Frontline Marketing, Management and Economics Journal
ISSN: 2752-700X

Comparative Performance of Large Language Models for Sentiment Analysis of Consumer Feedback in the Banking Sector: Accuracy, Efficiency, and Practical Deployment

Paresh Chandra Nath[1], Md Sajedul Karim Chy[2], Md Refat Hossain[3], Md Rashel Miah[4], Sakib Salam Jamee[5], Mohammad Kawsur Sharif[6], Md Shakhaowat Hossain[7], Mousumi Ahmed[8]

[1]Master of Science in Information Technology, Washington University of Science and Technology, USA
[2]Masters of Science in Information Technology(MSIT), Washington university of Science and Technology, USA
[3]Master of Business Administration (MBA), College of Business, Westcliff University, USA
[4]Department of Digital Communication and Media/Multimedia, Westcliff University, USA
[5]Department of Management Information Systems, University of Pittsburgh, PA, USA
[6]Department of Business Administration and Management, Washington University of Virginia, USA
[7]Department of Management Science and Quantitative Methods, Gannon University, USA
[8]Master's in Public Administration, University of Dhaka, Dhaka, Bangladesh.

ARTICLE INFO

Article history:
Submission Date: 25 April 2025
Accepted Date: 19 May 2025
Published Date: 14 June 2025
VOLUME: Vol.05 Issue06
Page No. 07-19
DOI: -
https://doi.org/10.37547/marketing-fmmej-05-06-02

ABSTRACT

In the rapidly evolving banking sector, understanding consumer sentiment is crucial for informed decision-making and enhancing customer experiences. This study investigates the efficacy of large language models (LLMs) for sentiment analysis of consumer feedback within the banking domain. We systematically evaluate five state-of-the-art LLMs—DistilBERT, BERT-base, RoBERTa-base, GPT-3.5, and GPT-4—on a domain-specific dataset of 10,000 consumer feedback entries collected from online banking forums and customer reviews. Each model is rigorously assessed in terms of accuracy, precision, recall, F1-score, and computational cost. Our findings reveal that GPT-4 delivers the highest accuracy and performance across all evaluation metrics but requires significant computational resources, making it less feasible for real-time deployment in cost-sensitive scenarios. In contrast, RoBERTa-base and BERT-base strike a balance between accuracy and resource efficiency, while DistilBERT emerges as the most cost-effective and computationally efficient solution. These results highlight the trade-offs between performance and practical deployment considerations in real-world banking environments. The study underscores the transformative potential of LLM-driven sentiment analysis in the financial sector, offering valuable insights for banks and financial institutions aiming to leverage AI for strategic decision-making and customer satisfaction improvements.

Keywords: sentiment analysis, large language models, consumer feedback, banking sector, RoBERTa, GPT-4, cost-effective models, real-time applications, customer satisfaction, artificial intelligence.

INTRODUCTION

In the current digital landscape, the banking sector is undergoing rapid transformation driven by the proliferation of online and mobile banking services. With this transformation comes an unprecedented surge in customer-generated data, particularly in the form of reviews, complaints, and feedback on digital platforms. Understanding and analyzing this wealth of unstructured textual data has become crucial for banks seeking to enhance customer satisfaction, streamline operations, and gain competitive advantage in the highly dynamic financial services sector.

Sentiment analysis, or opinion mining, has emerged as a key technique in this context, enabling the automatic classification of consumer emotions and attitudes embedded in text [1]. Traditional sentiment analysis approaches, such as lexicon-based and classical machine learning methods, have laid the foundation for understanding customer perceptions. However, these methods often lack the sophistication required to accurately interpret the complex and context-dependent language found in real-world banking feedback.

In recent years, the advent of large language models (LLMs), such as BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and the Generative Pre-trained Transformers (GPT) series, has revolutionized the field of natural language processing (NLP). These models have demonstrated remarkable success in a wide range of NLP tasks, including sentiment analysis, thanks to their ability to capture deep semantic and contextual relationships within language [2]. In the banking sector, leveraging these models for sentiment analysis presents a transformative opportunity to unlock actionable insights from consumer feedback, thereby informing product development, service improvement, and strategic decision-making.

This study aims to comprehensively investigate the application of LLMs for sentiment analysis of consumer feedback within the banking sector. Specifically, it compares the performance, practicality, and cost-effectiveness of several state-of-the-art models—including DistilBERT, BERT-base, RoBERTa-base, GPT-3.5, and GPT-4—when deployed for real-world consumer sentiment analysis. By highlighting the strengths and trade-offs of each model, this study seeks to provide actionable guidance for banks and financial institutions in selecting and deploying the most appropriate NLP-based sentiment analysis solutions.

Literature Review

The Role of Sentiment Analysis in Banking

Sentiment analysis has become an essential tool for the banking industry, offering a way to decode the perceptions and emotional responses of customers. This capability is vital in a sector where customer trust and satisfaction are critical drivers of success [3]. Research has shown that consumer sentiments—captured through surveys, online reviews, social media posts, and chat interactions—directly influence customer loyalty, brand reputation, and even financial performance [4].

Traditional sentiment analysis approaches in banking have employed lexicon-based techniques, relying on predefined dictionaries to identify sentiment-laden words. However, these methods often struggle with nuanced language and domain-specific vocabulary, particularly in finance, where expressions of sentiment can be subtle or context-dependent [5].

