Abstract

Examining the opinions and sentiments conveyed in reviews shared by customers provides businesses with insights into product quality and customer preferences and helps pinpoint areas that may require enhancement. For e-commerce players, analyzing customer reviews could prove beneficial for competitive analysis, assisting these businesses in comprehending their standing within the market. Furthermore, it could play a pivotal role in their decision-making processes by guiding product development, shaping effective marketing strategies, and facilitating improvements in customer service. This study conducted sentiment analysis on customer reviews about eBay and Amazon. The reviews used for the analysis were collected by web scraping the Trust Pilot website using the Beautifulsoup library in Python. The eBay reviews collected were for the period from 2017 to 2023, while for Amazon, the data collected was for the period from 2020 to 2023. For both dataset groups, the models were trained on reviews for the periods before 2023 and tested using reviews shared in 2023. Five models were used for the experiments conducted, including logistic regression, support vector machines, bi-LSTM, CNN, and BERT. In addition, TF-IDF and BOW techniques were used for feature extraction in the logistic regression and support vector machines. On the other hand, Glove embedding was used for experiments conducted on the Bi-LSTM and CNN models. The BERT model achieved the best performance, demonstrating a good ability to limit the number of negative reviews incorrectly predicted as positive (FP). For the eBay dataset, the model recorded a recall score of 75% for the positive class, while this was 84% for the positive class in the Amazon dataset.

1 Introduction

Over the last two decades, the e-commerce landscape has undergone a profound period of transformation, aided primarily by access to smartphones, high-quality internet services, secure online payment systems, and the convenience of doorstep delivery. (AlQahtani, 2021). According to Insider Intelligence (2021), e-commerce markets across the world raked in an estimated US$4.3 trillion in sales in 2020, experiencing more than 230.8% growth versus US$1.3 trillion in 2014 (Statista, 2023). For most of these sales, it is posited that customer purchase decision-making was primarily influenced by reviews shared by other customers. These online reviews helped rake in an estimated US$3.8 trillion in sales (Cavazos, 2019), given that about 89% of customers who shop online are believed to have done so only after reviewing opinions about a product or service shared online (Trustpilot, 2020), with this number increasing from 82% of customers in 2013 (Think Tank European Parliament, 2015). Reviews shared by other customers help potential new customers mitigate concerns about potential deception and enhance trust between customers and businesses within the e-commerce industry. (Aljuhani & Alghamdi, 2019; Singh et al., 2022).

Carefully examining reviews shared by customers can provide businesses with insights into product quality and customer preferences and help pinpoint areas that may require enhancement. Hence, for e-commerce players, analyzing customer reviews could prove beneficial for competitive analysis, assisting these businesses in comprehending their standing within the market. Furthermore, it could play a pivotal role in their decision-making processes by guiding product development, shaping effective marketing strategies, and facilitating improvements in customer service (Huang et al., 2023).

Sentiment analysis (SA) is a field of Natural Language Processing (NLP) that involves the use of big data and computational methods including Machine Learning (ML) algorithms to uncover emotions in text, which could be particularly valuable to players in the e-commerce industry (Hidayat et al., 2022; Singh et al., 2022). In recent years, NLP has gained significant attention for analyzing textual data among corporates and researchers. Large fast food brands including Mcdonald's, Pizza Hut, and KFC have implemented NLP technologies to analyze customer reviews allowing them to improve products and services (AI Multiple, 2023). NLP techniques enable the analysis of large-scale data as opposed to having humans manually analyze the data. Various SA techniques can be used to extract insights from customer reviews including lexicon-based, ML, and DL models. Lexicon models use rules or sentiment lexicons, ML techniques employ algorithms to understand patterns in labeled data, while DL techniques excel at capturing complex patterns and context from customer reviews (Tammina & Annareddy, 2020; AlQahtani, 2021; Huang et al., 2023).

1.1 Research Objectives

Amazon and eBay are some of the biggest e-commerce players in the world offering a vast range of products on their platforms. Given their status as global leaders in the e-commerce market, their platforms ordinarily receive many reviews about the products sold. In addition, their operational activities and financial performance typically serve as key indicator of the direction of the e-commerce market globally. Collectively, both companies account for an estimated 15% of the global ecommerce market (Channelsight, 2023). This makes both players compelling subjects for research.

The primary goal of the study is to analyze the reviews shared about products on these platforms to understand whether they contain positive or negative sentiments. The research question to address is:

i. Which ML or DL models perform best in classifying reviews posted about Amazon and eBay into positive or negative sentiments?

2 Related Work

Several studies have been conducted using datasets containing customer reviews across different industries and about several companies. These studies have used various ML and DL NLP techniques, in addition to experimenting with different text processing and feature extraction techniques. Using ML models, Haque et al. (2018) in their study conducted sentiment analysis using six classifiers, and their analysis was carried out on reviews about 3 different product categories including cell phones and accessories, musical instruments, and electronics data. The models considered included Stochastic Gradient Descent (SGD), Linear Support Vector Machine (SVM), RF, Multinomial Naïve Bayes (MNB), Decision Tree (DT) and Logistic Regression (LR). Of all the models, and across all 3 dataset groups, the SVM models achieved the best accuracy scores with 93.6% on the cell phones and accessories data, 94.0% on the musical instruments data, and 93.52% on the electronics data, with F1-Score of 97%, 98% and 98% respectively. Their analysis showed the efficacy of SVM models for sentiment analysis.

