Optimization Techniques for Dealing with Small Dataset for Sentiment Analysis

Isfaque AL Kaderi Tuhin *, Zhengkui Wang, Xiaorong Li, and Wei Zhang

Information and Communications Technology, Singapore Institute of Technology, Singapore, Singapore
Email: tuhin.kaderi@singaporetech.edu.sg (I.A.K.T.); zhengkui.wang@singaporetech.edu.sg (Z.W.);
xiaorong.li@singaporetech.edu.sg (X.L.); wei.zhang@singaporetech.edu.sg (W.Z.)
*Corresponding author

Abstract*—Sentiment analysis is crucial for many organizations, including those in the transportation industry which use it to gain insights into current issues and improve services provided by public transport operators. However, industries such as transportation face difficulties in fully utilizing AI tools due to the lack of annotated, domain-specific datasets. This scarcity often stems from challenges such as the sensitive nature of the data and a shortage of manpower dedicated to data annotation. Although many sentiment analysis technologies exist, including state-of-the-art transformer-based models, typically require access to large, annotated datasets. This creates a gap in solutions for scenarios characterized by limited and imbalanced data. Our research aims to address this gap by systematically exploring strategies for optimizing sentiment analysis with small, imbalanced datasets for multi-class sentiment classification tasks. We consider constraints posed by data privacy and resource limitations, proposing methodologies that enhance sentiment analysis accuracy without the need for large datasets or extensive annotation efforts. Using RoBERTa, a transformer-based pre-trained model designed for sentiment analysis, and a combination of optimization and data augmentation techniques, we aim to extend the capabilities of sentiment analysis models to perform effectively in data-sparse situations. Our approach addresses the challenges of small datasets and contributes to the broader field of sentiment analysis by offering scalable solutions that can be adapted to various domain-specific environments. Our experimentation has achieved significant improvements in prediction accuracy, demonstrating the feasibility and effectiveness of our approach. By integrating theoretical insights with practical applications, our study sheds light on the untapped potential of small datasets in sentiment analysis. It provides a roadmap for leveraging advanced optimization techniques and innovative data processing strategies to achieve high levels of accuracy, thus opening new avenues for research and application in areas where data is inherently limited.

***Keywords*—Natural Language Processing (NLP), sentiment analysis, small data, imbalance data, transformers**

I. Introduction

In the digital age, sentiment analysis has emerged as a key tool across various industries, allowing organizations to decipher the emotional undertones of large volumes of textual data. This computational technique is particularly beneficial in the transportation sector, where understanding passenger sentiment can directly lead to enhanced service delivery and customer satisfaction. Analyzing commuter feedback not only provides insights into ongoing situations but also improves customer experience by enabling real-time analysis and timely responses. Additionally, it facilitates the identification of emerging trends, allowing for timely adjustments in business strategies, operations, and maintenance. Studies such as Mansor and Abri [1] highlight the importance of real-time data analysis in improving urban transport systems, while research by Randheer *et al.* [2] discusses the impact of receiving travel feedback on commuter behavior and the quality of public transport services.

Despite its potential, the adoption and effective use of sentiment analysis in such domains are hindered by significant challenges. One of the primary obstacles is the scarcity of domain-specific annotated datasets. In fields like transportation, where data can be highly sensitive, compiling and annotating large volumes of text poses both privacy concerns and logistical difficulties. For example, the Personal Data Protection Act (PDPA) in Singapore imposes strict regulations that complicate the data collection process, underscoring the need for methodologies that respect privacy while still extracting valuable insights. References to data sensitivity in the transportation domain can be found in [3], which discusses the challenges of data annotation, and in articles on the GDPR that explore the complexities of data privacy in AI applications. Collecting such labeled data is a notoriously expensive and time-consuming process, as noted by Ein-Dor *et al.* [4].

Most advanced sentiment analysis technologies, including state-of-the-art transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and its variants, rely heavily on large, richly annotated datasets to achieve high accuracy. These models have revolutionized the field of natural language processing due to their ability to understand context and nuance far beyond the capabilities of earlier systems. However, their reliance on extensive datasets renders them less effective in scenarios where such resources are scarce or imbalanced. Our research addresses

Manuscript received February 3, 2025; revised March 4, 2025; accepted May 29, 2025; published September 5, 2025.

this gap by deploying RoBERTa, using hyperparameter tuning, including adjustments to maximum token length, learning rate, and epoch settings, and incorporating innovative data augmentation techniques aimed at enhancing model training under constrained data conditions. Foundational work on BERT by Devlin *et al.* [5], along with subsequent studies by Hassani *et al.* [6] on the limitations imposed by data availability for transformer models, provides a robust framework for our methodologies.

Our systematic exploration and experimentation have led to significant advancements. For example, our modified RoBERTa model, fine-tuned on our specific datasets and further enhanced, achieved an F1-Score of 0.73, a substantial improvement from the 0.49 achieved using the unoptimized pre-trained model. This demonstrates the efficacy of our approach in data-sparse situations. In this study, "data-sparse situations" refer specifically to environments where access to large, annotated datasets is restricted, making it challenging to train data-hungry models like transformers without employing data-efficient techniques. Studies on sentiment analysis performance in limited data scenarios by Tan *et al.* [7] further validate our findings and underscore the potential for significant improvements in accuracy through targeted optimizations.

By integrating theoretical insights with practical applications, our study highlights the untapped potential of small datasets in sentiment analysis. Specifically, our work addresses a multi-class sentiment classification task, categorizing sentiments into five distinct classes based on commuter feedback. It offers a roadmap for leveraging advanced optimization techniques and innovative data processing strategies to achieve high levels of accuracy, opening new avenues for research and application in areas where data are inherently limited. Beyond transportation, sectors such as hospitality and food and beverage can also apply these insights to significantly enhance their services, demonstrating the versatile applicability of our findings. The broader applications of optimized sentiment analysis techniques can be adapted to various service industries [8].

