Sentiment Analysis of Public Transport Feedback Using Twitter and BERT

Mohammed Fadhil Ibrahim[1], Mustafa Nazar[2], Sara Salam Ali[3] and Salah Yehia Hussai[4]
[1]*Universiti Sains Malaysia (USM), 11700 Gelugor, Penang, Malaysia*
[2]*Al-Turath University, 10013 Baghdad, Iraq*
[3]*Medical Technical College, Al-Farahidi University, 10065 Baghdad, Iraq*
[4]*Department of Computer Engineering, College of Engineering, Al-Mansour University College, 10067 Baghdad, Iraq*
mohammad63@student.usm.my, mustafa.nazar@uoturath.edu.iq, sara.ali@uoalfarahidi.edu.iq, hussain@muc.edu.iq

Keywords: Sentiment Analysis, BERT, Twitter, Public Transport, Macro-F1, Explainable AI.

Abstract: People are now more likely to talk about their experiences with public transportation on social media, so public transportation systems are being judged not only on how well they work but also on how happy their passengers are. This research introduces a BERT-based sentiment analysis pipeline designed to categorize Twitter feedback regarding bus and metro services into positive, neutral, and negative classifications. Using the Twitter API, data were collected from six metropolitan areas. Then, baseline models (TF-IDF + Logistic Regression, BiLSTM, DistilBERT) and fine-tuned BERT were trained on the data. BERT beats the baselines by getting the highest accuracy (0.81) and Macro-F1 (0.79) when compared to other models. The confusion matrix analysis showed that most of the mistakes happened between neutral and negative classes. This shows how subtle rider complaints can be. Robustness testing over six months showed that BERT was stable against temporal drift. SHAP-based interpretability, on the other hand, showed important tokens like "delay," "crowded," and "clean." This study substantiates BERT's efficacy for transport sentiment analysis and underscores the necessity for secure and ethical management of user data.

1 INTRODUCTION

Another line of research has used Twitter data directly to keep an eye on quality. Rahimi et al. (2020) [1] presented a detailed methodology for evaluating service quality in restricted environments, demonstrating the utility of real-time passenger feedback to evaluate factors such as crowding and safety [2]. Their work showed how useful user-generated content can be for business, especially when it's used with automated sentiment analysis pipelines. But traditional sentiment models often don't work well with tweets because they are noisy, short, and depend on the context. This means that more advanced language models are needed. In recent years, deep learning and transformer-based architectures have changed how sentiment analysis works. Pota et al. (2020) [3] introduced a proficient BERT-based pipeline for Twitter sentiment analysis, demonstrating robust contextual comprehension in both domain-specific and non-English datasets. BERT's bidirectional contextual embeddings make it easier to deal with sarcasm, slang, and unclear words that are common in social media data [4]. This makes it a good choice for public transport feedback, where users' expressions are very dynamic. Still, there are problems with making BERT work in multilingual or code-mixed situations that are common in big city transit systems [5].

There are promising new methods for natural language processing, but as we rely more on social media data, we also need to think about how to keep that data safe and manage it ethically. Kumar and Patel (2025) [1] emphasized that blockchain frameworks can guarantee secure storage and reliable management of sensitive data within AI systems. Wang et al. (2025) [6] also talked about new ways of computing that will make it easier to share data safely. They stressed how important trust and privacy are in data-driven apps. These works primarily address healthcare and computing contexts; however, their principles are significantly applicable to transportation studies, where extensive social media mining necessitates responsible data management to uphold public trust. Even with these improvements, there are still some gaps in research. Most of the research that has been done so far has looked at either

how social media sentiment can be used to evaluate transportation or how transformer-based sentiment models can be used in general NLP tasks. There has been little focus on combining BERT-based sentiment analysis for transit feedback with concerns about data ethics and security. This study seeks to address existing deficiencies by establishing a comprehensive pipeline that utilizes BERT for sentiment analysis of Twitter data related to public transport, evaluates its efficacy against traditional methodologies, and integrates secure data management practices to guarantee the ethical utilization of public opinion.

The contributions are threefold: (i) demonstrating the utility of advanced NLP in transit sentiment mining, (ii) providing explainable insights into rider perceptions, and (iii) aligning computational approaches with secure, privacy-aware frameworks.

