Analysing the performance of word embeddings on model performance: a sentiment analysis of Malaysian restaurants customers reviews

1 Introduction

In recent years, the number of individuals opting to dine at restaurants has seen a substantial rise owing to increased financial capacity, diverse food preferences, easily accessible delivery services, and enhanced lifestyles. As a result, the restaurant industry has gained noteworthy popularity in recent times. The surge in digital content related to restaurants and food on the internet has led to a growing inclination among people to consult reviews before patronizing any dining establishment. Therefore, the practice of evaluating restaurants based on written comments has become a widespread occurrence (Sharif et al., 2019; Junaid et al., 2022).

Textual customer review data continues to experience significant growth annually and this can be attributed to the easy accessibility of internet services and the adoption of social networking platforms as a means of communication. To choose a restaurant, potential customers often find themselves sifting through a plethora of feedback, a task that involves examining numerous reviews to gauge the quality of the restaurant and its services. This phenomenon arises from individuals expressing their opinions, emotions, and sentiments about products or services through tweets, Facebook posts, status updates, blog articles, and reviews. The profound impact of customer reviews becomes apparent when understanding their role in gauging customer attitudes toward restaurants. Consequently, the automated analysis of sentiment from food services reviews holds numerous advantages for the stakeholders of food, beverages and restaurants services, enabling them make informed decisions which can help improve their products and services delivery (Marine-Roig & Clave, 2015; Sharif et al., 2019; Hossain et al., 2020a; Zahoor et al., 2020; Başarslan & Kayaalp, 2021; Li et al., 2021).

For these reviews shared online, the sheer volume of feedback presents a challenge for stakeholders who wish to comprehensively assess opinions and product quality. Therefore, Artificial Intelligence (AI) technologies and techniques such as sentiment analysis offer an automated solution for categorizing emotional tones within review content. This technique involves extracting insights from text, aiming to transform vast, unorganized datasets into tangible sentiment indicators (e.g., happy, sad, or neutral). These acquired insights can be condensed into summarized viewpoints, represented through numerical data or graphical representations. As a result, individuals with vested interests, such as managers or customers, can efficiently and swiftly access the necessary information. This underscores the motivating concept behind sentiment analysis, fostering heightened enthusiasm for this area of research (Alamoudi & Alghamdi, 2021; Matlatipov et al., 2022).

The growing significance of customer reviews is evident as both individuals online and businesses alike now prioritize them. When making purchasing decisions, online consumers factor in past customer experiences. Simultaneously, the perception of the public regarding businesses holds a crucial role in marketing strategies, unlocking novel prospects, and predicting sales trends. Furthermore, corporate leadership relies on evaluating this digital feedback to gauge the extent of customer contentment (Alamoudi & Alghamdi, 2021).

The rapid evolution of e-commerce has given rise to a wealth of consumer-generated reviews. These reviews hold immense value from both economic and societal standpoints, as sentiment analysis applied to product reviews enables us to distill user sentiments. The insights thus derived not only assist potential buyers in their decision-making process but also furnish manufacturers with the strengths and weaknesses of their offerings, facilitating further enhancements in their products or services (Zhao et al., 2021).

1.1 Study objectives and questions

Accurately predicting the sentiments in reviews shared by customers who visit restaurants in Malaysia is the core objective of this study and in addition to this, the study will explore other objectives including:

- Comparing the performance of the Bi-LSTM model with and without Word2Vec and Glove word embedding techniques.
- Comparing the performance of RNN models with various word embedding techniques with the BERT transformer model.

2 Background

A variety of techniques have been considered by several researchers who have investigated the adoption of Machine and Deep Learning methods to extract insights from textual reviews shared by customers. Using ML techniques such as Naïve Bayes Sharif et al. (2019) analysed the sentiments from a dataset of 1,000 Bengali-authored reviews of restaurants with the goal of understanding if the model can be used to extract sentiments from these reviews. Their experiments achieved relatively good results with accuracy score of 80.5%. In their research, Zahoor et al. (2020) also used ML techniques such as Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM). Their analysis was based on a dataset size of 4,000 reviews collected for restaurants in Karachi. The NB model achieved a classification accuracy score of 91% and this was higher compared to the result obtained by Sharif et al. (2019), however, the RF model achieved the a high accuracy score of 95% and outperformed all other models used by the researchers.

