A Comparative Study of Large Language Models and Traditional Sentiment Analysis Methods for Sentiment Analysis under Data Imbalance Scenario

Yi Li

Department of Computer Science, Tianjin University of Technology, Tianjin, China

Abstract: Abstract Sentiment analysis is crucial for extracting insights from user-generated text, but its real-world application is often hindered by the pervasive challenge of data imbalance, where majority sentiment classes dominate. This study presents a controlled empirical comparison between traditional sentiment analysis methods (including dictionary-based approaches, classical machine learning models such as Decision Tree, Random Forest, SVM, and a deep learning model LSTM) and a Large Language Model (DeepSeek API with prompt engineering) for sentiment analysis on highly imbalanced datasets. Experiments were conducted on the Twitter US Airline Sentiment dataset with three constructed imbalance ratios (65%:25%:10%, 80%:15%:5%, and 90%:8%:2%) and an additional general sentiment analysis dataset with a fixed 65%:25%:10% ratio, using class-specific precision, recall, F1-score, and Macro-average F1-score as primary metrics. The results reveal a dramatic divergence in model robustness. Traditional methods, despite achieving reasonable overall accuracy in milder scenarios, showed significant limitations in recognizing minority sentiment classes under severe imbalance. In contrast, the LLM consistently achieved the highest Macro-average F1-scores across all Twitter imbalance scenarios (0.800, 0.743, and 0.720), maintained strong minority class performance (e.g., Positive class F1-score of 0.62 and recall of 88% under extreme 90:8:2 imbalance, compared to the best traditional model's F1 of 0.52 and LSTM recall of 20%), and demonstrated superior cross-dataset generalizability (Macro F1 of 0.800 on Twitter and 0.660 on the general dataset). We conclude that the inherent prior knowledge and superior contextual understanding of LLMs, activated through simple prompt engineering, confer a significant advantage over traditional models that learn solely from the imbalanced training data. Our findings strongly suggest that LLMs offer a more robust and effective solution for sentiment analysis in realistic, imbalanced scenarios.

Keywords: Sentiment Analysis; Data Imbalance; Large Language Models; Traditional Sentiment Analysis Methods; Comparative Study

1. Introduction

Sentiment analysis, a fundamental NLP task[1], plays a vital role in extracting opinions and emotions from text, with wide applications in social media monitoring and market research. While traditional methods[2] (including dictionary-based approaches, classical machine learning models, and deep learning models) have long dominated this field, the emergence of Large Language Models (LLMs)[3-7] epresents a paradigm shift, offering superior contextual understanding through pre-training on massive corpora.

However, most studies rely on balanced datasets, failing to explore the class imbalance prevalent in real-world data. In practical scenarios like product reviews, sentiment distributions are often skewed, leading models to favor majority classes at the expense of minority classes.

This prevalent challenge motivates a critical investigation into whether the emergent capabilities of LLMs provide a decisive advantage in imbalanced scenarios. LLMs possess a key potential strength: their performance is not solely derived from the limited, skewed task data. Instead, they can leverage vast, world knowledge and nuanced semantic understanding acquired during pre-training. This may allow them to better recognize subtle cues, sarcasm, and context for minority sentiment classes-precisely where traditional models, which must infer patterns primarily from the imbalanced training distribution, often fail. A direct, empirical comparison is therefore essential to validate this hypothesis and to provide practitioners with clear, evidence-based guidance on model selection when deploying sentiment analysis systems in real-world, skewed data environments.

Therefore, to investigate whether Large Language Models hold a decisive advantage over traditional methods in imbalanced sentiment analysis, this study makes the following contributions:

It designs and conducts a controlled, empirical comparison between a suite of traditional sentiment analysis methods and a state-of-the-art LLM across multiple, artificially constructed data imbalance scenarios.

It systematically evaluates the robustness of each model, with a focused analysis on their capability to recognize minority sentiment classes.

It provides a theoretical interpretation of the LLM's superior performance based on the concept of prior knowledge activation and contextual understanding, offering insights beyond mere empirical results.

Our findings indicate that LLMs offer more reliable and comprehensive discriminative capability in such scenarios.