The use of DL models has gained popularity for SA in recent years. In their study, AlQahtani (2021) conducted sentiment analysis on Amazon reviews and compared various feature extraction techniques and ML models including Bag-of-Words (BOW), TF-IDF, GLOVE, LR, RF, NB, Bi-LSTM, and BERT. The dataset used for the experiments contained 413,840 reviews with ratings ranging from 1 to 5. The ratings in the dataset were used to create 2 categories of

labels with one group categorized into binary labels of positive and negative and another group with 3 labels of positive, negative, and neutral. For the first set of experiments with the multiclass labels, the BERT model had the best accuracy score of 94.7% and F1 score of 94.6%. For the second group of experiments with the binary labels, the BERT model also recorded the best accuracy and F1 scores of 98.4% and 98.4% respectively. The BERT model outperforming the other models compared validated the case of superiority over ML models. Aljuhani and Alghamdi (2019) in their research used techniques such as BOW, TF-IDF, Glove, and Word2Vec, and compared the results of these techniques on 4 models including LR, NB, SGD, and CNN. The researchers used the Amazon reviews dataset which consisted of 400,000 reviews and also conducted analysis on the original imbalanced data, in addition to the balanced data. The CNN with Word2Vec achieved the best accuracy score across the imbalanced and balanced dataset with scores of 92.72% and 79.6% respectively.

In their study, Singh et al. (2022) explored how different word embedding techniques impact ML models for classifying customer reviews. Their analysis was conducted on customer reviews of Amazon which contained 34,660 reviews. The researchers used 4 models including LSTM, CNN, Multi-Channel CNN, and RMDL. The word embedding techniques compared in the study were Elmo, Fast text, GLOVE, and BERT. The Multi-Channel CNN with Fast text embedding recorded the best accuracy score of 79.83%. Zhao and Sun (2022) in their study used the BERT model to predict review scores assigned to reviews provided by customers about food products on Amazon. The dataset used for the analysis contained 568,454 customer reviews. The BERT model used for the experiment scored an accuracy of 79.8%.

An analysis of the previous work in the field shows that for researchers who have focused solely on ML approaches, LR and SVM models are popular ML models used for SA. Also, in experiments that have involved the use of DL and ML models, researchers typically aim to assess the performance of these DL models against baseline ML models. Furthermore, there is a growing body of work on the use of BERT models for SA as these models have been proven by several researchers to outperform RNN and ML models as in the case of AlQahtani (2021).

For most of the previous work reviewed and over the course of this study, there is limited research on SA for eBay. Therefore, this study will aim to compare the results of models developed with those already achieved by other researchers for Amazon. In addition to this, this study will conduct SA on eBay reviews and also assess the performance of models used to do this against the performance of results achieved by other researchers for other datasets. Unlike previous studies which split datasets for training and testing without considering to the periods, the analysis in this study will aim to create a holdout test set mimicking what is

obtainable in industry and real-life situations where models built are expected to take new chronological data as inputs.

3 Methodology

3.1 Data Acquisition

Two groups of datasets were collected for this study - one for eBay reviews and the other for Amazon reviews. The reviews were collected by web scraping the Trustpilot website, using the Beautifulsoup library in Python. For each of these companies, 10,000 reviews were collected each. The eBay reviews collected were for the period from 2017 to 2023, while for Amazon, the data collected was from the periods between 2020 to 2023. Each of these datasets contained 16 columns. Table 1 presents the data web scraped from Trustpilot for eBay.

Table 1: eBay data description

Feature	Description
Query	URL link to company page on Trustpilot
Total reviews	Total number of reviews about eBay
Review rating	Rating assigned by the author
Review title	Review the title as provided by the author
Review text	Content: actual review provided
Review likes	Review likes
Review date	Date of the review
Review date UTC	Date of the review in UTC
Review ID	The review ID
Author Title	Author name
Author ID	I assigned to the author
Author reviews number	Number of all reviews shared by the author
The author reviews number same domain	Number of reviews shared by the author
Author country code	Country of the author
Owner answer	Answer from the company
Owners answer date	Date company answered

3.2 Data and Reviews Cleaning Steps

Classic data cleaning was first done on the data before conducting text cleaning. This step involved checking for duplicates in the reviews and removing rows that are duplicated texts. The text cleaning actions involved removing characters that are considered noise when

implementing NLP models. These actions included removing characters such as emojis, punctuations, non-ASCII characters, URLs, HTMLs, numbers, and English language stop words. In addition, English language contractions were expanded whereby words such as "I'm" become "I am". Also, words were lemmatized which involves returning them to their base forms otherwise known as lemmas (Singh et al., 2022). After data cleaning, the dataset size for eBay was 9,984 reviews while the dataset size for Amazon was 9,989. Table 2 presents the characteristics of the dataset after data and text cleaning.

Table 2: Characteristics of the datasets after preprocessing

Data Group	Count of labels	Max. review length	Min. review length	Mean review length	Train data size	Test data size
eBay	2	609	2	47	8,067	2023 reviews: 1917
Amazon	2	454	2	49	8,023	2023 reviews: 1966

3.3 Sentiment Analysis

Following data and text cleaning, the dataset was analyzed in an effort to gather relevant insights. For this analysis, the ML models used are Support Vector Machines and Logistic Regression. Bidirectional Long Short-Term Memory (Bi-LSTM), Convolutional Neural Network (CNN), and the Bidirectional Encoder Representations from Transformers (BERT) transformer model were the DL models used. Bag of Word (BOW) and Term Frequency – Inverse Document Frequency (TF-IDF) were the text representation and feature extraction methods used in the ML models. For the DL models, GLOVE was also used as a word embedding technique. Word embeddings transform words into compact, dense vectors, capturing both syntactic and semantic information (Zhang et al., 2018; Khatri, 2020). The hyperparameters used for these models are discussed in the Experiments section.