The novelty of this study lies in the integration of synthetic data augmentation using ChatGPT with stratified sampling and fine-tuning strategies to enhance multi-class sentiment classification on small and imbalanced datasets. Unlike previous research, which typically focuses on either data augmentation or model optimization in isolation, our approach uniquely combines both techniques to maximize performance gains. This combined strategy addresses the limitations of small, annotated datasets and imbalanced class distributions, offering a scalable and replicable solution for domain-specific sentiment analysis tasks.

II. Literature Review

In the field of sentiment analysis, numerous methods for analyzing and evaluating sentiment analysis have been offered in the past. The approaches, on the other hand, are constrained in a variety of ways, which will be discussed in this section.

A. Challenges in Sentiment Analysis with Limited Data

Sentiment analysis has attracted extensive research attention, yet it remains challenging under data-scarce and imbalanced conditions. Several studies illustrate these limitations while offering valuable insights that inform our proposed approach.

Zhang *et al*. [9] evaluated the performance of pre-trained transformer models for sentiment analysis in the software engineering domain. While their study demonstrated that transformer-based models outperform traditional baselines, it also revealed significant annotation inconsistencies, with over 18% disagreement among human annotators. This highlights the subjective nature of sentiment labelling, a weakness that can introduce noise and reduce model reliability. Our work acknowledges this challenge and aims to improve model robustness in the presence of such label noise, particularly in small datasets.

Prabhu *et al*. [10] applied BERT-based models for multiclass classification in a logistics domain using over 44,000 labeled samples. The strength of their work lies in showing how transformer models scale with sufficient data. However, they also noted that acquiring such large, labeled corpora is often infeasible for organizations with limited resources. This limitation reinforces the relevance of our research, which is designed to optimize model performance under constraints of dataset size and annotation effort.

In the context of financial sentiment analysis, Omarkhan *et al*. [11] proposed a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model. While their architecture successfully captured both spatial and temporal features of text, its performance degraded significantly in the presence of class imbalance. This weakness exposes a critical shortcoming in deep learning models when faced with skewed distributions. Addressing this, our approach integrates stratified sampling and synthetic augmentation to counteract imbalance and enhance minority class learning.

Similarly, Lu *et al.* [12] conducted a comparative analysis of deep learning models on unstructured medical texts with varying class distributions. Their findings confirm that even advanced models struggle to generalize when minority classes are underrepresented. Although their study is situated in the healthcare domain, the methodological implications are directly relevant: it supports our argument that balanced training data is essential for reliable sentiment classification, regardless of domain.

Stantic *et al*. [13] and Ghanem *et al.* [14] investigated sentiment classification on noisy and imbalanced datasets from social media and news. While both studies provided valuable insight into the challenges of generalizability in real-world data, their models suffered from performance drops due to inconsistencies and skewed distributions. This underscores the importance of preprocessing and augmentation strategies, the two pillars of our methodology, to stabilize training and improve class-level performance.

B. Use of Transformer Models and Transfer Learning

Transformer-based models have significantly advanced sentiment analysis by enabling deep contextual understanding of language. However, their effectiveness often hinges on the availability of large, well-annotated datasets, posing challenges in low-resource scenarios like ours.

Ein-Dor *et al*. [15] conducted an empirical study on active learning applied to BERT for sentiment analysis. Their work demonstrated a strength in reducing the need for exhaustive annotation by intelligently selecting training samples. However, the method's effectiveness was still dependent on having access to a large and diverse seed dataset. This limitation reinforces the relevance of our approach, which seeks to reduce dependence on both seed data and manual annotation through augmentation and stratified sampling.

Serna *et al*. [16] applied the XLM-RoBERTa model to analyze sentiment in the context of sustainable transport, using 2,000 manually annotated reviews. The study successfully adapted transformer models to a domain-specific task with relatively high accuracy, which is a key strength. However, they highlighted that manual annotation was labor-intensive and introduced potential biases. Our methodology addresses this shortcoming by incorporating synthetic augmentation through language generation to increase both dataset size and diversity without manual effort.

Gao *et al*. [17] explored target-dependent sentiment classification using BERT, achieving state-of-the-art results. The strength of their work lies in effectively capturing sentiment in context-specific settings. Yet, they acknowledged the approach's dependence on large, labeled corpora, making it less viable in domains with limited data availability. This aligns directly with our motivation to adapt transformer models for performance in data-sparse environments.

Wu *et al*. [18] developed a BERT-based framework for sentiment classification on Twitter data. Their model achieved strong performance by leveraging a large, annotated corpus of social media text, a clear demonstration of BERT's scalability. However, as with the previous studies, the reliance on substantial labeled data presents a barrier in low-resource applications. Our work builds upon this by introducing an efficient augmentation strategy to compensate for annotation constraints, while still benefiting from the representational power of pre-trained transformers.

C. Optimization Techniques and Their Gaps

Recent studies have attempted to adapt sentiment analysis methods to severely limited datasets, offering partial solutions while revealing persistent gaps that our work aims to address.

Nugumanova *et al*. [19] applied transfer learning for sentiment classification using just 30 Kazakh-language reviews. The strength of their approach lies in demonstrating that transfer learning can be applied even in highly resource-constrained languages and contexts. However, their reliance on a very small sample size and non-expert annotations introduced significant risks of overfitting and label noise. These limitations affected the generalizability of the model, underscoring the challenges of achieving reliable performance in extremely low-data scenarios. Our research builds on this insight by addressing label inconsistency and data scarcity through scalable data augmentation strategies that do not rely on manual annotation.

