WEB MINING

COMPARISION-BASED MODELLING FOR SENTIMENT ANALYSIS OF SMALL CORPUS AIRLINE DATASET

António Pratas de Oliveira
Jean Zuosheng Yeo
Shiyi Xu
Sudhir Joon
Zhe Ming Chang

MAY , 2024

Table of Contents

Comparison-based Modelling for Sentiment Analysis of Small Corpus Airline Dataset

António Pratas de Oliveira[1,3] , Jean Zuosheng Yeo[1,2], Shiyi Xu[1], Sudhir Joon[1] and Zhe Ming Chang[1,2]
Comparison-Based Modelling for Analysis of Small Corpus Airline Dataset

[1] Universität Mannheim, Universität Mannheim, 68131 Mannheim, Germany
[2] National University of Singapore, Lower Kent Ridge Road. 21, 119077, Singapore
[3]Universidade NOVA de Lisboa, 1099-085 Lisbon, Portugal

Abstract

This research paper carries out a comparative analysis of various sentiment analysis models and their effectiveness in determining the polarity of a small dataset of airline customer reviews. Machine Learning (ML) models of Random Forest (RF) and Gradient Boosting (GB) Classifiers were pitted against deep-learning BERT and LSTM models to assess their reliability in classifying a small corpus of airline customer sentiments based on textual reviews. The performance of these ML models were determined through F1-Score. Our results show that LSTM outperformed pre-trained BERT models given a small corpus dataset. LSTM and BERT models also outdo ML models like Logistic Regression, RF and GB, due to their ability to determine context from textual documents.

Keywords: Sentiment Analysis, Airline Quality, Natural Language Processing, BERT, LSTM, Gradient Boosting, Random Forest

1 Introduction

1.1 Problem Formulation

Text-based airline reviews provide a key source of information for airline companies to gather opinions on service standards and customer satisfaction[1]. Through the mining of online customer feedback with sentiment analysis tools, airline organisations can better comprehend the shortfalls and strong points of their services[1]. In turn, sentiment analysis provides an opportunity to better comprehend customer expectations and improve service offerings to suit customer needs.

Bidirectional Encoder Representation from Transformers (BERT) model is a transformer-based model that processes input text in parallel. BERT is able to consider the context of a given text sentence due to the dynamic representation of text in reference to words before and after a given text token. It is pre-trained on a large amount of text data and can be fine-tuned to suit specific text analysis task goals.

Long Short-Term Memory Networks (LSTM), on the other hand, is a recurrent neural network model. Input, forget and output gates control the flow of information in and out of memory cell state. In turn, information from past states are considered in the existing state, making it an ideal model for understanding sequential textual inputs for sentiment analysis[2].

Both BERT and LSTM models were deemed to outperform ML models like Random Forest(RF) and Gradient Boosting (GB) classifiers due to ML models inability to consider textual inputs apart from sequential siloes[1].

Airline companies frequently face the difficulty of working with a small corpus of customer reviews[3]. Ezen-Can suggests that complex, pre-trained BERT models are not necessary and can underperform on small corpus datasets due to their tendency to overfit training data[2]. In turn, simple LSTM architectures would be sufficient in reliably predicting customer sentiment for small datasets.

In contrast, Sanwal et al. conducted sentiment analysis on a hostel textual dataset and determined BERT models to outperform LSTM, even on a small corpus dataset of 1000 hostel reviews[2]. Contextual comprehension by BERT triumphed over the exploitation of the sequential nature of text by LSTM in the research findings[2].

1.2 Research Questions and Objective

This research paper considers the findings taken from a small corpus of 8099 airline reviews. With below 8000 data points, the repository is deemed to be a small dataset[3]. It utilises machine learning methods like Logistic Regression,

Random Forest, GB and lexicon-based models like VADER to provide a baseline for comprehending the effectiveness of LSTM and BERT in classifying the textual sentiment of airline reviews into positive, neutral and negative classes. The comparison of LSTM and BERT in classifying sentiments of airline dataset are then conducted.

Our research question is as follows: "Given a small airline corpus dataset, is there a need for usage of pre-trained BERT architecture or would LSTM architecture be sufficient?" It attempts to contribute to the scholarly understanding of sentiment analysis techniques for corporate usage given the limitations of sufficient procurable textual datasets for analysis in the corporate workplace[2].

2 Methodology

In this section of the report we will describe the necessary steps taken to solve our Research Questions. We will start with Data Collection, followed by Data Preprocessing and Data Exploration. Finally, the models used will be briefly described.

2.1 Data Collection

The dataset contains reviews of the top 10 airlines on Airline Quality, an independent website set up by Skytrax Research. The reviews are provided by customers of the respective airlines. The data is accessed through the public CSV file provided on Kaggle by Sujal Suthar. It can be obtained via https://www.kaggle.com/datasets/sujalsuthar/airlines-reviews[4].

2.2 Data Preprocessing

Python and Jupyter Notebook were the primary tools used for model building and experimentation. Key libraries included the Natural Language Toolkit (NLTK), scikit-learn, and HuggingFace, which were instrumental in data preprocessing, model architecture development, and evaluation.

The first steps consisted of an Initial Exploration, where methods such as *describe* and *info* were used, as well as the count as display of some unique values. Then, moving on to the Exploratory Data Analysis, we started by addressing Missing Values, which were non-existent. Similarly, no outliers were found from the boxplots. With these matters addressed, we moved on to Feature Engineering, where we dropped the only observation that contained emojis. We also converted our target to 3 different classes: (Rating 1-4, 5-6, 7-10 as negative (0), neutral (1) and positive (2) classes). Additionally, the Reviews column was converted into a list and the year of each review was extracted by converting Review Date to *datetime* and only extracting the year.