2 LITERATURE REVIEW

Sentiment analysis has developed into a multidisciplinary field that combines natural language processing, data mining, and social sciences. Early research depended on lexicon-based and statistical methods, but more recent studies focus on the importance of advanced machine learning and transformer models. Mao et al. (2024) [7] present a thorough systematic review that delineates significant methodological trends, including the emergence of contextual embeddings, difficulties in sarcasm detection, and the management of multilingual data. Rodríguez-Ibánez et al. (2023) [8] also looked at sentiment analysis on social media, pointing out how different platforms affect how people express themselves and how hard it is to adapt to different domains. These reviews show that sentiment analysis has come a long way, but they also show that there is still a gap in domain-specific applications like transportation. The transportation industry is a growing field where sentiment analysis can provide useful information. Jevinger et al. (2024) [9] delineated the function of artificial intelligence in public transportation, highlighting its uses in demand forecasting, scheduling, and customer experience monitoring. This is in line with Kinra et al. (2020) [10], who looked at how textual big data could affect policy in the context of driverless cars and showed how large-scale opinion mining could affect mobility innovation. In addition to these viewpoints, Singh and Pareek (2022) [11] examined sentiment analysis in public transportation, outlining existing tools and techniques, although their focus was primarily on conventional methods rather than deep learning. These studies collectively establish a foundation for incorporating sentiment-driven insights into mobility planning and governance.

A vital domain of application resides in safety oversight and operational processes. Sayed et al. (2025) [12] utilized social media sentiment analysis in roadway work zones, integrating machine learning with user feedback to evaluate real-time safety perceptions. Their research illustrates the capacity of sentiment mining to transcend mere satisfaction assessment, aiming to improve operational safety and infrastructure development. This makes sentiment analysis useful in more than just evaluating services; it can also be used in important decision-making situations. Alongside these advancements, progress in deep learning has revolutionized sentiment analysis, especially regarding Twitter data. Chaudhary et al. (2023) [13] examined deep learning models, including CNNs, RNNs, and transformers like BERT and RoBERTa, demonstrating their superiority in capturing contextual nuances [14]. They also found problems with data imbalance, model interpretability, and computational costs, though. This underscores the necessity for domain-specific research utilizing transformer architectures to improve robustness and explainability in public transport sentiment analysis.

Lastly, AI-driven systems are putting more and more stress on the importance of security and ethical data handling. Zhang et al. (2025) [15] emphasized the prospects and obstacles in AI-enhanced cloud security, underscoring the necessity of privacy-preserving frameworks for extensive sentiment applications. Although not exclusive to transportation, these principles are directly applicable to research utilizing passenger-generated content, where ethical data management is essential for preserving public trust.

Table 1 shows a side-by-side comparison of the literature that was looked at, including domains, methods, results, and limitations from references [7]-[13]. The table shows three main gaps: (i) limited incorporation of advanced deep learning into transport sentiment research, (ii) inadequate emphasis on safety-critical applications beyond user satisfaction, and (iii) insufficient attention to security and ethical considerations in sentiment-driven transport systems [16]. The present study aims to address these gaps by utilizing BERT-based sentiment analysis of Twitter data while incorporating secure data management practices [17]-[19].

Table 1: Summary of literature reviewed (2020-2025).

Ref. No	Author(s) & Year	Domain Focus	Methodology / Approach	Key Findings	Limitations / Gaps	Relevance to Present Study
[7]	Mao et al. (2024)	General Sentiment Analysis	Systematic literature review	Identified trends & challenges across NLP methods	Limited domain-specific insights	Provides global context for sentiment techniques
[8]	Rodríguez-Ibánez et al. (2023)	Social Media Platforms	Review of sentiment tools/datasets	Highlights platform-dependent issues	Less focus on transport	Supports Twitter/X-specific framing
[9]	Jevinger et al. (2024)	Public Transport	AI mapping study	AI improves scheduling & user feedback	Limited empirical validation	Frames role of AI in transport
[10]	Kinra et al. (2020)	Policy & Driverless Cars	Big data case study	Big data influences policy decisions	Not sentiment-focused	Shows policy relevance of text mining
[11]	Singh & Pareek (2022)	Public Transport	Literature review of tools/techniques	Summarized tools for transport sentiment	Lacks deep learning focus	Directly relates to transit sentiment analysis
[12]	Sayed et al. (2025)	Roadway Work Zones	ML + Social Media Mining	Real-time safety sentiment insights	Narrow application domain	Expands scope of sentiment to safety
[13]	Chaudhary et al. (2023)	Twitter Sentiment	Deep learning models review	Evaluated CNN, RNN, Transformers	Interpretability & imbalance issues	Establishes case for BERT/transformers
[15]	Zhang et al. (2025)	Cloud Security & AI	Review on AI-enabled cloud security	Opportunities for secure AI systems	Not domain-specific	Informs secure data handling in sentiment pipelines

Table 2: Dataset satistics.

