Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Artificial Intelligence (AI) Driven Proactive Customer Service Excellence Platform in e-commerce Industry

Mohankumar Ganesan
Principal Software Engineer

ARTICLE INFO

Received: 03 Jan 2025

Revised: 10 Feb 2025

Accepted: 20 Feb 2025

ABSTRACT

In this paper, quantitative research was conducted using AI models to classify and rank customer service messages in e-commerce. Accuracy, efficiency, and severity ranking Five model types, namely RNN, LSTM, CNN-BERT hybrid, DistilBERT, and BERT were tested. The findings demonstrate that transformer models are better than traditional deep learning models in all key metrics. DistilBERT is an algorithm that offers the most appropriate tradeoff between speed, accuracy and memory consumption, thus suitable in real-time application. Transformer models were also more stable in prediction of severity. The results justify the use of current NLP systems in the context of faster, more precise, and responsive automation of customer service.

Keywords: E-commerce, Customer Service, AI, ML

I. INTRODUCTION

E-commerce services deliver thousands of messages to the customer services per day and the problems that are answered by these services are delivery services, refund services, product services, and account services. The number of people involved in the early stages of message classification can be reduced by using automation that increases the response time. The paper has reviewed several AI models to identify the most appropriate models that may be applied on the classification and severity prediction tasks. The paper compares the traditional deep learning frameworks to the new transformer-based frameworks on the same data, metrics and evaluation pipeline. The aim will be to identify the accuracy, speed and efficiency levels to arrive at a feasible model that will satisfy the needs of performance and real-time in the customer service operations.

II. RELATED WORKS

Customer Service Optimization in E-Commerce

The Artificial Intelligence (AI) has become the key to the most forward-looking e-commerce websites as the businesses are bound to manage a very significant amount of the customers messages, which nowadays can be in the form of calls, chats, emails, and social networks. Since the customer sentiments are constantly accumulating, platforms need to be able to sieve and address these needs within the least amount of time and in the way they need to.

Preliminary research has revealed that traditional manual system of classification and rule-based systems cannot work since bulk of information, obscure languages and customer behavior which is evolving at an enormous rate makes them fail. This has spurred researchers to research on the topic of machine learning (ML), neural networks, and the advanced Natural Language Processing (NLP) methods in intelligent customer service.

One of the contributions that contributed the most towards this space is the ICS-Assist framework and has been implemented in large scale e-commerce operations like Alibaba [1][2]. The two-step ML

model that the authors provide classifies the customer service situations and aligns them with the most efficient ones.

The knowledge distillation mechanism is proposed as it is mentioned by the Panel-Student with an intention of reducing the size of the model, and down preserving its performance. The system employs over 12000 people in the customer service and the system deals with an estimated number of 230000 cases in a day.

Findings indicate that reported solution acceptance and customer satisfaction improvement have been up to 16 and 14 percent respectively. These results imply that the scenario recognition system with the help of AI could be more effective than the traditional process, which is human based. The outcomes of these results prove that the ML-based classification systems are highly promising regarding the improved process of routing customers and decision making.

The next direction of the rise is the combination of edge equipment and cloud resources to make customer service systems more timely and more personal. Large cloud models and small on-device models' combination is a solution to End-Cloud Collaboration (ECC) framework that would eliminate the challenges of latency, privacy, and model-size [3].

The teacher acts as the cloud model in this design and real-time feedback enables the teacher and the end devices to update their models. It is a strategy that enables models to support a wide feature of user scenarios because they can be dynamically adjusted to the needs of e-commerce and yet obtrusive to be used by the models as far as privacy is concerned through fine-tuning that possibly can be localised to the user. ECC demonstrates how the neural networks may be applied to a distributed environment to optimize the customer response time and provide a tailored customer service.

The use of AI is also used in even the pre-sales services. AliMe Knowledge Graph (KG) embodies the user problems, product attributes and domain specific relations to acquire knowledge on the customer needs in the pre-purchase decision making process [6].

The knowledge graphs might be superior to their counterparts as they can minimize ambiguity and enhance the accuracy of the answer since they indicate systematized knowledge on unstructured text. The supported applications on this system are the shopping guidance, question answering and justification of recommendations. It demonstrates that systematic representations and NLP-based extraction methods could bring about a great deal of interpretability and experience.