After the regular preprocessing, we moved on to the preprocessing required to perform Sentiment Analysis, which includes addressing Capitalisation, Punctuation, Stopword Removal, Stemming and Lemmatization and Tokenization. A function was used to make the preprocessing required for Sentiment Analysis using Lexicon-Based Models, and another one was used in LSTM. In the case of BERT, tokenization and encoding was done in the respective notebook.

The preprocessing used in the BERT model included the removal of punctuation and texts were converted to lower-case, further decreasing complexity of model inputs. FOR LSTM, the preprocessing is similar to the BERT model. However, while the BERT tokenizer was used in the former, word vectors for the LSTM model were generated using the GloVe algorithm. The LSTM model also differs in that it can only accept inputs of fixed sizes. Hence, shorter inputs were padded while longer inputs were trimmed such that all the reviews had a fixed length of 200 vectors. In the case of Lexicon Based models, the preprocessing just removed stopwords.

The extracted dataset was split into a 80% train, 20% test set.

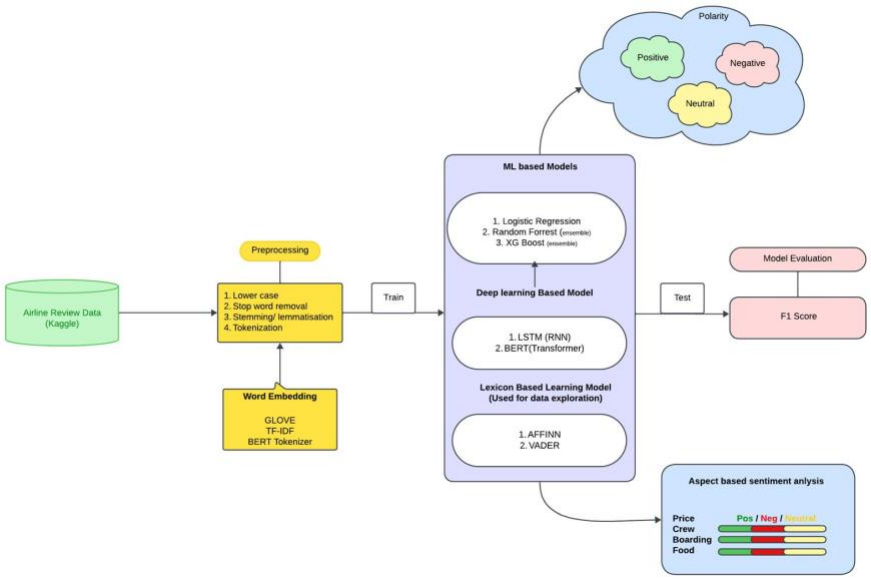


Fig. 5. Pipeline for Sentiment Analysis

Above figure describes the project outline the different steps in the sentiment analysis project.

2.3 Vectorization and Embedding Techniques

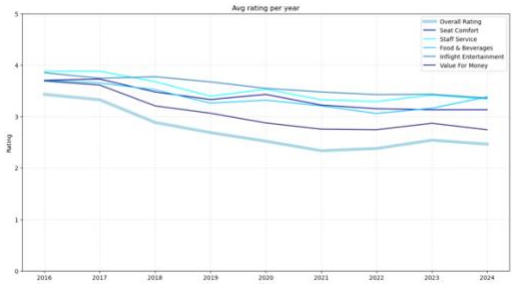
For vectorization, we employed TF-IDF, GloVe, and BERT embeddings. TF-IDF was used to assign weights to terms based on their importance, calculated by the frequency of a term's appearance in a document relative to its prevalence across all documents.

GloVe operates as an unsupervised learning algorithm that generates vector representations for words by leveraging global word-word co-occurrence statistics from a corpus. The resulting vectors often reveal fascinating linear substructures within the word vector space. For this project, the glove.6B.300d model is used.

BERT embeddings were used to manage out-of-vocabulary words and enhance contextual comprehension by considering the context of words before and after a given text token. These techniques allowed us to effectively convert textual data into numerical representations suitable for machine learning and deep learning models

2.4 Data Exploration and Visualisation

The dataset consist of 17 columns ranging from Title and Review Date to customer ratings of staff service and seat comfort. Minimal connection was observed between factors such as review length and customer rating. Similarly, whilst average ratings per year were observed to be on a downward trend, the trend was constant across airlines and metrics. The drop in overall ratings might be attributed to the period of COVID-19.