Recently, advancements in Deep Learning have spurred research using Recurrent Neural Networks (RNN) and transformer-based models. In their paper, Hossain et al. (2020a) used the Bi-LSTM model to classify reviews with a dataset size of 8,425 observations. The researchers aimed to classify the data into negative and positive sentiments and their model recorded 91.4% accuracy. Junaid et al. (2022) conducted experiments using eight models of LR, RF, NB, Decision Tree (DT), SVM, GRU, RNN and LSTM and experimented with five feature extraction techniques. For the DL models, the feature extraction techniques used were Glove and Word2sequence. The analysis was implemented on a dataset containing 1,100 reviews collected from various platforms for reviews in Bangla language and manually labelled. Of the three DL models used, the LSTM with Word2Sequence achieved the best models of 90.9%.

Bhuiyan et al. (2020) evaluated the impact of adding an attention mechanism to the CNN model comparing the results of their proposed model with that of CNN and LSTM. The study was based on experiments conducted using data collected from Foodpandas and with a size of 1,600 reviews. The models used for the analysis also included Word2Vec models with 50 dimensions. The proposed CNN + Attention mechanism model achieved the best accuracy score 98.5%, outperforming the LSTM and CNN models. Hossain et al. (2020b) conducted a similar experiment however the researchers used the CNN + LSTM model applied on a dataset containing 1,000 reviews and extracted from Shohoz and Foodpandas. The researchers also applied the Word2vec model, however they used a vector size of 300 dimensions. Their model achieved an accuracy of 75.0%, performing significantly lower compared to the results achieved by the CNN + Attention model proposed by Bhuiyan et al. (2020).

A more popular variation of DL that has gained popularity among researchers in recent years are transformer-based DL models such as the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). Transformer models and it’s variations have been applied to various categories of datasets from reviews about food, services and products. Başarslan and Kayaalp (2021) conducted several experiments using various word embedding techniques which were applied to ML and DL models. For their study, BERT was used as word embedding technique and its results were compared with approaches such as Word2Vec, Glove, TF-IDF and BOW. The models used for the analysis included SVM, NB, CNN, RNN and LSTM. These models and word embedding techniques were applied on two groups of datasets including restaurant reviews consisting of 598,000 observations and movies reviews dataset consisting of 50,000 observations. Across all the models implemented, the BERT model had higher scores among all the embedding techniques used. The LSTM and BERT text representation achieved the highest accuracy and precision scores of 94% each for the movies reviews dataset. For the restaurant reviews the LSTM and BERT technique had a classification accuracy of 89%. Similarly, Mutinda et al. (2023) conducted experiments comparing the performance of different word embedding approaches with their proposed model including a combination of N-grams, sentiment lexicon and BERT. CNN was used as the classifier for the analysis and other word embedding techniques considered included Glove, Word2Vec, BERT, Le-Glove and Le-Word2Vec. The techniques were then deployed on 3 different datasets and included 70k reviews about Amazon, 50k reviews on IMDB and 300k reviews about restaurants from Yelp. The CNN model proposed by the researchers named LeBERT achieved the highest accuracy scores across all 3 datasets with 88.2% for the restaurant's reviews dataset, 86.1% for the IMDB dataset and 82.4% for the Amazon reviews dataset.

3 Methodology

3.1 Restaurant Reviews Data Collection

139,764 customer reviews about Malaysian restaurants were collected from Kaggle (Ng, 2022) to be used for the experiments conducted in this study. These reviews had been originally extracted from Google reviews and TripAdvisor and are for leading Malaysian restaurants based in different cities of the country. These reviews were labelled on a scale of 1 to 5 with 5 being the highest and meaning positive and 1 being the lowest and meaning negative.