Tong *et al*. [20] proposed a multimodal deep learning framework incorporating DistilBERT and LSTM for short-text sentiment classification on a small, highly imbalanced dataset. Their model's strength lies in its hybrid architecture, which captures both semantic and sequential features. However, their results were constrained by limited training diversity, resulting in inconsistent performance across classes. This highlights a common weakness in small-data environments: even sophisticated architectures cannot compensate for inadequate data representation. Our methodology tackles this limitation directly by enhancing both data diversity and label balance through stratified sampling and synthetic augmentation.

These studies collectively demonstrate the potential of transfer learning and advanced model architectures in low-resource settings. However, they also make clear that without strategies to compensate for data scarcity, such as augmentation or refined sampling, performance gains remain limited. Our approach addresses these gaps by combining synthetic data generation with model-level optimization, enabling robust performance without increasing the annotation burden.

Furthermore, as most existing studies rely on large datasets, often comprising tens of thousands of labeled samples, to fine-tune pre-trained transformer models, they implicitly assume access to abundant resources. This assumption is problematic for specialized domains or low-resource languages where such datasets are unavailable. We counter this constraint through augmentation techniques that expand training data synthetically, offering both diversity and representativeness without the high cost of manual labeling.

Finally, model optimization techniques such as learning rate scheduling and hyperparameter tuning are essential to maximizing performance under constrained settings. These methods allow pre-trained transformer models to adapt more effectively to small datasets, improving generalization without additional data collection. By integrating these strategies, data augmentation, stratified sampling, and parameter optimization, our work provides a practical and scalable solution for enhancing sentiment analysis in resource-limited environments.

III. Proposed Methodology

This section outlines the methodology employed to enhance sentiment analysis accuracy in the context of limited, annotated, domain-specific datasets. Our approach began with the collection of a manually annotated dataset from domain users, which served as the ground truth. We then cleaned and preprocessed the dataset to prepare it for model evaluation.

We initially evaluated several recent pre-trained transformer models for sentiment analysis using transfer learning without fine-tuning. This step allowed us to identify the most promising model to establish a performance baseline. Subsequently, we fine-tuned the selected model and evaluated its accuracy on the ground truth dataset. To address challenges related to dataset scarcity and class imbalance, we explored several optimization strategies in conjunction with fine-tuning. These included data augmentation, such as generating synthetic samples using ChatGPT, and data splitting techniques like stratified sampling [21].

Our methodology, summarized in the process diagram (Fig. 1), followed these steps: (1) collect a manually annotated ground truth dataset; (2) perform data cleaning and preprocessing; (3) evaluate pre-trained models via transfer learning without fine-tuning to select the best-performing model; (4) fine-tune the selected model and assess accuracy; and (5) apply optimization strategies, including data augmentation and stratified sampling to mitigate issues of data scarcity and imbalance.

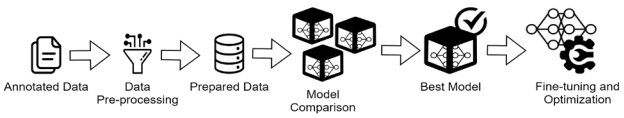

Fig. 1. Optimization processes.

To validate our methodology under operational conditions, we replicated the process using an open-source dataset that is dedicated to reviewing transportation services. This aided in confirming the effectiveness and generalizability of our approach. We will go into further detail in the following subsections.

A. Dataset

Our sentiment analysis task is framed as a multi-class classification problem involving five sentiment classes. To validate our methodology, we employed two datasets: a real-world dataset from a public transport system in Singapore and an open-source dataset for testing and validation. The Transportation Feedback Corpus (TFC) was derived from the system's Customer Relationship Management platform and comprises commuter feedback on transportation services collected from various sources such as websites, emails, and WhatsApp. This feedback is stored in a data lake, where operators manually analyze and report on it monthly. As the dataset initially lacked annotations, operators manually labeled the sentiment of each feedback item based on their understanding of the linguistic context and their domain experience.

The TFC dataset was normalized and transformed into a set of 500 samples categorized into five distinct sentiment classes (1–5) for the multi-class classification task. Samples were randomly selected from each class label to reflect an imbalanced distribution, with a skew toward certain sentiment levels, representing the typical challenges encountered in real-world natural language processing tasks [22]. The distribution and class breakdown are shown in Fig. 2.

To further validate our methodology with diverse, real-world datasets, we included two open-source datasets that exhibited a similar imbalanced distribution as the TFC dataset, as illustrated in Fig. 3. The first dataset was the Gold Standard Corpus (GSC), a sustainable transport sentiment dataset previously explored by Serna [16]. This corpus was compiled from User-Generated Content (UGC) on TripAdvisor, focusing on various modes of transport in Croatia. It includes reviews collected between 2007 and 2020, totaling 117,458 sentences, each rated on a 1–5 star sentiment scale in a multi-class classification format. The second dataset consisted of customer reviews [23] extracted from Amazon product reviews, featuring detailed sentiment ratings from customers based on their post-purchase experiences, also on a 1–5 scale.

All datasets were fully anonymized prior to their use in this research to ensure that no personally identifiable information was included. The data collection and annotation processes adhered to relevant data privacy regulations, including Singapore's Personal Data Protection Act (PDPA). As the real-world dataset was collected as part of operational feedback management and anonymized before analysis, explicit user consent was not required. No ethical concerns were identified in the use of the open-source datasets employed for validation.

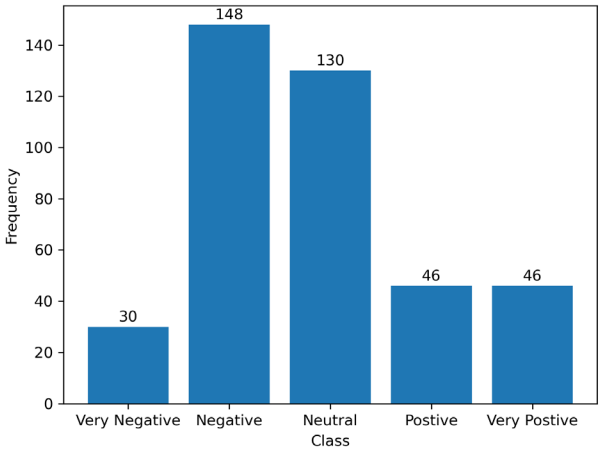

Fig. 2. Class distribution graph for TFC datasets.