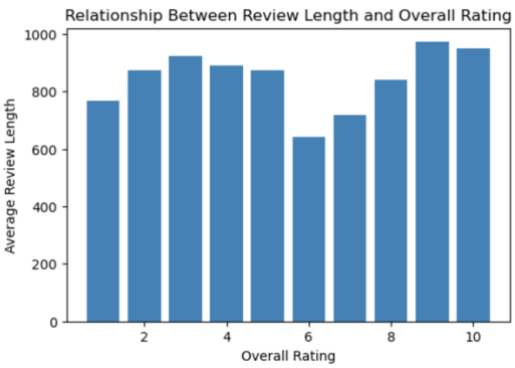


Fig 1 & 2: Average Overall Ratings of Various Features from 2016 to 2024 (Over an Eight-Year Period) and Relationship Between Average Review Length (Number of Words) and Overall Rating

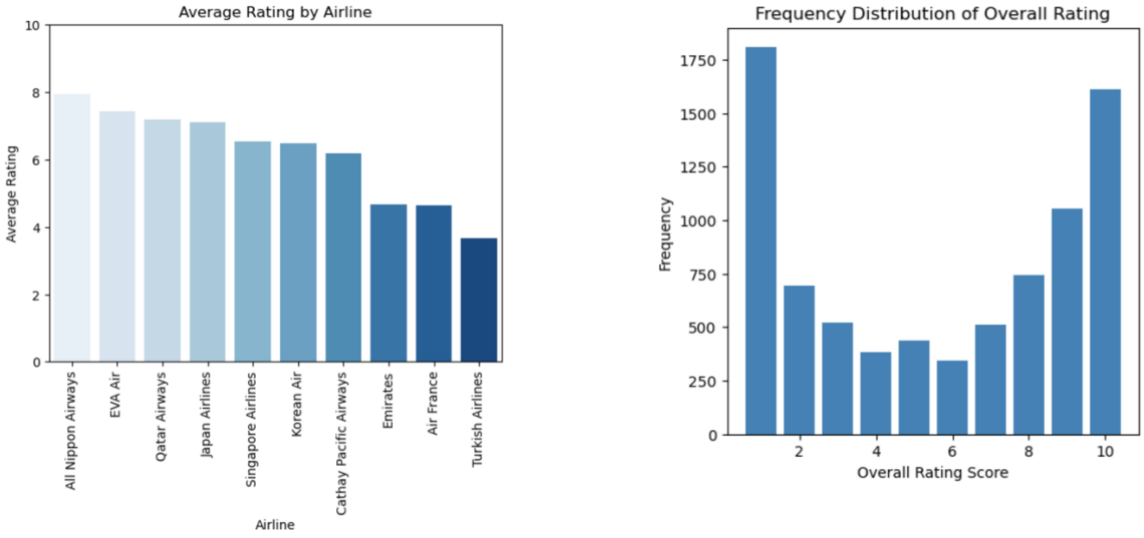


Fig 3 & 4: Average Overall Ratings of Eleven Airline Companies | Count of overall Customer review rating against number of ratings

As demonstrated in Figure 3, the average overall ratings differ significantly across various airline companies. All Nippon Airways exhibits the highest average rating at 7.9, while Turkish Airlines records the lowest in the dataset.

Overall ratings range from 1 to 10, with the majority of reviews clustered around the extremes. Sentiments were binned according to negative, neutral and positive sentiments. Based on customer reviews and overall ratings, prebuilt dictionaries (lexicons) were used to have a better understanding of the polarity of texts. They work by assigning a Sentiment Score, which can be negative, neutral or positive, to each word or sentence. Based on the sentiment scores of the identified words, the models assign an Overall Sentiment to the sentences.

Our main goal was to better understand our data, and since no predictions are being made, it does not make sense to evaluate these models with the usual metrics, such as f1-score. Consequently, we decided to compare the 3 models used by using several visualisations, as well as by looking at some specific observations to see how the attributed score matched our opinion on how negative or positive said review was.

Lexicon-based approaches such as AFINN, TEXBlob and VADER were utilised to provide a better background understanding for our dataset. By comparing the average Polarity Compound computed by each algorithm for each Sentiment, one can clearly spot some irregularities: on one hand, TexBlob concludes that the reviews which received low scores (Sentiment = 0) have an average of approximately 0 mean polarity. On the other hand, AFINN shows that the reviews that correspond to the most positive Sentiment (2) still transmit a negative feeling. After exploring this issue with AFINN, we concluded that these values may be a consequence of the scaling. Since the maximum value of Afinn_polarity_compound is 108, even observations with a final polarity compound of 5 when scaled are attributed a low normalised polarity. Hence, when observing these values, one must analyse them always keeping this in mind.

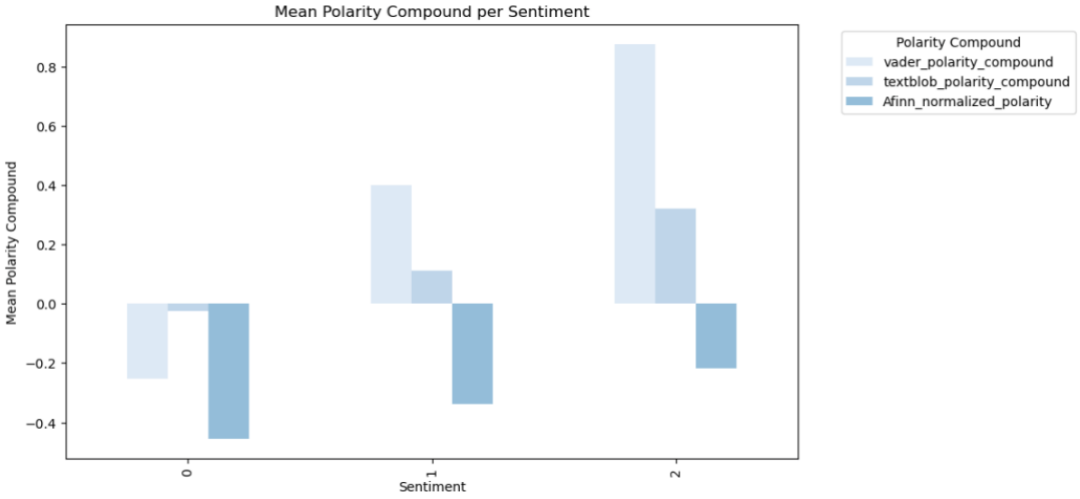


Fig. 5. Mean Polarity Per Sentiment

When taking a look at the polarity change across the years in each sentiment, there are no significant positive or negative spikes, being the mean values of each year close to each other. However, when looking at the polarity variation across the years per travel class (Economy, Premium Economy, Business and First), one can spot a pitfall in 2021, where all the used algorithms detected a decrease in polarity for most classes. Since 2021 was a year highly affected by the Covid-19 pandemic, there were issues with flights being cancelled or delayed, which may explain the customer's discontentment in that particular year.