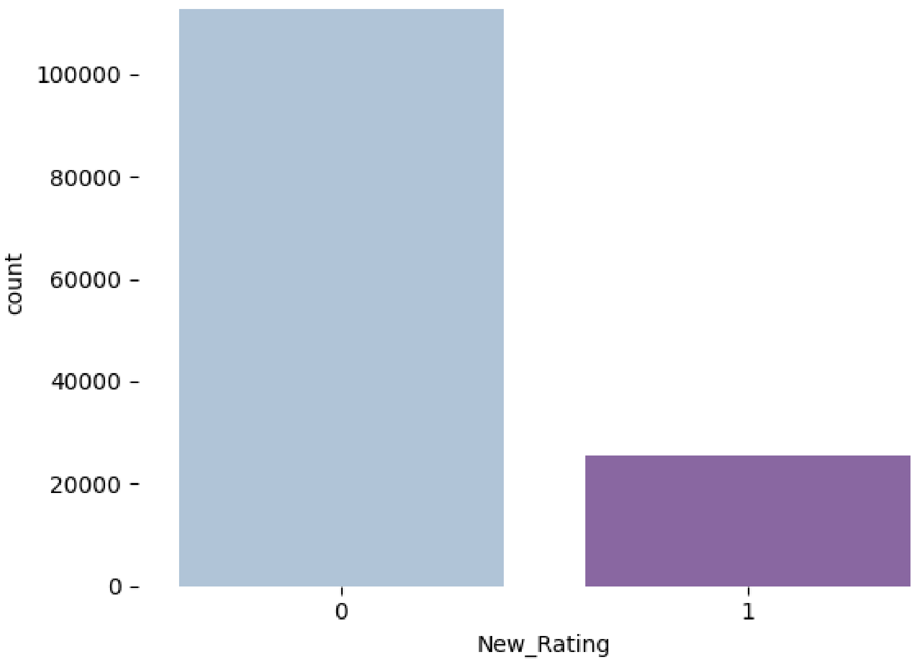
3.2 Text Cleaning

The necessary text cleaning carried was done given consideration for the types of models to be implemented in this study. The cleaning was done to remove errors and noises which might influence the performance of RNN models used for the experiments conducted. In addition, consideration was given to the type of the cleaning required by transformer-based models which typically do not require extensive text preprocessing. For the RNN model, the text cleaning included removing duplicates, expanding contractions, making sure all the text are in lower case format, removing URLS and HTMLs, taking out emojis, special characters, punctuations, numbers and stopwords. After this was carried out, the text was lemmatized which meant returning words to their base forms.

3.3 Remapping Text Labels

Following text cleaning, the size of the dataset was reduced to 138,505 observations. Subsequently, given that the labels in the dataset were between a range of 1 to 5, this was mapped into two labels of positive and negative. Rating scores of between of one to three were labelled as 1 and denotes negative reviews as presented in Figure 1 while rating scores of four and five were labelled as 0 and denotes positive.

Figure 1: Labels in the dataset, with 0 meaning positive and 1 meaning negative.



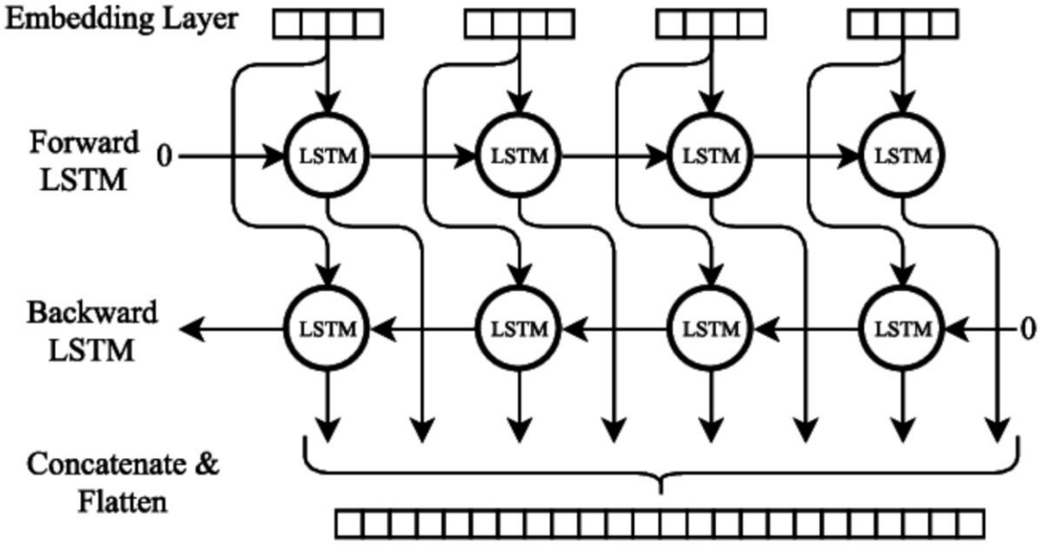
3.4 Modelling

Two DL models were implemented for the experiments in this study. The Bi-LSTM and BERT models were used for the analysis conducted. Furthermore, word embedding techniques of GLOVE and Word2Vec were used for the experiments conducted with the Bi-LSTM model. The sections below provide summarised descriptions of the different techniques.