2. Related Work

2.1 Sentiment Analysis Evolution

The evolution of sentiment analysis has progressed through distinct phases. Initial lexicon-based methods relied on sentiment dictionaries. Subsequently, machine learning approaches[2] treated it as a classification task, heavily dependent on feature engineering. The paradigm then shifted towards deep learning models like LSTMs and CNNs, which used word embeddings to automatically learn features. This progression culminates in the current era of powerful Large Language Models[3][4] setting the context for our performance comparison under data imbalance.

2.2 Data Imbalance in Sentiment Analysis

Data imbalance severely impacts sentiment analysis in real-world applications across multiple domains. Social media platforms exhibit extreme skewness in airline service discussions, public responses to economic news, and low-resource language content. Product review systems show dominant positive or negative opinion distributions, while domain-specific applications such as travel reviews feature significantly more positive samples. Various techniques have been proposed to address data imbalance, including oversampling methods like SMOTE [8-10] and loss function modifications like focal loss[9]. Even benchmark datasets demonstrate bias when processed, and federated learning environments with non-IID data further exacerbate imbalance issues[11].

2.3 LLM-based Approaches

The emergence of Large Language Models (LLMs) such as BERT[4] and GPT-3[3] has introduced new paradigms for sentiment analysis. Key strategies include in-context learning (ICL) [3][12] which leverages pre-trained knowledge for few-shot classification but demonstrates high sensitivity to demonstration quality; parameter-efficient fine-tuning (PEFT), which offers a balance between performance and computational cost; and fine-tuning with imbalance-aware loss functions, which enhances minority class recognition at the expense of substantial computational resources.

While traditional methods struggle with issues such as feature space overfitting, a systematic comparison between conventional imbalance mitigation techniques and LLM-based approaches across diverse sentiment analysis scenarios remains absent from the literature. This gap underscores the critical need for the present comparative study.

3. Evaluation for Imbalanced Sentiment Recognition

3.1 Dataset and Imbalance Introduction

3.1.1 Datasets description and selection rationale

This study utilizes two publicly accessible sentiment analysis datasets from Kaggle to ensure a robust and generalizable comparison.

The "Twitter US Airline Sentiment Dataset" consists of tweets regarding major U.S. airlines, with raw text as input and ternary sentiment labels (negative, neutral, positive) as output. It represents a real-world social media scenario with inherent sentiment imbalance, characterized by short, informal, and often noisy text. The "General Sentiment Analysis Dataset" contains user reviews from various domains such as products and services, with structured and lengthier text as input and the same ternary sentiment labels. This dataset allows us to examine whether model performance trends under imbalance are consistent across different text styles and domains.

3.1.2 Construction of imbalance scenarios

To systematically evaluate model robustness, we constructed controlled imbalance scenarios from a sample of 10,000 instances from the Twitter US Airline Sentiment dataset. Three distinct distributions were created to simulate varying levels of imbalance severity, as summarized in Table 1. For the general sentiment dataset, a fixed 65%:25%:10% imbalance ratio was maintained to facilitate a fair cross-dataset comparison.

The following bar chart (Figure 1) illustrates the three imbalance distributions, visually emphasizing the increasing dominance of the negative class and the diminishing proportion of neutral and positive sentiments.

Table 1. Constructed Imbalance Scenarios for the Twitter Dataset

Scenario	Negative	Neutral	Positive	Description
Moderate	65%	25%	10%	Common real-world skew
High	80%	15%	5%	Severe underrepresentation
Extreme	90%	8%	2%	Tests model limits on rare classes

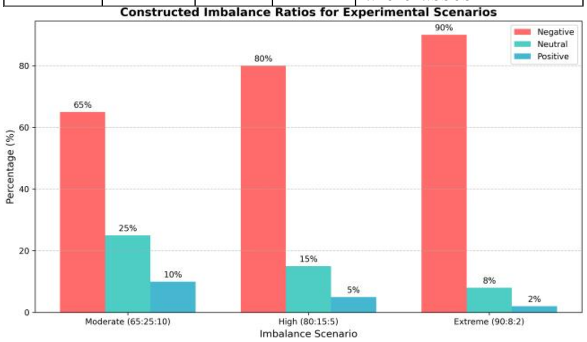


Figure 1. Distributions of Constructed Imbalance Scenarios for the Twitter

3.2 Evaluation Metrics

To comprehensively assess model performance a suite of evaluation metrics that move beyond overall accuracy. Relying solely on accuracy is misleading in such scenarios. Therefore, our evaluation focuses on class-specific metrics derived from the confusion matrix.