4 Experiments

4.1 Data Labels

The ratings available in the dataset as provided on the Trustpilot website were between the range of 1 to 5 with a rating of 1 meaning bad and a rating of 5 meaning great. For this study, ratings 2 and 1 were taken together as negative, and ratings 4, 3, and 5 were combined to be positive. Figure 1 presents the count of labels and classes in the eBay data while Figure 2

presents the count of labels and classes in the data after combining the ratings in the Amazon data.

4.2 Class Imbalance: Upsampling with SMOTE

For this study, SMOTE was chosen as the technique for upsampling the data. Other methods were considered including random over-sampling which simply randomly takes and duplicates reviews from the minority class, and random under-sampling simply randomly eliminates some samples in the majority class. However, the decision against using these other approaches was based on the highly imbalanced data structure, in addition to the conclusion from Jiang (2003) which noted that randomly over-sampling sometimes may lead to overfitting while under-sampling takes out relevant data.

Figure 1: Reviews sentiment class and number of reviews in the eBay data

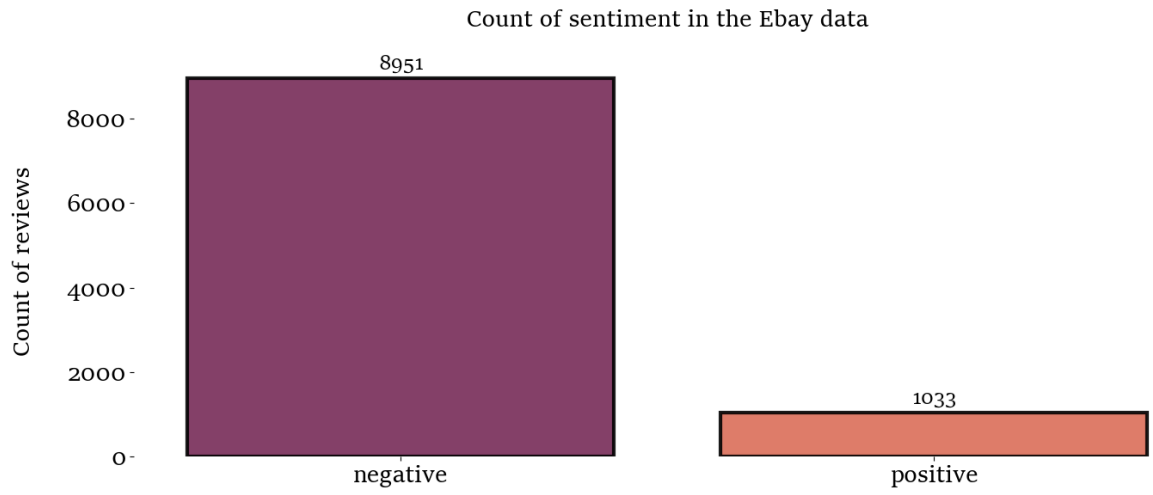
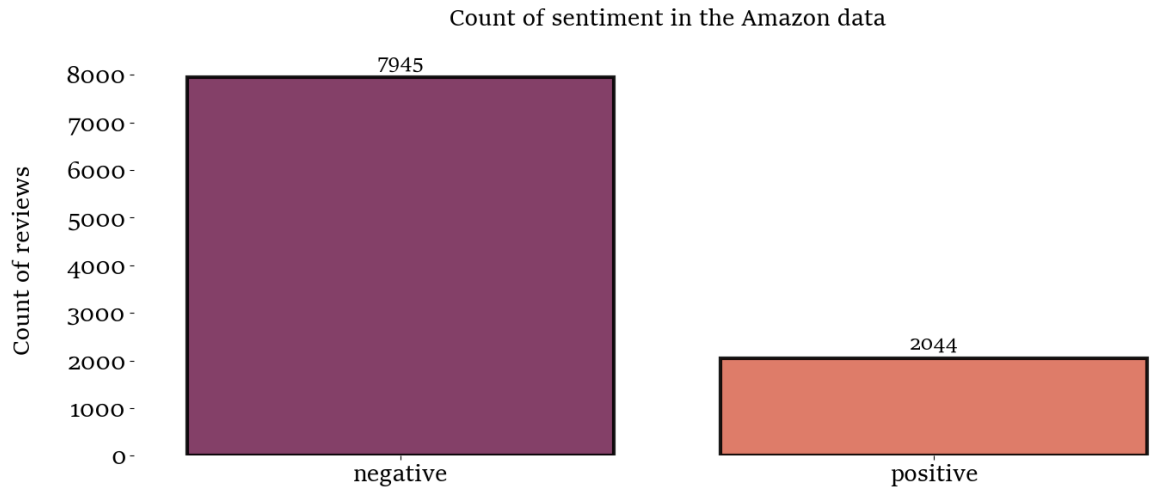


Figure 2: Reviews sentiment class and number of reviews in the Amazon data



4.3 Models Parameter Settings

For the experiments, the dataset was separated into training and validation groups using chronological order. For both datasets, reviews provided in the 2023 were taken as the test data while reviews any time before were taken as the training data. For the ML models, Random Search hyperparameter tuning was implemented to look for the optimal parameters for the LR and SVM models. Table 4 presents the hyperparameters used for the different ML models.