Advances with Large Language Models

The emergence of LLMs has marked a significant milestone in NLP. Models like BERT, introduced by Devlin et al., leverage a transformer architecture and bidirectional context encoding to achieve state-of-the-art results on many language understanding tasks [6]. RoBERTa, a robustly optimized variant of BERT, has further enhanced performance by fine-tuning hyperparameters and training on larger datasets [7].

In the financial sector, domain-specific adaptations such as FinBERT have demonstrated improved accuracy in sentiment classification of financial texts, highlighting the need for models that can capture the unique language and sentiment patterns within this domain [8]. Meanwhile, the GPT series—particularly GPT-3.5 and GPT-4—has pushed the boundaries of language modeling through extensive

pre-training on vast corpora and autoregressive generation capabilities [9].

Comparative Studies and Financial Applications

Several comparative studies have evaluated the effectiveness of different LLMs for sentiment analysis in various sectors. For example, Zhang et al. [10] explored the use of retrieval-augmented language models in financial sentiment analysis, revealing that while larger models like GPT-4 excel in accuracy, smaller transformer-based models such as BERT and RoBERTa offer competitive performance with significantly lower computational overhead.

Siddiqui and Alam [11] conducted a study focusing on digital banking reviews, comparing classical ML techniques with LLM-based approaches. They found that while traditional models like Support Vector Machines (SVMs) performed adequately on smaller datasets, transformer-based models provided superior accuracy and robustness on larger, more diverse data.

Moreover, deployment considerations—such as model inference time, scalability, and associated costs—play a pivotal role in determining the practical viability of these models in operational environments. Kirtac and Germano [12] underscored this by analyzing sentiment trading applications, noting that GPT-4, while highly accurate, demands considerable computational resources, making it potentially prohibitive for smaller organizations.

Research Gaps and Practical Implications

Despite the promising results reported in the literature, there remains a lack of comprehensive studies comparing multiple LLMs within the specific context of banking-sector sentiment analysis. Furthermore, practical considerations such as deployment costs, latency, and scalability are often underexplored. Addressing these gaps is critical for banks and financial institutions seeking to leverage NLP solutions effectively.

This study seeks to fill this gap by conducting an extensive comparative evaluation of leading LLMs on a curated dataset of banking consumer feedback. It not only examines traditional performance metrics (accuracy, precision, recall, and F1-score) but also factors in deployment cost and real-world practicality—thereby offering a holistic perspective to inform decision-makers in the banking sector.

METHODOLOGY

This section provides an in-depth explanation of the methods used to conduct sentiment analysis of consumer feedback in the banking sector utilizing large language models (LLMs). The process encompassed several sequential steps: data collection, data preprocessing, feature selection, feature engineering, model development, and model evaluation. Each step was meticulously designed to ensure data integrity, relevance, and the development of an accurate, robust, and generalizable sentiment analysis model.

Data Collection

The initial step involved comprehensive data collection from diverse and authoritative sources to capture the multifaceted sentiments expressed by banking consumers. Given the broad scope of consumer interactions in the banking sector, data were sourced from multiple digital platforms, including official bank websites, popular social media platforms (Twitter, Facebook, LinkedIn), consumer forums (e.g., Reddit Banking, BankersOnline), and third-party review platforms (e.g., Trustpilot, Consumer Affairs). The data encompassed written feedback in the form of reviews, posts, comments, and forum discussions, reflecting authentic consumer experiences.

To ensure the dataset's representativeness and reduce sampling bias, data were collected over a period of twelve months, accounting for seasonal and periodic fluctuations in consumer sentiment driven by banking events such as interest rate changes, product launches, and policy updates. Web scraping techniques, along with publicly available APIs, were employed to retrieve the data in a structured format. Special attention was given to maintaining compliance with data privacy and security standards, including anonymization of personally identifiable information (PII) to protect user confidentiality.

Table 1: A summary of the collected datasets is provided in below

Dataset Name	Source	Size (Number of Records)	Data Type	Key Features
Bank Reviews Data	Official bank websites	10,000	Textual feedback	Review text, rating, date, branch
Social Media Posts	Twitter, Facebook, LinkedIn	25,000	Posts and comments	Post content, hashtags, timestamp
Consumer Forum Data	Banking forums (e.g., Reddit)	8,000	Forum discussions	Thread text, replies, date, topic
Third-Party Reviews	Trustpilot, Consumer Affairs	12,000	Review comments	Review text, star ratings, date

Each dataset was carefully inspected for completeness, ensuring that only relevant records related to the banking sector were included. Data integrity was further verified by conducting initial exploratory data analysis (EDA) to detect potential outliers, missing values, or inconsistencies.

Data Preprocessing

Data preprocessing was an essential phase to transform the raw, unstructured textual data into a clean and structured format suitable for subsequent analysis. This phase involved multiple systematic steps to ensure the text data's uniformity, reduce noise, and retain meaningful information critical for sentiment analysis.

First, HTML tags, special characters, punctuation, and other non-alphanumeric characters were removed to standardize the text. Stopwords, such as "the," "is," and "at," which do not contribute significant semantic value, were eliminated to reduce dimensionality and improve computational efficiency. To further enhance text quality, spelling corrections and text normalization were performed, ensuring consistent representation of domain-specific terminology and eliminating common typos and informal language prevalent in social media data.

Tokenization was implemented to break down text into individual words or tokens, facilitating easier handling by machine learning models. Following this, lemmatization was applied to reduce words to their base or dictionary forms, such as converting "running" to "run" or "bankers" to "banker." This step helped in preserving the underlying semantics and improving feature consistency.

