahmoudi and Salem [10] demonstrated that the BalBERT framework significantly enhances minority-class performance by performing data balancing within a contextual representation space. Similarly, Taskiran et al. [11] reported that advanced SMOTE variants integrated with transformer-based text representations can improve both F1-Score and Balanced Accuracy. Furthermore, Al-harbi and Alghieth [12] emphasized that, despite the strong performance of BERT-based models in low-resource languages, class imbalance remains a major challenge that limits classification effectiveness.

In recent years, the synthetic data generation capabilities of LLMs have also attracted considerable attention in the context of imbalanced learning. Comprehensive surveys have shown that LLMs have fundamentally transformed text data augmentation by supporting not only synthetic data generation but also data labeling, data reformation and teacher-assisted learning paradigms. Furthermore, recent studies have

emphasized that the quality, diversity and evaluation of LLM-generated data are critical factors affecting downstream task performance [13, 14]. Belbekri et al. [15] semantic-based oversampling approaches have been shown to improve minority-class learning in tasks such as named entity recognition. Recent theoretical studies have further demonstrated that LLM-based synthetic oversampling provides a principled alternative to conventional feature-space oversampling by generating semantically meaningful samples that improve minority-class learning while mitigating class imbalance [16]. Cegin et al. [17] systematically compared LLM-based and conventional text augmentation methods, showing that the benefits of LLM-generated data are most pronounced in low-resource settings, while traditional augmentation often provides comparable performance.

Saad et al. [18] employed GPT-4o, Gemini and T5 to generate synthetic data for imbalanced text classification. Their findings demonstrated that LLM-based data augmentation combined with transformer fine-tuning improved Macro F1 performance, particularly for minority classes. Aubaidan et al. [2] further highlighted that class imbalance has critical implications for the reliability and accuracy of artificial intelligence systems in healthcare applications. Arık et al. [19] proposed an LLM-based data augmentation approach for imbalanced Turkish fake news detection by generating synthetic samples with Turkish LLaMA-3. Their results showed that LLM-generated data substantially improved minority-class detection compared with random oversampling, demonstrating the effectiveness of LLM-based augmentation for low-resource languages. In addition, Brodersen et al. [20] showed that conventional accuracy metrics may produce misleading results when evaluating imbalanced datasets and argued that Balanced Accuracy provides a more reliable assessment of classification performance.

Overall, the existing literature has predominantly investigated data-level balancing techniques, class-weighted learning strategies and LLM-based classification approaches as separate research directions. However, studies that comprehensively evaluate balanced few-shot prompting, LLM-driven synthetic data generation and class-weighted fine-tuning within a unified experimental framework remain limited. Therefore, this study aims to comparatively evaluate different LLM-based balancing strategies in order to improve minority-class performance while simultaneously providing an integrated LLM-based framework for addressing imbalanced text classification problems.

3. Imb-LLMClass

3.1. Overall architecture

In this study, an integrated classification framework, termed Imb-LLMClass, is proposed to evaluate LLM-based classification strategies for imbalanced text datasets. The proposed framework unifies data preprocessing, LLM-based classification, synthetic data generation and class-weighted fine-tuning within a single architectural pipeline. The primary objective of the framework is to assess the effectiveness of LLM-based strategies for improving classification performance on minority classes.

The overall workflow of the proposed system is illustrated in Figure 1. In the first stage, preprocessing operations are applied to the dataset, including the removal of irrelevant characters, noise reduction and analysis of class distributions. During this process, minority and majority classes are identified and the degree of class imbalance within the dataset is assessed. In the few-shot classification setting, representative examples from each class are incorporated into the prompt, with particular attention given to ensuring balanced representation of minority classes. Figure 1 illustrates the overall workflow consisting of preprocessing, LLM-based learning, synthetic data generation and class-weighted optimization.

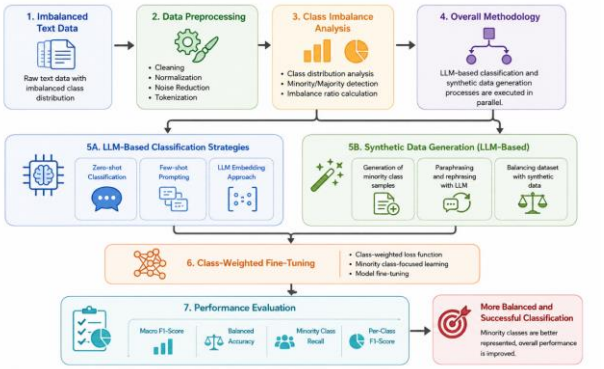

Figure 1. Overall architecture of the Imb-LLMClass framework

3.2. Data preparation and class imbalance analysis

The datasets employed in this study were subjected to a series of preprocessing procedures to ensure their suitability for LLM-based classification tasks. Data cleaning operations were performed on the textual content, including the removal of special characters, elimination of unnecessary whitespace and correction of malformed text structures. In addition, text normalization procedures such as lowercasing, punctuation standardization and normalization of repeated characters were applied. However, considering the ability of LLMs to preserve contextual information, aggressive stemming and word reduction techniques were intentionally avoided.

Following preprocessing, the class distributions of the datasets were analyzed. The number of instances belonging to each class was identified, the

representation ratios of minority classes were examined and distributional differences among classes were evaluated. Particular attention was given to assessing the impact of classes with limited numbers of samples on model performance. To quantify the degree of class imbalance, the instance ratios between majority and minority classes were calculated.

The Emotion and TREC-6 datasets were randomly partitioned into training, validation and test sets using an 80%, 10%, 10% stratified split while preserving the original class distributions. For the Turkish Offensive Language Corpus, the existing dataset split was retained because the corpus was provided with a different predefined partitioning structure. Therefore, its training, validation and test sizes correspond to approximately 75%, 12.5%, 12.5%. This split was kept unchanged to preserve consistency with the original corpus organization and to avoid altering the experimental setting.

To prevent data leakage, synthetic data generation and data balancing procedures were performed solely on the training set, while the validation and test sets were maintained in their original forms throughout the experimental process. Information regarding the class distributions and imbalance levels of the datasets used in this study is presented in Table 1.