The parameters used in the DL and BERT models are presented in Table 3. The CNN and Bi-LSTM models use the Keras embedding layer in addition to also using the Glove embedding layer for separate experiments. For these models, the 100-dimensional vector pre-trained word embeddings were used. These glove embeddings were downloaded from the website of Jeffrey Pennington et al. (2014).

Table 3: DL and BERT model parameters

Dataset	**Bi-LSTM**	**CNN**	**BERT**
Learning rate	0.0001	0.0001	0.00001
Optimizer	Adam	Adam	Adam
Batch Size	128	128	32
Epochs	30	30	5
Embedding	Keras and Glove	Keras and Glove	BERT
Maximum length	200	200	200
Vocab size	16,000	16,000	-
Output dimension	100	100	-
Others	2 Bi-LSTM layers of 128 and 32 units, a dropout of 0.4 in between them. The second Bi-LSTM layer contained the ReLu activation. The last layer, a Dense layer of one unit given the binary classification problem uses the Sigmoid activation.	1D convolutional layer with 128 filters, each of size 5, next was a GlobalMaxPooling1D layer. A fully connected dense layer with 32 and 16 units and ReLu activation functions were followed. A one-unit dense layer was the last with the activation Sigmoid.	BERT model with a fully connected layer with 64 units hidden Dense layer and tanh activation function. The model also includes a 20% dropout applied to the first output layer and uses Sigmoid activation in the final Dense layer

Table 4: ML model parameters

Dataset	Feature extraction	Models	Hyperparameter	Value
eBay	**BOW**	LR	Penalty	L2
			C	100
		LR with SMOTE	Penalty	L2
			C	10
		SVM	Kernel	Linear
			Gamma	Scale
			Scale	10,000
		SVM with SMOTE	Kernel	Linear
			Gamma	Auto
			Scale	10,000
	TF-IDF	LR	Penalty	L2
			C	10,000
		LR with SMOTE	Penalty	L2
			C	100
		SVM	Kernel	Linear
			Gamma	Scale
			Scale	10,000
		SVM with SMOTE	Kernel	Linear
			Gamma	Auto
			Scale	10,000
Amazon	**BOW**	LR	Penalty	L2
			C	10,000
		LR with SMOTE	Penalty	L2
			C	10
		SVM	Kernel	Rbf
			Gamma	Scale
			Scale	10
		SVM with SMOTE	Kernel	Linear
			Gamma	Auto
			Scale	10
	TF-IDF	LR	Penalty	L2
			C	10,000
		LR with SMOTE	Penalty	L2
			C	100
		SVM	Kernel	Linear
			Gamma	Scale
			Scale	10,000
		SVM with SMOTE	Kernel	Linear
			Gamma	Scale
			Scale	10,000

4.4 Evaluation Metrics

An assessment of the quality of models developed in this study was evaluated using metrics including precision, accuracy, F1-score, and recall. Accuracy gauges overall correctness. Precision assesses positive prediction accuracy, while recall evaluates a model's ability to identify positive samples. F1-score, considering both precision and recall, is an important metric for imbalanced datasets. In addition to this, the model evaluation will be carried out using a confusion matrix which shows the predictions of the model regarding actual positive classes predicted correctly called True Positives (TP), wrongly predicted positive samples called False Positives (FP), correctly predicted negative class called True Negatives (TN) and wrongly predicted negative class called False Negatives (FN) (Hameed & Garcia-Zapirain, 2020; Kumar et al., 2020; Tusar & Islam, 2021). The formulas for these metrics are presented in Equations 1 to 4.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Neg + True\ Pos}}{\boldsymbol{True\ Neg + True\ Pos + False\ Neg + False\ Pos}} \qquad \text{Equation 1}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Pos}}{\boldsymbol{True\ Pos + False\ Pos}} \qquad \text{Equation 2}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{True\ Pos}}{\boldsymbol{True\ Pos + False\ Neg}} \qquad \text{Equation 3}$$

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

5 Results

Across both datasets (eBay and Amazon) 16 experiments were conducted using LR and SVM ML models. These experiments involved the use of different feature extraction techniques and the original and balanced dataset using SMOTE. Similarly, for the DL models, 16 experiments were conducted across both datasets with some experiments involving the use of the Keras embedding layer while some experiments included a Glove embedding layer. 4 experiments were conducted for the BERT models. Also, experiments were conducted using the original and datasets balanced with SMOTE. Earlier, it was noted that the 2023 reviews across the eBay and Amazon dataset were used to test the models. Figure 3 shows the split of sentiment across the years with 2023 indicating an imbalanced data for eBay. Figure 4 presents the split of sentiments over the years for Amazon.