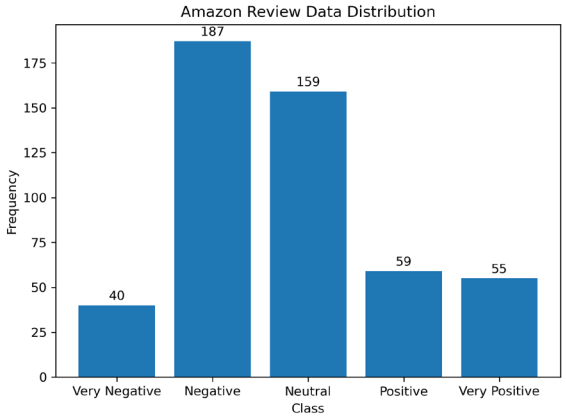

Fig. 3. Class distribution graph for GSC and Amazon Review datasets.

B. Data Preprocessing

Data preprocessing is a crucial step in transforming raw text into a format that pre-trained models can effectively interpret [24]. The preprocessing pipeline was carefully designed to align with the requirements of the pre-trained BERT-based model [25]. This process involved removing extraneous elements such as unnecessary line breaks, URLs, and non-English words, which could introduce noise into the model's input [26]. However, original casing and special characters were retained, as these features can provide important contextual cues that help the transformer model capture nuanced sentiment expressions. BERT-based models typically perform best with minimal preprocessing [18], given that their architecture is designed to process raw text with rich linguistic variation. This preprocessing step ensures the dataset is optimized for RoBERTa's powerful language parsing capabilities, which depend on high-quality input to operate effectively.

C. Pre-trained Model Selection

Given the state-of-the-art performance of BERT models in sentiment analysis tasks [18, 27], our research focused on BERT-based architectures. Their ability to capture contextual information and understand nuanced meanings in text [5], combined with pre-training on large corpora, makes them highly suitable for accurate sentiment analysis via transfer learning, consistently outperforming traditional approaches [28].

To establish a baseline, we first evaluated the dataset using various pre-trained BERT-based models without fine-tuning, as transfer learning is particularly effective for limited datasets [29]. This allowed us to assess their out-of-the-box performance and select the most promising candidate for further optimization.

The pre-trained models considered were:

- **BERT**: The original bidirectional Transformer model developed by Google [5].
- **DistilBERT**: A lighter, more efficient version of BERT [30].
- **XLNet**: An autoregressive Transformer designed to capture longer-range dependencies [31].
- **RoBERTa**: A robustly optimized version of BERT trained on a larger and more diverse corpus [25]. Two variants of RoBERTa were used: **RoBERTa-21**, trained on 58 million Twitter records, and **RoBERTa-22**, trained on 124 million records.

To leverage transfer learning effectively, we selected models that had been pre-trained on sentiment analysis tasks and made available through the Hugging Face open-source platform, as models typically perform better when applied to tasks similar to those they were originally trained on [32].

Table I below lists the pre-trained models used along with their relevant properties.

TABLE I. MODEL TRAINING DETAILS

Model	Max Token	Params	Data	Records	Dataset
RoBERTa-21	512	125M	Twitter	58 M	Wiki, books
RoBERTa-22	512	125M	Twitter	124 M	Wiki, books
BERT	512	110M	Reviews	150 K	Wiki, books
DistilBERT	512	135M	Twitter	1.84 K	Twitter
XLNet	512	117M	Unstructured	Unknown	Wiki, books

D. Fine Tuning and Optimisation

Fine-tuning is the process of taking a pre-trained large language model and further training it on a specific task and dataset to adapt its learned representations [33]. This involves updating the model's parameters for the new domain-specific task while retaining previously learned features and patterns from pre-training. According to Serna *et al*. [16], fine-tuning a pre-trained language model on a small amount of annotated data is an effective strategy for automatically processing large volumes of opinions related to transport. The study also found that models fine-tuned on relevant datasets consistently outperformed those that were not, across all six datasets evaluated.

To address these challenges and further enhance model performance, we employed several optimization techniques in conjunction with fine-tuning:

1. **Data Augmentation with ChatGPT**: We utilized ChatGPT's language generation capabilities to synthesize realistic text samples based on our dataset, effectively expanding the training set and balancing the distribution of sentiment labels. This approach enriched linguistic diversity, improving the model's robustness and generalization.
2. **Stratified Sampling**: To ensure representative subsets for training and validation, we applied stratified sampling to the preprocessed data. This method preserved the proportion of samples in each class, maintaining the original class distribution and reducing bias toward majority classes.
3. **Hyperparameter Optimization**: We optimized key model hyperparameters, such as learning rate, batch size, and number of training epochs, to improve accuracy and generalization.

The integration of fine-tuning with these optimization strategies was designed to enhance sentiment analysis accuracy under the constraints of limited and imbalanced transportation data. The following sections present the specific impacts of these strategies, supported by figures, tables, and discussions on performance improvements and the broader implications of our findings.

IV. EXPERIMENTS AND RESULTS

In this section, we detail the experiments done on selecting the best-performing pre-trained model without fine-tuning and then following the fine-tuning and optimization techniques to further enhance the accuracy to

address the limitations posed by the dataset's size and skewed distribution.

A. Dataset Splitting and Experimental Setup

All experiments in this study were conducted on a workstation equipped with an Intel Core Ultra 9 185H processor (2.50 GHz), 32 GB RAM, and an NVIDIA GeForce RTX 4070 Laptop GPU with 8 GB dedicated VRAM. The model fine-tuning and optimization processes were implemented using the Pytorch deep learning framework and the Hugging Face Transformers library.