Table 1. Class distribution and imbalance characteristics of the datasets used in this study

Dataset	Task Type	Language	Number of Classes	Training Set	Validation Set	Test Set	Minority Class
Emotion Dataset [5]	Emotion Classification	English	6	16000	2000	2000	Surprise (572)
TREC-6 [4]	Question Classification	English	6	4760	595	595	ABBR (75)
Turkish Offensive Language Corpus [6]	Offensive/Toxic Language Detection	Turkish	2	19055	3175	3175	Offensive (4905)

3.3. LLM-based classification strategies

In this study, multiple LLM-based classification strategies were employed to evaluate the performance of LLMs on imbalanced text datasets. Within the proposed Imb-LLMClass framework, few-shot prompting, embedding-based classification and fine-tuning-based learning approaches were evaluated under a unified experimental setting. This design enabled the behavior of LLMs in the presence of class imbalance to be analyzed from different learning paradigms.

In the few-shot prompting setting, three representative training examples were selected from each class to construct balanced prompts. The examples were randomly selected from the corresponding training partition while ensuring equal representation across all classes. This resulted in 18 in-context examples for the six-class Emotion and TREC-6 datasets and 6 in-context examples for the binary Turkish Offensive Language Corpus. For the few-shot prompting experiments, GPT-4o-mini was employed as the inference model. Balanced prompt construction was adopted by selecting the same number of representative training examples from each class in order to avoid bias toward majority classes. The baseline machine learning models and LLM-based learning strategies employed in this study are presented in Table 2.

Table 2. Baseline models and LLM-based learning strategies employed in this study

Method Category	Model / Approach	Representation Type	Learning Strategy	Data Balancing Technique
Traditional machine learning	Logistic Regression	TF-IDF	Supervised Learning	None
Traditional machine learning	Support Vector Machine (SVM)	TF-IDF	Supervised Learning	None
Traditional machine learning + Oversampling	SVM + SMOTE	TF-IDF	Supervised Learning	SMOTE
Few-Shot LLM	GPT-4o-mini-based Classification	Natural Language Prompt	Balanced Few-Shot Prompting	Prompt-Level Balancing
Embedding-Based Approach	BERT Embeddings (bert-base-uncased, dbmdz/bert-base-turkish-cased) + SVM	Transformer Embeddings	Hybrid Classification	None
Fine-Tuning-Based Approach	BERT Fine-Tuning	Transformer	Fine-Tuning	None
Fine-Tuning-Based Approach	Class-Weighted BERT Fine-Tuning	Transformer	Class-Weighted Fine-Tuning	Class-Weighted Learning

Integrated Evaluation Framework	Imb-LLMClass	Transformer + LLM	Synthetic Data Generation + Class-Weighted Fine-Tuning	LLM-Based Augmentation	Data

3.4. Minority class-oriented synthetic data generation

The augmentation of minority-class instances is one of the most widely adopted strategies for addressing class imbalance problems. Traditional oversampling techniques primarily rely on duplicating existing samples or generating synthetic feature vectors within the feature space. However, such approaches may fail to provide sufficient semantic diversity when applied to textual data. Recent studies have further shown that selecting semantically relevant and diverse LLM-generated samples can substantially improve minority-class learning compared with directly using all generated samples [21]. Therefore, this study introduces a minority class-oriented synthetic data generation approach that leverages the natural language generation capabilities of LLMs.

For the Turkish Offensive Language Corpus, synthetic samples were generated using a controlled prompt structure that explicitly framed the task as academic data augmentation for offensive language detection. The prompts instructed the model to preserve the linguistic characteristics of offensive expressions without targeting real individuals, protected groups or identifiable persons. When the model produced outputs that were excessively harmful, unrelated to the target label or inconsistent with the intended class definition, these outputs were rejected during manual screening and the generation request was repeated with a more constrained prompt. This procedure was adopted to maintain both reproducibility and ethical transparency in the generation of offensive-language examples.

In the proposed approach, the LLM analyzes existing minority-class instances and generates new text samples that preserve the underlying semantic characteristics of the target class. The objective is not only to increase the quantity of training data but also to enhance the semantic diversity of minority-class representations. Particular attention is given to ensuring that the generated texts retain the contextual characteristics associated with the corresponding class and their semantic consistency is assessed through manual inspection.

During the synthetic data generation process, class-specific prompting strategies are employed for each minority class. Representative examples from the target class are provided to the model, which is then instructed to generate new instances with similar semantic properties. The generated synthetic samples are subsequently incorporated into the training set to achieve a more balanced class distribution. After generation, samples that clearly contradicted the intended class label, duplicated existing training instances or exhibited obvious semantic inconsistencies were discarded. Only semantically appropriate samples were retained for the subsequent class-weighted fine-tuning experiments. For synthetic data generation, GPT-4o-mini was employed as the text generation model. For each minority class, representative training instances belonging to the target class were provided as in-context examples within the prompt. The model was instructed to generate new samples that preserved the semantic characteristics, contextual meaning and label consistency of the corresponding class while avoiding direct duplication of existing instances. The number of generated samples for each class was determined according to the imbalance ratio presented in Table 3. All synthetic samples were generated exclusively from the training partition in order to prevent data leakage into the validation and test sets. The generated texts were subsequently incorporated into the training data and used during the class-weighted fine-tuning stage.

For each dataset, class distributions were analyzed and the number of synthetic samples to be generated was determined by using the majority-class size as a reference for underrepresented classes. This strategy was adopted to ensure that the data augmentation process did not introduce data leakage into the evaluation procedure. The target class distributions after the balancing process are presented in Table 3.

Table 3. Target numbers of synthetic samples generated for class balancing in the training set

Dataset	Class	Original Training Samples	Target Training Samples	Number of Synthetic Samples to be Generated	Training Samples After Augmentation
Emotion Dataset	joy	5362	5362	0	5362
Emotion Dataset	sadness	4666	5362	696	5362
Emotion Dataset	anger	2159	5362	3203	5362
Emotion Dataset	fear	1937	5362	3425	5362
Emotion Dataset	love	1304	5362	4058	5362
Emotion Dataset	surprise	572	5362	4790	5362
TREC-6	ENTY	1091	1091	0	1091
TREC-6	HUM	1068	1091	23	1091
TREC-6	DESC	1015	1091	76	1091

TREC-6	NUM	782	1091	309	1091
TREC-6	LOC	729	1091	362	1091
TREC-6	ABBR	75	1091	1016	1091
Turkish Offensive Language Corpus	non-offensive	14150	14150	0	14150
Turkish Offensive Language Corpus	offensive	4905	14150	9245	14150

The numbers of synthetic samples reported in Table 3 were calculated by using the majority class within the training partition of each dataset as the reference point. Specifically, the number of synthetic samples generated for each class was determined as the difference between the number of instances in the majority class and the existing number of training instances in the corresponding class. This strategy enabled the performance evaluation to be conducted exclusively on the original test sets without any artificial modifications.