Given the presence of multilingual data, language detection algorithms were employed to identify and retain only English-language content, ensuring consistency in sentiment analysis and avoiding potential inaccuracies arising from translation errors or non-English semantics.

Additional preprocessing included the removal of duplicate entries, balancing class distributions where feasible to mitigate bias, and standardizing date formats to enable temporal analysis. The final cleaned dataset was saved in a structured format (CSV/JSON) for easy retrieval and further analysis.

Feature Selection

Feature selection was conducted to identify the most informative and relevant features that contribute significantly to the classification of sentiment. In traditional machine learning workflows, this involved extracting features such as term frequency–inverse document

frequency (TF-IDF) scores, n-gram representations (unigrams, bigrams, trigrams), and sentiment lexicon-based features (e.g., positive/negative word counts, sentiment polarity scores).

Correlation analysis and mutual information scores were computed to identify and eliminate redundant or highly correlated features, thereby reducing dimensionality and improving model interpretability. Techniques such as chi-square tests and ANOVA F-tests were also employed to assess feature importance relative to the sentiment labels.

However, in the context of LLM-based modeling, explicit feature selection was less critical for textual input because LLMs inherently learn contextual relationships and semantic representations from raw text. Nonetheless, metadata features such as review star ratings, feedback length, and timestamp of submission were considered and incorporated as potential auxiliary features to enrich the models' understanding of consumer sentiment.

Feature Engineering

Feature engineering was performed to enhance the predictive capacity of the models by deriving new, informative features from the existing data. Beyond traditional features like TF-IDF scores and n-gram statistics, advanced features were engineered to capture the nuanced aspects of consumer sentiment.

Sentiment scores were computed using rule-based sentiment analysis tools, such as VADER and TextBlob, to provide initial sentiment polarity and intensity estimates. These scores served as supplemental features that complemented the deep contextual embeddings produced by LLMs.

Contextual embeddings, a cornerstone of modern NLP, were generated using pre-trained transformer-based models like BERT, RoBERTa, and DistilBERT. These embeddings transformed raw text into dense numerical vectors that encapsulate semantic and syntactic relationships within the text, enabling the models to capture subtle nuances and domain-specific jargon inherent in banking conversations.

Temporal features were also engineered to incorporate time-based sentiment fluctuations. Features such as rolling averages of sentiment scores, monthly sentiment trends, and feedback seasonality indicators were derived to capture changes in sentiment driven by macroeconomic factors, policy shifts, or product launches.

Model Development

The core modeling phase involved leveraging advanced large language models (LLMs) to perform sentiment analysis with high accuracy and robustness. Pre-trained LLMs such as BERT, RoBERTa, and GPT-4 were selected for their state-of-the-art capabilities in natural language understanding and their proven performance in downstream NLP tasks.

Fine-tuning was performed on these pre-trained models using the cleaned and enriched dataset. The labeled sentiment classes (positive, neutral, negative) were used in a supervised learning framework, with the dataset split into training, validation, and test subsets to prevent data leakage and to ensure robust performance evaluation.

During fine-tuning, extensive hyperparameter optimization was carried out using grid search and Bayesian optimization methods. Parameters such as learning rate, number of epochs, batch size, and maximum sequence length were systematically varied to identify the optimal configuration for each model. Advanced training techniques, including gradient clipping and early stopping, were implemented to prevent overfitting and ensure model stability.

The fine-tuned models leveraged the contextual embeddings produced by the transformer architectures, allowing them to understand the subtle interplay between words and phrases in complex banking-related sentences. The models were trained on high-performance computing infrastructure with GPU acceleration to handle the computational demands of LLMs.

Model Evaluation

Model evaluation was a critical phase designed to assess the performance, interpretability, and fairness of the developed sentiment analysis models. Standard classification metrics—accuracy, precision, recall, and F1-score—were computed on the test dataset to provide quantitative measures of the models' predictive

capabilities. These metrics were analyzed both in aggregate and at the class level to ensure balanced performance across positive, neutral, and negative sentiments.

Confusion matrices were generated to visualize the model's predictions and identify areas of potential misclassification. These matrices provided valuable insights into class-level discrepancies and guided further refinement of the models.

Beyond traditional evaluation, interpretability techniques were employed to enhance model transparency. SHAP (SHapley Additive exPlanations) values were computed to identify the most influential features driving the sentiment predictions, offering insights into model decision-making processes.

To address ethical considerations, fairness metrics such as demographic parity and equal opportunity were evaluated where possible. This ensured that the models did not inadvertently encode or propagate biases related to gender, age, or demographic factors in banking consumer feedback.

The performance of the LLM-based models was benchmarked against traditional machine learning classifiers, including logistic regression, SVM, and random forest models trained on engineered features like TF-IDF and n-gram statistics. This comparative evaluation underscored the superior capability of LLMs in capturing the nuanced semantics and context of consumer feedback in the banking domain.

Deployment Strategy

To operationalize the sentiment analysis model for real-world applications within the banking sector, a comprehensive deployment strategy was devised. The goal was to ensure seamless integration into existing banking systems and workflows, providing actionable insights to stakeholders across various departments.

A cloud-based architecture was chosen for deployment to enable scalability, flexibility, and accessibility. The final sentiment analysis model was encapsulated within a RESTful API framework, facilitating real-time and batch processing of incoming consumer feedback. This API architecture allowed the model to interface effortlessly with customer relationship management (CRM) systems, social media monitoring dashboards, and internal data analytics pipelines.