Figure 3: Split of sentiments across the years for eBay

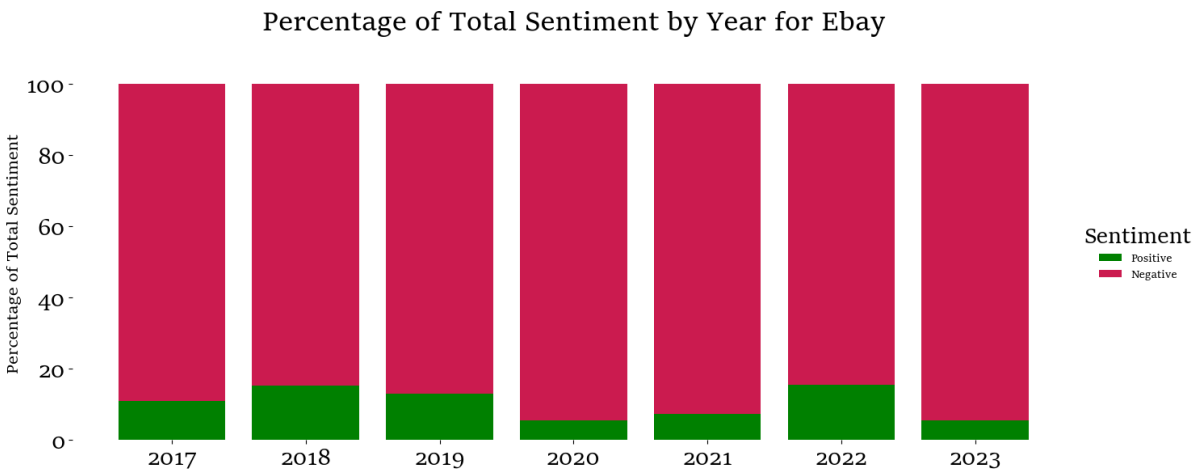
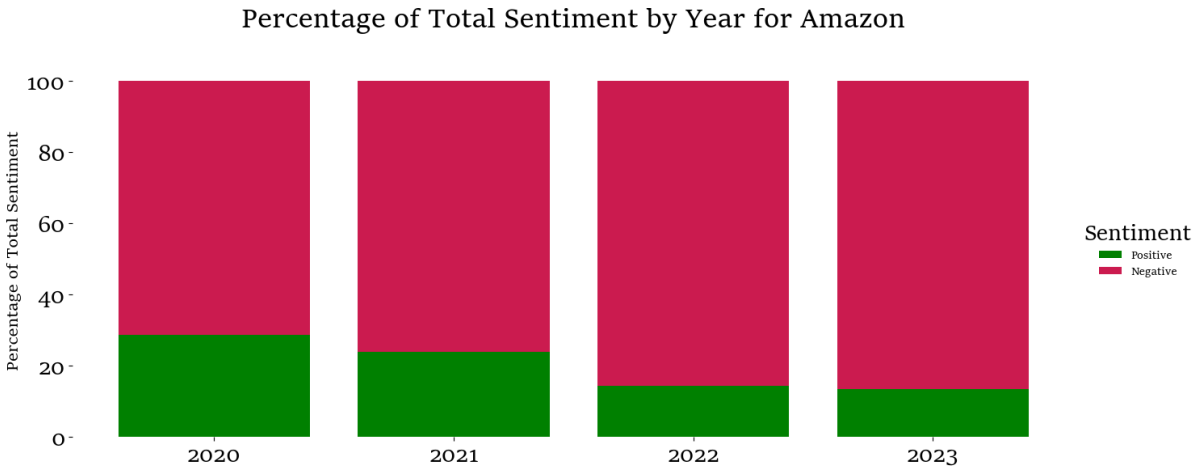


Figure 4: Split of sentiments across the years for Amazon



5.1 Results of ML Models

Of the 8 ML experiments conducted for eBay, all but one of the models recorded accuracy scores of 97%. The LR model with BOW and the dataset balanced using SMOTE recorded an accuracy score of 96%. Given the nature of this study whereby the aim is to address customer needs especially focusing on negative reviews, this means attention should be paid to negative reviews which are classified as positive reviews, known as False Positives. When this is considered, the best ML model for the eBay data is the LR model with the data balanced with SMOTE and with BOW feature extraction. As presented in Table 4, this model recorded the

highest balanced recall score of 79%. Also, as presented in Figure 5 the model had the lowest TP reviews with 42 (of the 106 positive reviews) - a recall score of 60% for the positive class.

For the Amazon data, all 8 models had accuracy scores of 95% as shown in Table 5. Just as with the eBay data, the LR with BOW and the SMOTE dataset had the lowest number of FP of 61 reviews, and a recall score of 77% for the positive class. Figure 6 presents the confusion matrix for the LR model with BOW and SMOTE.