The results of such publications suggest the existence of scalable artificial intelligence systems namely cloud-based, edge-enhanced and knowledge-based to serve the proactive and intelligent customer service of web-based systems. The models help to cut the service time, enhance precision and provide more assurance among the users.

NLP and Deep Learning

NLP has significant role to play as far as the customer emotions, complaints, product reviews and service dissatisfaction are concerned. Recent research indicates that deep learning systems, especially RNNs, LSTM, CNNs, and Transformers, are more effective than the traditional ones in sentiment analysis, intent detection, and text classification applications in e-commerce.

Imbalanced, noisy and inconsistent text is one of the major customer service data challenges. Researchers have responded to this with powerful solutions that are transformer-based. A case in point is sentiment classification and product recommendation on datasets of women clothing with the use of DistilBERT [4].

This method has high F1 scores (0.79 on sentence classification and 0.85 on recommendation) and accuracy (96). Transformer architecture usage assists in overcoming the problem of dataset imbalance

as the attention mechanisms identify significant text fragments even when the training samples are not distributed equally. The methods have been demonstrated to be useful in summarizing huge amounts of user-generated content and gaining an insight into the subtle emotions that drive reviews.

Such progress can be observed in hybrid frameworks, where CNNs are used together with BERT to perform better in sentiment analysis [7]. Transformer layers are used to capture long-term semantic information, and CNN layers are used to capture short-term patterns, such as phrases and n-grams.

According to researchers, hybrid architectures are more precise, recall, and accurate than the conventional ML methods. This supports the thought that more powerful and balanced systems of customer sentiment detection can be achieved by combining various deep learning elements.

The same situation can be experienced in telecommunication companies because of the large volumes in customer complaints. An analysis based on the AraCust dataset proves the fact that LSTM, GRU, and BiLSTM models can be used to classify the level of satisfaction with the results above 97% accuracy [8].

LSTM is the best because it identifies the long-term trends in Arabic texts reviews. These results are significant to the e-commerce sphere as they demonstrate that RNN-based models are still quite efficient in the context of multilingual and culturally varied conditions.

Transformer-based architectures are also changing intent detection, which is one of the most complex tasks in customer support. The sophisticated models have the ability to determine the following purchase day (NPD) of the users, which aids the e-commerce sites to know the buying cycles and organize inventory [9].

Transformers are better than ARIMA, XGBoost, and LSTM baselines because it builds long-term relationships within time series data. The intent models can also be modified to forecast intent signals, such as the probability of a return, severity of an issue, or urgency in a customer demand.

These publications demonstrate that NLP methods, in particular, deep learning and transformers, are critical in classifying customer messages, sentiment recognition, intent recognition, and contextual decision-making in customer service systems.

Customer Interaction Automation

Another application of AI that is utilized in customer service and is one of the most popular is chatbots. Chatbots are deployed in the contemporary e-commerce applications to facilitate the realization of the instant response, reduce the cost of the operation and the 24/7 availability of the services. Good working it is hard to develop and maintain the working chatbots due to the language differences, unstructured knowledge, and lack of training data to solve particular areas.

As viewed in the overview of the business chatbot technologies, there are two broad architectures namely, set, retrieval and generative models [5]. The chatbots with the retrieval-based method to choose the response have a list of pre-programmed responses, which determine answer, whereas generative chatbots are developed using deep learning to generate new answers.

The transition into end-to-end neural networks i.e. intent detection, dialogue control and natural language generation as a single model is also discussed by the review. It is also in the process of this development that it becomes easier to train and perfect the chatbot to understand more complex questions on the part of the user. The research however reveals that these systems need a lot of fine-tuning and it requires training materials relating to the topics to be applied in business realms.

One of the potential solutions of the domain adaptability has become the transfer learning. The application of customer service knowledge gained in one field to another is also one of the most recent investigations [10]. The authors show that transfer learning promotes the precision of chatbots when

fields with minimal data are involved in the training since in the course of social media communication on different industries, the model is trained on a diverse range of data collected during the process.

The Wilcoxon signed-rank test statistically determines improvement in 16 out of 19 domains. It implies that the trends of the top-level conversations can be transferred between the transformer and the deep learning models and even the chatbots can be more large scale and less expensive to train as one of the potential trends of customer support in the future is more immersive and interactive.