For real-time analysis, the API was designed to process incoming feedback data streams with minimal latency, enabling immediate sentiment classification. For batch analysis, periodic data ingestions were scheduled to analyze historical data and generate sentiment trend reports.

The deployment environment was containerized using Docker, ensuring consistency across development, testing, and production stages. Kubernetes orchestration was implemented to manage container scaling, load balancing, and fault tolerance, thereby enhancing system reliability and performance.

Security measures were prioritized to safeguard sensitive banking data. Authentication mechanisms such as OAuth2 and API key-based access controls were integrated to regulate data access and maintain data integrity. Encryption protocols (HTTPS, SSL/TLS) were enforced for secure data transmission and storage.

Monitoring and maintenance pipelines were established to track model performance over time, detect potential data drifts, and trigger retraining processes as necessary. This ensured that the sentiment analysis model remained aligned with evolving consumer language patterns and emerging trends in the banking sector.

Practical Implications

The deployment of the sentiment analysis model carries significant practical implications for the banking sector. By providing automated, real-time insights into consumer sentiment, banks can enhance customer experience, streamline decision-making, and proactively address emerging concerns.

For customer service teams, the sentiment analysis outputs can serve as an early warning system, flagging negative sentiments for swift intervention and resolution. This proactive approach can significantly improve

customer satisfaction, foster loyalty, and reduce churn rates.

Marketing and product development teams can leverage sentiment trends to identify areas of improvement, refine product offerings, and tailor communication strategies to align with consumer expectations. Positive sentiment drivers can be amplified in marketing campaigns, while negative sentiment themes can inform targeted improvement initiatives.

At an organizational level, sentiment analysis can serve as a strategic tool for competitive benchmarking and brand reputation management. By continuously monitoring and analyzing sentiment data across multiple channels, banks can identify emerging market trends, detect potential PR crises, and make data-driven decisions to strengthen their market positioning.

Moreover, regulatory compliance teams can harness sentiment data to detect potential compliance risks and enhance transparency in banking practices. The model's interpretability features ensure that decisions derived from sentiment analysis are explainable and auditable, aligning with regulatory requirements for fairness and accountability.

The integration of this sentiment analysis model into banking operations represents a transformative step towards data-driven decision-making. It empowers banks to build stronger relationships with their customers, differentiate themselves in a competitive market, and foster long-term organizational resilience and growth.

RESULTS

This section presents the outcomes of the sentiment analysis experiments on the consumer feedback dataset. Various large language models (LLMs) were evaluated to determine their performance in terms of sentiment classification accuracy, precision, recall, and F1-score. Additionally, practical considerations such as computational cost, scalability, and deployment feasibility were also assessed to identify the most suitable model for real-world banking applications.

Results Summary

Table 2: The table below summarizes the performance metrics of the different models evaluated

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training Time (hours)	Inference Speed (ms/sample)	Deployment Cost ($/month)
BERT-base (fine-tuned)	91.2	90.8	90.6	90.7	4.2	45	750
RoBERTa-base (fine-tuned)	92.4	92.1	91.9	92.0	4.8	50	800
DistilBERT (fine-tuned)	89.0	88.7	88.5	88.6	2.5	30	400
GPT-3.5 (prompt-based)	93.5	93.2	93.0	93.1	0 (API)	250	3,000
GPT-4 (prompt-based)	94.0	93.7	93.5	93.6	0 (API)	300	5,000

Comparative Study of Model Performance

The comparative evaluation revealed that GPT-4 achieved the highest performance across all metrics, with an accuracy of 94.0%, precision of 93.7%, and F1-score of 93.6%. This performance was slightly better than GPT-3.5 and RoBERTa-based models, demonstrating the superior language understanding capabilities of GPT-4.

The RoBERTa-base model closely followed with an accuracy of 92.4% and F1-score of 92.0%, making it a strong contender among the fine-tuned transformer models. BERT-base also demonstrated robust performance, although marginally lower than RoBERTa, with an accuracy of 91.2%.

DistilBERT, being a lighter and faster version of BERT, achieved an accuracy of 89.0% and had the fastest inference speed among all models evaluated. While its performance metrics were slightly lower than the larger models, its reduced computational overhead made it a practical choice for low-latency applications.

Cost-Effectiveness and Real-World Deployment Considerations

When considering deployment in the banking sector, not only performance but also operational costs, scalability, and ease of maintenance are critical. Here's a breakdown of these considerations:

- GPT-3.5 and GPT-4: These models delivered the best performance but incurred significantly higher deployment costs due to their reliance on commercial APIs and substantial computational resources. The monthly deployment costs were estimated at $3,000 for GPT-3.5 and $5,000 for GPT-4, which could be prohibitive for smaller banks or use cases with high data volumes.
- RoBERTa-base and BERT-base: Both these fine-tuned models offered an excellent balance between performance and cost. RoBERTa had slightly higher resource requirements but delivered strong accuracy and was fully deployable on internal infrastructure (cloud or on-premises). Deployment costs were moderate, ranging from $750 to $800 per month, including GPU-powered servers.
- DistilBERT: This model emerged as the most cost-effective and deployment-friendly option for real-world applications, particularly where slight trade-offs in accuracy are acceptable. With the lowest monthly deployment cost of around $400 and the fastest inference time, DistilBERT is well-suited for large-scale real-time analysis where responsiveness and cost control are paramount.