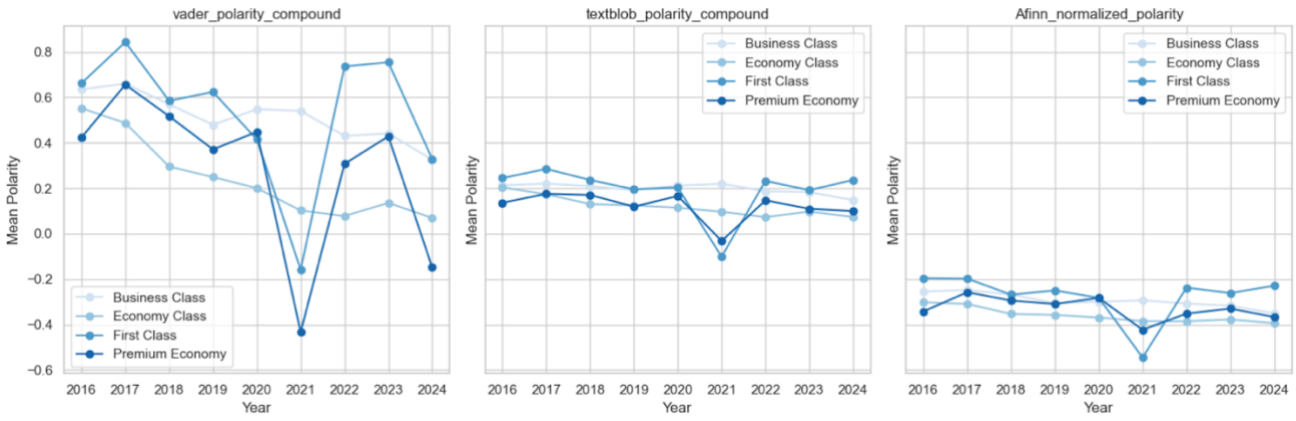


Fig. 6. Polarity Change across the years for each Lexicon-Based Model per class Travelled

As expected, we also concluded that passengers are, in general, more satisfied with their experience when they give high ratings to the Seat Comfort, Staff Service, Food & Beverages, Inflight Entertainment and Value for Money variables.

With this analysis, we concluded that our algorithms perform very differently in the same data. For example, AFINN appears to always attribute negative final scores, even when the Sentiment is maximal (2). Since Vader is tuned for non-formal data, we expected it to be the most accurate. However, after taking a look at observations with Sentiment = 2 and Vader polarity < 0, we can see that there are still 79 observations that Vader did not classify correctly. Taking a closer look at one of these observations with index 64, it says '(...) the experience was second to none with outstanding cabin crew (...)', to which Vader attributed a final score of -0.2778. When evaluating what could be causing this issue, we reached the conclusion that even though Vader is trained on a somewhat large dataset, it does not include all existing words, which may lead to failures. The same applied to the other algorithms, which indicates that the human touch is still required in this area of Sentiment Analysis.

2.5 Models

This subsection will focus on describing the models used for the Lexicon-Based Analysis and the polarity Predictions.

Lexicon-Based Models

We utilised VADER, Textblob and Afinn as our lexicon-based models. VADER, designed for social media text, was effective at capturing sentiment in informal text. Textblob computes the Polarity and Subjectivity of a text. Afinn assigned sentiment scores to individual words, which were then summed to produce a final score, facilitating straightforward interpretation of the text.

Vader

Vader, in the context of Sentiment Analysis, refers to a specific lexicon-based model -Valence Aware Dictionary and Sentiment Reasoner - rather than the Star Wars character. It uses a list of 7500 curated lexical features that are labelled as very positive (1) or very negative (-1), where 0 stands for neutral, based on their semantic orientation. Each lexicon can be attributed a range of values from -1 to 1.

TextBlob

TextBlob is a python library used for various NLP tasks, including Sentiment Analysis. As all algorithms in this section, it relies on a pre-built dictionary that assigns sentiment scores to words. It functions by computing the Polarity, which ranges from -1 to 1, and the Subjectivity of each review, which indicates how personal or biased the used language is. Subjectivity ranges from 0 (highly fact-based) to 1 (highly subjective, very opinionated). It has difficulty capturing sarcasm, but it is good for basic sentiment analysis tasks.

AFINN

AFINN is another lexicon-based model which relies on a dictionary that contains words with assigned scores. These scores, unlike the previous ones, range from -5 (very negative) to 5 (very positive).It attributes a score to each word and then sums up all the scores in the same review to compute a final, overall score. Since this score can have very positive or very negative numbers, we decided to scale the overall score from -1 to 1, to make it easier to compare with the remaining models

Machine Learning Models

For machine learning, we employed Logistic Regression, Random Forest, and XGBoost. Logistic Regression was chosen for its simplicity and interpretability in binary and multi-class classification tasks. Random Forest was selected for its robustness in mitigating overfitting and providing insights into feature importance. XGBoost was used for its efficiency and ability to handle missing data well, incorporating regularisation techniques to prevent overfitting. To differentiate them from the deep learning techniques, the term Machine Learning (ML) models will be used to refer to these three methods exclusively.