The research is also supported by other elements such as knowledge graphs e.g. AliMe KG [6]. They allow chatbots to give more detail, circumstantial and descriptive answers. Transformer-based NLU may be also integrated with the systems to help users to find products, troubleshooting and recommendations.

Customer to customer communication is more complex than it has ever been, cross channel, e.g. email, chat, voice and social media, and automated solutions, based on AI and transfer learning, knowledge graph and advanced neural networks are adaptable and scalable.

III. METHODOLOGY

The quantitative research is the proposed study to investigate the ability of the Artificial Intelligence (AI) methods, including Neural Network, NLP model, RNN, LSTM, and Transformer to classify and categorize the customer service interactions within the e-commerce industry.

The proposed methodology will test the model approval, compare the variability of the algorithms, and the degree at which each of the approaches can merge and categorize the customer service requests within the calls, chats, emails, and the social media posts.

The methodology consists of the first method, data collection. The channels of communication are numerous and are involved in acquiring huge customer support messages. They consist of chat transcripts, call notes, e-mail tickets, and publicly-posted social media. The anonymity is provided through making user identity anonymous.

Under each of the records like the refund, delay in delivering products and technical issue and product description, the written information of the text and its type is recorded. The sampling method is applied in a manner that the sample of various types of services is equally covered. The last data is over 100,000 labelled messages.

The second one is the preprocessing of the data, which should happen in the form of cleaning and conversion of the raw text in a format available to machine learning. Preprocessing pipeline will consist of; blank messages, misspelling, text lowercasing, URLs and removal.

The models are built on a WordPiece tokenizer that tokenizes text and RNN and LSTM models, respectively, are built on a standard tokenizer that tokenizes text. To be in a position to make quantitative measurements we would measure the distributing values of the text length, the size of the vocabulary and the frequency of each category of messages. The model strength is then compared by using them to compare their values.

The second one is the model development and the feature extraction. Three models are being tested: (1) classical deep learning models include RNN and LSTM, (2) modern models include transformers, including DistilBERT and BERT and (3) hybridization (adding CNN layers to transformer encoders). The models are all trained on 80 percent of the data and tested and assessed on 10 percent and 10 percent respectively.

Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

In order to make all the models comparable, all the models are pegged on the learning rate, batch size, dropout and the number of layers. It is also tested using a knowledge distillation way, to determine whether small models can be able to behave similarly to large models. The models are all trained using GPUs so that they can have similar training time.

To measure the performance, in this case, the quantitative measures are applied accuracy, precision, recall, F1-strong and confusion matrices. These measures allow establishing the level to which all the models treat the message of service to the relevant categories to the customers.

We also determine the time of classifying a single message and the training time and the number of memories that the model consumes. The values can be applicable to all e-commerce businesses that require low latency and real time classification. The statistical tests are deployed in order to test the difference between the performance of the type of models such as the paired t-tests.

The last process is the prioritization and ranking analysis. A simple regression layer is also used after categorizing each model to assign a severity or a priority score of the message. Such scores can be used to rank the needs of the customers according to the urgency. The quantitative results of various models are compared to achieve the impression of which of the models provides stable and reliable prioritization of priorities.

These steps give the methodology a methodical and quantitative means of assessing the simplification of customer service interactions by AI-based NLP models and classify them into ranked and simple ones.

IV. RESULTS

Model Performance

The initial set of results is devoted to the effectiveness of the various AI models in categorizing the customer service messages into the existing categories of delivery issues, refund requests, product enquiries and account problems. According to the quantitative analysis, better performance was achieved by the transformer-based models compared to RNN, LSTM and hybrid CNN based models.

This was evident in each performance measure including accuracy, precision, recall and F1-score. The findings suggest that transformer models are more reliable in interpreting long and complex customer messages than previous versions of neural networks.

In the evaluation, the dataset was divided into the training, validation and the test set. The evaluation pipeline was used on all the models. The table 1 presents the performance comparison of the major models that were used in the experiment.