3.5 Bidirectional Long Short-Term Memory (BI-LSTM)

The Bi-LSTM consists of units of LSTM operating bidirectionally, enabling the assimilation of historical and future contextual data. Bi-LSTM possesses the capacity to capture long-term dependencies without the redundancy of preserving duplicate context data (Liang & Zhang, 2016). The model has shown good performance in addressing sequential modelling challenges and is extensively employed in text classification tasks. The Bi-LSTM network employs dual parallel layers, enabling both forward and reverse passes to capture dependencies within two distinct contexts. (Jang et al., 2020). Figure 2 presents the Bi-LSTM process.

Figure 2: The Bi-LSTM process (Paperswithcode, 2023)



3.6 Word Embeddings

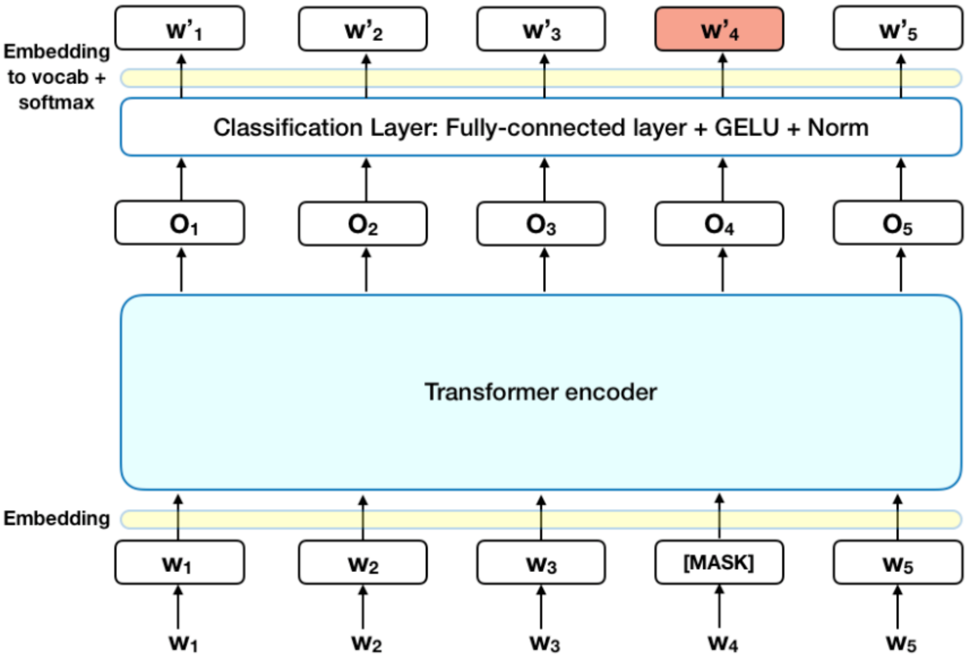
The word embedding techniques used during the experiments are GLOVE and Word2Vec. GLOVE as an algorithm for creating semantic vector space representations dependent on the co-occurrence of words within a context and adopting a count-based modelling approach. In contrast, Word2Vec follows a prediction-based modelling approach. GLOVE's effectiveness stems from its utilization of global matrix factorization and local context window, making it superior to other word embeddings. In addition, GLOVE's superior performance is attributed to its focus on elements other than zero and a portion of the corpus, as opposed to processing the entire corpus or a distinct window within it. This advantage is particularly evident in applications such as word similarity, analogy, and named entity recognition, where GLOVE consistently outperforms its counterparts (Pennington et al., 2014; Mohammed et al., 2021).

Word2Vec stands as a widely used sequence embedding technique, converting natural language into distributed vector representations (Liu, 2017). It has the capability to capture intricate contextual relationships between words within a multidimensional space, making it a commonly employed initial step for predictive models in tasks related to semantics and information retrieval (Jang et al., 2020).