For our model performance comparison, we will use the F1-Scoring system. The F1-Score is a performance metric used in machine learning and information retrieval to evaluate the accuracy of a classification model. It is a harmonic average of the model's precision and recall and provides a balanced measure of the model's performance [34]. We are evaluating a multiclass classification task using a weighted F1-Score with the given formula where the weighted F1-Score is the harmonic mean of weighted precision and recall:

$$\text{Weighted F1} - \text{Score} = 2 \cdot \frac{\text{Weighted Precision} \cdot \text{Weighted Recall}}{\text{Weighted Precision} + \text{Weighted Recall}} \quad (1)$$

This is derived from Weighted precision calculated by taking the average of the precision for each class, weighted by the number of true instances in each class. The formula is:

$$\text{Weighted Precision} = \frac{\sum_{i=1}^{n} w_i \cdot TP_i}{\sum_{i=1}^{n} w_i \cdot (TP_i + FP_i)} \quad (2)$$

The weighted recall is the average recall per class, weighted by the number of true instances for each class. It is calculated as:

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{n} w_i \cdot TP_i}{\sum_{i=1}^{n} w_i \cdot (TP_i + FN_i)} \quad (3)$$

where *wi*: The weight associated with class i. This reflects the relative importance of each class in the dataset. Higher weights can be assigned to classes that are more significant or less represented. True Positives (TPi): These are instances where the model correctly predicted the positive class (i.e., it predicted "yes" when the actual label was also "yes") for a specific class *i*. False Positives (FPi): These are instances where the model incorrectly predicted the positive class (i.e., it predicted "yes" when the actual label was "no") for a specific class *i*.

The weighted F1-Score is particularly useful when dealing with imbalanced datasets as it provides a more balanced measure of the model's performance across all classes, rather than being dominated by the majority class [35].

We split the TFC dataset as mentioned in the previous section by the 80:20 rule where 80% of the data is the training set and 20% of the data is a test set as shown in Fig. 4.

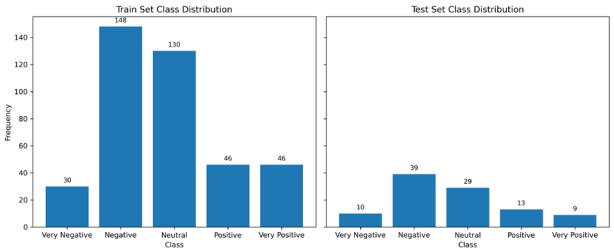

Fig. 4. Class distribution graph for selected datasets.

B. Baseline Model Evaluation

After setting up the dataset and experimental setup, we trained the selected pre-trained models that we selected as stated in the previous section (Table I) without any optimization and fine-tuning to see the accuracy of transfer learning on domain data. We then selected the pre-trained model RoBERTa by hugging face [36] due to it being the best performing of the rest, as shown in Fig. 5. This choice also makes sense since it was trained on 124M tweets from January 2018 to December 2021 and fine-tuned for sentiment analysis with the TweetEval benchmark. This model has been shown to achieve state-of-the-art results on several benchmark NLP tasks and is particularly adept at understanding and generating nuanced textual interpretations, an essential feature for analyzing the varied and complex sentiments typically found in customer feedback data [37].

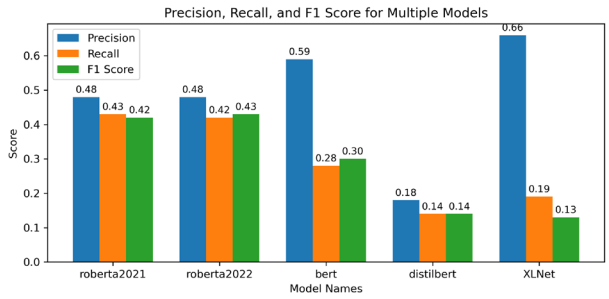

Fig. 5. Comparison of performance across different pre-trained models.

C. Fine-Tuning and Optimization Results

Based on the pre-trained model selected, we then proceeded to experiment while fine-tuning the model using optimization techniques as mentioned below to find the best optimization methods.

For our hyperparameters, we set it as follows, which provided the best configuration of the model. Increasing the batch size reduced training time but led to a slight decrease in accuracy and F1-Score. The best batch size was found to be 16, outperforming batch sizes of 8 and 32 [38]. Regarding the maximum sequence length, longer lengths increased training time. The optimal value was 256, surpassing 200 and 512 [39]. For the learning rate and optimizer, the best learning rate was 1e−6 with the Adam optimizer, which performed well. Due to the low learning rate, we also increased our training epoch to 40.

D. Experiment 1—Stratified Splitting

The first experiment is to use a splitting technique called stratification as compared to random splitting [21]. Stratification is a crucial technique used in data splitting for machine learning models, especially when dealing with imbalanced data distribution.

The key idea behind stratification is to ensure that the training and testing sets have a similar distribution of the target variable or other key features. In the case of imbalanced data, where one class (e.g., the minority class) is significantly underrepresented compared to the other class(es), stratification helps maintain the relative proportions of the classes in the training and testing sets. This is important because it prevents the model from being biased towards the majority class and ensures that it can learn to accurately predict both the majority and minority classes [40]. The data distribution after stratification with the 80:20 rule is shown in Fig. 6.

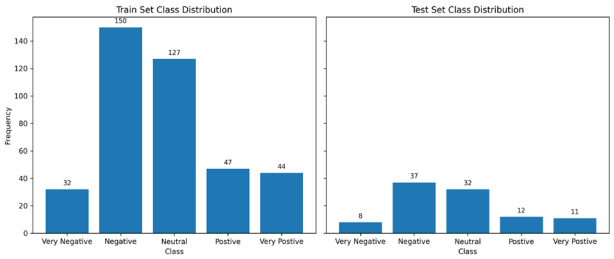

Fig. 6. Stratified split of train and test data.