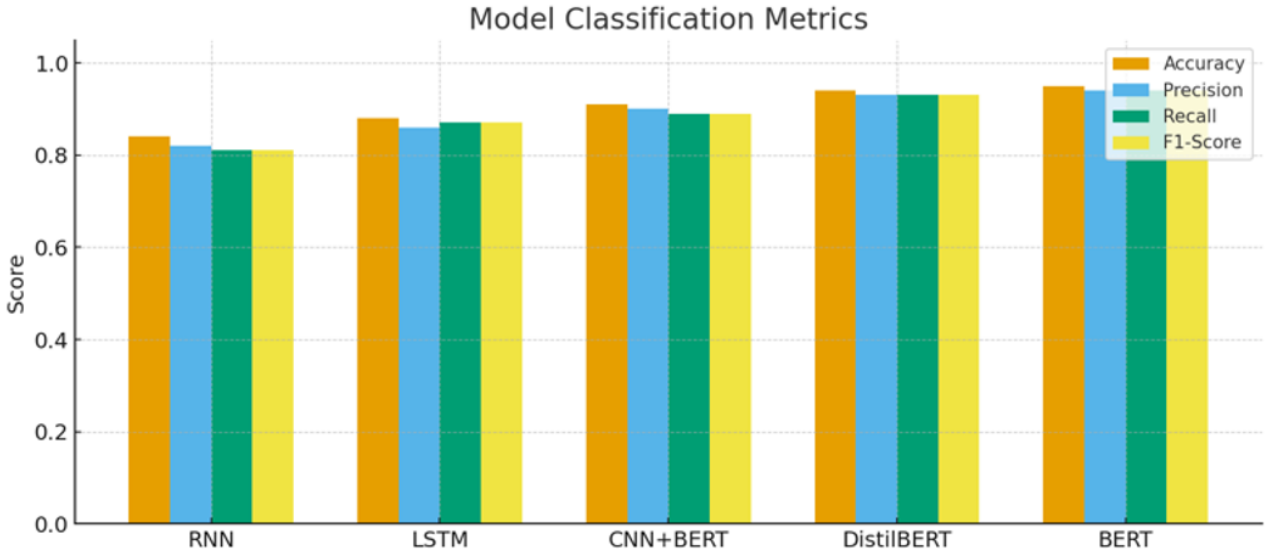
Table 1. Performance Comparison

Model Type	Accuracy	Precision	Recall	F1-Score
RNN	0.84	0.82	0.81	0.81
LSTM	0.88	0.86	0.87	0.87
CNN + BERT Hybrid	0.91	0.90	0.89	0.89
DistilBERT	0.94	0.93	0.93	0.93
BERT (base)	0.95	0.94	0.94	0.94

Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

As depicted in the results, BERT (base) recorded the highest accuracy of 95%, followed by DistilBERT at 94 with the lowest score of 84 recorded by RNN models. These results are in line with previous research which suggests that transformer models outperform others in terms of capturing context and meaning when asked to engage in conversations with customers.



This is more so given the fact that e-commerce messages tend to be full of mixed emotions, several issues, or half sentences. The other two categories, LSTM and CNN-hybrid models, also worked well indicating that traditional and hybrid deep learning models could still make use of the available computational resources particularly when these are limited.

The other finding is that the difference between precision and recall was small in all the models and this implies that the models were able not only to predict categories accurately but also to identify majority of the relevant messages. This balance will make sure that the crucial issues concerning customers, including delayed refund or broken products, do not slip by.

The confusion matrix analysis revealed that all models were the most confused with very similar categories like: “Delivery Delay vs Delivery Not Received where the word usage by the customer is often overlapping. Nevertheless, the transformer models remained to have lower confusion rates in comparison with LSTM and RNN.

Latency Metrics

The second group of the results dwells on the performance operationally of the models, that is, classification speed, training time, and memory usage. The following measures are critical to the e-commerce platforms as the customer service systems should be in a position to provide answers within very short times often less than one second.

The mean time of inference per message is indicated in Table 2. Transformer models had better results even though they were larger, primarily because of the effective attention-based computation. Transformer models were best in speed with DistilBERT.