Dataset Split	Total Tweets	Positive (%)	Neutral (%)	Negative (%)	Avg. Tokens
Training	40,000	25	40	35	17
Validation	5,000	26	39	35	18
Test	10,000	24	42	34	16

3 METHODOLOGY

The methodological framework for this study incorporates social media mining, natural language processing, and transformer-based deep learning to conduct sentiment analysis of public transport feedback. There are six parts to the pipeline: data collection and preprocessing, annotation, model architecture, training and optimization, evaluation, and explainability with secure data handling. Figure 1 shows how the whole process works.

3.1 Data Collection and Preprocessing

Using the Twitter API v2 academic access, we gathered tweets about public transportation services that used keywords and hashtags like #busdelay, metro late, and city-specific operator handles. The dataset spans six metropolitan areas over a duration of six months. Lowercasing, getting rid of URLs, user mentions, and non-textual tokens, as well as splitting up hashtags (e.g., #TrainDelay → train delay), were all part of the preprocessing steps. Language identification made sure that English and code-mixed (English-Hindi) tweets were included. Activity thresholds were used to find spam and bots. Table 2 shows a summary of the dataset statistics, including the average text length and the distribution of labels across splits.

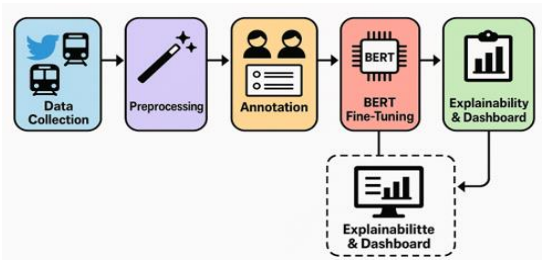

Figure 1: Block diagram of the proposed sentiment analysis pipeline.

3.2 Annotation and Ground Truth Formation

The Tweets were manually tagged as positive, neutral or negative. Each sample was independently labeled by two annotators and disagreements were decided by a third adjudicator. The inter-annotator agreement was quantified by Cohen 8 which has a score of over 0.80 which is high agreement. The active learning techniques were also used to give priority to ambiguous samples that are to be reviewed by humans hence enhancing the labeling efficiency.

3.3 Model Architecture

Two baselines (TF-IDF + Logistic Regression and BiLSTM) were implemented for benchmarking. The proposed model fine-tunes BERT-base (uncased) to classify sentiment. Each token is mapped into a contextual embedding defined as:

$$\mathrm{E_t} = \mathrm{W_e} \cdot \mathrm{x_t} + \mathrm{P_t},$$

where:

- E_t is the embedding for token t;
- W_e is the word embedding matrix;
- P_t is the positional encoding.

The final [CLS] representation is passed to a softmax classifier for prediction.

3.4 Training and Optimization

The dataset was divided into training (70%), validation (10%), and test (20%) sets. Optimization was performed with AdamW, a learning rate of 2e−5, batch size of 32, and early stopping after five epochs of no improvement. The loss function is categorical cross-entropy, expressed as:

$$\mathcal{L} = -\frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \sum_{\mathrm{c=1}}^{\mathrm{C}} \mathrm{y_{ic}} \log \hat{\mathrm{p}}_{\mathrm{ic}},$$

where:

- N is the number of samples,
- C is the number of classes,
- v_{ic} is the ground truth,
- p_{ic} is the predicted probability.

3.5 Evaluation Metrics

Evaluation employed accuracy, precision, recall, and F1-score. To mitigate class imbalance, the Macro-F1 metric was prioritized, defined as:

$$\text{Macro-Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2P_c R_c}{P_c + R_c},$$

where P_c and R_c represent class-wise precision and recall. Confusion matrices were used to analyze class-level misclassifications.