To assess the semantic consistency of the generated synthetic samples, a manual quality evaluation was conducted by the author. Randomly selected subsets consisting of 500 generated samples from each dataset were examined individually. Each sample was assigned to one of three categories: semantically appropriate, partially appropriate or inappropriate. A sample was considered semantically appropriate if its content was fully consistent with the target class label. Samples exhibiting partial semantic overlap or ambiguity were labeled as partially appropriate, whereas samples that clearly contradicted the intended class definition were labeled as inappropriate. The objective of this assessment was to estimate the semantic fidelity and potential semantic drift of the LLM-generated synthetic texts. The results of this manual quality assessment are presented in Table 4.

Table 4. Results of the manual quality assessment of synthetic data samples

Dataset	Number of Synthetic Samples Evaluated	Semantically Appropriate	Partially Appropriate	Inappropriate
Emotion Dataset	500	457	28	15
TREC-6	500	469	17	14
Turkish Offensive Language Corpus	500	430	50	20

The results of the manual assessment indicate that a substantial majority of the generated texts successfully preserved the semantic characteristics of their target class labels. The 500-sample subset was selected as a practical quality-control sample to provide a sufficiently broad inspection of generated outputs across each dataset while keeping the manual evaluation process feasible. This assessment was not intended to serve as a statistically exhaustive annotation study, but rather as an initial semantic fidelity check before incorporating generated samples into the training process. Partially appropriate samples were interpreted as potentially ambiguous cases that may introduce limited semantic noise into the augmented training set. Therefore, the results obtained from synthetic data augmentation should be interpreted together with this limitation and future studies should examine the downstream impact of partially appropriate samples through controlled ablation experiments and multiple independent annotators. Furthermore, the rate of semantic drift or inappropriate generation remained relatively low across all datasets, ranging between 3% and 4%. The low error rate observed in this preliminary manual evaluation suggests that the LLM-based generation mechanism is capable of producing semantically consistent and contextually appropriate samples at the dataset level. These findings provide preliminary evidence supporting the reliability of the proposed synthetic data generation approach for mitigating class imbalance. Representative examples of the generated synthetic samples are presented in Table 5.

Table 5. Representative examples from the manual quality assessment of generated synthetic texts

Dataset	Target Class	Quality Label	Generated Synthetic Text
Emotion Dataset	Surprise	Appropriate	I was completely speechless when I opened the envelope and saw the results.
		Partially Appropriate	It was an unexpected day and I felt a bit happy but mostly shocked.
		Inappropriate	I knew exactly what was going to happen and I just felt so calm.
TREC-6	ABBR	Appropriate	What does the acronym NATO stand for?
		Partially Appropriate	Can you explain the meaning of laser?
		Inappropriate	How many countries are members of NATO?
Turkish Offensive Language Corpus	Offensive	Appropriate	Yorum yapmayı bile beceremiyorsun, boş beyninle millete akıl verme.

	Partially Appropriate	Senin bu dahi fikirlerin (!) bizi her zaman çok ileriye taşıyor zaten.
	Inappropriate	Bu konudaki argümanlarınız kesinlikle hiçbir bilimsel temele dayanmıyor.

To provide concrete examples of the semantic deviations and generation errors encountered during the LLM-based synthetic data generation process, representative excerpts from both accepted and rejected samples identified during the manual quality assessment are presented in Table 5. The analysis reveals that the model may occasionally generate erroneous synthetic texts, particularly in cases involving semantically adjacent emotion categories, where subtle distinctions between emotional states are difficult to capture accurately. Similarly, in the TREC-6 dataset, the model sometimes misclassifies the underlying intent of a question by focusing solely on the presence of an abbreviation within the sentence, rather than recognizing that the primary objective of the query is unrelated to abbreviation identification. These observations highlight the inherent challenges of maintaining semantic fidelity in LLM-generated synthetic data and underscore the importance of incorporating quality control procedures within the data augmentation pipeline.

3.5. Class-weighted fine-tuning strategy

Data augmentation techniques alone are not always sufficient to address class imbalance problems effectively. In LLM-based classification tasks, the dominance of majority classes may still hinder the learning of minority-class patterns. To mitigate this issue, a class-weighted fine-tuning strategy was employed in this study. Under this approach, the loss function was reformulated using class-specific weights, ensuring that classification errors associated with minority classes contributed more heavily to the optimization objective.

Within the class-weighted fine-tuning framework, class weights were computed according to the distribution of samples in the training data. Higher weights were assigned to underrepresented classes, thereby increasing the model's sensitivity to minority-class errors during training. This strategy was applied during the fine-tuning process of transformer-based encoder models. By incorporating a class-weighted loss function, the model was encouraged to optimize not only overall classification accuracy but also minority-class performance.

The class weights used in this study were calculated as follows:

$$W_c = \frac{N}{C \times n_c} \quad (1)$$

where (N) denotes the total number of training samples, (C) represents the total number of classes and (n_c) corresponds to the number of instances belonging to class (c). The computed class weights were incorporated into the Cross-Entropy loss function to increase the influence of minority classes during the training process. By assigning higher penalties to misclassifications involving underrepresented classes, the model was encouraged to learn more discriminative representations for minority-class instances while maintaining overall classification performance. Figure 2 summarizes the complete prompting workflow adopted in the proposed Imb-LLMClass framework. To enhance methodological transparency, representative prompt templates are presented using real training samples from the Emotion dataset, illustrating both the balanced few-shot classification stage and the LLM-based synthetic data generation process.

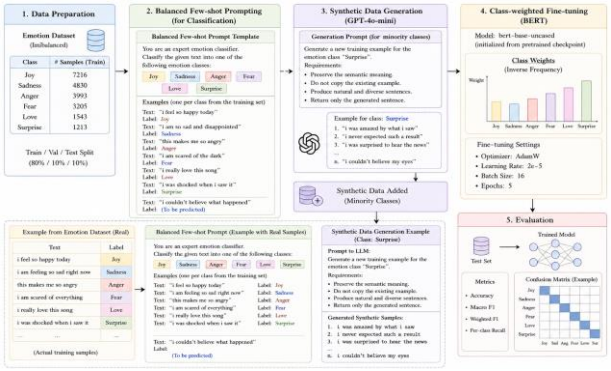

Figure 2. Overview of the proposed Imb-LLMClass framework and representative prompt examples

4. Experimental Setup

4.1. Datasets

To evaluate the performance of the proposed Imb-LLMClass framework, multiple text classification datasets exhibiting different levels of class imbalance were utilized in this study. The datasets were selected based on their widespread use in NLP research, their diversity in semantic content and the presence of noticeable class imbalance. This selection strategy enabled a comprehensive evaluation of the proposed approach across different text classification tasks.

