Student Name: Kelvin Obruche Idogun

Student Number: 202250185

Supervisor: Dr. Lawrence Bilton

Course: Msc Project Thesis

Sentiment Analysis of Customer Reviews: A Comparative Study of BERT, RNN, and LSTM using the model evaluation metrics for Business Performance Evaluation

Abstract

This study delves into examining how effective different deep learning models, Bidirectional Encoder Representations from Transformers (BERT), Long Short-term Memory Network (LSTM) and the Simple Recurrent Neural Network (RNN) are when it comes to evaluating sentiments through Customer review texts for the purpose of estimating business performance. By combining data from Trustpilot and an out of domain (OOD) dataset from Kaggle, the study investigates how well these models can classify sentiments into negative, neutral, and positive categories by measuring precision, recall and f1 scores. The BERT Model with its prediction accuracy of 96% stands out as compared to the Prediction accuracy values of the LSTM and RNN models which obtained metrics of 90% and 88% respectively. Despite the largely imbalanced nature of the dataset with an approximate 74% of the total reviews being rated as positive, the BERT model showed great prowess in predicting the minority classes as seen from the precision, recall and f1 scores with an improvement by 19%, 23% and 21% respectively when compared with the combined minority class metrics obtained from the classification report of the other two models. Although the LSTM and RNN models show decline in performance, they demonstrate promising potential in detecting sentiments. However, the susceptibility to overfitting during validation highlights the complexity associated with sentiment analysis. This research underscores the significance of adapting models to contexts while emphasizing the trade-off between accuracy and generalization in real world applications of sentiment analysis.

1.0 Introduction

The success of a product can be greatly impacted by customer reviews posted online, sentiment analysis has become a vital tool for both businesses and researchers. This computational technique, which involves analysing and categorizing emotions conveyed in text has revolutionized how organizations comprehend consumer behaviour and preferences. Online reviews serve as a platform for individuals to express their emotions and experiences ranging from happiness and satisfaction to frustration and disappointment. They offer insights for potential buyers who rely on these opinions to make informed decisions about their purchases (Zhang, et al., 2014).

The rise of social media platforms and online forums has led to an exponential growth in user generated content that serves as a valuable source of data for sentiment analysis. Additionally, word of mouth channels like user generated reviews on e-commerce websites and microblogging sites have gained popularity. Prominent e-commerce platforms such as Amazon and Trustpilot have implemented systems for customers to share their experiences with products. This shift in consumer behaviour has had an impact, on their purchasing decisions (Chevalier & Mayzlin, 2006). According to a study conducted by (Liu & Zhang, 2012), 80% of data are unstructured and consists mostly of text. This presents an opportunity for sentiment analysis to extract valuable insights from this vast amount of data, which traditional methods struggle to handle. These traditional methods include but not limited to the manual review and analysis of reviews, the use of surveys and questionnaires e.t.c.

The motivation behind this study is rooted in the evolving landscape of sentiment analysis methodologies. Initially, sentiment analysis relied heavily on rule-based and lexicon approaches, which, while effective in certain contexts, showed limitations in handling the nuances and complexities of human language (Ahmad & Edalati, 2022). The advent of machine learning and deep learning techniques brought significant progress to this field. Models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks demonstrated abilities in capturing contextual information in text resulting in more accurate sentiment predictions.

However, the recent introduction of Bidirectional Encoder Representations from Transformers (BERT) has brought about a revolution in sentiment analysis due to its performance in natural language processing tasks. BERTs exceptional capability to comprehend the meaning of a word based on its relationship with words in a sentence (Devlin, et al., 2018) using a transformer architecture has established new standards in the field.

This study aims to conduct an analysis of these advanced approaches BERT, RNN and LSTM specifically, as a way of assessing business performance through customer reviews. The data set will be extracted from Trust pilot and the model's ability to generalize will be evaluated using an out of domain dataset obtained from Kaggle. By examining their effectiveness and accuracy in real world situations this study seeks to contribute insights for both academic research and practical applications, in various domains advancing the field of sentiment analysis techniques.

2.0 Background/Literature Review

Sentiment analysis at its core involves using methods to study the opinions, sentiments and emotions expressed in text. Its objective is to understand how speakers or writers feel about a topic or the overall sentiment conveyed in a document (Anjaria & Guddeti, 2014). The importance of sentiment analysis extends across fields, including business intelligence and social science research. As described by (Pang & Lee, 2008) sentiment analysis helps us comprehend the subjective aspects of textual data but also enables us to predict market trends, political outcomes, and public opinions.

The applications of sentiment analysis are diverse and widespread. In the business domain, it is employed to assess customer sentiments towards products or services. By doing so, it informs marketing strategies and product development initiatives (Liu, et al., 2020). For example, a study conducted by (Hu & Liu, 2004) showcased how sentiment analysis can effectively mine and analyse customer reviews to extract product features and summarize opinions. Their approach involved using part of speech tagging to identify adjectives in customer reviews, as indicators of sentiment. They then extracted opinion words and determined their semantic orientation using WordNet. Although their approach was innovative at the time, they had some limitations in capturing the full range of customer opinions.

In the field of politics and public policy, sentiment analysis helps us understand opinion on social issues. This was demonstrated by (Tumasjan, et al., 2010), who used Twitter data to analyse sentiments and predict election outcomes. They examined over 100,000 tweets related to the German federal election focusing on how often political parties and politicians were mentioned, as well as analysing the sentiment expressed in those tweets. Interestingly, they discovered a correlation between the number of tweets mentioning a party and that party’s success in the elections.

The introduction of machine learning brought about a positive change in sentiment analysis. Unlike rule-based systems, machine learning approaches learn from data improving their accuracy over time. As mentioned in a research study conducted by (Pang, et al., 2002), sentiment classification has been effectively carried out using machine learning models, like Naïve Bayes, Support Vector Machines (SVM) and Random Forests. However, these models often require extensive feature engineering and may not fully capture the sequential nature of text.

2.1 Deep Learning Advances: RNNs and LSTMs

The field of sentiment analysis has seen progress with the emergence of deep learning techniques such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM). According to a study by (Zhang, et al., 2018), these models excel at handling sequences and capturing long-range dependencies in text. This makes them particularly suitable for sentiment analysis as they can better grasp the context and subtleties of language compared to machine learning models. It is worth noting that these models can be more complex and computationally demanding especially when dealing with large datasets. A recent study by (Singh, et al., 2022) highlighted this challenge while proposing an LSTM-RNN based network with attention layers that improved feature weighting for sentiment analysis specifically focused on COVID 19 related tweets. Their approach showed improved performance in accuracy and precision compared to machine learning models like SVMs and Random Forests. However, a comprehensive understanding of the models generalization capabilities can be gained by evaluating its performance across domains.