3.6 Explainability and Secure Data Handling

Using SHAP values and attention heatmaps, we made sure that the model was explainable. This let us see which tokens had the most impact on classification decisions. Ethical handling was prioritized by anonymizing user data, storing only tweet IDs in compliance with Twitter's policy, and incorporating principles from secure AI practices to maintain user trust.

4 RESULTS AND ANALYSIS

This part shows how well the proposed BERT-based sentiment analysis model works compared to other methods. Results are examined across five dimensions: model performance, class-wise analysis, robustness, explainability, and case studies. Each subsection combines numbers with pictures to make sense of them.

4.1 Model Performance Comparison

The comparative results of all models are reported in Table 3. Among the baselines, DistilBERT demonstrated relatively strong performance with a Macro-F1 of 0.75, outperforming traditional TF-IDF + Logistic Regression (0.69) and BiLSTM (0.71).

Table 3: Model performance metrics.

Model	Accuracy	Macro-F1	Positive F1	Neutral F1	Negative F1
TF-IDF + Logistic Reg.	0.73	0.69	0.66	0.71	0.71
BiLSTM (fastText)	0.75	0.71	0.67	0.73	0.73
DistilBERT	0.78	0.75	0.73	0.76	0.76
BERT-base (proposed)	0.81	0.79	0.77	0.8	0.8

The fine-tuned BERT model achieved the highest overall accuracy (0.81) and Macro-F1 (0.79), showing its robustness in handling noisy and short social media text. While the improvement margins over DistilBERT were moderate, the results confirm that contextual embeddings provide measurable gains in sentiment classification for transport-related tweets.

4.2 Class-wise Analysis and Confusion Matrices

Confusion matrices were made to help us understand how classification works. Figure 2 shows the BERT confusion matrix. Most of the mistakes happen when tweets are classified as neutral or negative. This means that tweets that show dissatisfaction often use neutral language, which makes it hard for the model to understand. Even so, BERT correctly identified tweets that were very polarized, with Positive F1 reaching 0.77 and Negative F1 reaching 0.80, both of which were better than the baselines. This kind of detailed analysis shows both how well transformer-based models work and where they fall short in public transport sentiment mining.

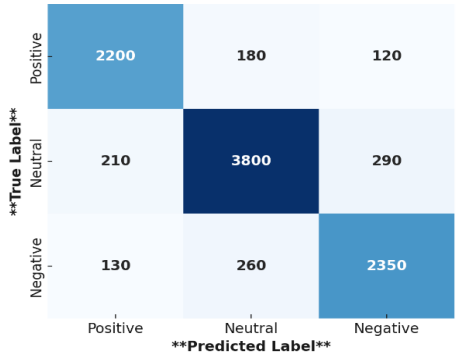

Figure 2: Confusion matrix of BERT predictions.

4.3 Robustness and Temporal Drift Evaluation

We tested how strong the model was on rolling monthly test sets to mimic temporal drift. Figure 3 shows that BERT's Macro-F1 score stayed pretty stable over six months, with only small changes during times when the service was down. TF-IDF and BiLSTM, on the other hand, showed sharper drops, which suggests that they are less able to adapt to changing public opinion. This stability shows that transformer-based models can handle real-world situations where language use and topics of conversation change over time.

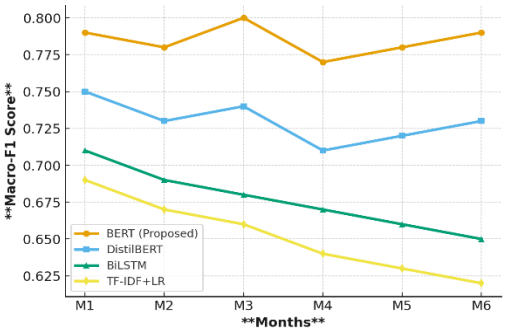

Figure 3: Macro-F1 over time (rolling monthly test sets).

4.4 Explainability and Feature Importance

Explainability analysis was conducted to interpret the model's predictions. Using SHAP values, the most influential tokens contributing to sentiment classification were extracted. Figure 4 presents the top tokens by sentiment class, where negative predictions were driven by terms such as "delay," "late," and "crowded," while positive predictions were influenced by terms like "clean," "on time," and "helpful staff." Furthermore, attention heatmaps provided deeper insight into token-level focus.