E. Experiment 2—Data Augmentation Using GPT4 Model

In the second experiment, we focused on using data augmentation with recent popular Large Language Models (LLMs) ChatGPT with model GPT4 with the help of using prompting to synthetically generate more data samples [1]. We employed ChatGPT's language generation capabilities to create realistic text samples from our dataset, effectively enlarging the training data and balancing the distribution of sentiment labels. For each class, we generated an additional 50 samples using ChatGPT's commands for classes 1, 4, and 5 respectively that had the lowest number of samples. Fig. 7 shows the input samples from the dataset and Fig. 8 shows the generated output from the ChatGPT GPT-4 model. This approach not only augments the quantity of training data but also enriches the diversity of linguistic expressions, which is critical for improving the robustness of the pre-trained model when finetuning it.

Text	Label
I would to like to express my extreme displeasure on the staff attitude of {person}. My friend ...	1
Dear Sir/Mdm I am writing in with regards to your staff at the station at {location} on ...	2
Hello I am resident at {location} the lift for the disabled people can not be used for at at least one ...	3
Before {location} - Compliment to Staff Cleaner (Making my day and a pleasant Journey!) incident happen at about to hrs ...	4
Dear Madame/Sir, I would like to thank the staff at {location}, namely	5

Fig. 7. Original data: Input text samples before augmentation.

Text	Label
Commuting by your trains has become a daily challenge. They're either too late or too packed, and the cleanliness is a joke.	1
Why is there never any working air conditioning on the bus during a heatwave? It's unbearable and unsafe for us commuters.	1
I appreciate the cleanliness of the metro stations lately. It makes traveling much more enjoyable.	4
I love the new express service. It cuts down my travel time significantly.	4
Riding the new express line for the first time was a game-changer for me. It cut my commute in half! Honestly, couldn't be happier with this efficient service.	5
The collaboration with local artists to decorate the stations is brilliant. It not only beautifies the space but also supports the local art community.	5

Fig. 8. Synthetic text samples generated with ChatGPT GPT4 model.

F. Experiment 3—Combined Strategy and Results

This experiment evaluated the combined impact of data handling and augmentation strategies on sentiment classification across three datasets: SMRT, Transport Feedback Corpus (TFC), and Amazon Reviews (AMZ). A consistent pipeline was applied for data preparation, model fine-tuning, and optimization. The experiments were conducted in an environment running CUDA 12.4, with peak GPU memory usage of approximately 3835 MB. The model was fine-tuned for 40 epochs. Training took an average of 473.98 seconds, with a throughput of 45.57 samples per second. Inference was completed in 0.57 seconds, achieving 175.95 samples per second.

Figs. 9 to 11 illustrate performance comparisons across three configurations: Random Split, Stratified Split, and Stratified Split with ChatGPT-based augmentation.

Using random splits as a baseline, stratified splitting produced consistent improvements in performance. For instance, on the SMRT dataset, F1-Score improved from 0.56 (Random Split) to 0.68 (Stratified Split), as shown in Fig. 11. Similar gains were observed in the Transport dataset (from 0.62 to 0.66) and the AMZ dataset (from 0.64 to 0.67). Precision and recall also showed corresponding improvements (Figs. 9 and 10).

Further gains were achieved through ChatGPT-based augmentation, which expanded the training data with synthetic but contextually relevant samples. This led to significant increases in all performance metrics. F1-Score reached 0.73 for SMRT, 0.73 for Transport, and 0.73 for AMZ (Fig. 11), confirming the effectiveness of this strategy. Precision also improved notably, with Transport reaching 0.76 and AMZ matching that value (Fig. 9).

The results validate the benefits of combining stratified sampling with prompt-based augmentation, particularly under data-limited and imbalanced scenarios. This approach enhances model robustness and generalizability, offering a scalable solution for domain-specific sentiment analysis.

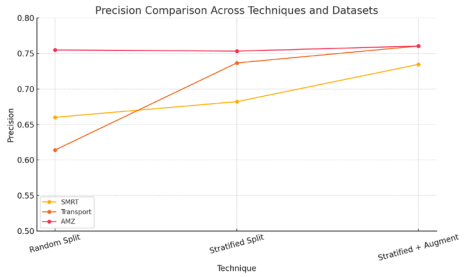

Fig. 9. Precision across all three datasets and techniques.

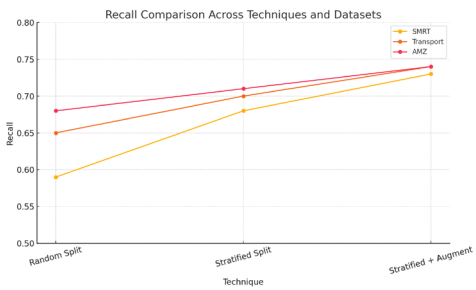

Fig. 10. Recall across all three datasets and techniques.

Fig. 11. F1-Score across all three datasets and techniques.

Fig. 12 presents the confusion matrix for the SMRT dataset using stratified split combined with ChatGPT-based augmentation, our most optimized configuration for data-scarce conditions. The results show that the model handles sentiment classification effectively, with most predictions aligning along the diagonal. Misclassifications primarily occur between adjacent sentiment classes (e.g. class 0 misclassified as 1, or class 2 as 1 or 3), indicating difficulty in separating fine-grained sentiment levels under limited data scenarios. However, critical misclassifications between opposing sentiments (e.g. very negative vs. very positive) are rare, demonstrating the model's robustness in preserving sentiment polarity. These patterns validate our approach: the combined use of stratified sampling and prompt-based augmentation improves overall prediction reliability while maintaining precision even in nuanced sentiment contexts. This highlights the feasibility of achieving strong sentiment classification performance without requiring large, fully annotated datasets.