The selected metrics are:

Precision: precision=TF/(TF+FP)

precision=TF/(TF+FP) Measures the reliability of a model's positive predictions for a specific class.

Recall: Recall=TP/(TP+FN)

Recall=TP/(TP+FN) Measures the model's ability to identify all relevant instances of a specific class. Recall for the minority class is a primary indicator of model robustness.

F1-score

F1=2×(Precision×Recall)/(Precision+Recall)

F1=2×(Precision×Recall)/(Precision+Recall)

The harmonic mean of precision and recall, providing a balanced summary metric.

Overall Accuracy:

Accuracy=Total Correct Predictions/Total Predictions

Accuracy=Total Correct Predictions/Total Predictions

Reported for context but interpreted cautiously alongside class-specific metrics.

For each sentiment class (Negative, Neutral, Positive), we calculate precision, recall, and F1-score. By examining these metrics across progressively severe imbalance ratios, we can directly observe which models maintain a healthy balance and retain high recall for minority classes.

4. Experiment

4.1 Experimental Setup

4.1.1 Traditional sentiment analysis models

The traditional models evaluated included dictionary-based methods, classical machine learning models (Decision Tree, Random Forest, SVM), and a deep learning model (LSTM). The datasets were split into 70% training and 30% test sets using stratification. Standard text preprocessing was applied, and features were extracted using TF-IDF for classical models.

4.1.2 LLM approach (DeepSeek API)

The LLM-based analysis was performed using the DeepSeek API. The prompt was:" You are a professional sentiment analysis assistant.”

The temperature parameter was set to 0.1 to ensure highly deterministic outputs.

4.2 Model Performance Under Varying Imbalance Ratios

We constructed three scenarios (65%:25:10%, 80%:15:5%, and 90%:8:2%) from the Twitter dataset. The Macro-average F1-score was employed as the primary metric to avoid evaluation bias caused by majority class dominance.

LLM consistently achieved the highest Macro-average F1-scores across all imbalance scenarios. The LLM consistently achieved the highest Macro-average F1-scores across all three scenarios (0.800, 0.743, and 0.720), demonstrating remarkable stability. In contrast, traditional models displayed significant performance degradation under severe conditions. Under extreme imbalance (90:8:2), the LSTM dropped to a Macro F1 of 0.547, and the dictionary-based method fell to 0.327, as shown in Figure 2.

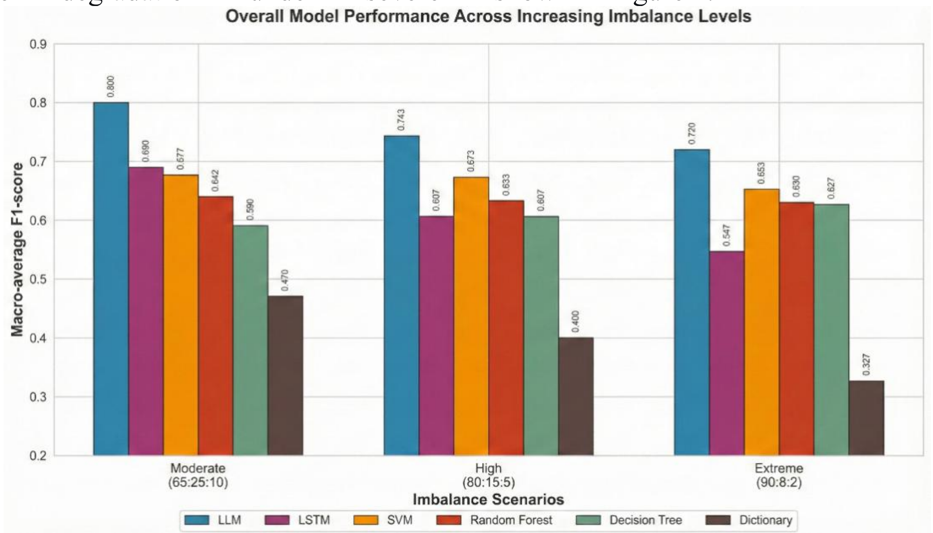
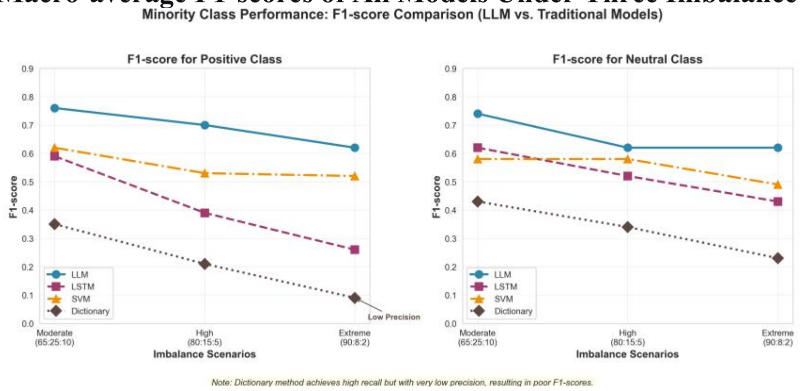