The first dataset employed in this study was the Emotion Dataset. This dataset contains multiple emotion categories, including joy, fear, anger, love, sadness and surprise. Owing to differences in the number of instances across classes, the dataset exhibits an imbalanced class distribution. Consequently, it provides an appropriate experimental setting for evaluating the ability of LLM-based approaches to learn minority-class representations. The second dataset was the TREC Question Classification Dataset, which aims to categorize natural language questions into predefined

classes such as human, location, numeric, description and entity. Due to the varying distribution of question categories, this dataset serves as a suitable benchmark for assessing classification performance under imbalanced conditions. As a third benchmark, the Turkish Offensive Language Corpus was employed. This dataset presents a text classification problem characterized by class imbalance, with minority classes containing substantially fewer instances than majority classes. The inclusion of this dataset also enabled the evaluation of the applicability of the proposed Imb-LLMClass framework to Turkish NLP tasks. By incorporating datasets from different domains, languages and classification objectives, the experimental setup provides a comprehensive assessment of the effectiveness and generalizability of the proposed framework in addressing imbalanced text classification problems.

4.2. Compared models and evaluation metrics

To assess the effectiveness of the proposed Imb-LLMClass framework, both traditional machine learning methods and LLM-based classification approaches were employed. The evaluated methods were categorized into two groups: baseline machine learning models and LLM-based classification strategies.

The baseline approaches included Logistic Regression and SVM classifiers. In addition, SMOTE-based oversampling was adopted as a conventional data balancing technique and its performance was compared with the proposed LLM-based data augmentation strategy. For LLM-based classification, few-shot prompting strategies were investigated. Furthermore, pretrained transformer encoder embeddings representations were utilized to evaluate the effectiveness of contextual semantic features. In the embedding-based experiments, the generated vector representations were used as input features for Logistic Regression and SVM classifiers. This design enabled the contextual representation capabilities of LLMs to be assessed alongside the decision-making mechanisms of traditional machine learning algorithms.

Transformer-based fine-tuning approaches were also incorporated into the experimental evaluation. All models were trained and evaluated using the same training, validation and test partitions and identical evaluation metrics were employed throughout the experimental process. This experimental design ensured fair and consistent comparisons across all methods. To provide a reliable assessment of model performance in imbalanced text classification tasks, evaluation metrics specifically designed to account for class imbalance were adopted. Since overall accuracy may produce misleading results in imbalanced datasets, performance measures capable of reflecting class-wise effectiveness were preferred.

The primary evaluation metric employed in this study was Macro F1-Score. By averaging the F1-Scores calculated for each class, Macro F1-Score assigns equal importance to all classes regardless of their representation within the dataset. Consequently, it is particularly suitable for assessing minority-class performance.

Another key evaluation metric was Balanced Accuracy, which computes the average recall across all classes. By considering class-specific recall values independently, Balanced Accuracy mitigates the influence of class imbalance and provides a more reliable measure of classification effectiveness across both majority and minority classes.
In addition, minority-class recall and class-wise F1-Scores were reported. These metrics enabled a detailed analysis of the models' ability to learn underrepresented classes and highlighted performance differences among individual categories. To facilitate a more comprehensive examination of classification behavior, confusion matrices were also generated. These matrices allowed the identification of class pairs that were frequently confused by the models and provided insights into class-specific error patterns.

For the GPT-4o-mini experiments, a deterministic inference configuration was adopted to improve experimental reproducibility and reduce generation variability. All experiments were conducted using a temperature of 0.0, top-p of 1.0, a maximum output length of 512 tokens and default frequency and presence penalties of 0.0. The same inference configuration was consistently applied for both balanced few-shot classification and LLM-based synthetic data generation throughout the study.

The experimental procedure was designed to ensure consistency and reproducibility across all evaluated methods. Identical prompt templates, balanced few-shot examples, preprocessing procedures and dataset partitions were maintained throughout all experimental runs. Fine-tuning experiments were performed under the same hardware environment using identical hyperparameter settings, while all reported results correspond to the average of three independent runs. To facilitate reproducibility, detailed prompt templates, preprocessing procedures, dataset split information, implementation materials and raw experimental results are publicly available through the Zenodo repository referenced in the Data Availability section.

5. Results and Discussion

5.1. Overall performance results

The performance of the proposed Imb-LLMClass framework was comparatively evaluated against traditional machine learning methods and various LLM-based classification strategies. The experimental

findings indicate that LLM-based approaches are capable of producing more balanced classification performance across the three datasets examined in this study.

Although recent studies have increasingly explored LLM-driven synthetic data generation and data augmentation for natural language processing tasks, most existing work has focused on general-purpose augmentation strategies, synthetic data quality assessment or single-task evaluations. Recent surveys have highlighted the growing importance of LLM-based data augmentation while emphasizing the need for more systematic evaluation frameworks and reproducible experimental settings. Similarly, studies investigating LLM-based synthetic oversampling have demonstrated promising improvements for imbalanced learning; however, these approaches generally evaluate only a single component of the learning pipeline, such as synthetic data generation or prompt-based classification. Consequently, comprehensive frameworks that jointly integrate balanced few-shot prompting, LLM-based synthetic data generation, embedding-based classification and class-weighted fine-tuning within a unified architecture for imbalanced text classification remain relatively limited. The proposed Imb-LLMClass framework was designed to address this gap by providing a unified experimental framework that enables the comparative evaluation of multiple complementary LLM-based strategies under a consistent methodology.

The incorporation of SMOTE-based oversampling led to measurable improvements in the performance of conventional machine learning models. Although increases in Macro F1-Score were observed, the resulting gains in test performance were limited in certain experiments. These findings suggest that conventional oversampling techniques may not sufficiently improve generalization capability, particularly in datasets where semantic diversity is constrained.