2.2 BERT and Transformers: A Revolutionary Change.

In a research project conducted by (Biswas, et al., 2020), they explored the impact of BERT and Transformers. They compared BERT with RNN and showcased its superiority in terms of precision and recall, for both negative and positive sentiments in the Software Engineering domain. BERTs remarkable ability to understand the nuances of context makes it a powerful tool for sentiment analysis although it may pose resource related challenges due to its demanding requirements.

2.3 Comparative Analysis of Sentiment Analysis Methods.

In sentiment analysis studies, researchers often compare variations of models to establish performance benchmarks. (Biswas, et al., 2021) conducted a study involving five algorithms (Multinomial Naive Bayes, Support Vector Machine, Hidden Markov Model, Long Short-Term Memory, and BERT) for sentiment analysis purposes. This study assessed these models based on factors like training time, prediction time and accuracy levels. The findings indicated that while deep learning models such as LSTM and BERT offer better accuracy levels, they require more training time. This research provides insights into the trade-offs that arise when considering approaches to sentiment analysis. It emphasizes the balance between computational resources and model performance. These findings also align with the trend in sentiment analysis, which favors deep learning methods known for their efficient handling of complex language patterns and improved accuracy.

This study takes inspiration from the works of (Biswas, et al., 2020) and (Biswas, et al., 2021). While (Biswas, et al., 2020) focused on utilizing BERT for sentiment analysis in the software engineering domain, demonstrating its superiority over RNN models, (Biswas, et al., 2021) conducted a broader comparative analysis of various algorithms including BERT, LSTM, and others, evaluating them on aspects like training time and accuracy. This research diverges in its specific focus on customer reviews in the business/product context, aiming to provide a deeper understanding of consumer sentiments using these advanced models which in turn could be extrapolated to the performance of a product or business entity.

3.0 Methodology

Using real world data from platforms like Trustpilot[1] is crucial for several reasons. Firstly, it ensures that the analysis of sentiments accurately reflects the experiences and perceptions of consumers. Secondly it brings in a range of linguistic expressions, including both formal feedback and informal endorsements or criticisms which enhances the reliability of sentiment analysis. Twenty-six (26) businesses in the Money and insurance section were selected and their review pages numbered 1- 500 were scraped each. The data collection process aimed not to gather qualitative data but also to convert it into measurable metrics for sentiment analysis.

3.1 Data Collection Process

The primary tool utilized for web scraping was Beautiful Soup[2], a Python library renowned for its ability to pull data out of HTML and XML files. It provides simple methods for navigating, searching, and modifying the parse tree, making it an ideal tool for the task. Together with the requests[3] library, which was used to make HTTP requests to the review pages, these tools formed the foundation of the data scraping architecture.

The selection criteria for the businesses on Trustpilot were based on factors such as the number of available reviews and the average overall ratings. The scraping procedure was automated to allow for efficient management of the large data involved in this process.

The script was designed to go through the pages of the review website one after the other accessing and collecting data, also adhering to the robots.txt files by maintaining a reasonable request rate to avoid overwhelming the websites server and periodically refreshing sessions to prevent detection of non-human activity. These measures allowed for the collection of a dataset while upholding ethical scraping standards. The collected data includes metrics essential for sentiment analysis such as review titles, dates, ratings, and the actual reviews themselves. This comprehensive dataset served as the foundation for the subsequent research analysis.

[1] Trustpilot
[2] Beautiful Soup Documentation
[3] Requests Library Documentation

3.2 Data Cleaning & Preprocessing

A thorough cleaning process was implemented to improve dataset quality for further analysis. This involved dropping duplicate entries; and instances with missing critical information like review texts were replaced with their corresponding review titles. Techniques such as forward and backward filling were used to maintain data continuity—especially regarding temporal attributes, like review dates and dates of experience which had missing entries.

Additionally, the data was subjected to standardization by converting all review text to lowercase and transforming emojis into text format. This step played a role in addressing any potential biases that may arise due to differences in text formatting. Regular Expressions (Regex) were used to clean up the textual data by removing unnecessary characters that could potentially affect the results of sentiment analysis.

Lemmatization was another key preprocessing step, streamlining the textual data by reducing words to their base or dictionary form. This not only aids in the consolidation of various forms of a word but also significantly contributes to the efficiency of the sentiment analysis, as it trims down the complexity of the data (Plisson, et al., 2004).

To measure the sentiment conveyed in the reviews, a sentiment polarity score was computed for each entry using the Textblob[4] library. Based on the predefined threshold existing within textblob, this score was then categorized into three labels: neutral, negative, and positive. These labels formed the basis for the quantitative evaluation of consumer emotions and opinions during sentiment analysis.

Furthermore, the preprocessing phase included transforming the review texts into a numerical format suitable for deep learning models. The text was tokenized, sequences were generated, and padding was applied to ensure consistent input sizes. Class imbalance was addressed through random oversampling, resulting in a fair representation of all sentiment classes. Figure 1 below shows the distribution of the sentiment labels before the random oversampling technique was employed.

[4] Text blob Documentation

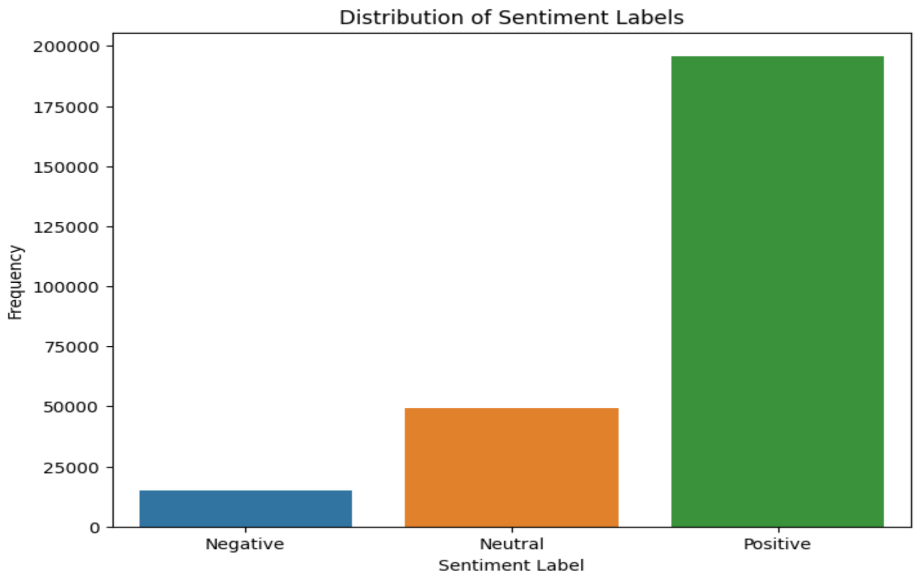


Figure 1: Distribution of Sentiment Labels

By following these steps, a refined dataset was formed, and it created a foundation, for accurate sentiment classification using the deep learning techniques. Figure 2 below shows a graphical flow of the several steps taken to prepare the Dataset obtained.