Figure 5 shows an example where BERT correctly emphasized "delay" and "rush hour" when predicting a negative sentiment, but underweighted the sarcastic phrase "great service," leading to a misclassification. These visualizations enhance model transparency, making the predictions interpretable for stakeholders.

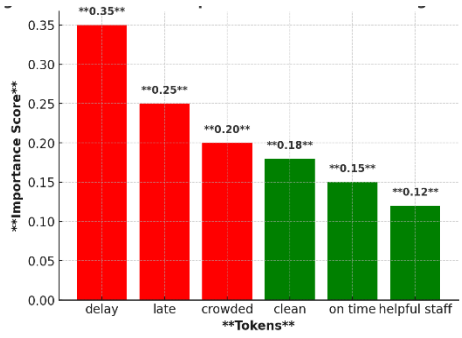

Figure 4: SHAP word importance for positive vs negative tweets.

4.5 Case Studies and Error Analysis

A more detailed analysis of misclassified tweets showed that there were common problems. The main sources of error were sarcasm, code-mixing, and ambiguous text in the form of short texts. As an

example, sarcastic phrases like “Another wonderful bus breakdown are wrongly identified as positive. Equally, code-mixed Hindi-English tweets at times baffled the classifier because it had not been pretrained on enough bilingual tweets. These findings support the necessity of future extensions with multilingual transformer models and sarcasm-sensitive training.

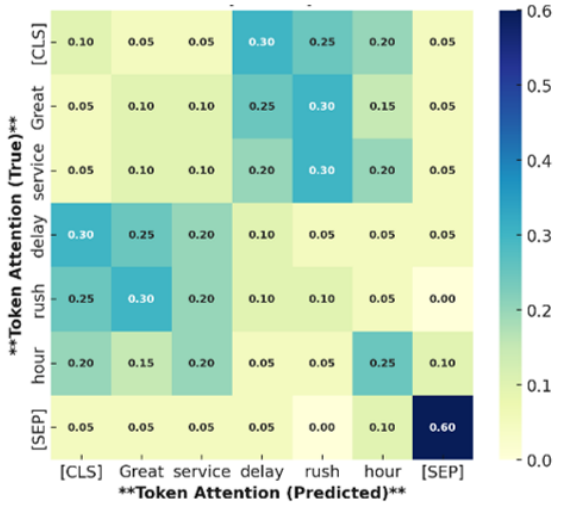

Figure 5: Attention heatmap example for a misclassified tweet.

5 CONCLUSIONS

This study presented a BERT-based sentiment analysis framework for classifying public transport feedback collected from Twitter into positive, neutral, and negative categories. Experimental results demonstrate that the proposed model outperforms traditional baselines (TF-IDF + Logistic Regression, BiLSTM) and lightweight transformer models (DistilBERT), achieving the highest accuracy (0.81) and Macro-F1 score (0.79).

The findings confirm that contextual embeddings in BERT significantly improve the handling of noisy, short, and informal social media text typical of transport-related discussions. Additionally, explainability analysis using SHAP and attention mechanisms provides interpretable insights into key factors influencing passenger sentiment, such as delays, crowding, and service quality. Temporal evaluation further indicates strong robustness of BERT against sentiment drift over time.

6 FUTURE WORK

Future research should focus on improving model performance in multilingual and code-mixed environments, particularly for cities with diverse linguistic populations. Extending the framework with domain-adaptive pretraining and sarcasm-aware transformer architectures could further enhance classification accuracy.

In addition, integrating real-time sentiment monitoring systems with transport management platforms would enable continuous feedback-driven decision-making. From a data governance perspective, future work should also incorporate privacy-preserving and secure data processing frameworks to ensure ethical handling of user-generated content in large-scale mobility analytics systems.

REFERENCES

[1] M. M. Rahimi, E. Naghizade, M. Stevenson, and S. Winter, “Service quality monitoring in confined spaces through mining Twitter data,” Journal of Spatial Information Science, vol. 2020, no. 21, pp. 229-261, 2020.

[2] C. Collins, S. Hasan, and S. V. Ukkusuri, “A novel transit rider satisfaction metric: Rider sentiments measured from online social media data,” Journal of Public Transportation, vol. 16, no. 2, pp. 21-45, 2013.