3.7 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is an advanced embedding layer that excels in training deep bidirectional representations from unlabelled document. It accomplishes this by considering the right and left setting within all its layers. BERT undergoes pretraining on substantial unsupervised text data sources, like Wikipedia dumps or Book Corpus, focusing on two primary goals of masking predictions of words whereby 15% of words within an input sequence are masked, and the entire sequence is processed through a deep bidirectional Transformer encoder. Its task is to predict the concealed words, and then it learns sentence relationships by taking two sentences, A and B, as inputs and classifying whether B logically follows A or is randomly paired. Unlike traditional sequential models, BERT's attention architecture concurrently processes the entire input sequence, enabling parallel processing of all input tokens (Horev, 2018; Munikar et al., 2019). The model comprises an encoder featuring 12 Transformer blocks, each equipped with 12 self-attention heads, and boasts a hidden size of 768. When fed with a sequence of no more than 512 tokens, BERT generates a representation of that sequence. Notably, the sequence begins with the special token [CLS], housing a unique classification embedding, and employs [SEP] to separate segments. When tackling the task of classifying text, the model utilizes the last hidden state (h) of the initial [CLS] token as the representation encapsulating the entire sequence (Sun et al., 2019). An example of a BERT architecture is presented in Figure 3.

Figure 3: An example of BERT model architecture



w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

4 Experimental Set-up

The experiments for this study were conducted using Python and Jupyter notebook and was carried out in the Google Colab environment. Following text cleaning and mapping the labels, this cleaned and pre-processed dataset was divided into training dataset of 80% and of which 20% of this was used to validate the models while training, while 20% was used to test the models after training. For the Bi-LSTM models, the dataset was then tokenized whereby sentences are broken down into individual words. After this these tokens are converted into integers with each word replaced as numbers. Furthermore, sequence padding was carried to make sure sequences fed into the model have the same length with the maximum size defined at 500.

The Bi-LSTM model used for experiments with and without the Word2Vec and Glove embedding consisted of a Bidirectional layer with 100 units, with the output layer consisting of a single neuron given the required binary task and using the Sigmoid activation function. The models also included a 20% dropout to help manage overfitting and were trained for 20 epochs. Adam was used as the Optimizer and the models had a batch size of 64. For the model without Glove or Word2vec, the maximum features considered was 5,000 and an embedding length of 32. For the Word2Vec model, 100-dimension vectors were used. Similarly, for the Glove model, the 100-dimension vectors were used as well.

For the BERT model, the word embeddings capture a vocabulary of 30,522 words in 768 dimensions, with the model using a BERT encoder consisting of 12 layers. Following the encoder is the pooling layer which uses a dense linear transformation, followed by a hyperbolic tangent (Tanh) activation function. Finally, the last layer is a linear layer which maps the 768-dimensional vector to a two-dimensional output, considering the binary nature of the of the classification task to be carried out.

4.1 Results Evaluation Measures

The measures used to assess the results after implementing the models include F1-score, precision, accuracy, and recall. Equation 1 shows the formula for accuracy, and it is defined accuracy as the split of precisely classified predictions relative to the sum of all predictions, with values ranging between 0.00 and 1.0, where 1.0 indicates perfect accuracy and 0.00 denotes the lowest achievable accuracy. Equation 2 presents the formula for precision, calculated as the split of precisely classified instances within a class label to all the instances classified as that class. Equation 3 presents recall, as the number of classified samples that are correct relative to the overall total actual samples for that class. Finally, Equation 4 presents the formula for F1-score. F1-Score is computed as harmonic average capturing recall and precision, and gives model performance a measure of balance (Mutinda et al., 2023).

$$Accuracy = \frac{Correct\ prositve\ and\ negative\ predictions}{All\ samples} \qquad \textbf{Equation 1}$$

$$Precision = \frac{Correct\ positive\ predictions}{All\ positive\ samples} \qquad \textbf{Equation 2}$$

$$Recall = \frac{Correct\ positive\ predictions}{Correct\ positive\ predictions + Wrongly\ predicted\ negatives} \qquad \textbf{Equation 3}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \textbf{Equation 4}$$

5 Results and Discussion

Four experiments were conducted using RNN and transformer models. These models were implemented to address the core goal of this study which is to precisely classify reviews shared by customers of restaurants across different cities of Malaysia. In addition to this goal, the study aimed to answer questions including whether word embedding techniques such as Word2Vec and Glove have any impact on improving RNN model performance, and whether BERT transformer model is better at classifying reviews when compared to RNN models.