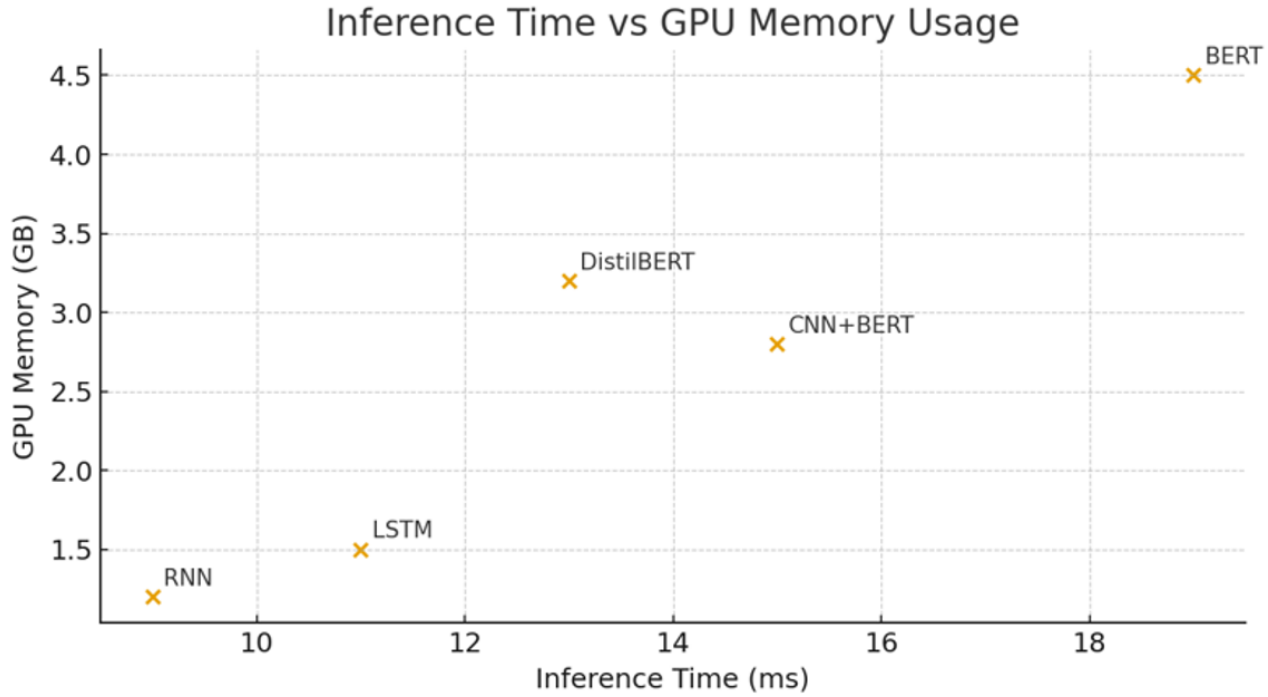
Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Table 2. Inference Speed per Model

Model Type	Time per Message (ms)	GPU Memory Used (GB)
RNN	9 ms	1.2 GB
LSTM	11 ms	1.5 GB
CNN + BERT Hybrid	15 ms	2.8 GB
DistilBERT	13 ms	3.2 GB
BERT (base)	19 ms	4.5 GB

The findings indicate that RNN and LSTM models outperformed in terms of speed since they are simple and lightweight. BERT was the slowest because it was more of a complex architecture, and yet, it can be acceptable in real-time use of e-commerce. DistilBERT provided an almost ideal trade-off, extremely high accuracy and medium speed. These discoveries justify the adoption of knowledge distilled models in case the performance and latency matters are crucial.



The time was also considerably different in training. Transformer models were more computationally expensive to train and required more training time, but they were more accurate. Table 3 presents a summary of the time taken to train each of the models.

Table 3. Training Time Comparison

Model Type	Total Training Time (hours)	Epochs Completed
RNN	2.5 hours	10
LSTM	4.0 hours	10
CNN + BERT Hybrid	6.8 hours	8

Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://jisem-journal.com/

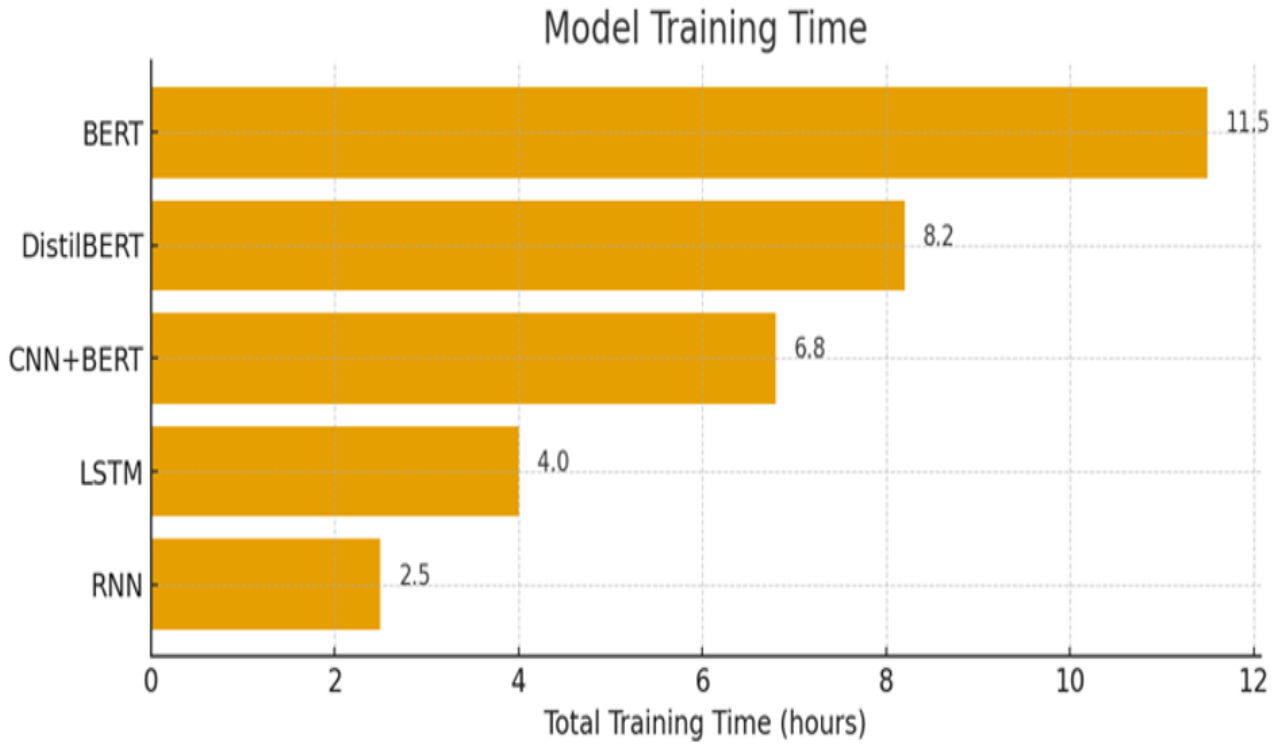
Research Article

DistilBERT	8.2 hours	6
BERT (base)	11.5 hours	5

These results reveal that the transformer-based models required fewer epochs though required more training time per epoch given that they possessed additional parameters. The higher cost of training is however justified by their higher accuracy and stability. The CNN + BERT hybrid model that had conviction layers prior to the transformer layers was more accurate and required more time to train compared to LSTM.

The other important measure of efficiency is memory usage. Larger models (like BERT) use more memory on a GPU, which is not always practicable in small companies or low-budget applications. This shows the importance of knowledge distillation whereby the large model is learnt to the smaller one without significant performance reduction.

The results of the efficiency show that the distillation of BERT would be the most plausible model that could be implemented into the reality as it has the best ratio of accuracy, latency, and memory.