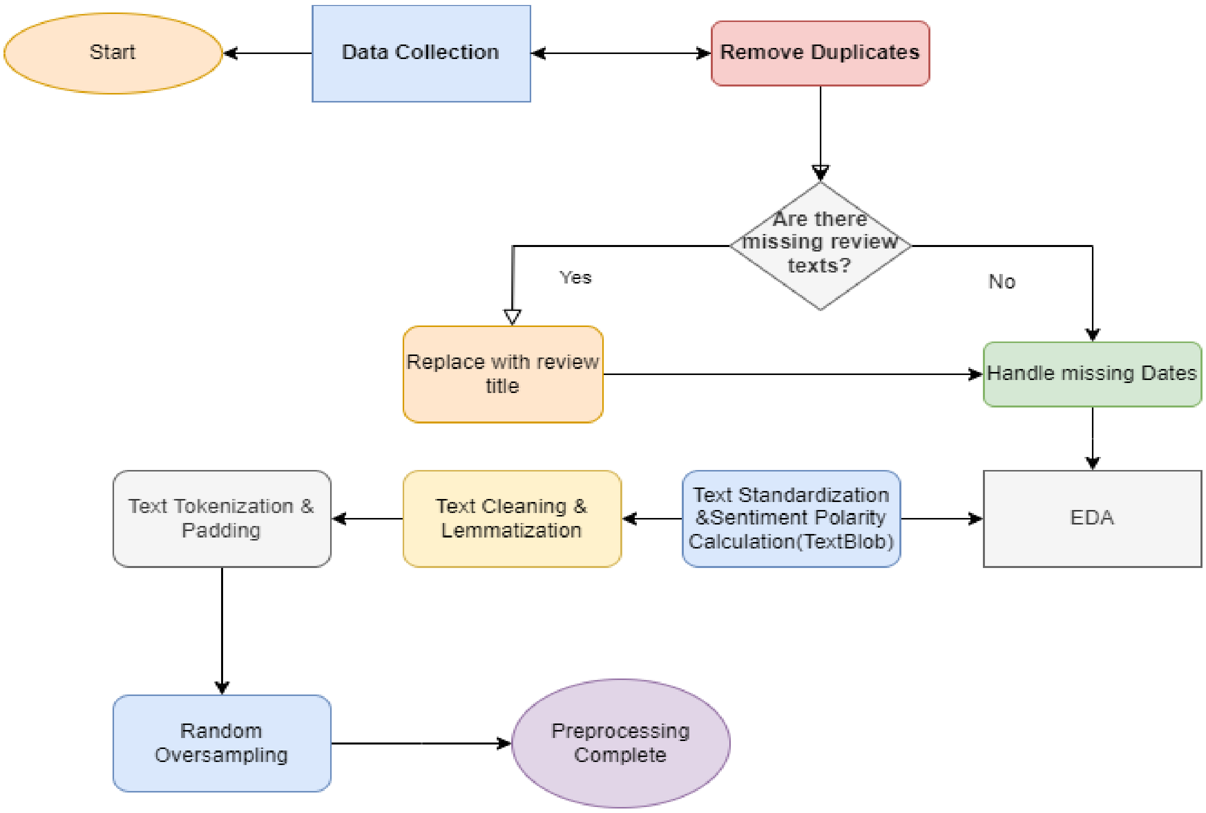

Start
Data Collection
Remove Duplicates
Are there missing review texts?
Yes
No
Replace with review title
Handle missing Dates
EDA
Text Standardization &Sentiment Polarity Calculation(TextBlob)
Text Cleaning & Lemmatization
Text Tokenization & Padding
Random Oversampling
Preprocessing Complete

Figure 2: A flowchart of the data handling procedure

3.3 **BERT Model**

BERT's architecture as described by (Devlin, et al., 2018) allows it to understand the context of a word in a sentence by looking at the words that come before and after it, which is quite different from other models that process words in order one after the other. The bi-directional training of BERT is its distinguishing characteristic enabling it to grasp language nuances effectively.~~.~~

This implementation involved using BERT's pre-trained model bert-base-uncased[5], which has been trained on an extensive dataset comprising ~~(~~‘Wikipedia’ and the ‘Book Corpus’~~)~~ and can understand the English language in its lowercase form. The BERT tokenizer was employed to convert text data into a format that the BERT model can understand. This process involved splitting the text into tokens, converting these tokens into IDs that BERT is familiar with, and creating the necessary attention masks to ignore padding in the input sequences.

The BERT model was fine-tuned to classify sentiments into three categories: negative, neutral, and positive which involved creating a feature from each review text that included input IDs, attention masks, and the sentiment label, which was converted into an integer format. The training data were oversampled to address class imbalance and ensure that the model is exposed to a fair distribution of sentiment labels during training.

Throughout this process, the validation loss was constantly monitored and early stopping mechanisms had already been implemented to combat overfitting, also saving the best model weights.

[5] bert-base-uncased

3.4 **LSTM Model**

The Long Short-Term Memory (LSTM) model, a type of recurrent neural network (RNN), has made significant strides in handling sequences and time series data. LSTMs were specifically designed to overcome the limitations of RNNs when it comes to capturing long range dependencies in data. This makes them ideal for tasks like sentiment analysis on data, where understanding the context is vital (Li & Qian, 2016).

In this study, the LSTM model was defined using the Keras Sequential API, indicating a linear stack of layers. The model architecture included an Embedding layer to convert text tokens into dense vectors of fixed size, followed by an LSTM layer with 64 units. The LSTM layer captured the sequential dependencies in the text data. A Dropout layer was added with a rate of 50% to prevent overfitting by randomly setting a fraction of input units to 0 at each update during training. The last Dense layer uses a SoftMax activation function to generate the probability distribution, for the three sentiment classes: Negative, Neutral and Positive.