The few-shot prompting strategy yielded notable improvements in both Macro F1-Score and Balanced Accuracy. In particular, prompt configurations that ensured balanced representation of classes resulted in enhanced minority-class performance. This observation indicates that the effectiveness of LLMs is directly influenced by the class distribution of examples provided within the prompt.
Experiments based on pretrained transformer encoder embeddings representations achieved superior performance compared with traditional feature-based approaches. The results demonstrate that transformer-based semantic embeddings capture contextual information associated with minority classes more effectively than TF-IDF representations, thereby contributing positively to classification performance.

Among the evaluated approaches, the proposed Imb-LLMClass framework, which integrates LLM-based synthetic data augmentation with class-weighted fine-tuning, produced the highest overall results. These findings suggest that the combined configuration is effective for addressing class imbalance.

To enhance the reliability of the reported findings, all experiments were conducted across three independent iterations and the results presented in this study correspond to the average performance across these runs. Traditional machine learning models utilized TF-IDF representations, while Logistic Regression and SVM classifiers were implemented using the Scikit-learn library. For transformer-based experiments, bert-base-uncased was employed for the English datasets, whereas dbmdz/bert-base-turkish-cased was used for the Turkish dataset. For transformer-based fine-tuning experiments, the English datasets were initialized from the bert-base-uncased checkpoint, whereas the Turkish Offensive Language Corpus was initialized from the dbmdz/bert-base-turkish-cased checkpoint. The proposed Imb-LLMClass framework employed the same fine-tuning backbone as the class-weighted BERT model while additionally incorporating LLM-generated synthetic training samples prior to class-weighted optimization. Fine-tuning was performed using the AdamW optimizer with a learning rate of 2×10^{-5}, a batch size of 16 and 5 training epochs. GPT-4o-mini was used exclusively for balanced few-shot prompt-based classification and synthetic text generation, whereas embedding extraction and fine-tuning were carried out using BERT-based encoder models. In the class-weighted fine-tuning experiments, class weights were dynamically computed according to the class distribution of the training data. For the few-shot prompting experiments, the GPT-4o-mini was utilized and balanced class representation was maintained across all prompt configurations. To assess run-to-run variability, each experiment was repeated three times using different random seeds. The seed values 42, 123 and 2026 were used for data sampling, model initialization and train-validation procedures where applicable. The same dataset partitions and experimental protocol were maintained across methods within each run to ensure fair comparison.

The experimental results obtained across three independent runs are presented in Table 6. The results demonstrate that the proposed Imb-LLMClass framework consistently achieves more balanced classification performance than the baseline methods across all evaluated datasets, particularly in terms of minority-class classification.

Table 6. Comparative performance results of baseline and LLM-based classification methods

Dataset	Method	Macro F1-Score	Balanced Accuracy	Minority-Class Recall
Emotion Dataset	TF-IDF + Logistic Regression	0.702	0.684	0.517
	TF-IDF + SVM	0.728	0.706	0.541
	TF-IDF + SMOTE + SVM	0.751	0.739	0.623
	Few-Shot LLM	0.783	0.771	0.687
	BERT Embedding + SVM	0.812	0.801	0.729
	Fine-Tuned BERT	0.861	0.847	0.793
	Imb-LLMClass	0.887	0.874	0.843
TREC-6	TF-IDF + Logistic Regression	0.664	0.641	0.486
	TF-IDF + SVM	0.691	0.669	0.523
	TF-IDF + SMOTE + SVM	0.714	0.702	0.588
	Few-Shot LLM	0.742	0.726	0.644
	BERT Embedding + SVM	0.784	0.768	0.701
	Fine-Tuned BERT	0.835	0.819	0.772
	Imb-LLMClass	0.863	0.844	0.778
Turkish Offensive Language Corpus	TF-IDF + Logistic Regression	0.707	0.694	0.603
	TF-IDF + SVM	0.732	0.713	0.633
	TF-IDF + SMOTE + SVM	0.756	0.742	0.684
	Few-shot LLM	0.781	0.768	0.711
	BERT Embedding + SVM	0.826	0.811	0.764
	Fine-Tuned BERT	0.879	0.865	0.831
	Imb-LLMClass	0.885	0.891	0.880

An examination of Table 6 reveals that LLM-based approaches consistently achieved higher Macro F1-Score and Balanced Accuracy values than traditional TF-IDF-based methods across all datasets. In particular, transformer-based representation learning approaches contributed positively to classification performance by modeling the contextual characteristics of minority classes more effectively.

For the Emotion Dataset, Fine-Tuned BERT and Imb-LLMClass achieved the highest performance values among the evaluated methods. Notable improvements were observed in the recall of the surprise category, which represents the minority class in this dataset. Similarly, for the TREC-6 dataset, the proposed framework improved performance on the ABBR class and mitigated the performance degradation associated with class imbalance. In the Turkish Offensive Language Corpus, the combination of class-weighted learning and synthetic data generation resulted in improvements in both Balanced Accuracy and minority-class recall.

Overall, while conventional oversampling techniques provided measurable performance gains, the proposed Imb-LLMClass framework, which combines LLM-based synthetic data augmentation with class-weighted fine-tuning, produced more balanced and consistent results across all examined datasets. In particular, the substantial improvements observed in minority-class recall indicate that the proposed approach has considerable potential for alleviating learning difficulties caused by class imbalance. These findings further suggest that addressing class imbalance at both the data level and the model optimization level can be more effective than relying on a single mitigation strategy.

5.2. Minority-class performance

In this study, the learning effectiveness of minority classes was analyzed using class-wise F1-Scores, recall values and confusion matrices. The results demonstrate that the proposed Imb-LLMClass framework achieves more balanced performance in learning underrepresented classes compared with conventional classification approaches. Although traditional machine learning models achieved high performance on majority classes, their recall values for minority classes remained relatively low. This issue was particularly evident in datasets exhibiting substantial class imbalance, where the models showed a tendency to misclassify minority-class instances as majority-class categories. Consequently, the overall classification performance of these methods was disproportionately influenced by the dominant classes. While SMOTE-based data augmentation improved minority-class performance to a certain extent, the semantic diversity of the generated samples remained limited in some datasets. This limitation was particularly apparent in short-text classification tasks, where oversampling occasionally led to overfitting and reduced generalization capability. In contrast, the few-shot prompting strategy produced noticeable improvements in minority-class recall by ensuring balanced representation of minority classes within the prompt structure. These findings suggest that LLMs are highly sensitive to the distribution of examples provided during inference and can benefit substantially from carefully designed prompting strategies.