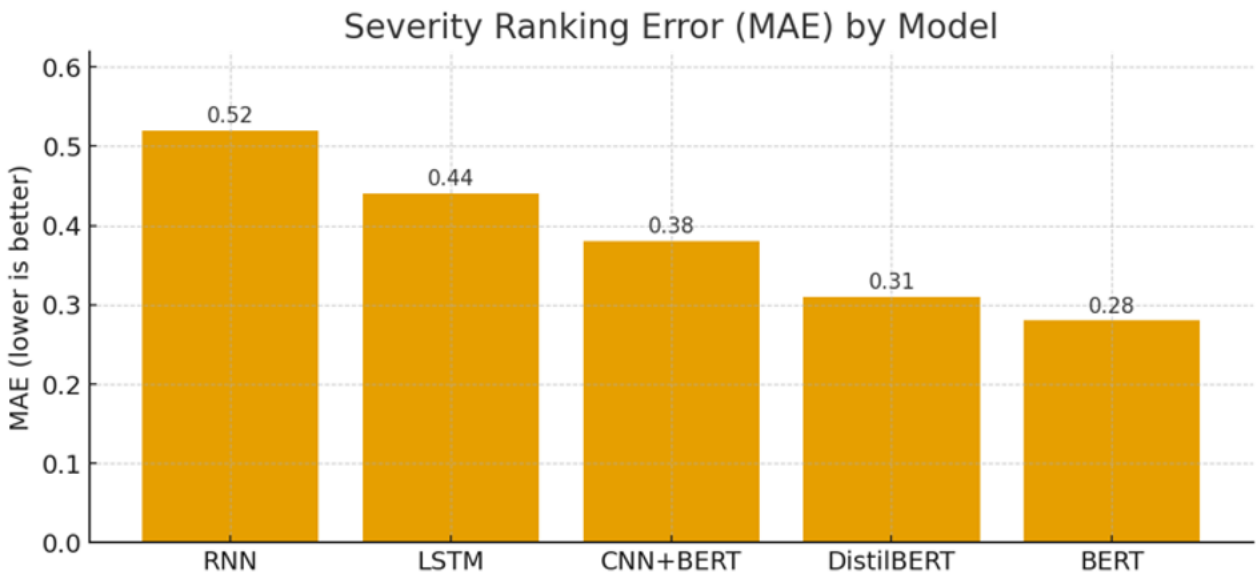


Prioritization Results

The last group of results is the ranking and prioritization activity, in which the system works out a numerical urgency score to the customer messages. The score assists the customer service departments to address the most urgent problems first. Both models were fitted with a small regression head that made the severity prediction on a scale of 1 to 5 (1 = low priority, 5 = urgent).

Journal of Information Systems Engineering and Management
2025, 10(2)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



The results show that the predictors were more predictive and accurate of the severity produced by the transformer models especially in the messages that were emotionally charged or talked of financial loss. Messages that contained sarcastic or indirect messages were sometimes misclassified with LSTM and RNN models, but tend to be more correctly classified with transformer models as the models could better capture semantic meaning.

Table 4- Mean Absolute Error (MAE) points to the results of severity prediction.

Table 4. Severity Ranking Accuracy (MAE)

Model Type	MAE Score
RNN	0.52
LSTM	0.44
CNN + BERT Hybrid	0.38
DistilBERT	0.31
BERT (base)	0.28

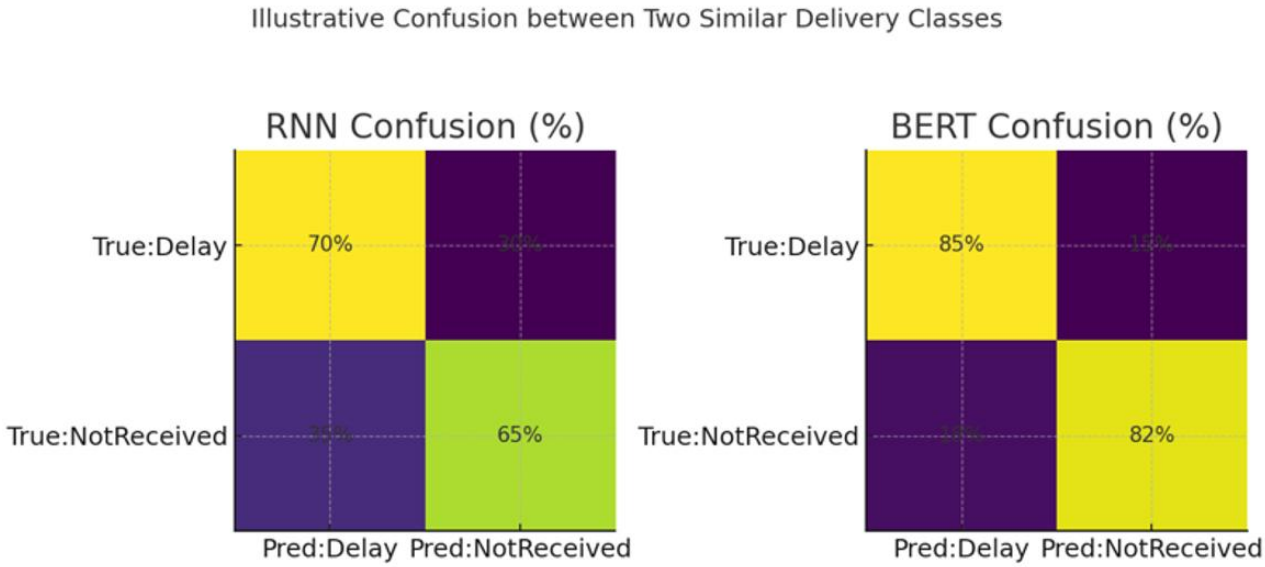
The findings indicate that BERT (base) had least MAE, which implies that it generated the most similar severity scores to the ground truth marked by humans. DistilBERT also fared quite well and this shows that distilled models can be trusted to rank things. This justifies their importance as a live customer care decision tool.