The model was compiled with the Adam optimizer, which is known for its ability to adjust the learning rate as needed. Also, to handle class classification tasks, categorical cross entropy was chosen as the loss function. Early stopping was implemented to stop training when the validation loss no longer decreased as a measure to combat overfitting. The best-performing model was saved and evaluated on a test set.

The LSTM model was also subjected to K-fold cross-validation to validate its robustness and generalizability. This involved dividing the dataset into K consecutive folds and ensuring each fold was used once as a validation set while the k-1 remaining folds formed the training set. This process was essential in assessing the model's performance across different subsets of data, providing a comprehensive view of its predictive capabilities.

3.5 **RNN Model**

The Recurrent Neural Network (RNN) model shares several setup similarities with the previously discussed LSTM model. RNNs are renowned for their ability to process sequential data, making them highly suitable for text analysis tasks like sentiment analysis (Lipton, et al., 2015). The RNN model was constructed using a Sequential model from Keras, beginning with an Embedding layer to convert text tokens into dense vectors. This approach is standard in text processing (Goodfellow, et al., 2016). The core of the RNN model was a SimpleRNN layer with 64 units, analogous to the LSTM's setup but with simpler internal dynamics. This layer captures the temporal characteristics of the text data. To combat overfitting, a Dropout layer with a 50% rate was employed, mirroring the LSTM model's approach. In terms of compilations and call backs, the method applied in the LSTM model was replicated here.

These techniques are widely recognized for improving neural network training outcomes. (Goodfellow, et al., 2016)

3.6 **Out of Domain Testing**

To assess the generalizability and robustness of the sentiment analysis models (BERT, LSTM, and RNN), an out-of-distribution testing strategy was adopted, using a distinct dataset obtained from Kaggle[6], specifically targeting hotel reviews. This approach aligns with the recommendations by (Teney, et al., 2020) on the importance of external validation to evaluate model performance in diverse contexts. The dataset was processed and classified based on the ratings provided, employing a mapping system like the one used for the primary dataset. Afterwards, the trained models were employed to predict sentiments on this new dataset. The predictions were then compared with the actual ratings to calculate the accuracy. This step was crucial in examining the models' adaptability and accuracy in varying scenarios, thereby offering a comprehensive assessment of their applicability in real-world situations beyond the initial training domain (Teney, et al., 2020).

[6] Kaggle Dataset

4.0 Results & Discussion

4.1 **BERT Model**

In this section, we examine the performance of the BERT model trained using a Tensor processing unit (TPU) on Google Colab to leverage accelerated computing, reducing training time significantly. The model's evaluation on the validation set produced a strong performance across the precision, recall and F1 score as seen in Table 1 below. The overall accuracy stood at 96%, indicating high predictive power. The macro-average precision, of 91%, demonstrated a balanced performance across classes despite class imbalances.

Table 1: BERT model Classification report

	Precision	**Recall**	**F1-score**
Negative	0.84	0.94	0.89
Neutral	0.89	0.91	0.90
Positive	0.99	0.98	0.98
Accuracy **OOD Accuracy** **Training Time**	0.96 0.77 4 hrs and 13 min		

The confusion matrix shown below in Table 2 further solidifies the model's robustness, showing a high true positive rate for each class, especially for positive sentiments.

Table 2: BERT model Confusion Matrix

		Predicted Label		
		Negative	*Neutral*	*Positive*
True Label	*Negative*	**4099**	244	16
	Neutral	760	**13416**	640
	Positive	38	1343	**57414**

The model accuracy and loss curves shown in Figure 3 below indicate good convergence behaviour, with validation accuracy closely following the training accuracy, and a consistent decrease in validation loss over epochs. These curves suggest that the model generalizes well and is not overfitting.

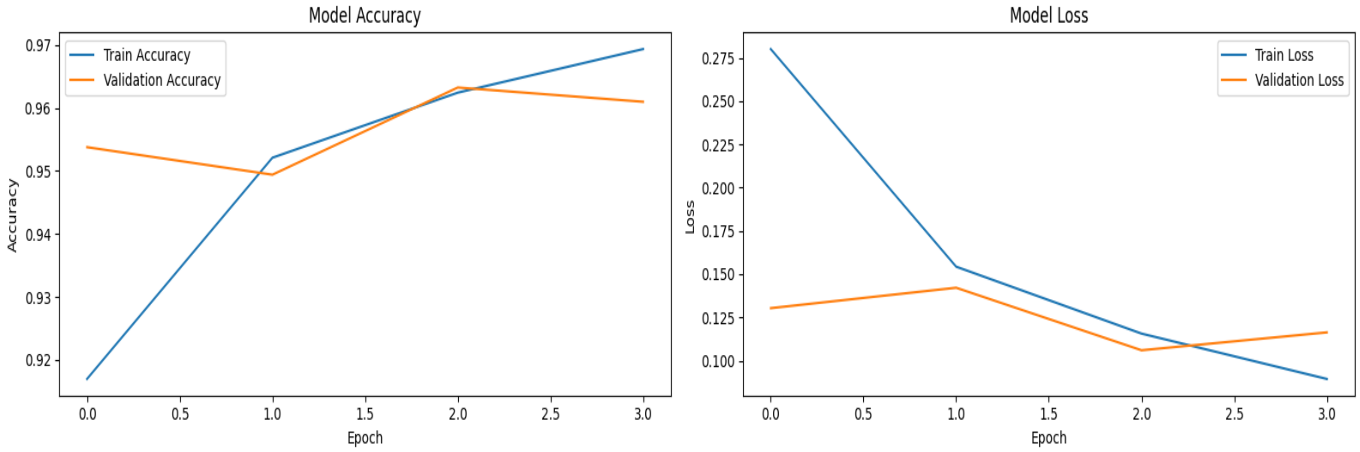


Figure 3: BERT Training Curves

Moreover, the ROC (Receiver Operating Characteristic) curves shown in figure 4 below for all classes are near the ideal point with an area of 1.00 for classes 0 and 2, and 0.99 for class 1, which suggests excellent separability of the classes by the model.