Figure 5: LR - BOW - SMOTE for eBay dataset: Confusion Matrix

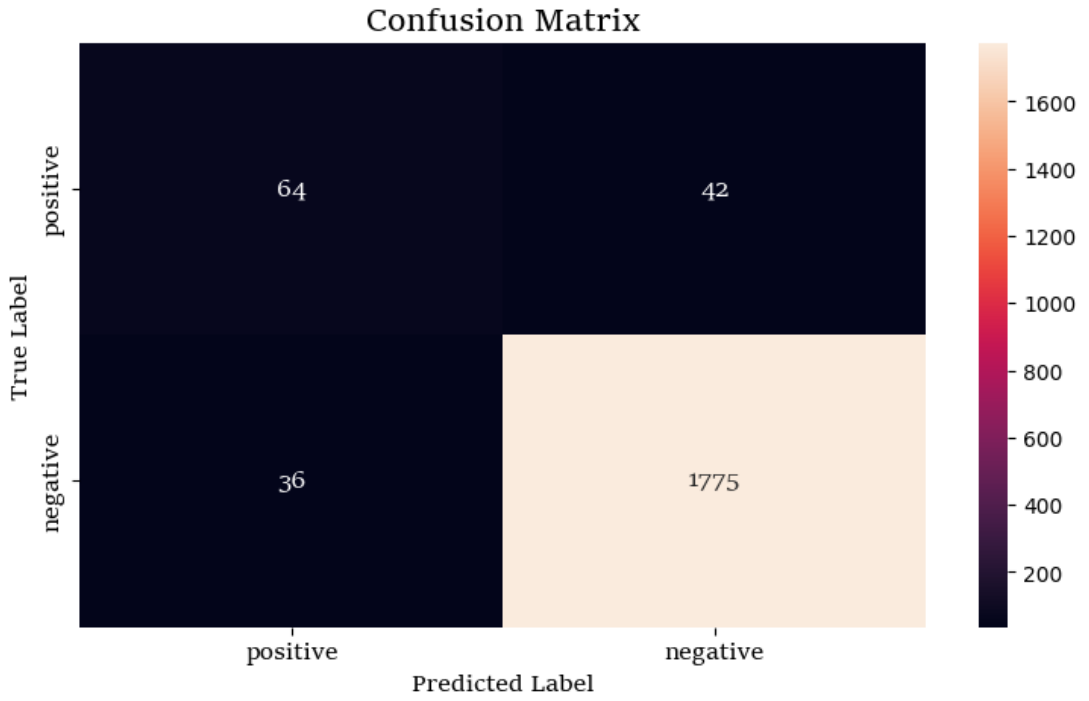


Figure 6: LR - BOW - SMOTE for Amazon dataset: Confusion Matrix

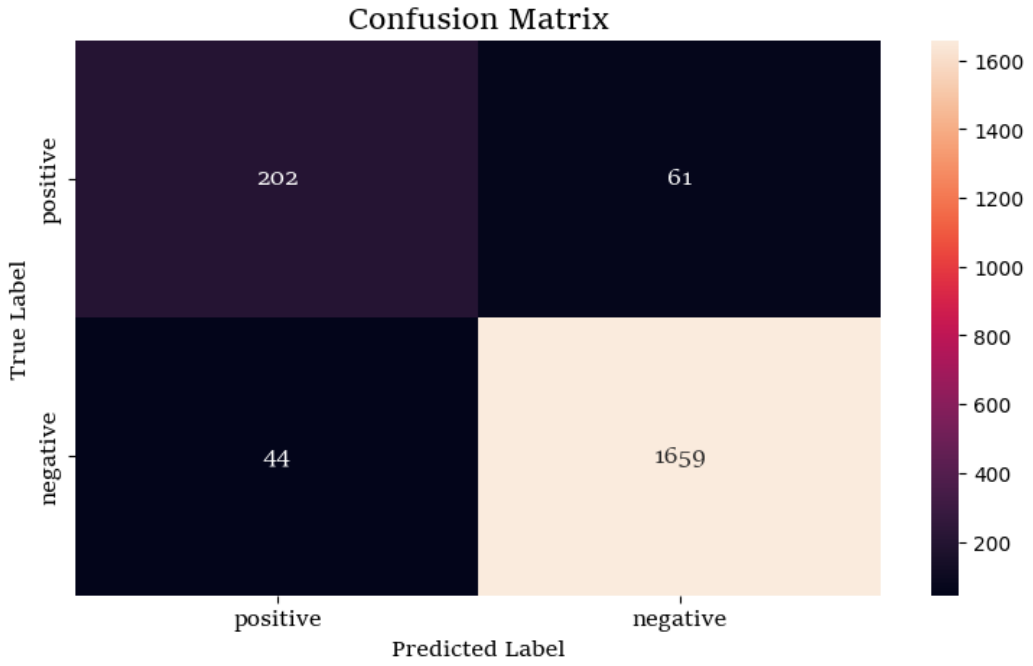


Table 5: Results of the ML models

Model	Company	Feature Extraction	Model	Acc.	Precision	Recall	F1-Score
Imbalanced	**eBay**	**BOW**	LR	97%	87%	76%	81%
			SVM	97%	87%	76%	81%
		TF-IDF	LR	97%	92%	76%	82%
			SVM	97%	90%	77%	82%
Balanced		**BOW**	LR	96%	81%	79%	80%
			SVM	97%	87%	76%	81%
		TF-IDF	LR	97%	87%	77%	81%
			SVM	97%	90%	77%	82%
Imbalanced	**Amazon**	**BOW**	LR	95%	91%	86%	88%
			SVM	95%	93%	84%	88%
		TF-IDF	LR	95%	93%	86%	89%
			SVM	95%	92%	86%	89%
Balanced		**BOW**	LR	95%	89%	87%	88%
			SVM	95%	90%	86%	88%
		TF-IDF	LR	95%	92%	87%	89%
			SVM	95%	92%	86%	89%

5.2 Results of DL Models

For the DL models, the best accuracy score for the eBay dataset was 96% recorded before balancing the dataset with SMOTE. The Bi-LSTM and CNN models with Keras embedding and

the CNN model with Glove embeddings recorded these results. Table 6 presents the results of the DL models. Across the models, as seen in the confusion matrix in Figure 7, the lowest number FP of 33 was achieved by the Bi-LSTM with Keras embedding and on the balanced data – a recall score of 69% for the positive class. For the Amazon dataset, the Bi-LSTM and CNN models with Keras embedding and imbalanced data had the highest accuracy scores of 94%, whereas the CNN model in this case had the best precision of 88%. Table 5 presents the results of the DL models for the Amazon dataset. Overall, the best model when FP is considered is the Bi-LSTM model with Glove embedding and balanced data had the lowest number of FP of 50, a recall score of 81% for the positive class. Figure 8 presents the confusion matrix for this model.