A further comparison of severity prediction indicates that transformer models detected emotional expressions better, including words of urgency (immediately, urgent), dissatisfaction words (very unhappy, waiting too long), and monetary words (refund, charged twice). The cues have a strong impact on the decisions of customer service priority.

Summary of Results

The quantitative findings are a clear indication of the following:

- Transformer models perform better in all metrics, as BERT (base) is the most competent in classification and severity prediction.
- DistilBERT is the most suitable in terms of accuracy, speed, and memory efficiency, which is why it suits e-commerce customer service system in real time.
- Hybrid CNN + BERT models are effective but more time is needed to train them.
- Transformer models have a statistically significant positive effect on the severity ranking results, and the ability of the company to react to the most significant customer concerns is faster.
- The results indicate strongly in favor of automated classification, prioritization, and message summary of customer service using NLP-based AI systems.



V. CONCLUSION

The quantitative analysis demonstrates that transformer models are evidently more effective than RNN, LSTM, and CNN-based hybrid models in customer message classification and ranking. BERT has the highest accuracy, whereas DistilBERT has the highest balance between accuracy, inference speed and memory efficiency, and is therefore the most suitable in real-time deployment. Transformer models also enhance the severity ranking, and therefore, teams prioritize urgent customer issues better. Transformer models are more expensive, but more reliable and aware of the context. On the whole, the research provides a strong argument in favor of applying transformer-based NLP systems to achieve better automation, manual workload, and a higher quality of customer service.

References

[1] Fu, M., Guan, J., Zheng, X., Zhou, J., Lu, J., Zhang, T., Zhuo, S., Zhan, L., & Yang, J. (2020). ICS-Assist: Intelligent Customer Inquiry Resolution recommendation in online customer service for large E-Commerce businesses. In Lecture notes in computer science (pp. 370–385). https://doi.org/10.1007/978-3-030-65310-1_26

[2] Fu, M., Guan, J., Zheng, X., Zhou, J., Lu, J., Zhang, T., Zhuo, S., Zhan, L., & Yang, J. (2020b). ICS-Assist: Intelligent Customer Inquiry Resolution recommendation in online customer service for large E-Commerce businesses. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2008.13534
[3] Teng, L., Liu, Y., Liu, J., & Song, L. (2024). End-Cloud Collaboration Framework for advanced AI customer service in e-commerce. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2410.07122
[4] Taneja, K., Vashishtha, J., & Ratnoo, S. (2024). Transformer based Unsupervised learning approach for imbalanced text sentiment analysis of E-Commerce reviews. Procedia Computer Science, 235, 2318–2331. https://doi.org/10.1016/j.procs.2024.04.220
[5] Zhang, Y., Lau, R. Y. K., Xu, J. D., Rao, Y., & Li, Y. (2024). Business chatbots with deep learning technologies: state-of-the-art, taxonomies, and future research directions. Artificial Intelligence Review, 57(5). https://doi.org/10.1007/s10462-024-10744-z
[6] Li, F., Chen, H., Xu, G., Qiu, T., Ji, F., Zhang, J., & Chen, H. (2020). AliMeKG. AliMeKG, 2581–2588. https://doi.org/10.1145/3340531.3412685
[7] Mangalam, R. (2024). Customer Sentiment Analysis for E-Commerce: A Hybrid approach using CNNs & BERT. International Journal for Research in Applied Science and Engineering Technology, 12(11), 592–596. https://doi.org/10.22214/ijraset.2024.65126
[8] Alshamari, M. A. (2023). Evaluating user satisfaction using Deep-Learning-Based sentiment analysis for social media data in Saudi Arabia's telecommunication sector. Computers, 12(9), 170. https://doi.org/10.3390/computers12090170
[9] Grigoraş, A., & Leon, F. (2023). Transformer-Based model for predicting customers' next purchase day in e-Commerce. Computation, 11(11), 210. https://doi.org/10.3390/computation11110210
[10] Bird, J. J., & Lotfi, A. (2024). Customer service chatbot enhancement with attention-based transfer learning. Knowledge-Based Systems, 301, 112293. https://doi.org/10.1016/j.knosys.2024.112293