Deep Learning Models

In the deep learning phase, we used BERT and LSTM models. The pre-trained BERT tokenizer and model from the HuggingFace library were employed to process text in parallel, considering the context of surrounding words. Tokenization of text was conducted using the BERT tokenizer, and the tokenized datasets were stored in tensor form. The LSTM model utilised GloVe word vectors, with inputs standardised to a fixed length by padding shorter inputs and trimming longer ones. LSTM was particularly chosen for its ability to capture long-term dependencies in sequential data, making it ideal for understanding the flow of text.

Overall, the integration of these diverse techniques allowed us to leverage the strengths of each method. Preprocessing with NLTK ensured the data was clean and standardised. Embedding techniques like TF-IDF, GloVe, and BERT enabled effective vectorization of text. Lexicon-based models provided straightforward sentiment scores, while machine learning models offered robust classification capabilities. Finally, deep learning models like BERT and LSTM captured complex contextual and sequential information, enhancing the overall performance of our sentiment analysis.

3 Experimental Settings

This section will focus on how each model was used and which were the parameters that lead to better results.

3.1 Logistic Regression

The Logistic Regression model, optimised through grid search and TF-IDF vectorization for embedding, achieved a mean validation F1 weighted score of 0.781 and an accuracy of 0.80. Although the model performed adequately overall, it struggled with accurately predicting neutral reviews. Despite these challenges, it showed a slight improvement over the Random Forest model in this area, achieving an F1 score of 0.30 for neutral reviews.

3.2 Random Forest Classifier

The Random Forest model was also trained on TF-IDF embedding.. This model achieved a mean validation F1 weighted score of 0.728 and an accuracy of 0.78. It exhibited strong performance in predicting both positive and negative reviews, demonstrating high recall for positive reviews. However, the model struggled significantly with the neutral class, yielding an F1 score of only 0.21 for this category

3.3 Gradient Boosting Classifier

When applying the XGBoost model to the GloVe embeddings, a baseline accuracy of 0.79 and a weighted F1 score of 0.75 are achieved. Subsequently, GridSearch is utilised for hyperparameter optimization, setting cv=3 to optimise the balance between computational efficiency and runtime. The final result is shown in Table.1.

3.4 BERT Model

A total of 10 epochs were carried out on the train dataset. The scheduler utilised was linear scheduler with warm up. The optimizer method was AdamW and the loss function was Cross Entropy Loss. The train dataset was fed into the BERT model via Data Loaders in batch sizes of 10 data points per instance. The BERT model was evaluated using the test dataset via Data Loaders.

3.5 LSTM Architecture

The simple LSTM model that was employed consisted of an LSTM layer, followed by a linear layer.. The LSTM cells consolidate language information from the reviews (given in vector form), while the linear layer then classifies them into the three target classes of “positive”, “neutral” and “negative”.

With regards to hyperparameters, 2 LSTM cells were used, with a dropout probability of 0.1. The hidden layer has a size of 64.

4 Results

4.1 Result Evaluation and Discussion

Table 1. Model F1-score by class and Overall F1-score

Model	Weighted F1 Score (Positive Class)	Weighted F1 Score (Neutral Class)	Weighted F1 Score (Negative Class)	Model Weighted F1 Score
Logistic Reg	0.84	0.30	0.89	0.77
Random Forest	0.81	0.01	0.84	0.74
XG Boost	0.83	0.05	0.86	0.77
BERT	0.90	0.25	0.91	0.84

LSTM	0.92	0.45	0.94	0.89

The LSTM Model was found to outperform pre-trained BERT. Ezen-Can noted a similar pattern with a small corpus dataset of 23000 chat-bot utterance dataset, with a close to 5% increase in accuracy of LSTM models as opposed to a pre-trained BERT model[3]. Samir et al. similarly carried out a study on a corpus of 65947 airline customer reviews, with a result of LSTM architecture outperforming BERT by one percentage point[1]. This phenomenon was attributed to the BERT model's tendency to overfit training data as opposed to simple LSTM architecture[3].

BERT and LSTM models outperformed ML-based models, indicated by the higher weighted F1-score for both models as opposed to the ML counterparts. Sanwal et al. noted BERT's usage of contextual embeddings as well as LSTM model's exploitation of textual sequence as key factors in the improved predictability of both models[3]. Both methods provide a more comprehensive understanding of textual inputs, allowing for better classification of sentiment.

The readings indicate the superior performance of pre-trained language models like BERT as well as deep learning models like LSTM in prediction of textual sentiments as opposed to other machine learning based models. In addition, for small corpus datasets, LSTM models are viewed to outperform BERT models.

4.2 Error Analysis

Logistic Regression & Random Forest Classifier

Both the Random Forest and Logistic Regression models were used as baseline approaches for text classification or sentiment analysis, with TF-IDF vectorization outperforming the Bag of Words approach. Despite this improvement, both models struggled with accurately predicting neutral reviews, likely due to the TF-IDF approach not preserving semantic meaning. The baseline steps included word vectorization and the Bag of Words approach. To improve performance, we employed sentence vectorization to capture more semantics, TF-IDF, grid search cross-validation, and SMOTE to manage class imbalance. The poor performance of the models on the neutral class could be attributed to class imbalance, out-of-vocabulary words, and a lack of semantic understanding.