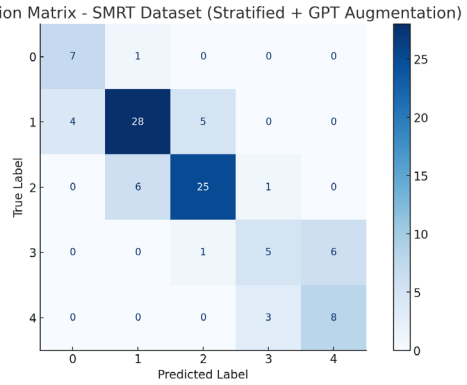

Fig. 12. Confusion Matrix—SMRT (Stratified + GPT Augmentation).

To evaluate the effectiveness of our approach, we compared our results with recent studies in sentiment analysis that addressed small or imbalanced datasets. Table II summarizes the comparison in terms of dataset size, classification type, model used and achieved F1-Score.

TABLE II. COMPARISON WITH RECENT STUDIES

Study	Dataset	Classification Type	Model	F1-Score	Remarks
Omarkhan *et al.* [11]	Financial news, 5000 samples	Binary	CNN + LSTM	0.68	Imbalanced data, small dataset
Prabhu *et al.* [10]	Pickup & Delivery, 44K samples	Multi-Class	BERT + Active Learning	0.80	Large dataset, active learning applied
Tong *et al.* [20]	Short text, 60 samples	Multi-Class	BERT + BI-LSTM	0.61	Small dataset, complex model architecture
Our Study	Transportation CRM, 500 samples	Multi-Class	RoBERTa + Optimization (Fine-tuning + Stratified Sampling + ChatGPT Data Augmentation)	0.73	Small, imbalanced dataset. No manual annotation or active learning overhead

As shown in Table II, our proposed approach achieved an F1-Score of 0.73, which is comparable or superior to recent studies under small dataset constraints, while avoiding the overhead of active learning and manual annotation efforts.

V. CONCLUSION

This research aimed to address the significant challenges posed by limited and imbalanced domain-specific datasets in sentiment analysis tasks. By leveraging advanced transformer-based models such as RoBERTa, we systematically explored optimization strategies to improve sentiment classification accuracy without relying heavily on extensive data annotation efforts. Our experiments demonstrated that fine-tuning the pre-trained RoBERTa model on transportation-specific datasets significantly improved performance over the non-fine-tuned baseline. However, fine-tuning alone was insufficient to fully overcome limitations associated with data scarcity and skewed class distribution.

To mitigate these challenges, we employed a combination of optimization techniques. Stratified sampling during data splitting ensured representative subsets for training and testing, reducing bias toward majority classes and enabling the model to learn all sentiment categories effectively. Data augmentation using ChatGPT's language generation capabilities played a key role in synthesizing diverse and realistic text samples, effectively expanding the training dataset and helping to balance the sentiment label distribution. Together, these strategies enhanced the model's linguistic understanding and robustness, resulting in substantial improvements in F1-Score. Specifically, the F1-Score improved from 0.62 to 0.73 on the TFC dataset, from 0.56 to 0.73 on the SMRT dataset, and from 0.64 to 0.73 on the Amazon Review dataset. Validation using additional open-source transportation datasets further supports the generalizability and real-world applicability of our approach.

By integrating theoretical insights with practical implementation, this study highlights the untapped potential of small datasets in sentiment analysis and

provides a roadmap for leveraging advanced optimization techniques to achieve high accuracy in data-sparse environments. Looking forward, there is significant potential to refine and expand the proposed methodology. Future work could focus on improving prompting strategies to generate more realistic and domain-appropriate synthetic data using ChatGPT and other language models. Further research may also explore the application of these techniques in other data-scarce domains, such as healthcare or finance, and investigate the integration of domain-specific knowledge or semi-supervised and unsupervised learning approaches to further enhance model performance and adaptability.

One limitation of this study lies in the reliance on synthetic data generated by ChatGPT for data augmentation. While this approach effectively increased dataset size and addressed class imbalance, it may also introduce biases or sentiment expressions that do not fully reflect real-world distributions. This issue was not critically examined in the current study. Future work will investigate the quality, representativeness, and potential biases of synthetic data to further improve the reliability and applicability of sentiment analysis models.

In conclusion, this study demonstrates both the feasibility and effectiveness of optimizing sentiment analysis models for limited and imbalanced datasets. It also lays the foundation for developing more robust, scalable solutions in natural language processing. By addressing the constraints of data scarcity, our research contributes to the broader goal of democratizing sentiment analysis across various industries, enabling more informed decision-making and enhanced service delivery.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Isfaque AL Kaderi Tuhin conducted the research, performed the data collection and analysis, and drafted the manuscript. Zhengkui Wang, Li Xiaorong, and Zheng Wei provided critical revisions, contributed to refining the methodology, and offered valuable insights to enhance the interpretation of results. All authors reviewed and approved the final version of the manuscript.

ACKNOWLEDGMENT

We gratefully acknowledge SMRT for their support and contribution to this research.

REFERENCES

[1] T. S. Mansor and R. Abri, "Data-driven optimization of urban traffic using AI and real-time analysis," in *Proc. Int. Conf. Pioneer Innov. Stud.*, Jun. 2023, vol. 1, pp. 507–514

[2] K. Randheer, A. A. AL-Motawa, and P. J. Vijay, "Measuring commuters' perception on service quality using SERVQUAL in public transportation," *International Journal of Marketing Studies*, vol. 3, no. 1, Jan. 2011.

[3] I. Lana *et al*., "From data to actions in intelligent transportation systems: A prescription of functional requirements for model actionability," *Sensors*, vol. 21, no. 4, p. 1121–1121, Feb. 2021.