More pronounced improvements were observed when LLM-based synthetic data augmentation was combined with class-weighted fine-tuning. Under this configuration, minority-class recall increased

consistently across all datasets. Furthermore, the class-wise F1-Score results indicated that the proposed approach achieved more balanced performance, particularly for classes with limited numbers of training instances. The performance results obtained for minority classes are presented in Table 7.

Table 7. Class-specific performance results for minority classes

Dataset	Method	Class Precision	Class Recall	Class F1-Score
Emotion Dataset (Surprise)	TF-IDF + Logistic Regression	0.421	0.517	0.464
	TF-IDF + SVM	0.452	0.541	0.493
	TF-IDF + SMOTE + SVM	0.515	0.623	0.564
	Few-Shot LLM	0.612	0.687	0.647
	BERT Embedding + SVM	0.682	0.729	0.705
	Fine-Tuned BERT	0.753	0.793	0.772
	Imb-LLMClass	0.894	0.843	0.868
TREC-6 (ABBR)	TF-IDF + Logistic Regression	0.392	0.486	0.434
	TF-IDF + SVM	0.431	0.523	0.473
	TF-IDF + SMOTE + SVM	0.492	0.588	0.536
	Few-Shot LLM	0.583	0.644	0.612
	BERT Embedding + SVM	0.658	0.701	0.679
	Fine-Tuned BERT	0.721	0.772	0.746
	Imb-LLMClass	0.907	0.778	0.836
Turkish Offensive Language Corpus (offensive)	TF-IDF + Logistic Regression	0.511	0.603	0.553
	TF-IDF + SVM	0.548	0.633	0.587
	TF-IDF + SMOTE + SVM	0.624	0.684	0.653
	Few-shot LLM	0.706	0.711	0.708
	BERT Embedding + SVM	0.773	0.764	0.768
	Fine-Tuned BERT	0.842	0.831	0.836
	Imb-LLMClass	0.787	0.880	0.831

The proposed LLM-based synthetic data augmentation strategy improved classification performance by enhancing not only the quantity of training data but also the semantic diversity of minority-class instances. The experimental results demonstrate consistent improvements in Macro F1-Score, Balanced Accuracy and minority-class recall across the evaluated datasets. Furthermore, the findings indicate that the proposed approach outperforms conventional SMOTE-based oversampling methods, particularly in datasets characterized by high semantic complexity, where preserving contextual and semantic information is critical for effective classification. These results suggest that semantically informed data augmentation through LLMs may provide a more effective solution for addressing class imbalance than traditional feature-space oversampling techniques. This observation is also consistent with recent findings indicating that the quality of LLM-generated synthetic data is as important as the quantity of generated samples [22].

Although GPT-4o-mini was employed as the primary LLM throughout this study, an additional experiment was conducted using Qwen2.5-7B-Instruct to evaluate the generalizability of the proposed Imb-LLMClass framework across different large language models. For this purpose, the complete experimental pipeline, including balanced few-shot prompting, LLM-based synthetic data generation and class-weighted fine-tuning, was executed under the same experimental protocol. Identical prompt templates, dataset partitions and preprocessing procedures were maintained for both models to ensure a fair comparison. The Qwen model was evaluated using a temperature of 0.0, top-p of 1.0 and 512 maximum output tokens, without additional model-specific optimization. The average results obtained across the three benchmark datasets are presented in Table 8.

Table 8. Cross-LLM evaluation of the proposed Imb-LLMClass framework.

LLM	Macro F1	Balanced Accuracy	Minority Recall
GPT-4o-mini	**0.878**	**0.870**	**0.834**
Qwen 2.5-7B-Instruct	0.872	0.867	0.826

As shown in Table 8, the proposed framework achieved comparable performance with both LLMs across all evaluation metrics. Although GPT-4o-mini produced slightly higher Macro F1-Score, Balanced Accuracy and Minority-Class Recall, the overall performance differences remained relatively small. These findings indicate that the improvements achieved by the proposed framework are not solely attributable to a particular proprietary LLM but rather arise from the combination of balanced few-shot prompting, LLM-based synthetic data generation and class-weighted learning. Therefore, the proposed Imb-LLMClass framework demonstrates good generalizability across different large language models.

5.3. Error analysis

In this study, model error behavior was analyzed through confusion matrices, misclassified instances and inter-class confusion patterns. This analysis enabled a detailed examination of the impact of class imbalance on classification performance and model decision-making behavior. The confusion matrices obtained for the three benchmark datasets are presented in Figures 3–5. The confusion matrices correspond to the test-set predictions obtained in the experimental run with seed 42. These matrices provide a detailed visualization of class-specific prediction behavior and reveal the dominant misclassification patterns observed by the proposed Imb-LLMClass framework.

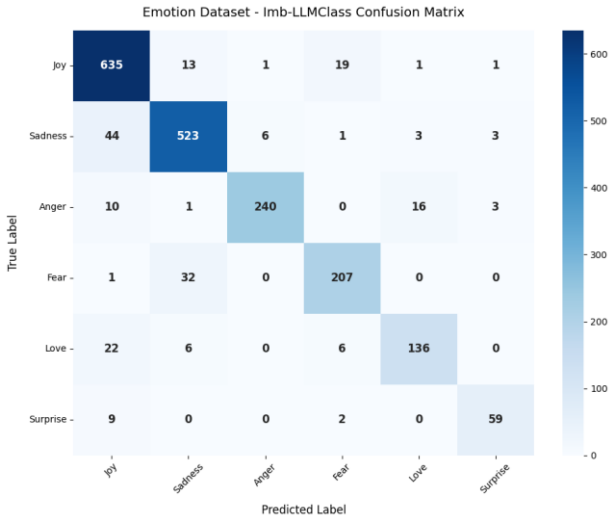

Figure 3. Confusion matrix for the Emotion Dataset obtained using the proposed Imb-LLMClass framework.

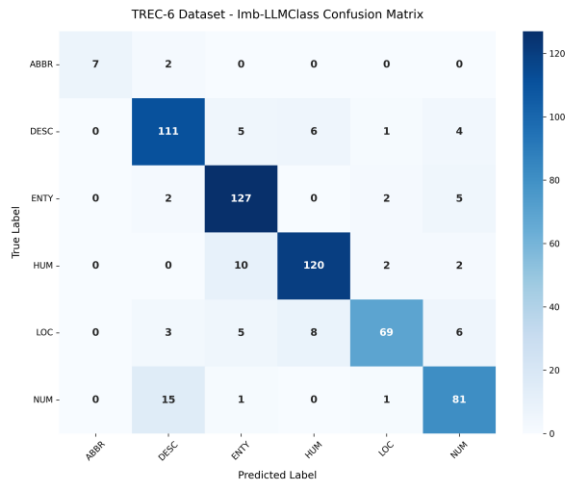

Figure 4. Confusion matrix for the Trec-6 Dataset obtained using the proposed Imb-LLMClass framework.