Gradient Boosting Classifier

One of the most prominent problems is the low F1 score for neutral reviews. One reason could be the dataset imbalance. SMOTE (Synthetic Minority Over-sampling Technique) was utilised to oversample the minority class. SMOTE works by selecting instances that are close in the feature space, drawing a line between the minority class instances and creating new instances along this line. The F1 score for neutral classes was improved to 0.24.

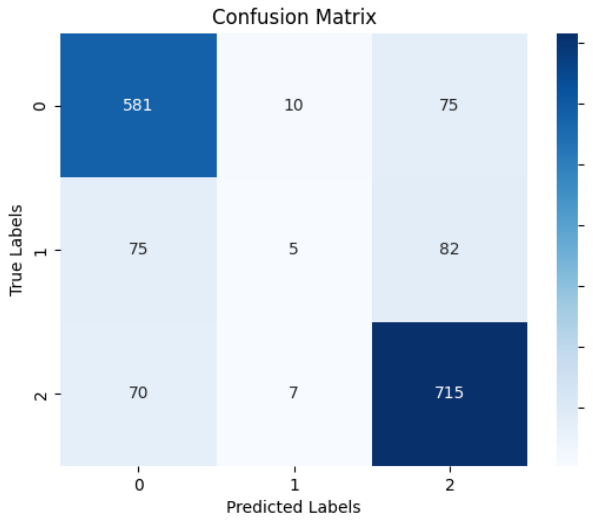


Fig. 7. Confusion Matrix for XGBoost

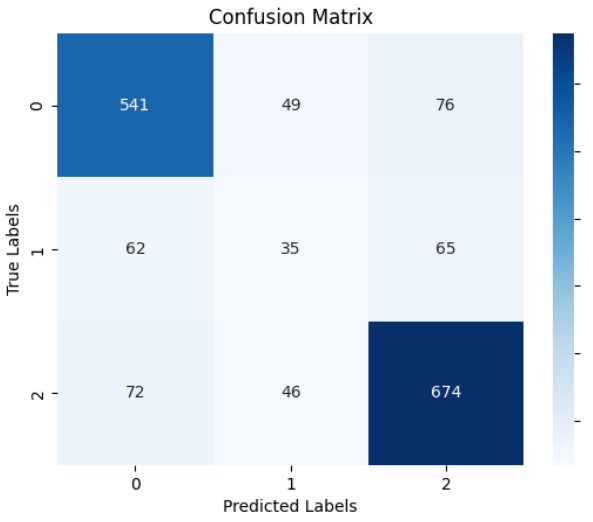


Fig. 8. Confusion Matrix for XGBoost after SMOTE

However, compared to other two polars' scores, 0.24 is still not satifying. An examination of the neutral categories revealed that most of the neutral categories misclassified as positive or negative did express emotional trends. For example

Table 2. Example of misclassification (1)

Review	True Label	Predicted Label
This was my first time flying on a regional Singapore Airlines flight on a narrow-bodied aircraft since the old subsidiary Silk Air was absorbed by the parent carrier, and I was impressed. Everything went well on the flight, from boarding, to exemplary behaviour and courtesy shown by the cabin crew, to the food and boarding process at Changi airport. The inflight entertainment provided was the full product used for their long haul flights — amazing!	1	2
We used this airline recently for our trip to India and will not recommend this airline . Both ground service and cabin crew service not very friendly neither very welcoming. For a 16 hour straight flight you get two full meals and the snacks are just some nuts. No sandwiches are offered unless you ask for it. Meal services are also different, kids are not served first always.(sometimes yes they do). Meals also has no varieties like others like no fruits or yogurt. We were very excited when we got the tickets as it was our first time using this airline but will not use in the future.	1	0

Besides, since we only consider overall rating, our model has difficulty dealing with the situation that users have opposing opinions on different aspects.

Table 3. Example of misclassification(2)

Review	True Label	Predicted Label
Flew Singapore Airlines from Delhi to Singapore. The flight from Delhi to Singapore went well. However, the crew on the Singapore to Adelaide sector were terrible, I had to witness a dirty toilet early in the morning before landing. Seats were pretty uncomfortable as well.	1	0