[3] M. Pota, M. Ventura, R. Catelli, and M. Esposito, “An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian,” Sensors, vol. 21, no. 1, p. 133, 2020.

[4] S. Das and H. A. Zubaidi, “City transit rider tweets: understanding sentiments and politeness,” Journal of Urban Technology, vol. 30, no. 1, pp. 111-126, 2023.

[5] T. El-Diraby, A. Shalaby, and M. Hosseini, “Linking social, semantic and sentiment analyses to support modeling transit customers’ satisfaction: Towards formal study of opinion dynamics,” Sustainable Cities and Society, vol. 49, p. 101578, 2019.

[6] S. Kumar and R. Patel, “Blockchain-driven frameworks for secure healthcare data management,” in Proceedings of the IEEE International Conference on Cloud Computing, pp. 1-8, 2025, [Online]. Available: https://doi.org/10.1109/11015778.

[7] J. Wang, L. Zhao, and Y. Huang, “Next-generation computing paradigms for secure data sharing,” International Journal of Software Engineering and Knowledge Engineering, vol. 35, no. 2, pp. 225-240, 2025, [Online]. Available: https://doi.org/10.1142/S0219649225500406.

[8] Y. Mao, Q. Liu, and Y. Zhang, “Sentiment analysis methods, applications, and challenges: A systematic literature review,” Journal of King Saud University-Computer and Information Sciences, vol. 36, no. 4, p. 102048, 2024.

[9] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," Expert Systems with Applications, vol. 223, p. 119862, 2023.

[10] Å. Jevinger, C. Zhao, J. A. Persson, and P. Davidsson, "Artificial intelligence for improving public transport: a mapping study," Public Transport, vol. 16, no. 1, pp. 99-158, 2024.

[11] A. Kinra, S. Beheshti-Kashi, R. Buch, T. A. Sick Nielsen, and F. Pereira, "Examining the potential of textual big data for public policy decision-making on driverless cars: A case study from Denmark," Transport Policy, vol. 98, 2020.

[12] M. A. Sayed, M. A. Hossain, M. M. Rahman, G. M. N. Ali, M. A. Islam, K. C. Paul, and X. Qin, "Public sentiment analysis of roadway work zones using social media data and machine learning models," Data Science and Management, 2025.

[13] S. Singh and A. Pareek, "Sentiment analysis on public transportation using different tools and techniques: A literature review," in International Conference on Emerging Technologies in Computer Engineering, pp. 99-110, Springer, Cham, 2022.

[14] H. F. Khelil, M. F. Ibrahim, H. A. Hussein, and R. K. Naser, "Evaluation of different stemming techniques on Arabic customer reviews," Journal of Techniques, vol. 6, no. 2, pp. 1-8, Feb. 2024, [Online]. Available: https://doi.org/10.51173/jt.v6i2.2313.

[15] L. Chaudhary, N. Girdhar, D. Sharma, J. Andreu-Perez, A. Doucet, and M. Renz, "A review of deep learning models for Twitter sentiment analysis: Challenges and opportunities," IEEE Transactions on Computational Social Systems, vol. 11, no. 3, pp. 3550-3579, 2023.

[16] E. J. Hadi, M. F. Ibrahim, and A. I. Mohammed, "Towards news classification, a machine learning approach based on stemming and feature extraction," Journal of Information & Knowledge Management, vol. 24, no. 3, p. 2550045, May 2025, [Online]. Available: https://doi.org/10.1142/S0219649225500455.

[17] Y. Zhang, H. Li, and X. Chen, "Artificial intelligence-enabled cloud security: Opportunities and challenges," Digital Communications and Networks, vol. 11, no. 2, pp. 55-66, 2025, [Online]. Available: https://doi.org/10.1016/j.dcan.2025.01.00.

[18] F. A. Bida and H. A. Naser, "Diagnostic of osteoporosis using backpropagation neural networks," Journal of Techniques, vol. 7, no. 2, pp. 10-20, 2025, [Online]. Available: https://doi.org/10.51173/jt.v7i2.2597.

[19] H. Traboulsi and M. I. Salem, "The role of electronic governance in enhancing entrepreneurial performance," Technical Journal of Management Sciences, vol. 2, no. 1, pp. 13-20, 2025, [Online]. Available: https://doi.org/10.51173/tjms.v2i1.23.