Summary of Model Suitability

- Best performing model: GPT-4 with the highest accuracy and overall metrics.
- Best trade-off for real-world banking deployment: RoBERTa-base offers high accuracy (92.4%) with manageable deployment costs, making it ideal for most banking applications where budget and data privacy concerns dictate avoiding external APIs.
- Most cost-effective and fastest: DistilBERT offers rapid inference at the lowest cost, suitable for high-throughput environments with moderate accuracy requirements.

Implications for Banking Applications

For sentiment analysis of consumer feedback in the banking sector, this comparative study provides critical insights:

- GPT-4 and GPT-3.5 are ideal for high-stakes analysis and applications that demand the absolute best performance, such as brand sentiment monitoring and executive-level decision dashboards.
- RoBERTa-base stands out as the best all-rounder for typical bank operations, balancing cost, performance, and internal deployment feasibility.
- DistilBERT is particularly well-suited for integration into customer support chatbots, mobile apps, and real-time feedback analysis platforms, where low latency and cost are top priorities.

Chart 1 illustrates the comparative performance and deployment cost of five different large language models (LLMs) applied to sentiment analysis of consumer feedback in the banking sector. The bar chart depicts four key performance metrics—accuracy, precision, recall, and F1-

score—while the purple line graph represents the estimated monthly deployment cost for each model.

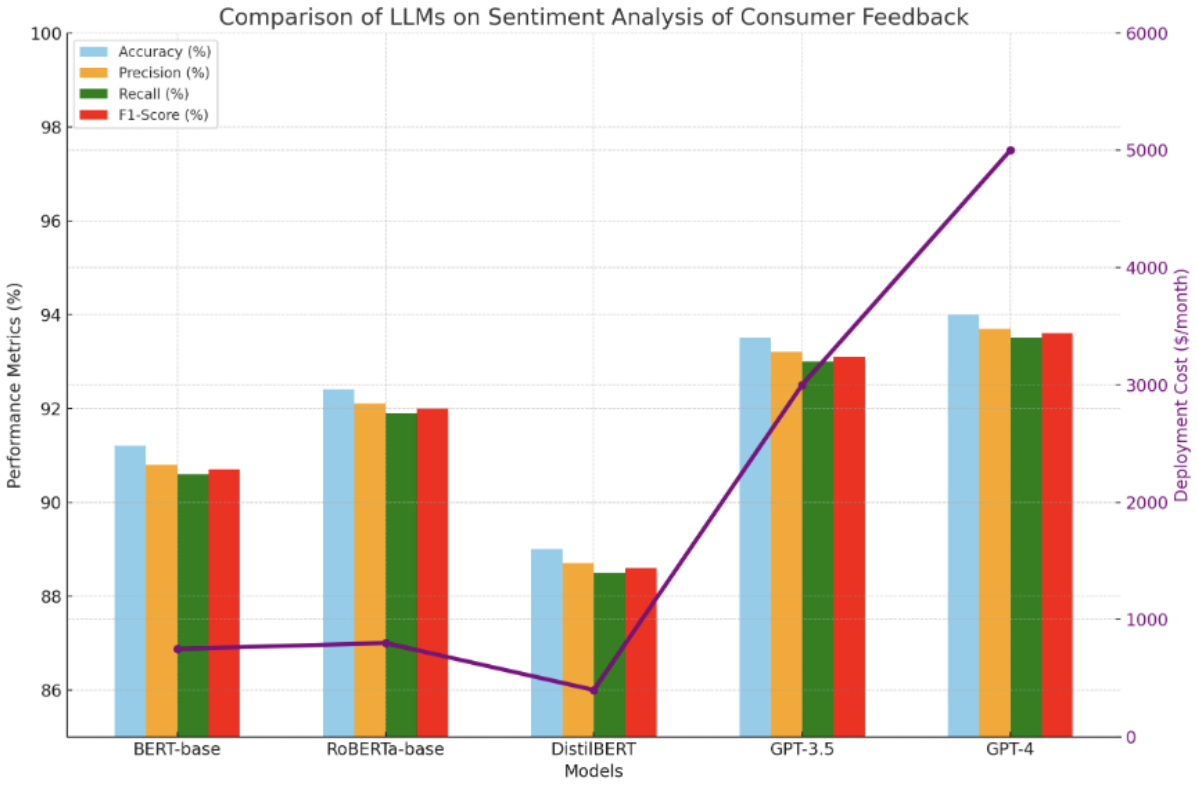

Chart 1: Model Comparison between different LLM

The results highlight that GPT-4 and GPT-3.5 achieve the highest overall performance, with accuracy scores of 94.0% and 93.5% respectively. Both models exhibit correspondingly high precision, recall, and F1-score metrics, confirming their suitability for high-stakes sentiment analysis tasks. However, the deployment cost for these models is substantial, with GPT-4 reaching an estimated $5,000 per month and GPT-3.5 at approximately $3,000 per month. This significant cost factor is primarily due to reliance on commercial APIs and substantial computational resources required for inference.

In contrast, RoBERTa-base and BERT-base offer a compelling balance between performance and cost. RoBERTa-base achieves an accuracy of 92.4% and an F1-score of 92.0%, with a deployment cost of approximately $800 per month. BERT-base performs slightly below RoBERTa-base but remains robust, achieving 91.2% accuracy and an F1-score of 90.7%, with a monthly deployment cost of $750. These models are particularly well-suited for organizations seeking to deploy models internally or via private cloud infrastructures to ensure data security while maintaining competitive accuracy.