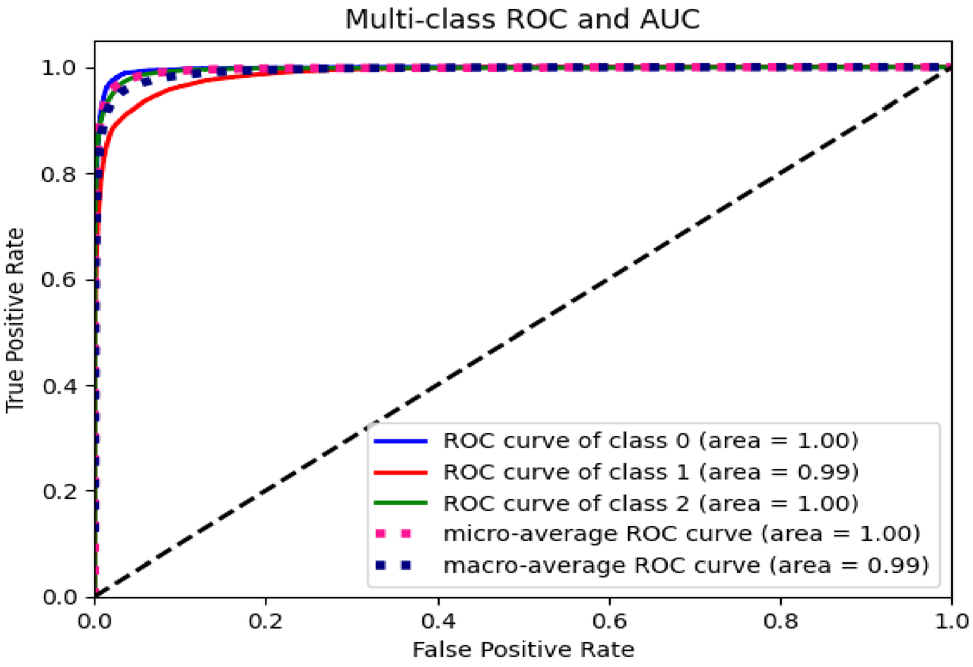


Figure 4: BERT MODEL ROC CURVE

This empirical evidence supports the effectiveness of BERT for sentiment analysis tasks, aligning with findings by (Devlin, et al., 2018) who introduced BERT and demonstrated its state-of-the-art performance on multiple NLP tasks. Upon its application to the OOD testing, the model demonstrated commendable generalizability, with a tendency to accurately identify positive sentiments, as indicated by the substantial number of true positives shown in Table 3 below. However, it also faced challenges with false negatives and false positives, particularly in distinguishing between negative and neutral sentiments.

Table 3: BERT model OOD Confusion Matrix

		Predicted Label		
		Negative	*Neutral*	*Positive*
True Label	*Negative*	**171**	636	314
	Neutral	17	**405**	768
	Positive	9	519	**7161**

4.2 **LSTM Model**

The model was trained using the Tesla V100 GPU available on Google Colab. There was a slight drop in performance metrics compared to BERT. The precision, recall, and F1-score for the LSTM model as shown in Table 4 below displays some respectable figures. However, the model struggled in its prediction of the minority classes as seen with its precision values which is highlighted further in Table 5 below. This is likely as a result of the dataset having a larger number of positive reviews. In the OOD testing, the model's accuracy dropped to 76%, indicating a potential overfit to the training data and highlighting the importance of generalization in model assessment.

Table 4: LSTM model classification report

	Precision	**Recall**	**F1-score**
Negative	0.76	0.56	0.65
Neutral	0.72	0.79	0.75
Positive	0.95	0.95	0.95
Accuracy	0.90		
OOD Accuracy	0.76		
Training Time	3 hours		

Table 5: LSTM Model Confusion Matrix

		Predicted Label		
		Negative	*Neutral*	*Positive*
True Label	*Negative*	**2849**	1350	251
	Neutral	841	**11563**	2357
	Positive	220	2934	**55605**

From the train test split learning curves in Figure 5 below, the model seemed to be performing quite well and achieving convergence with initial struggles with the validation loss and accuracy metrics but improved steadily with subsequent epochs. This improvement trajectory highlights the LSTM's capacity for learning and adaptation over time. The k fold cross-validation technique was then employed to verify the performance for which the model demonstrated signs of overfitting in early epochs as seen in figure 6, even after the use of class weights to penalize the model for wrong predictions.