Figure 7: Confusion matrix of the Bi-LSTM and balanced data for the eBay dataset

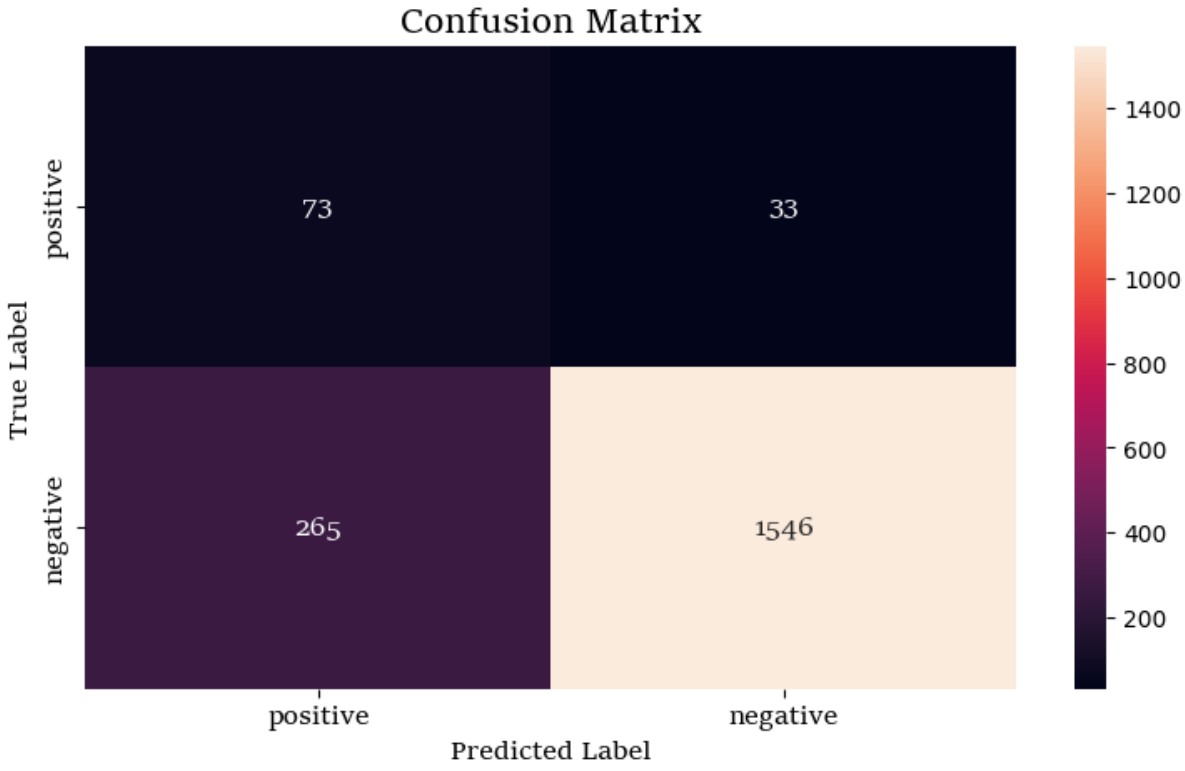
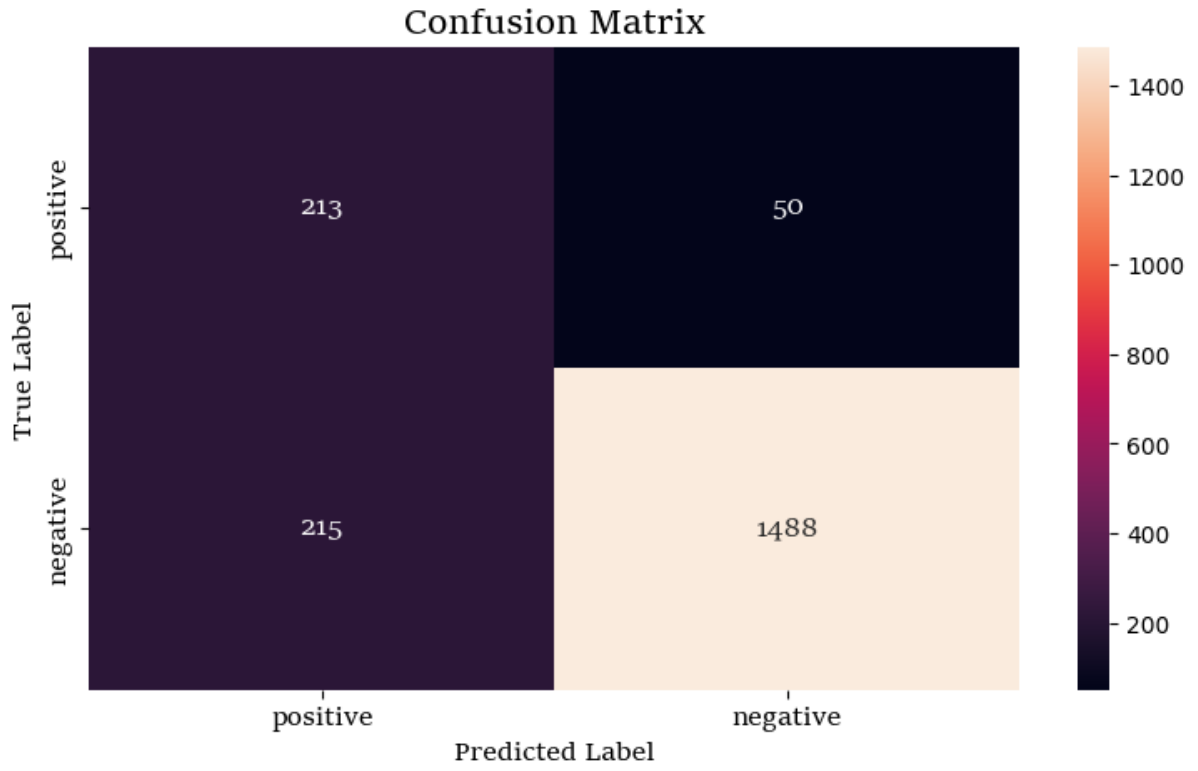