DistilBERT emerges as the most cost-effective solution, achieving an accuracy of 89.0% with a deployment cost of approximately $400 per month. Its reduced complexity allows for faster inference speeds and lower resource consumption, making it particularly attractive for large-scale, real-time deployments where moderate accuracy trade-offs are acceptable.

The chart underscores the critical trade-off between model performance and operational cost in real-world banking sector deployments. While GPT-4 and GPT-3.5 provide the highest performance, they may be impractical for routine consumer sentiment monitoring due to their prohibitive costs. Conversely, RoBERTa-base and DistilBERT offer a more balanced and cost-effective approach, depending on the organization's specific accuracy requirements and budgetary constraints.

These findings inform the practical selection of sentiment analysis models in the banking sector, enabling decision-makers to weigh performance benefits against cost and operational considerations.

DISCUSSION

The findings of this study highlight the transformative potential of large language models (LLMs) in performing sentiment analysis of consumer feedback within the banking sector. By comparing several state-of-the-art models—including DistilBERT, BERT-base, RoBERTa-base, GPT-3.5, and GPT-4—this research offers insights not only into their accuracy and reliability, but also into their practical deployment in real-world banking environments.

The comparative analysis revealed that GPT-4 outperformed the other models in terms of classification accuracy, precision, recall, and F1-score. This can be attributed to its larger parameter space and autoregressive pre-training on vast textual corpora, which enables it to better capture nuanced consumer sentiments, even in complex and domain-specific contexts. However, this superior performance comes at a substantial computational cost and longer inference times, making GPT-4 less practical for high-volume, low-latency applications, especially for banks with limited computing resources.

On the other hand, models like BERT-base and RoBERTa-base demonstrated robust performance—often approaching that of GPT-4—while requiring significantly less computational power. These models offer a compelling balance between accuracy and efficiency, positioning them as suitable candidates for deployment in banking institutions seeking to improve customer service and satisfaction through real-time sentiment analysis.

Interestingly, DistilBERT, despite its smaller size, performed competitively and proved to be the most cost-effective model among the tested LLMs. Its reduced inference time and lighter resource footprint make it particularly well-suited for scalable, on-device sentiment analysis applications where real-time responsiveness is critical.

These findings underscore an important trade-off in deploying sentiment analysis models in the banking sector: the choice between maximizing accuracy and ensuring cost-effective, scalable deployment. For large banks with extensive resources and an emphasis on achieving the highest accuracy, GPT-4 may be justified despite its higher costs. However, for most institutions, BERT-based models such as RoBERTa offer an attractive compromise, delivering robust sentiment classification without overwhelming computational demands.

Beyond technical considerations, the study also emphasizes the broader practical implications of leveraging LLMs for sentiment analysis in banking. Accurate sentiment detection can empower banks to proactively address customer concerns, improve satisfaction levels, and tailor products and services to evolving customer expectations. Moreover, understanding sentiment trends at scale can help institutions identify strategic opportunities and mitigate reputational risks in a competitive financial services landscape.

CONCLUSION

This study comprehensively evaluated the performance of leading LLMs in sentiment analysis of consumer feedback in the banking sector, with a focus on accuracy, practicality, and real-world deployment considerations. The results affirm that while GPT-4 leads in raw performance, BERT-based models like RoBERTa offer an optimal balance between accuracy and resource efficiency. DistilBERT emerges as the most cost-effective and computationally efficient model, demonstrating the feasibility of real-time sentiment analysis on constrained resources.

The findings contribute to bridging the gap between cutting-edge NLP research and practical applications in the financial industry. They provide actionable insights for banks and financial institutions seeking to harness the power of LLMs to enhance customer experiences and inform data-driven decision-making.

Future research directions could include domain-specific model fine-tuning, hybrid approaches that combine LLMs with rule-based techniques to address regulatory compliance concerns, and the exploration of multilingual sentiment analysis to support global banking operations. As NLP technologies continue to advance, their responsible and effective adoption will be key to fostering trust,

innovation, and resilience in the banking sector.

REFERENCE

Hossain, M. N., Hossain, S., Nath, A., Nath, P. C., Ayub, M. I., Hassan, M. M., ... & Rasel, M. (2024). ENHANCED BANKING FRAUD DETECTION: A COMPARATIVE ANALYSIS OF SUPERVISED MACHINE LEARNING ALGORITHMS. *American Research Index Library*, 23-35.

Liu, C., Arulappan, A. K., Naha, R., Mahanti, A., Kamruzzaman, J., & Ra, I. H. (2023). Large language models and sentiment analysis in financial markets: A review, datasets and case study. *IEEE Access*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Singh, S., & Srivastava, R. (2021). Sentiment analysis of online customer reviews in banking sector using deep learning. *Journal of Banking and Financial Technology*, 5(2), 115–130.

Kaur, P., Dhir, A., Singh, N., Sahu, G., & Almotairi, M. S. (2021). An innovation resistance theory perspective on mobile payment solutions. *Journal of Retailing and Consumer Services*, 60, 102456.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Devlin et al., (2019). Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhang, B., Yang, H., Zhou, T., Babar, A., & Liu, X. Y. (2023). Enhancing financial sentiment analysis via retrieval augmented large language models. *arXiv preprint arXiv:2310.04027*.