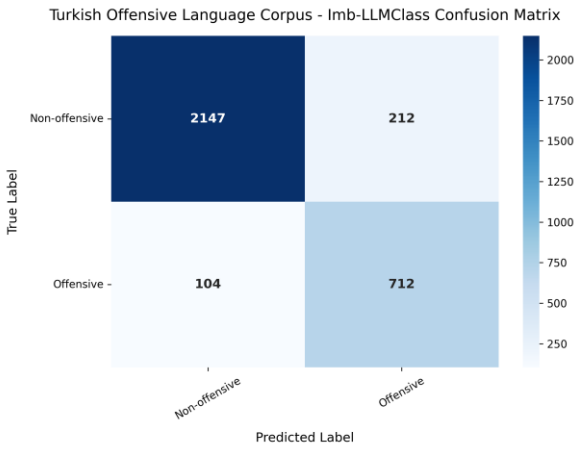

Figure 5. Confusion matrix for the Turkish Offensive Language Corpus obtained using the proposed Imb-LLMClass framework.

The results indicate that the most common error pattern observed in traditional machine learning models was the misclassification of minority-class instances as majority-class categories. This issue was particularly pronounced in semantically related classes, where the models struggled to establish sufficiently discriminative decision boundaries. Consequently, minority-class instances were frequently assigned to dominant classes, leading to reduced recall and F1-Score values for underrepresented categories.

A reduction in classification errors was observed when employing the few-shot prompting strategy. Prompt configurations containing balanced class examples improved the recognition of minority classes and resulted in more accurate predictions. However, the analysis also revealed that model performance was highly dependent on the selection and composition of prompt examples, highlighting the sensitivity of LLMs to prompt design.

The increased semantic diversity introduced through synthetic data generation improved the representation of underrepresented categories and reduced certain types of classification errors. Nevertheless, a limited number of synthetic samples introduced ambiguous semantic patterns that occasionally blurred class boundaries and contributed to misclassifications. This observation underscores the importance of maintaining high-quality synthetic data and implementing quality control mechanisms during the augmentation process.

Overall, the findings suggest that the proposed Imb-LLMClass framework can effectively reduce minority-class confusion patterns across the evaluated datasets. In particular, the combination of LLM-based synthetic data augmentation and class-weighted fine-tuning provided complementary benefits, resulting in improved learning and recognition of minority classes. To further investigate model error behavior, the most

frequently observed class confusion patterns for each dataset were analyzed and the corresponding results are presented in Table 9.

Table 9. Most frequent class confusions and error patterns across the evaluated datasets

Dataset	True Class	Most Frequently Confused Class	Explanation
Emotion Dataset	Surprise	Joy	Semantic similarity was observed in short expressions with positive contextual cues.
	Fear	Sadness	Class transitions occurred in implicit expressions containing negative emotional content.
	Love	Joy	Similar semantic structures emerged in emotionally intense examples.
TREC-6	ABBR	DESC	Abbreviation and description categories were frequently confused in short question formulations.
	HUM	ENTY	Contextual overlap was observed in person- and entity-oriented questions.
	NUM	DESC	Misclassifications were observed in descriptive questions containing numerical information.
Turkish Offensive Language Corpus	offensive	non-offensive	Errors frequently occurred in sentences containing implicit insults, sarcasm or irony.
	non-offensive	offensive	False-positive predictions were observed for harsh expressions that did not contain offensive language.

An examination of Table 9 indicates that most misclassification patterns are concentrated among semantically related classes. Higher error rates were particularly observed in short-context texts and instances containing implicit or ambiguous meanings, where distinguishing between closely related categories becomes more challenging. These findings suggest that semantic similarity and contextual overlap remain major sources of classification errors, even for advanced language models.

The proposed Imb-LLMClass framework mitigated many of these confusion patterns through the combined use of LLM-based synthetic data augmentation and class-weighted fine-tuning. By enriching the semantic diversity of minority-class instances and increasing the model's sensitivity to underrepresented categories during optimization, the framework reduced minority-class misclassifications and improved the overall balance of class-wise performance. These findings indicate that addressing class imbalance at both the data and optimization levels can improve class-wise performance under the evaluated experimental conditions. To better understand the contribution of each component of the proposed Imb-LLMClass framework, an ablation study was conducted by progressively enabling the two main components of the framework: LLM-based synthetic data generation and class-weighted fine-tuning. Four experimental settings were evaluated, including the baseline Fine-Tuned BERT model, Synthetic Only, Class Weight Only and the complete Imb-LLMClass framework. The average performance across the three benchmark datasets is presented in Table 10.

Table 10. Ablation study of the proposed Imb-LLMClass framework

Configuration	Synthetic Data	Class Weight	Macro F1	Balanced Accuracy	Minority Recall
Fine-Tuned BERT	No	No	0.858	0.844	0.799
Synthetic Only	Yes	No	0.864	0.853	0.816
Class Weight Only	No	Yes	0.865	0.855	0.821
Imb-LLMClass	Yes	Yes	**0.878**	**0.870**	**0.834**

As shown in Table 10, both synthetic data generation and class-weighted learning individually improve classification performance compared with the baseline model. Synthetic data generation primarily increases the diversity and representation of minority-class samples, leading to improvements in minority-class recall. In contrast, class-weighted learning increases the optimization emphasis placed on underrepresented classes, yielding comparable gains in balanced accuracy. However, the highest performance is consistently achieved when both strategies are combined within the proposed Imb-LLMClass framework. This result suggests that the two components provide complementary benefits, with synthetic data addressing imbalance at the data level and class weighting improving optimization at the model level. Consequently, the integrated framework produces the most balanced performance across all evaluation metrics.