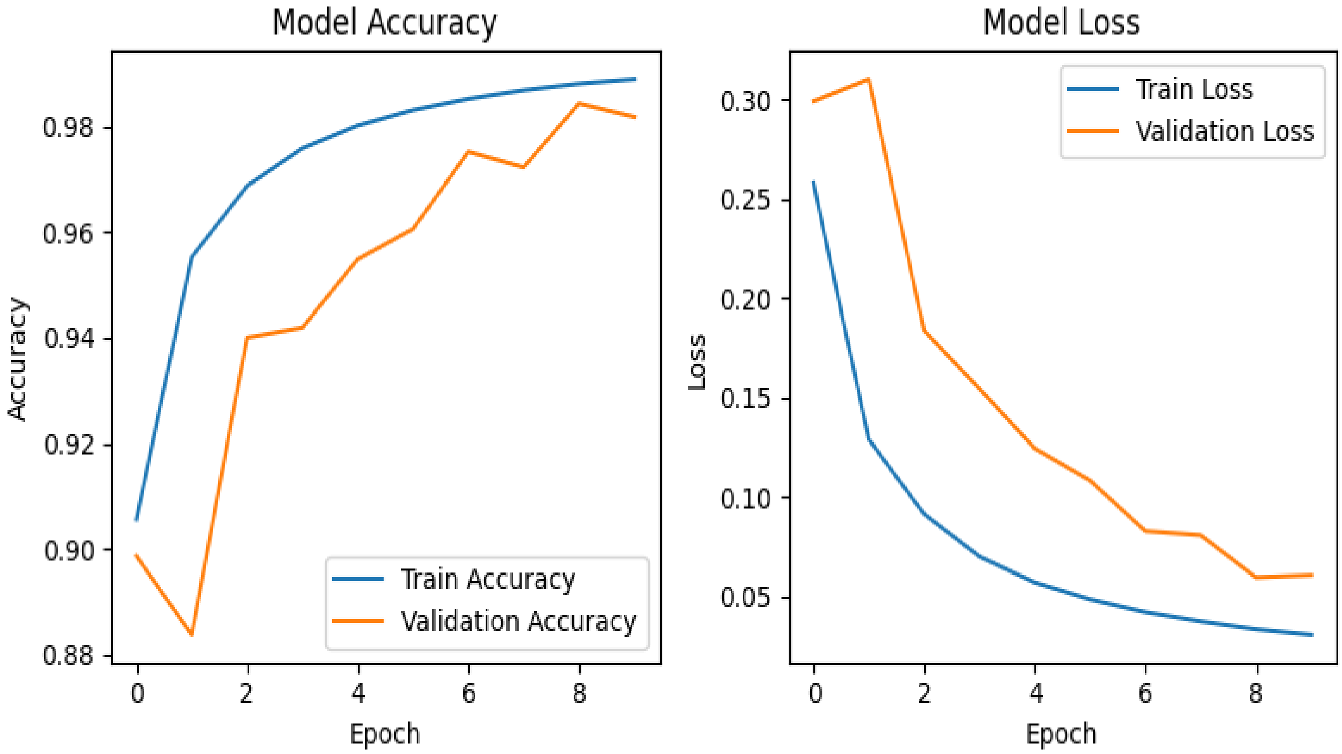


Figure 5: LSTM MODEL Train test split learning curves

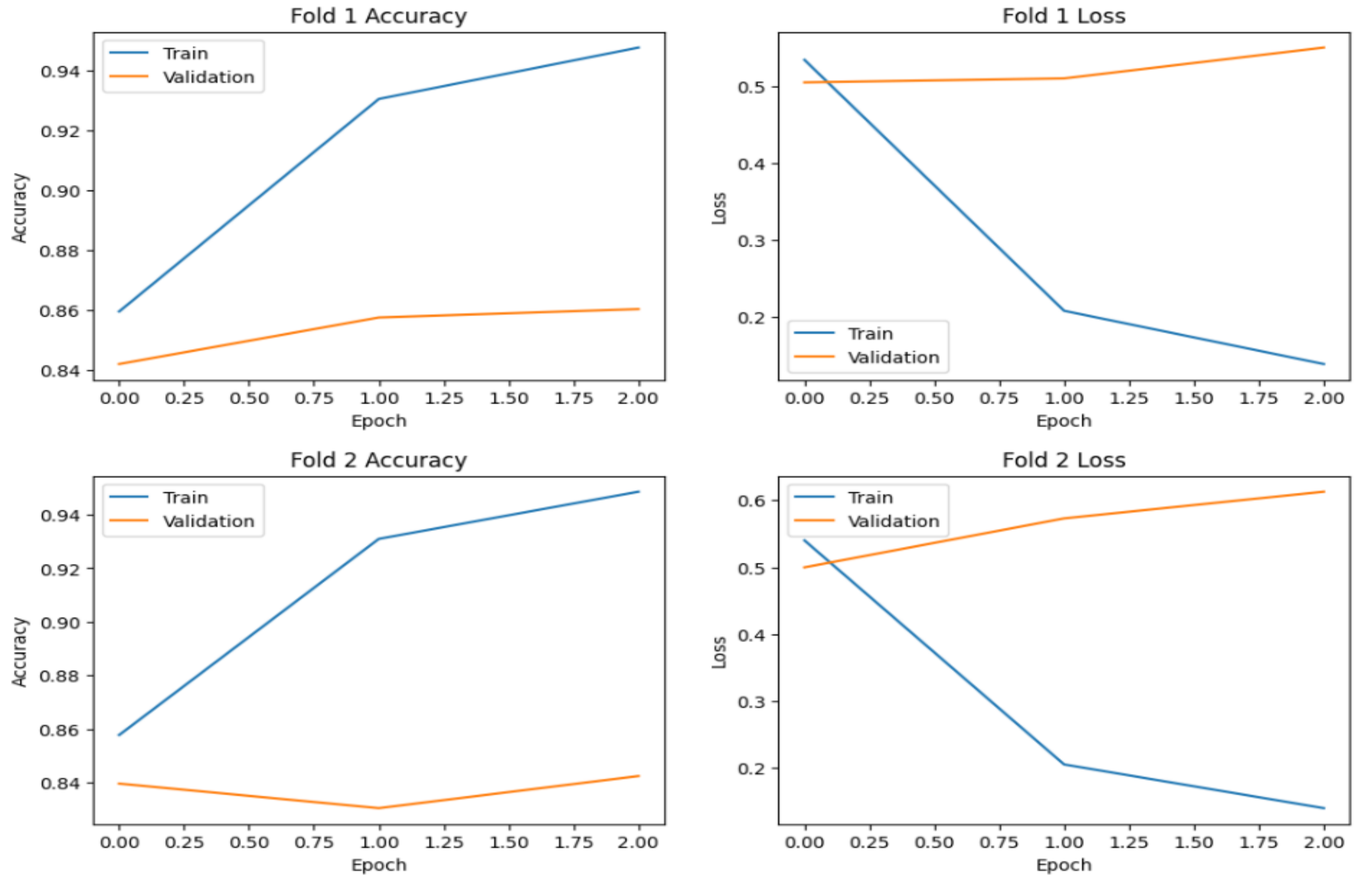


Figure 6: LSTM MODEL K fold Cross validation learning curves.

Despite these challenges, the LSTM model maintained a respectable AUC(Area under curve) score with micro-average and macro-average above 0.90 as seen in Figure 7 below.

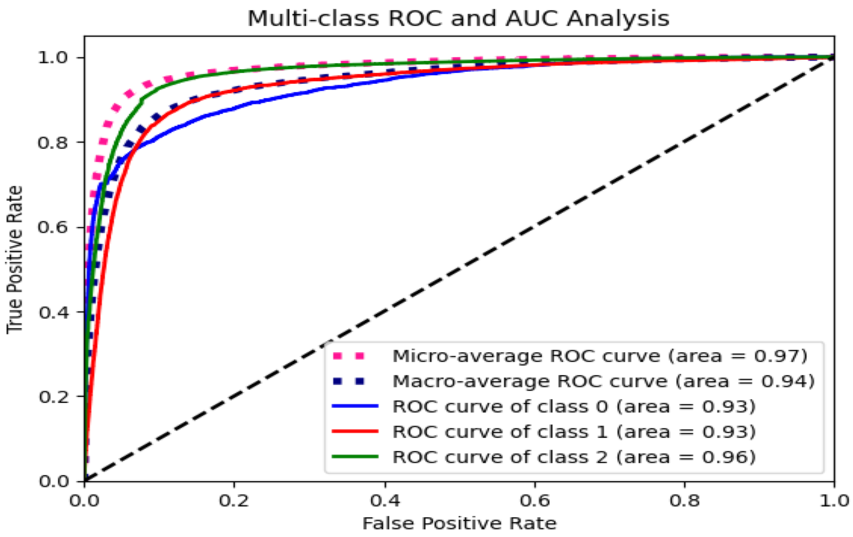


Figure 7: LSTM MODEL ROC Curves

4.3 **RNN Model**

The RNN model trained with the Tesla T4 GPU on google colab displayed a slightly more volatile training process, as evidenced by the fluctuations in the model accuracy and loss graphs in Figure 8 below. However, it still achieved a respectable overall accuracy of 0.87, with the classification report showing a balanced performance in detecting negative and neutral sentiments as shown in Table 6 below. The RNN's precision and recall for the positive class were particularly strong, reflecting the model's efficiency in identifying positive reviews.