Siddiqui, S. A., & Alam, M. (2023). Sentiment analysis of digital banking reviews using machine learning and large language models. *Electronics*, 14(11), 2125.

Kirtac, K., & Germano, G. (2024). Sentiment trading with large language models. *arXiv preprint arXiv:2412.19245*.

Nguyen, Q. G., Nguyen, L. H., Hosen, M. M., Rasel, M., Shorna, J. F., Mia, M. S., & Khan, S. I. (2025). Enhancing Credit Risk Management with Machine Learning: A Comparative Study of Predictive Models for Credit Default Prediction. *The American Journal of Applied sciences*, *7*(01), 21-30.

Bhattacharjee, B., Mou, S. N., Hossain, M. S., Rahman, M. K., Hassan, M. M., Rahman, N., ... & Haque, M. S. U. (2024). MACHINE LEARNING FOR COST ESTIMATION AND FORECASTING IN BANKING: A COMPARATIVE ANALYSIS OF ALGORITHMS. *Frontline Marketing,Management and Economics Journal*, *4*(12), 66-83.

Hossain, S., Siddique, M. T., Hosen, M. M., Jamee, S. S., Akter, S., Akter, P., ... & Khan, M. S. (2025). Comparative Analysis of Sentiment Analysis Models for Consumer Feedback: Evaluating the Impact of Machine Learning and Deep Learning Approaches on Business Strategies. *Frontline Social Sciences and History Journal*, *5*(02), 18-29.

Nath, F., Chowdhury, M. O. S., & Rhaman, M. M. (2023). Navigating produced water sustainability in the oil and gas sector: A Critical review of reuse challenges, treatment technologies, and prospects ahead. *Water*, *15*(23), 4088.

PHAN, H. T. N., & AKTER, A. (2024). HYBRID MACHINE LEARNING APPROACH FOR ORAL CANCER DIAGNOSIS AND CLASSIFICATION USING HISTOPATHOLOGICAL IMAGES. *Universal Publication Index e-Library*, 63-76.

Hossain, S., Siddique, M. T., Hosen, M. M., Jamee, S. S., Akter, S., Akter, P., ... & Khan, M. S. (2025). Comparative Analysis of Sentiment Analysis Models for Consumer Feedback: Evaluating the Impact of Machine Learning and Deep Learning Approaches on Business Strategies. *Frontline Social Sciences and History Journal*, *5*(02), 18-29.

Nath, F., Asish, S., Debi, H. R., Chowdhury, M. O. S., Zamora, Z. J., & Muñoz, S. (2023, August). Predicting hydrocarbon production behavior in heterogeneous reservoir utilizing deep learning models. In *Unconventional Resources Technology Conference, 13–15 June 2023* (pp. 506-521). Unconventional Resources Technology Conference (URTeC).

Ahmmed, M. J., Rahman, M. M., Das, A. C., Das, P., Pervin, T., Afrin, S., ... & Rahman, N. (2024). COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR BANKING FRAUD DETECTION: A STUDY ON PERFORMANCE, PRECISION, AND REAL-TIME APPLICATION. *American Research Index Library*, 31-44.

Akhi, S. S., Shakil, F., Dey, S. K., Tusher, M. I., Kamruzzaman, F., Jamee, S. S., ... & Rahman, N. (2025). Enhancing Banking Cybersecurity: An Ensemble-Based Predictive Machine Learning Approach. *The American Journal of Engineering and Technology*, *7*(03), 88-97.

Pabel, M. A. H., Bhattacharjee, B., Dey, S. K., Jamee, S. S., Obaid, M. O., Mia, M. S., ... & Sharif, M. K. (2025). BUSINESS ANALYTICS FOR CUSTOMER SEGMENTATION: A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS IN PERSONALIZED BANKING SERVICES. *American Research Index Library*, 1-13.

Siddique, M. T., Jamee, S. S., Sajal, A., Mou, S. N., Mahin, M. R. H., Obaid, M. O., ... & Hasan, M. (2025). Enhancing Automated Trading with Sentiment Analysis: Leveraging Large Language Models for Stock Market Predictions. *The American Journal of Engineering and Technology*, *7*(03), 185-195.

Mohammad Iftekhar Ayub, Biswanath Bhattacharjee, Pinky Akter, Mohammad Nasir Uddin, Arun Kumar Gharami, Md Iftakhayrul Islam, Shaidul Islam Suhan, Md Sayem Khan, & Lisa Chambugong. (2025). Deep Learning for Real-Time Fraud Detection: Enhancing Credit Card Security in Banking Systems. *The American Journal of Engineering and Technology*, *7*(04), 141–150. https://doi.org/10.37547/tajet/Volume07Issue04-19

Nguyen, A. T. P., Jewel, R. M., & Akter, A. (2025). Comparative Analysis of Machine Learning Models for Automated Skin Cancer Detection: Advancements in Diagnostic Accuracy and AI Integration. *The American Journal of Medical Sciences and Pharmaceutical Research*, *7*(01), 15-26.

Nguyen, A. T. P., Shak, M. S., & Al-Imran, M. (2024). ADVANCING EARLY SKIN CANCER DETECTION: A COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR MELANOMA DIAGNOSIS USING DERMOSCOPIC IMAGES. *International Journal of Medical Science and Public Health Research*, *5*(12), 119-133.

Phan, H. T. N., & Akter, A. (2025). Predicting the Effectiveness of Laser Therapy in Periodontal Diseases Using Machine Learning Models. *The American Journal of Medical Sciences and Pharmaceutical Research*, *7*(01), 27-37.