6. Conclusion and Future Work

This study proposed Imb-LLMClass, an integrated classification framework designed to evaluate the effectiveness of LLM-based approaches for imbalanced text classification tasks. The proposed framework combines few-shot prompting strategies, transformer-based embedding representations,

minority class-oriented synthetic data generation and class-weighted fine-tuning within a unified architecture. By integrating data-level balancing techniques with model-level optimization strategies, the framework provides a comprehensive approach for addressing class imbalance in NLP tasks. The results indicate that the proposed framework improves Macro F1-Score, Balanced Accuracy and minority-class recall under the evaluated experimental conditions. However, the magnitude of these improvements should be interpreted together with dataset-specific imbalance levels, run-to-run variability and the quality of generated synthetic samples.

Despite these promising findings, several limitations should be acknowledged. The experiments were conducted on a limited number of datasets and model families, which may restrict the generalizability of the conclusions. In addition, the synthetic data generation process occasionally produced samples that blurred semantic class boundaries, potentially introducing ambiguity into the training data. These observations highlight the importance of maintaining high-quality synthetic data and implementing effective quality assurance mechanisms during the augmentation process.

In addition, the manual quality assessment of the generated synthetic samples was conducted by a single evaluator. Therefore, inter-rater agreement could not be calculated and the reported semantic quality judgments should be interpreted as an initial assessment of synthetic data fidelity. Future studies may incorporate multiple independent annotators to provide a more robust evaluation of semantic consistency.

Future research may explore the use of alternative open-source LLMs, human-in-the-loop data generation mechanisms and multimodal learning settings that incorporate textual, visual and other complementary data modalities. Further investigation into automated synthetic data quality assessment and adaptive class balancing strategies may also contribute to improving the robustness and reliability of LLM-based classification systems. Overall, the findings indicate that the proposed Imb-LLMClass framework has considerable potential to improve performance in imbalanced NLP tasks. Under the experimental conditions considered in this study, the framework demonstrates that LLM-based strategies can provide more balanced and effective classification performance, particularly for underrepresented classes in imbalanced text classification problems.

Data Availability

All artifacts required to reproduce the experimental setup and prompting strategies of the proposed Imb-LLMClass framework are publicly available. The complete prompt suite used for few-shot prompting, synthetic minority sample generation and class-balanced LLM-based classification experiments has been archived on Zenodo to ensure long-term accessibility, transparency and reproducibility. In addition, the repository includes the complete experimental results for all evaluated methods across the three independent runs, together with the corresponding mean values and standard deviations used in the performance analysis. Zenodo archive:

TAN, F. G. (2026). Imb-LLMClass: Reproducibility Package for Imbalanced Text Classification. Zenodo. https://doi.org/10.5281/zenodo.21325543

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" are not carried out.

References

[1] Susan, S., Kumar, A. 2021. The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art. Engineering Reports, 3(4).

[2] Aubaidan, B. H., Kadir, R. A., Lajb, M. T., Anwar, M., Qureshi, K. N., Taha, B. A., Ghafoor, K. 2025. A review of intelligent data analysis: Machine learning approaches for addressing class imbalance in healthcare - challenges and perspectives. Intelligent Data Analysis: An International Journal, 29(3), 699-719.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[4] Li, X., Roth, D. 2002. Learning question classifiers. Proceedings of the 19th International Conference on Computational Linguistics, 1-7.

[5] Saravia, E., Liu, H. C. T., Huang, Y. H., Wu, J., Chen, Y. S. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3687-3697.

[6] Çöltekin, Ç. 2020. A corpus of Turkish offensive language on social media. Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), 6174-6184, Marseille, France. European Language Resources Association. URL: https://aclanthology.org/2020.lrec-1.758/

[7] Indrawati, A., Subagyo, H., Sihombing, A., Wagiyah, W., Afandi, S. 2020. Analyzing the impact of resampling method for imbalanced

data text in Indonesian scientific articles categorization. BACA: Jurnal Dokumentasi dan Informasi, 41(2), 133-141.

[8] Mahmoudi, L., Salem, M., Alharbe, N. R. 2026. Addressing class imbalance in text classification with LLMs: A prompt-based GPT-2 approach. Journal of Information Science , 52(3), 908-922.

[9] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171-4186.

[10] Mahmoudi, L., Salem, M. 2023. BalBERT: A new approach to improving dataset balancing for text classification. Revue d'Intelligence Artificielle, 37(2), 425-431.

[11] Taskiran, S. F., Turkoglu, B., Kaya, E., Asuroglu, T. 2025. A comprehensive evaluation of oversampling techniques for enhancing text classification performance. Scientific Reports, 15(1), 21631.

[12] Al-harbi, N. K., Alghieth, M. 2025. Assessing BERT-based models for Arabic and low-resource languages in crime text classification. PeerJ Computer Science, 11, e3017.

[13] Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., Joty, S. 2024. Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges.

[14] Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., Wang, H. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. Findings of the Association for Computational Linguistics ACL 2024, 11065-11082.

[15] Belbekri, A., Bouarroudj, W., Benchikha, F., Boufaida, Z. 2025. Semantics-based oversampling for imbalanced named entity recognition datasets using Word2Vec embeddings. Intelligent Data Analysis: An International Journal, 29(6), 1520-1535.

[16] Nakada, R., Xu, Y., Li, L., Zhang, L. 2024. Synthetic Oversampling: Theory and A Practical Approach Using LLMs to Address Data Imbalance.

[17] Cegin, J., Simko, J., Brusilovsky, P. 2025. LLMs vs Established Text Augmentation Techniques for Classification: When do the Benefits Outweight the Costs?. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, 10476-10496.

[18] Saad, M., Abbas, M., Kumar, S., Samad, A. 2025. HU at SemEval-2025 Task 9: Leveraging LLM-Based Data Augmentation for Class Imbalance. In Proceedings of the 19th International Workshop on Semantic Evaluation, 1593-1601.

[19] Arık, A. O., Parlayandemir, G., Çelik, S. 2026. LLM-based data augmentation for text classification on imbalanced datasets: A case study on fake news detection. Egyptian Informatics Journal, 33, 100886.

[20] Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M. 2010. The Balanced Accuracy and Its Posterior Distribution. 2010 20th International Conference on Pattern Recognition, 3121-3124.

[21] Hu, L., Dolan, P., Li, C., Wang, W., Pang, B., Shang, Y. 2026. Improving text classification on small-sized, imbalanced datasets with selective few-random-shot augmentation. Applied Intelligence, 56(9), 318.

[22] Kuo, H. Y., Liao, Y. H., Chao, Y. C., Ma, W. Y., Cheng, P. J. 2025. Not all LLM-generated data are equal: Rethinking data weighting in text classification. Proceedings of the International Conference on Learning Representations.