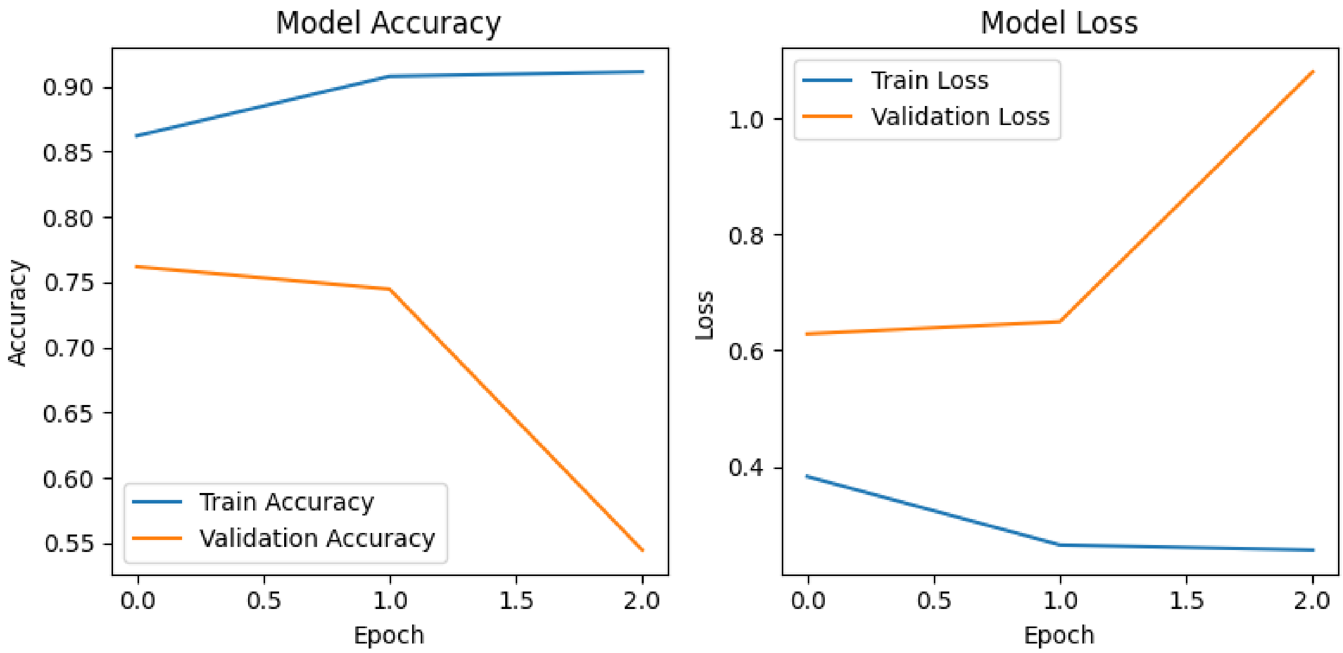


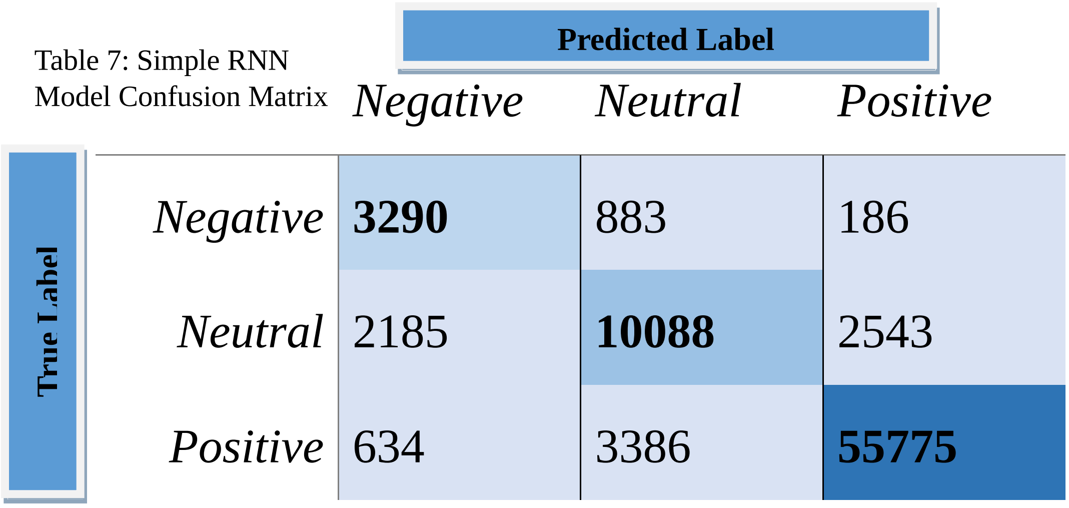
Figure 8: Simple RNN model learning curves

Table 6: Simple RNN Model Classification Report

	Precision	Recall	F1-score
Negative	0.54	0.75	0.63
Neutral	0.70	0.68	0.69
Positive	0.95	0.93	0.94
Accuracy	0.87		
OOD Accuracy	0.76		
Training Time	5 hrs		

The confusion matrix shown in Table 7 below further corroborates the RNN model's robustness, displaying substantial correct predictions across all sentiment labels. The AUC-ROC analysis also shown in Figure 9 below supports these findings, showcasing the model's good performance with a Macro average area under the curve of 0.94.