BERT Model

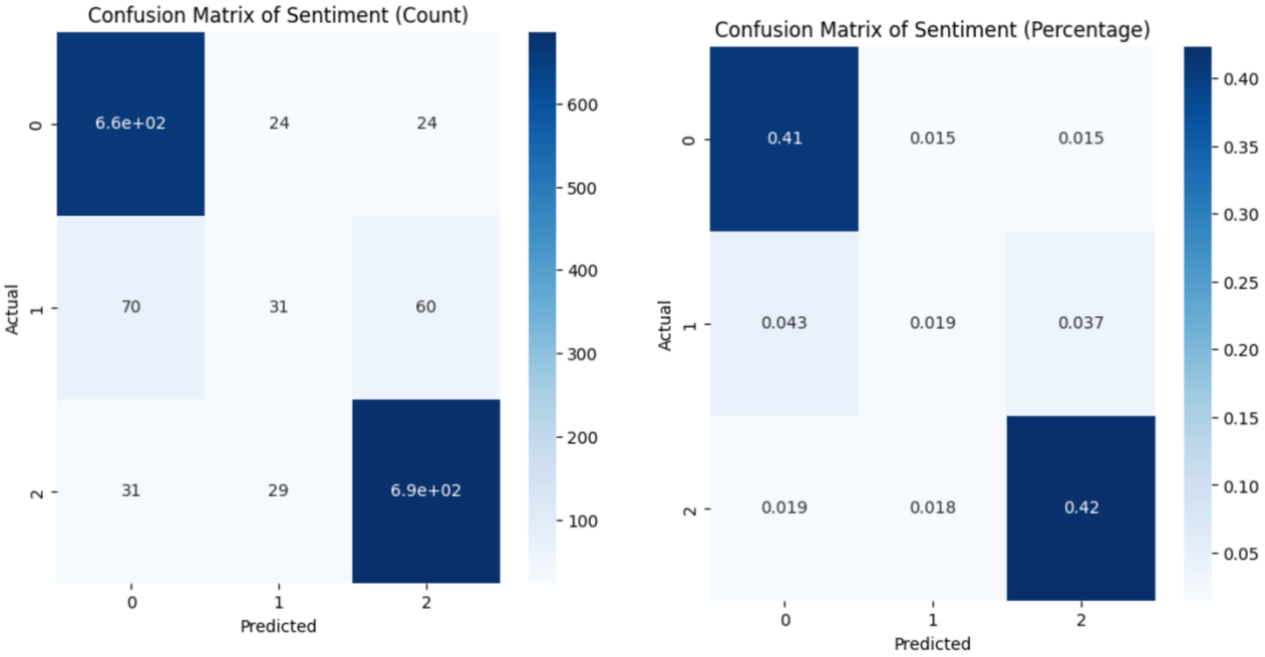


Fig. 9 and 10. Confusion Matrix for BERT Model on Test Set [0,1,2 - negative, neutral, positive]

The plot of confusion matrix of test set sentiments reveal high reliability in predicting negative and positive classes for the BERT model, but suboptimal predictions for the neutral class. Neutral sentiments have a classification F1-score of 25%, with the majority of neutral BERT predictions to be of positive and negative classes.

This observation can be attributed to the difficulty in predicting neutral sentiments, that include a mix of positive and negative textual phrasings. In addition, sarcasm can play a part in impacting the polarity of BERT classification.

LSTM Architecture

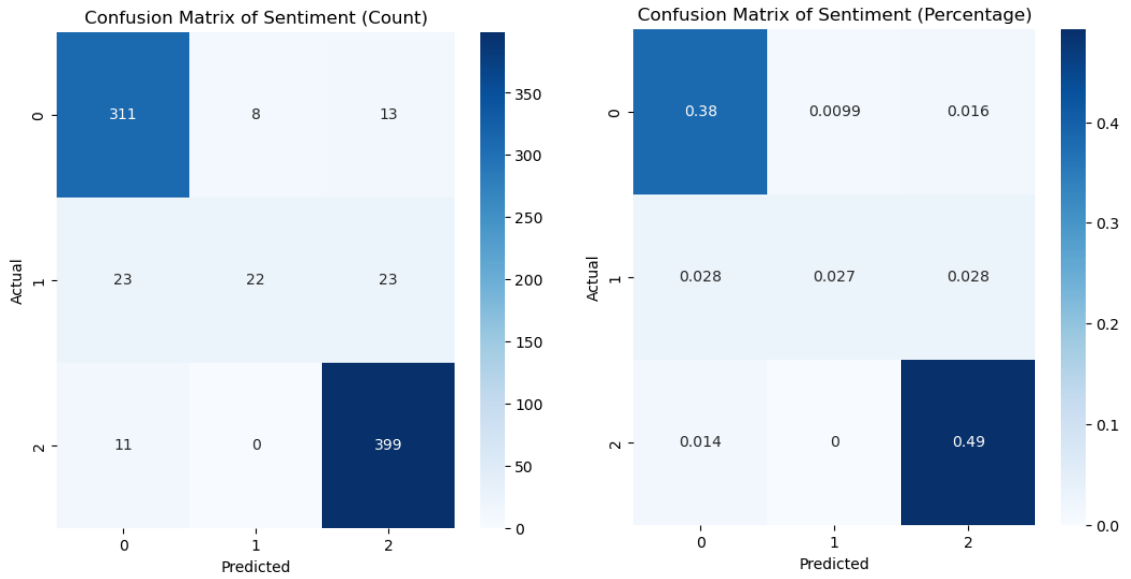


Fig.11 and 12. Confusion Matrix for LSTM Model on Test Set [0,1,2 - negative, neutral, positive]

Similarly, the LSTM model also performed better at predicting polar classes, and worse for the neutral class. This could be attributed to the nature of customers' reviews, which tend to focus on the best and worst aspects of their experience. For instance, a user who gave a neural score may only explain what dissatisfied them without mentioning the redeeming positive elements.

Table 4. Example of misclassification (3)

Review	User Rating	True Label	Predicted Label
Eva Air's slogan 'Striving for Perfection' is a strong strap line yet standards have fallen since Covid. I have used Eva on the London to Bangkok route since they started flying from Gatwick in 1994. Unfortunately my first post Covid flight there are a number of issues which need to be addressed. The In-Flight entertainment is not what it used to be with limited choice and a really difficult navigation issue where one has to select ones mood before a small choice appears. Food was mediocre and the vegetarian bread roll is served cold yet standard fayre comes with a hot roll. The evening meal was excellent with fresh salmon and potatoes but my breakfast hash was disgusting.	6	1	0

This possibly reveals an inherent flaw in the research design, as the user rating may not precisely reflect the sentiment of the review text.