5.1 Model Results

All 3 Bi-LSTM models, including models embedded with Word2Vec and Glove, achieved accuracy scores of 91%. However, they record varying precision recall and F1-scores. Considering the dataset used for this analysis is imbalanced, F1-score with its formula described in Equation 4 helps provide a balanced measure for evaluating the model. Of the 3 Bi-LSTM models, the Bi-LSTM + Glove embeddings achieved the highest F1-score of 85%. As seen in Table 1, the BERT model outperformed all the models in this study achieving the best accuracy score of 94% with the model also recording the highest F1-score of 90%.

Table 1: Results of the Bi-LSTM and BERT

Models for current study	Accuracy	Precision	Recall	F1-Score
Bi-LSTM Model	91%	87%	82%	84%
Bi-LSTM with GLOVE embeddings	91%	85%	85%	85%
Bi-LST with Word2Vec embeddings	91%	87%	82%	84%

BERT	94%	92%	88%	90%

When the results of the models across the different reviews classes are highlighted the BERT model also outperformed the other models when precision is considered, achieving a precision of 95% and 89% for the positive and negative classes respectively. Also, when recall is highlighted, as seen in Table 2, the BERT model also performed better than the other models with 98% recall score for the positive class and 79% for the negative class. This is validated by the results of the F1-score wherein the BERT model scored 97% and 84% for the positive and negative classes respectively. Strictly looking at the RNN models, when F1-Score is considered, the Bi-LSTM + Word2Vec achieved the best results for the positive class with a score of 95% vs the 94% each for the Bi-LSTM, and Bi-LSTM + Glove models. However, for the negative class, the Bi-LSTM + Glove embeddings achieved the best result of 75%.

Table 2: Bi-LSTM and BERT models results for reviews classes.

Model	Class	Precision	Recall	F1-Score
Bi-LSTM Model	Positive	93%	96%	94%
	Negative	80%	67%	73%
Bi-LSTM with GLOVE embeddings	Positive	94%	94%	94%
	Negative	75%	75%	75%
Bi-LST with Word2Vec embeddings	Positive	93%	86%	95%
	Negative	81%	68%	74%
BERT	Positive	95%	98%	97%
	Negative	89%	79%	84%

Taking in to a broader context and compared side by side with other works aimed at classifying reviews shared about restaurants, the results of the BERT model developed in this study outperforms the results of ML and DL models developed by Sharif et al. (2019) whose model achieved an accuracy of 80.5% and Junaid et al. (2022) whose LSTM + Glove recorded an accuracy of 87.5%. Also, the Bi-LSTM + Glove mode in this performed the model by Junaid et al. (2022). On the other hand the RF model implanted by Zahoor et al. (2020) achieved an accuracy score of 95% which is slightly better than the 94% achieved in this study. For other studies where the BERT embedding was used, the BERT model in this study outperforms the results achieved by Başarslan and Kayaalp (2021) whose LSTM and BERT model recorded an accuracy score of 89% for restaurant reviews.

6 Conclusion

This study was aimed at classifying customer reviews of restaurants across various cities in Malaysia, and a total of four experiments were conducted employing both RNN and transformer models. All three Bi-LSTM models implemented achieved accuracy score of 91%, however, the Bi-LSTM + Glove had an F1-score of 85% which was the highest, outperforming the other Bi-LSTM models. Overall, the BERT model achieved the best accuracy scores of 94% with an F1-Score of 90%. In conclusion, the research highlights the effectiveness of word embedding techniques in enhancing RNN model performance for sentiment analysis in restaurant reviews. However, it underscores the potential of the BERT transformer model, which performed better than the other models in terms of all 4 performance evaluation metrics assessed. For future works, other text preprocessing techniques such as stemming should be considered, in addition to variations in embedding vector sizes for the word embedding techniques used.