Figure 2. Macro-average F1-scores of All Models Under Three Imbalance Scenarios



Note: Dictionary method achieves high recall but with very low precision, resulting in poor F1-scores.

Figure 3. F1-scores for the Positive Class Across Imbalance Scenarios

LLM maintained robust discriminative capability for minority sentiment classes. For the Positive class, the LLM maintained robust F1-scores of 0.76, 0.70, and 0.62 across the three imbalance scenarios, significantly outperforming traditional models. Under extreme imbalance, the LLM's F1-score (0.62) was 1.2 times higher than the best traditional model, SVM (0.52). The dictionary method achieved high recall for the Positive class but suffered from exceptionally low precision, resulting in poor F1-scores. The LLM balanced high recall with strong precision. A similar trend was observed for the Neutral class, as shown in Figure 3.

Traditional models exhibited degraded performance on minority classes as imbalance increased. As imbalance worsened, traditional models showed improved accuracy on the majority class but suffered from degraded performance on minority classes. For instance, in the 90:8:2 scenario, the LSTM model achieved only 20% recall for the positive class, while the LLM maintained 88% recall. The LLM consistently demonstrated superior robustness, achieving the highest or competitive accuracy in all scenarios and maintaining strong performance across all sentiment classes.

The superior performance of LLM stems from its pre-trained semantic and contextual knowledge.The superior performance of the LLM can be attributed to its inherent architectural and training paradigm. Unlike traditional models that learn task-specific features directly from the limited and skewed training data, the LLM leverages vast, world knowledge acquired during pre-training on diverse and generally balanced corpora. This pre-existing knowledge base, activated through prompt engineering, allows the LLM to make inferences based on deep semantic understanding and contextual reasoning, effectively circumventing the bias towards majority classes.

4.3 Cross-Dataset Validation

To verify generalizability, we evaluated all models on an additional sentiment analysis dataset containing diverse user reviews, maintaining an imbalance ratio of 65%:25%:10%.

LLM demonstrated strong generalizability across datasets with different text styles. The LLM achieved the highest Macro-average F1-score on both datasets (Twitter: 0.800, General: 0.660), demonstrating its superior and consistent performance across different domains. In contrast, traditional models exhibited greater performance variance. The dictionary-based method showed the most pronounced inconsistency, underscoring its heavy reliance on lexical matches that may not transfer well between domains, as shown in Figure 4.