Again, the model struggled with reviews with mixed sentiments, classifying them as either polars instead of the neutral class.

Table 5. Example of misclassification (4)

Review	User Rating	True Label	Predicted Label
My wife and I flied JAL economy class back from Japan to Thailand early this month. The check-in was very good, The staff were very nice and efficient. We have made prior window seat booking to sit together, however, we were informed that JAl change the type of plane from B77 to B787 which is smaller, So we were not assigned to sit together by the window. Instead they assigned us to sit together in the middle row of 2-4-2 which means that we had other passengers on both side which made us very uncomfortable throughout the flight. The service is good, the food is OK..	6	1	2

In the above example, the customer's struggle with seating arrangements is not captured by the model. One workaround would be to further train the model with specific reviews on different aspects of the flight experience.

5 Conclusion

5.1 Takeaways

While the study initially sought to compare the performance of complex models that are prone to overfitting against simpler models, the experiments showed that a wide array of techniques can be used to perform sentiment analysis with respectable results (a weighted F1 score of above 0.7).
The most interesting finding in this study was the collective struggle to identify neutral sentiments. The presence of both positive and negative words in these neutral reviews proved to be misleading for the models trained in this study.

Furthermore, despite the neutral score, the reviews themselves would not be worded as such. Customers often bring attention to specific outstanding or problematic aspects of their experience, which may skew the language of the reviews such that it no longer accurately justifies the user rating score. This also revealed an inherent limitation of the experiment design in choosing to use reviews as the training corpus and user rating as the true sentiment score.

The project analysis demonstrates that LSTM models are highly effective for sentiment analysis of small corpus datasets, outperforming both BERT and traditional ML models like Logistic Regression, Random Forest, and XGBoost. This suggests that simpler architectures like LSTM can be more suitable for smaller datasets due to their ability to avoid overfitting and better handle sequential information.

The superior performance of LSTM over BERT, particularly in the context of small datasets, can be attributed to its architectural simplicity. While BERT's complexity is advantageous for large datasets with diverse contexts, it tends to overfit on smaller datasets, reducing its effectiveness. LSTM's recurrent structure allows it to maintain and utilise the sequential nature of text data more efficiently. This finding aligns with the conclusions drawn by Howard and Ruder[9] who emphasised the importance of model simplicity and the risk of overfitting with more complex architectures in specific scenarios.

Additionally, the study highlights the limitations of traditional ML models like Logistic Regression, Random Forest, and Gradient Boosting in handling sentiment analysis tasks. These models lack the ability to capture the deep contextual and sequential nuances present in textual data, which are crucial for accurate sentiment classification. The basic approaches used for vectorization, such as TF-IDF and Bag of Words, also struggle with out-of-vocabulary words. The results echo the findings of Yang et al.[10], who demonstrated that deep learning models significantly outperform traditional ML models in natural language processing tasks due to their superior ability to capture context.

5.2 Future Improvements

Numerous studies have shown that sentiment classifiers are often optimised to categorise text as either negative or positive, thus forcing balanced sentences into one of these two categories[13,14]. This is a common issue in sentiment analysis. Therefore, it is worthwhile to conduct further research in this area in the future.

On common review websites, such as Tripadviser and Rotten Tomatoes, reviews would be analysed from several aspects but not an overall score. This applies to our case as well. For example, customers from different classes might have different expectations - first class customers care more about the service experience than price, and so on. In light of this, the current sentiment analysis models can be further fine-tuned for specific aspects of the flight experience, like seat comfort, staff service, food and beverage, and value for money. Ratings for these aspects are also provided in the existing dataset.

Additionally, a more reliable method of scoring the sentiment of reviews should be explored if they are to be used as training data in the future, for instance manually labelling the reviews, which would be expensive but beneficial to the accuracy of the models.

Given the performance of the current models, however, they might still be useful for the airline companies. They can be used to detect negative comments from reviews with medium ratings, which in turn helps identify areas for improvement for the airline.

6 References

1. Samir HA, Abd-Elmegid L, Marie M: Sentiment analysis model for Airline customers' feedback using deep learning techniques. Sage Journals 15, 1-23(2023)
2. Sanwal M, Mazhar MM: Performance Comparison of Machine Learning and Deep Learning Models for Sentiment Analysis of Hotel Reviews. International Journal of Information Technology and Applied Sciences 5(1), 1-7(2023)
3. Aysu E: A Comparison of LSTM and BERT for Small Corpus. ArXiv abs/2009.05451, 1-12 (2009)
4. Kaggle, https://www.kaggle.com/datasets/sujalsuthar/airlines-reviews/data, last accessed 2024/7/4
5. Hutto, Clayton, and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the international AAAI conference on web and social media. Vol. 8. No. 1. 2014.
6. TexBlob, https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis, last access 2024/05/14.
7. Valdivia, Ana, et al.: Consensus vote models for detecting and filtering neutrality in sentiment analysis. Information Fusion 44, 126-135(2018)
8. Shelar, Amrita, and Ching-Yu Huang: Sentiment analysis of twitter data. 2018 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2018.
9. Howard J, Ruder S: Universal Language Model Fine-tuning for Text Classification. ArXiv abs/1801.06146, (2018)
10. Yang H, Luo L, Lap P.C, Ling D, Chin F.Y.L: Deep Learning and Its Applications to Natural Language Processing. Cognitive Computation Trends, (2019)