7 References

Alamoudi, E. S. & Alghamdi, N. S. (2021) Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *Journal of Decision Systems*, 30(2-3), 259-281.

Başarslan, M. S. & Kayaalp, F. (2021) Sentiment analysis on social media reviews datasets with deep learning approach. *Sakarya University Journal of Computer and Information Sciences*, 4(1), 35-49.

Bhuiyan, M. R., Mahedi, M. H., Hossain, N., Tumpa, Z. N. & Hossain, S. A. (2020) An Attention Based Approach for Sentiment Analysis of Food Review Dataset, *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 1-3 July 2020.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, 1810.04805.

Horev, R. (2018) *BERT Explained: State of the art language model for NLP*. Available online: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 [Accessed 31/08/2023].

Hossain, E., Sharif, O., Hoque, M. M. & Sarker, I. H. (2020a) Sentilstm: a deep learning approach for sentiment analysis of restaurant reviews, *International Conference on Hybrid Intelligent Systems*. Springer.

Hossain, N., Bhuiyan, M. R., Tumpa, Z. N. & Hossain, S. A. (2020b) Sentiment analysis of restaurant reviews using combined CNN-LSTM, *2020 11th International conference on computing, communication and networking technologies (ICCCNT)*. IEEE.

Jang, B., Kim, M., Harerimana, G., Kang, S.-u. & Kim, J. W. (2020) Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Applied Sciences*, 10(17), 5841.

Junaid, M. I. H., Hossain, F., Upal, U. S., Tameem, A., Kashim, A. & Fahmin, A. (2022) Bangla food review sentimental analysis using machine learning, *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE.

Li, L., Yang, L. & Zeng, Y. (2021) Improving sentiment classification of restaurant reviews with attention-based bi-GRU neural network. *Symmetry*, 13(8), 1517.

Liang, D. & Zhang, Y. (2016) AC-BLSTM: asymmetric convolutional bidirectional LSTM networks for text classification. *arXiv preprint arXiv:1611.01884*.

Liu, H. (2017) Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177*.

Marine-Roig, E. & Clave, S. A. (2015) A method for analysing large-scale UGC data for tourism: Application to the case of Catalonia, *Information and Communication Technologies in Tourism 2015: Proceedings of the International Conference in Lugano, Switzerland, February 3-6, 2015*. Springer.

Matlatipov, S., Rahimboeva, H., Rajabov, J. & Kuriyozov, E. (2022) Uzbek sentiment analysis based on local restaurant reviews. *arXiv preprint arXiv:2205.15930*.

Mohammed, S. M., Jacksi, K. & Zeebaree, S. (2021) A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 22(1), 552-562.

Munikar, M., Shakya, S. & Shrestha, A. (2019) Fine-grained sentiment classification using BERT, *2019 Artificial Intelligence for Transforming Business and Society (AITB)*. IEEE.

Mutinda, J., Mwangi, W. & Okeyo, G. (2023) Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network. *Applied Sciences*, 13(3), 1445.

Ng, C. K. (2022) *Malaysia Restaurant Review Datasets: Malaysia restaurant reviews collected from Google reviews and TripAdvisor*. Available online: https://www.kaggle.com/datasets/choonkhonng/malaysia-restaurant-review-datasets [Accessed 11/08/2023].

Paperswithcode (2023) *Bidirectional LSTM*. Available online: https://paperswithcode.com/method/bilstm [Accessed 24/08/2023].

Pennington, J., Socher, R. & Manning, C. D. (2014) Glove: Global vectors for word representation, *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Sharif, O., Hoque, M. M. & Hossain, E. (2019) Sentiment analysis of Bengali texts on online restaurant reviews using multinomial Naïve Bayes, *2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)*. IEEE.

Sun, C., Qiu, X., Xu, Y. & Huang, X. (2019) How to fine-tune bert for text classification?, *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*. Springer.

Zahoor, K., Bawany, N. Z. & Hamid, S. (2020) Sentiment analysis and classification of restaurant reviews using machine learning, *2020 21st International Arab Conference on Information Technology (ACIT)*. IEEE.

Zhao, N., Gao, H., Wen, X. & Li, H. (2021) Combination of convolutional neural network and gated recurrent unit for aspect-based sentiment analysis. *IEEE Access*, 9, 15561-15569.