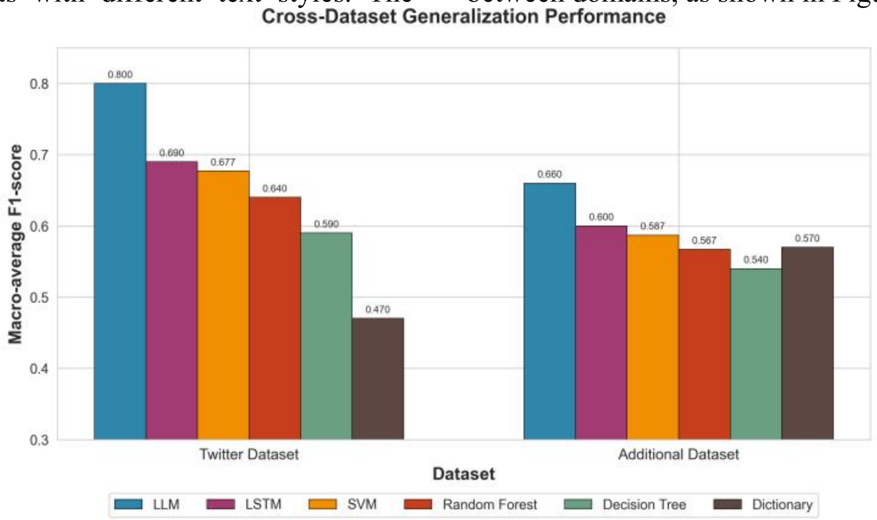


Figure 4. Macro-Average F1-Scores on Twitter and General Datasets (65:25:10)

LLM consistently outperformed traditional models in all experimental conditions. The LLM consistently demonstrated superior robustness across all experimental conditions, including cross-dataset validation. The results confirm that the LLM's advantage in handling imbalanced data is a transferable capability rooted in its foundational language understanding.

5. Discussion

5.1 Inferred Failure Mode of Traditional Models: Over-reliance on Lexical Cues

The drop in precision for the dictionary-based method and the declining recall of models like LSTM for minority classes suggest a common failure mode: over-reliance on surface-level keyword matching without contextual integration [2]. The LLM, with its pre-trained understanding of pragmatics and sarcasm, can correctly interpret complex constructs.

5.2 Inferred Strength of the LLM: Compositionality and Negation Handling

The LLM's consistent performance suggests a strong ability to handle semantic compositionality and negation-a known weakness for bag-of-words and even some sequence models. The LLM's ability to parse these complex structures is a key contributor to its robustness.

5.3 The "Robustness" as Prior Knowledge Activation

The LLM's advantage stems not from learning the training distribution, but from activating relevant prior knowledge. In a severely imbalanced setting, the LLM can map known semantic concepts to the correct label, effectively performing a form of few-shot or zero-shot inference that is largely insulated from the training set's bias.

6. Conclusion

This study conducted a rigorous comparison between traditional sentiment analysis methods and a Large Language Model under progressively severe data imbalance scenarios. Our experiments yield a clear conclusion: Large Language Models demonstrate fundamentally superior robustness for sentiment analysis in imbalanced data environments.

The core of this advantage lies in the divergent learning paradigms. Traditional models, confined to learning patterns exclusively from the limited and skewed training distribution, inevitably develop a bias towards the majority class. In stark contrast, the LLM bypasses this limitation by leveraging its vast, pre-existing world knowledge and semantic understanding. Consequently, the LLM maintained high and balanced performance across all sentiment classes, achieving exceptional recall for minority classes even under extreme imbalance.For practitioners, our findings offer decisive guidance: in real-world applications where perfectly balanced data is rare, LLMs present a more reliable and effective solution, potentially reducing the need for complex imbalance mitigation techniques.

7. Limitations and Future Work

Future work could compare LLMs against traditional models fortified with state-of-the-art imbalance-handling algorithms. The computational cost and inference latency of LLMs remain practical constraints; a cost-benefit analysis is a critical next step. Finally, validating these findings on larger, naturally imbalanced datasets from even more diverse domains would strengthen generalizability.

In summary, this research underscores that the paradigm shift brought by LLMs is transformative for challenging real-world NLP tasks like imbalanced sentiment analysis.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, pp. 79-86, 2002.

[3] T. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171-4186, 2019.

[5] W. Zhang et al., "Sentiment Analysis in the Era of Large Language Models: A Reality Check," in Findings of the Association for Computational Linguistics: NAACL 2024, pp. 3893-3909, 2024.

[6] J. Hartmann, J. Schwenzow, and M. Witte, "Sentiment Analysis in the Age of Generative AI," Customer Needs and Solutions, vol. 11, no. 1, pp. 1-14, 2024.

[7] H. Yang, X. Zeng, L. Xu, and T. Liu, "Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey," arXiv preprint arXiv:2406.08068, 2024.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 2980-2988, 2017.