[4] L. Ein-Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim, *Active Learning for BERT: An Empirical Study*, Association for Computational Linguistics, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Jan. 2019, vol. 1, pp. 4171–4186.

[6] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," arXiv preprint, arXiv:2104.05704, 2022.

[7] K. L. Tan, C. P. Lee, and K. M. Lim, "A survey of sentiment analysis: Approaches, datasets, and future research," *Applied Sciences*, vol. 13, no. 7, 4550, Apr. 2023.

[8] B. Yılmaz and C. Dilmegani, "Sentiment analysis methods in 2024: Overview, pros & cons," *AIMultiple*, 2024.

[9] T. Zhang, B. Xu, F. Thung, S. A. Haryono, D. Lo, and L. Jiang, "Sentiment analysis for software engineering: How far can pre-trained transformer models go?" in *Proc. IEEE Int. Conf. Softw. Anal. Evol. Reeng*., Sep. 2020, pp. 501–512.

[10] S. Prabhu, M. Mohamed, and H. Misra, "Multi-class text classification using BERT-based active learning," arXiv preprint, arXiv:2106.02596, 2021.

[11] M. Omarkhan, G. Kissymova, and I. Akhmetov, "Handling data imbalance using CNN and LSTM in financial news sentiment analysis," in *Proc. 2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, 2021, pp. 1–8.

[12] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, no. 1, 181, Jul. 2022.

[13] B. Stantic, R. Mandal, and E. Chen, *Target Sentiment and Target Analysis Technical Report*, Nathan, QLD, Australia: Griffith Univ, 2020.

[14] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles," *ACM Trans. Internet Technol*., vol. 20, no. 2, 15, 2019.

[15] L. Ein-Dor, A. Halfon, A. Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim, "Active learning for BERT: An empirical study," in *Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Nov. 2020, pp. 7949–7962.

[16] A. Serna, A. Soroa, and R. Agerri, "Applying deep learning techniques for sentiment analysis to assess sustainable transport," *Sustainability*, vol. 13, no. 4, 2397, Feb. 2021.

[17] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, Jan. 2019.

[18] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based Bert model in sentiment analysis," arXiv preprint, arXiv:2403.08217, 2024.

[19] A. Nugumanova, Y. Baiburin, and Y. Alimzhanov, "Sentiment analysis of reviews in Kazakh with transfer learning techniques," in *Proc. 2022 International Conference on Smart Information Systems and Technologies (SIST)*, Apr. 2022.

[20] J. Tong, Z. Wang, and X. Rui, "A multimodel-based deep learning framework for short text multiclass classification with the imbalanced and extremely small data set," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, Oct. 2022.

[21] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," *Lecture Notes in Computer Science*, pp. 145–158, Jan. 2011.

[22] T. Cai and X. Zhang, "Imbalanced text sentiment classification based on multi-channel BLTCN-BLSTM self-attention," *Sensors*, vol. 23, no. 4, 2257, Feb 2023.

[23] M. O. Peter. (2024). Sentiment analysis: Amazon product reviews. [Online]. Available: https://www.kaggle.com/datasets/miriamodeyianypeter/sentiment-analysis-amazon-product-reviews

[24] A. Joby. (2021). What is data preprocessing? 4 crucial steps to do it right. [Online]. Available: https://learn.g2.com/data-preprocessing

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized Bert pretraining approach," arXiv preprint, arXiv:1907.11692, 2019.

[26] P. Chauhan, N. Sharma, and G. Sikka, "On the importance of pre-processing in small-scale analyses of Twitter: A case study of the 2019 Indian general election," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 19219–19258, Jul 2023.

[27] H. Batra, N. S. Punn, S. K. Sonbhadra, and S. Agarwal, "BERT-based sentiment analysis: A software engineering perspective," *Adv. Intell. Syst. Comput.*, pp. 138–148, Jan 2021.

[28] F. D. Souza and J. Baptista, "BERT for sentiment analysis: Pre-trained and fine-tuned alternatives," *Adv. Intell. Syst. Comput.*, pp. 209–218, Jan. 2022.

[29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[30] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of Bert: smaller, faster, cheaper and lighter," arXiv preprint, arXiv:1910.01108, 2019.

[31] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 5753–5763, 2019.

[32] O. Coban, M. Yaganoglu, and F. Bozkurt, "Domain effect investigation for BERT models fine-tuned on different text categorization tasks," *Arabian Journal for Science and Engineering*, vol. 49, no. 3, pp. 3685–3702, Jul. 2023.

[33] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. the 56th Annual Meeting of the Association for Computational Linguistics*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339.

[34] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall, and f-score, with implication for evaluation," *Adv. Inf. Retr.*, pp. 345–359, Jan. 2005.

[35] D. Harbecke, Y. Chen, L. Hennig, and C. Alt, "Why only micro-F1? class weighting of measures for relation classification," in *Proc. NLP Power! The First Workshop on Efficient Benchmarking in NLP*, May 2022, pp. 32–41.

[36] CardiffNLP. (2022). Twitter RoBERTa base sentiment model. [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[37] L. Gao, L. Zhang, L. Zhang, and J. Huang, "RSVN: A RoBERTa sentence vector normalization scheme for short texts to extract semantic information," *Applied Sciences*, vol. 12, no. 21, 11278, Nov 2022.

[38] A. Lheureux. (2022). How to maximize GPU utilization by finding the right batch size. [Online]. Available: https://blog.paperspace.com/how-to-maximize-GPU-utilization-by-finding-the-right-batch-size/

[39] M. Bilal and A. A. Almazroi, "Effectiveness of fine-tuned Bert model in the classification of helpful and unhelpful online customer reviews," *Electronic Commerce Research*, vol. 23, no. 4, pp. 2737–2757, Apr 2022.

[40] D. Dablain, B. Krawczyk, and N. Chawla, "Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning," *Discover Data*, vol. 2, no. 1, Apr. 2024.