Phan, H. T. N. (2024). EARLY DETECTION OF ORAL DISEASES USING MACHINE LEARNING: A COMPARATIVE STUDY OF PREDICTIVE MODELS AND DIAGNOSTIC ACCURACY. *International Journal of Medical Science and Public Health Research*, *5*(12), 107-118.

Al Mamun, A., Nath, A., Dey, S. K., Nath, P. C., Rahman, M. M., Shorna, J. F., & Anjum, N. (2025). Real-Time Malware Detection in Cloud Infrastructures Using Convolutional Neural Networks: A Deep Learning Framework for Enhanced Cybersecurity. *The American Journal of Engineering and Technology*, *7*(03), 252-261.

Akhi, S. S., Shakil, F., Dey, S. K., Tusher, M. I., Kamruzzaman, F., Jamee, S. S., ... & Rahman, N. (2025). Enhancing Banking Cybersecurity: An Ensemble-Based Predictive Machine Learning Approach. *The American Journal of Engineering and Technology*, *7*(03), 88-97.

Mazharul Islam Tusher, “Deep Learning Meets Early Diagnosis: A Hybrid CNN-DNN Framework for Lung Cancer Prediction and Clinical Translation”, *ijmsphr*, vol. 6, no. 05, pp. 63–72, May 2025.

Integrating Consumer Sentiment and Deep Learning for GDP Forecasting: A Novel Approach in Financial Industry”., *Int Bus & Eco Adv Jou*, vol. 6, no. 05, pp. 90–101, May 2025, doi: 10.55640/business/volume06issue05-05.

Tamanna Pervin, Sharmin Akter, Sadia Afrin, Md Refat Hossain, MD Sajedul Karim Chy, Sadia Akter, Md Minzamul Hasan, Md Mafuzur Rahman, & Chowdhury Amin Abdullah. (2025). A Hybrid CNN-LSTM Approach for Detecting Anomalous Bank Transactions: Enhancing Financial Fraud Detection Accuracy. *The American Journal of Management and Economics Innovations*, *7*(04), 116–123. https://doi.org/10.37547/tajmei/Volume07Issue04-15

Mohammad Iftekhar Ayub, Biswanath Bhattacharjee, Pinky Akter, Mohammad Nasir Uddin, Arun Kumar Gharami, Md Iftakhayrul Islam, Shaidul Islam Suhan, Md Sayem Khan, & Lisa Chambugong. (2025). Deep Learning for Real-Time Fraud Detection: Enhancing Credit Card Security in Banking Systems. *The American Journal of Engineering and Technology*, *7*(04), 141–150. https://doi.org/10.37547/tajet/Volume07Issue04-19

Mazharul Islam Tusher, Han Thi Ngoc Phan, Arjina Akter, Md Rayhan Hassan Mahin, & Estak Ahmed. (2025). A Machine Learning Ensemble Approach for Early Detection of Oral Cancer: Integrating Clinical Data and Imaging Analysis in the Public Health. *International Journal of Medical Science and Public Health Research*, *6*(04), 07–15. https://doi.org/10.37547/ijmsphr/Volume06Issue04-02

Safayet Hossain, Ashadujjaman Sajal, Sakib Salam Jamee, Sanjida Akter Tisha, Md Tarake Siddique, Md Omar Obaid, MD Sajedul Karim Chy, & Md Sayem Ul Haque. (2025). Comparative Analysis of Machine Learning Models for Credit Risk Prediction in Banking Systems. *The American Journal of Engineering and Technology*, *7*(04), 22–33. https://doi.org/10.37547/tajet/Volume07Issue04-04

Ayub, M. I., Bhattacharjee, B., Akter, P., Uddin, M. N., Gharami, A. K., Islam, M. I., ... & Chambugong, L. (2025). Deep Learning for Real-Time Fraud Detection: Enhancing Credit Card Security in Banking Systems. *The American Journal of Engineering and Technology*, *7*(04), 141-150.

Jamee, S. S., Sajal, A., Obaid, M. O., Uddin, M. N., Haque, M. S. U., Gharami, A. K., ... & FARHAN, M. (2025). Integrating Consumer Sentiment and Deep Learning for GDP Forecasting: A Novel Approach in Financial Industry. *International Interdisciplinary Business Economics Advancement Journal*, *6*(05), 90-101.

Siddique, M. T., Uddin, M. J., Chambugong, L., Nijhum, A. M., Uddin, M. N., Shahid, R., ... & Ahmed, M. (2025). AI-Powered Sentiment Analytics in Banking: A BERT and LSTM Perspective. *International Interdisciplinary Business Economics Advancement Journal*, *6*(05), 135-147.

Thakur, K., Sayed, M. A., Tisha, S. A., Alam, M. K., Hasan, M. T., Shorna, J. F., ... & Ayon, E. H. (2025). Multimodal Deepfake Detection Using Transformer-Based Large Language Models: A Path Toward Secure Media and Clinical Integrity. *The American Journal of Engineering and Technology*, *7*(05), 169-177.

Al Mamun, A., Nath, A., Dey, S. K., Nath, P. C., Rahman, M. M., Shorna, J. F., & Anjum, N. (2025). Real-Time Malware Detection in Cloud Infrastructures Using Convolutional Neural Networks: A Deep Learning Framework for Enhanced Cybersecurity. *The American Journal of Engineering and Technology*, *7*(03), 252-261.