Table 7: Simple RNN Model Confusion Matrix



		Predicted Label		
		Negative	*Neutral*	*Positive*
True Label	*Negative*	**3290**	883	186
	Neutral	2185	**10088**	2543
	Positive	634	3386	**55775**

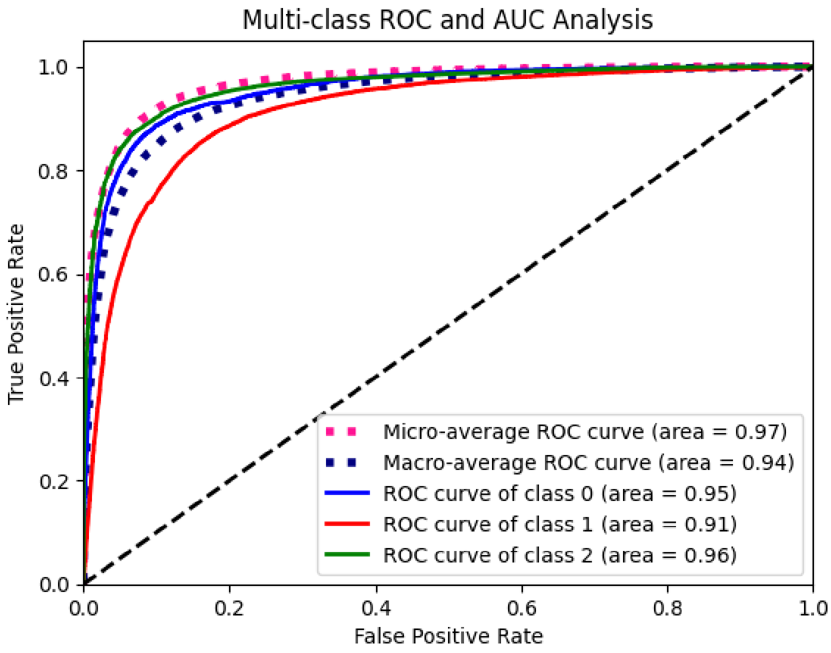


Figure 9: Simple RNN ROC Curve

The OOD accuracy for the RNN model was 0.76, indicating a consistent challenge across all models in generalizing beyond the training data.

Limitations and Future Directions

One possible drawback of this study is that it relies on data from Trustpilot, which may not fully represent the broader range of consumer opinions since it is limited to users who choose to leave reviews. This could lead to a bias in sentiment analysis, as people are often more likely to leave reviews when they have had either very positive or very negative experiences. Future studies could mitigate this by incorporating reviews from a variety of sources to capture a more representative sample of consumer sentiment. Additionally, the dataset's class imbalance posed challenges that affected the performance and generalization of the RNN and LSTM models on less represented classes, as evidenced by their respective classification reports. This led to overfitting and reduced model performance on minority classes. Future research could apply more sophisticated techniques for dealing with class imbalance such as varying resampling strategies, employing synthetic data generation techniques like SMOTE, or devising custom loss functions that penalize misclassification of minority classes more heavily. Due to time constraints and computational limitations, these approaches were not exhaustively explored in this study.

Conclusion

This study examines business performance through sentiment classification using BERT, LSTM and RNN models on customer reviews. The BERT model, optimized with TPU acceleration stands out for its precision and contextual understanding showing strong generalization beyond the known data. While the LSTM and RNN models may have displayed modest precision and accuracy scores, they are still quite effective in identifying positive sentiments. The study also highlights the challenge of overfitting during cross validation process emphasizing how complex sentiment analysis can be. Finding a balance between precision and generalization emerges as a recurring theme that reflects the task of tailoring sentiment analysis to diverse datasets and real-world scenarios. To some extent we can extrapolate the performance of these models to assess product or business performance although there are limitations involved (Soleymania, et al., 2017). These models analyse customer reviews to identify and classify sentiments, providing valuable insights into consumer opinions and attitudes towards a product or service. However, it's important to note that this analysis represents only one aspect of overall product performance. Other factors like product quality, customer service, and market trends also play crucial roles (Suchánek, et al., 2014). Therefore, while sentiment analysis can be a useful tool in assessing consumer perception, it should be integrated with other metrics for a comprehensive assessment of product or business performance.

Bibliography

Ahmad, W. & Edalati, M., 2022. Urdu Speech and Text Based Sentiment Analyzer. *arXiv preprint arXiv:2207.09163.*

Anjaria, M. & Guddeti, R. M. R., 2014. A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining,* 13 March, Volume 8, p. 181.

Biswas, E., Karabulut, M. E., Pollock, L. & Shanker, K. V., 2020. *Achieving Reliable Sentiment Analysis in the Software Engineering Domain using BERT.* s.l., IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 162-173.

Biswas, J. et al., 2021. Sentiment Analysis Using AI: A Comparative Study Comparative Study of 5 Different Algorithms and Benchmarking Them with A Qualitative Analysis of Training time, Prediction time, and Accuracy. *In 2021 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE 2021),* p. 373–377.

Chevalier, J. A. & Mayzlin, D., 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research,* Issue 43(3), pp. 345-354.

Devlin, J., Chang, M.-w., Lee, K. & Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805.*

Goodfellow, I., Bengio, Y. & Courvillle, A., 2016. *Deep learning.* s.l.:MIT press.

Hu, M. & Liu, B., 2004. *Mining and summarizing customer reviews.* s.l., s.n.

Li, D. & Qian, J., 2016. Text sentiment analysis based on long short-term memory. *First IEEE International Conference on Computer Communication and the Internet (ICCCI),* pp. 471-475.

Lipton, Z. C., Berkowitz, J. & Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019..*

Liu, B. & Zhang, L., 2012. A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal, C., Zhai, C. (eds) Mining Text Data. *Springer.*

Liu, L., Dzyabura, D. & Mizik, N., 2020. Visual listening in: Extracting brand image portrayed on social media. *Marketing Science,* 39(4), pp. 669-686.

Pang, B. & Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 1 January, 2(1-2), pp. 7-9.

Pang, B., Lee, L. & Vaithyanathan, S., 2002. *Thumbs up? Sentiment Classification using Machine Learning Techniques.* s.l., arXiv:cs/0205070 [cs.CL].

Plisson, J., Lavrac, N. & Mladenic, D., 2004. A Rule based Approach to. *In Proceedings of IS,* Volume 3, pp. 83-86.

Singh, C., Imam, T., Wibowo, S. & Grandhi, S., 2022. A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. *Applied Sciences,* Volume 12, p. 3709.

Soleymania, M. et al., 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing,* pp. 65, pp.3-14.

Suchánek, P., Richter, J. & Králová, M., 2014. Customer satisfaction, product quality and performance of companies. *Review of economic perspectives,* Issue 4, pp. 14 pp.329-344.

Teney, D. et al., 2020. On the value of out-of-distribution testing: An example of goodhart's law. *Advances in neural information processing systems,* Issue 33, pp. 407-417.

Tumasjan, A., Sprenger, T., Sandner, P. & Welpe, I., 2010. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* s.l., s.n., pp. 178-185.

Zhang, K. Z., Zhao, S. J., Cheung, C. M. & Lee, M. K., 2014. Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems,* Volume 67, pp. 78-89.

Zhang, L., Wang, S. & Liu, B., 2018. Deep Learning for Sentiment Analysis : A Survey. January.