[10] S. Havrylov and I. Titov, "A Survey of Methods for Addressing Class Imbalance in Deep Learning for Natural Language Processing," in Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), pp. 531-542, 2023.

[11] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intelligent Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.

[12] D.-Y. Kim and D.-J. Shin, "Enhancing Imbalanced Sentiment Analysis: A GPT-3-Based Sentence-by-Sentence Generation Approach," Applied Sciences, vol. 14, no. 2, p. 622, 2024.

Appendix A: Detailed Performance Tables

Table A1: Full Results for Twitter Dataset (65:25:10)

Twitter US Airline Sentiment dataset (neg: neu: pos=65%:25%:10%)										
method	accuracy	negative			neutral			positive		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
dictionary based	0.5134	0.9	0.49	0.63	0.44	0.43	0.43	0.22	0.88	0.35
Decision Tree	0.679	0.77	0.79	0.78	0.52	0.47	0.49	0.47	0.55	0.5

Random Forest	0.741	0.87	0.8	0.83	0.52	0.58	0.55	0.47	0.64	0.54
SVM	0.748	0.84	0.83	0.83	0.57	0.58	0.58	0.61	0.63	0.62
LSTM	0.7757	0.84	0.88	0.86	0.65	0.59	0.62	0.63	0.56	0.59
LLM-api	0.8427	0.95	0.85	0.90	0.68	0.82	0.74	0.70	0.82	0.76

Table A2: Full Results for Twitter Dataset (80:15:5)

Twitter US Airline Sentiment dataset (neg: neu: pos=80%:15%:5%)										
method	accuracy	negative			neutral			positive		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
dictionary based	0.503	0.95	0.5	0.65	0.29	0.42	0.34	0.12	0.85	0.21
Decision Tree	0.811	0.91	0.89	0.9	0.41	0.47	0.44	0.48	0.47	0.48
Random Forest	0.8396	0.95	0.88	0.91	0.4	0.58	0.47	0.45	0.6	0.52
SVM	0.8423	0.91	0.92	0.91	0.61	0.55	0.58	0.51	0.55	0.53
LSTM	0.8350	0.89	0.94	0.91	0.57	0.47	0.52	0.49	0.32	0.39
LLM-api	0.8403	0.97	0.86	0.91	0.51	0.79	0.62	0.65	0.76	0.70

Table A3: Full Results for Twitter Dataset (90:8:2)

Twitter US Airline Sentiment dataset(neg:neu:pos=90%:8%:2%)										
method	accuracy	negative			neutral			positive		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
dictionary based	0.4963	0.97	0.5	0.66	0.16	0.42	0.23	0.05	0.83	0.09
Decision Tree	0.8897	0.94	0.95	0.94	0.48	0.38	0.43	0.45	0.59	0.51
Random Forest	0.9203	0.98	0.95	0.96	0.42	0.56	0.48	0.33	0.69	0.45
SVM	0.9063	0.95	0.95	0.95	0.51	0.47	0.49	0.47	0.58	0.52
LSTM	0.9023	0.94	0.96	0.95	0.47	0.40	0.43	0.38	0.20	0.26
LLM-api	0.8597	0.99	0.86	0.92	0.54	0.88	0.62	0.58	0.88	0.62

Table A4: Full Results for Additional Dataset (65:25:10)

Sentiment Analysis Dataset(neg:neu:pos=65%:25%:10%)										
method	accuracy	negative			neutral			positive		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
dictionary based	0.6009	0.91	0.61	0.73	0.5	0.45	0.57	0.27	0.68	0.41
Decision Tree	0.6473	0.76	0.77	0.76	0.46	0.42	0.44	0.38	0.47	0.42
Random Forest	0.6657	0.79	0.76	0.78	0.42	0.43	0.42	0.45	0.56	0.5
SVM	0.6873	0.83	0.78	0.8	0.39	0.45	0.42	0.5	0.58	0.54
LSTM	0.6967	0.80	0.81	0.8	0.48	0.49	0.48	0.58	0.47	0.52
LLM-api	0.7690	0.90	0.73	0.81	0.50	0.71	0.59	0.53	0.65	0.58