Table 6: Results of the DL models

Model	Company	Embedding	Model	Acc.	Precision	Recall	F1-Score
Imbalanced	**eBay**	**Keras**	Bi-LSTM	96%	81%	76%	78%
			CNN	96%	83%	74%	78%
		GloVe	Bi-LSTM	95%	76%	78%	77%
			CNN	96%	85%	74%	79%
Balanced		**Keras**	Bi-LSTM	84%	60%	77%	62%
			CNN	92%	67%	80%	71%
		GloVe	Bi-LSTM	92%	67%	77%	70%
			CNN	92%	67%	78%	71%
Imbalanced	**Amazon**	**Keras**	Bi-LSTM	94%	87%	85%	86%
			CNN	94%	88%	85%	86%
		GloVe	Bi-LSTM	93%	86%	83%	84%
			CNN	92%	86%	75%	79%
Balanced		**Keras**	Bi-LSTM	87%	73%	83%	77%
			CNN	91%	79%	85%	81%
		GloVe	Bi-LSTM	87%	73%	84%	77%
			CNN	88%	75%	77%	76%

Figure 8: Confusion matrix of the Bi-LSTM with Glove and balanced data for the Amazon dataset



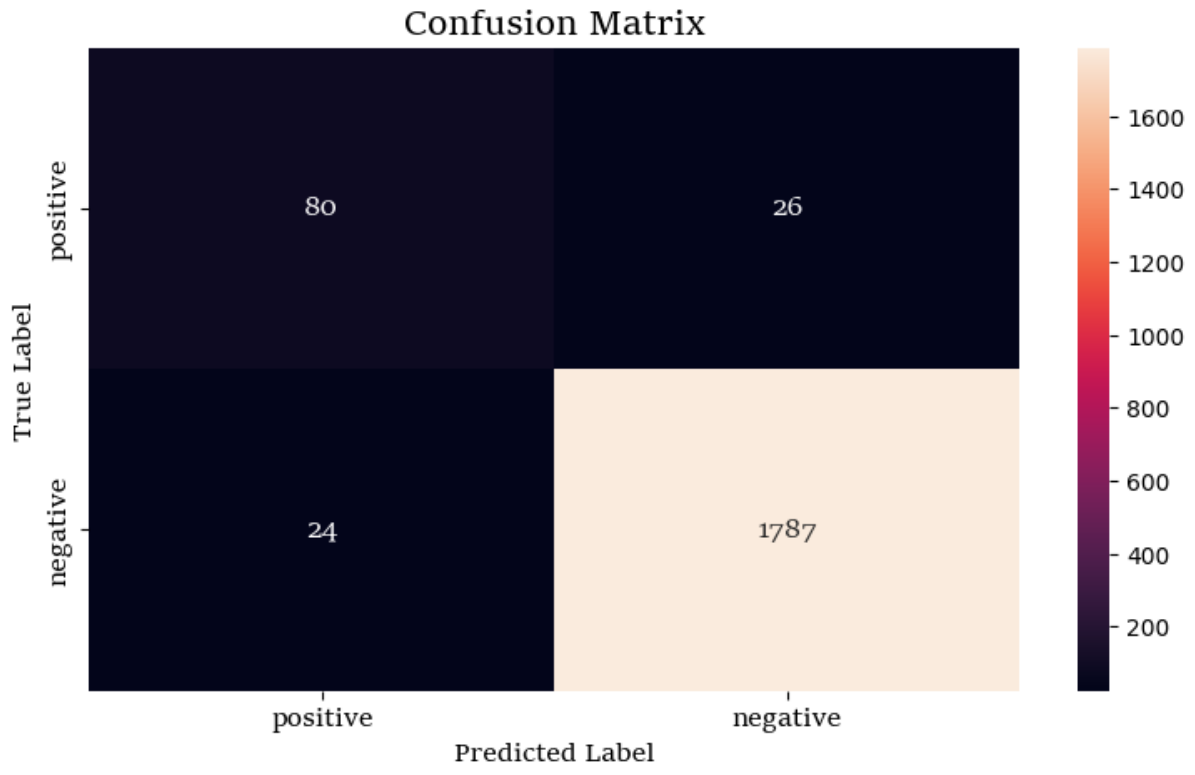
5.3 Results of the BERT models

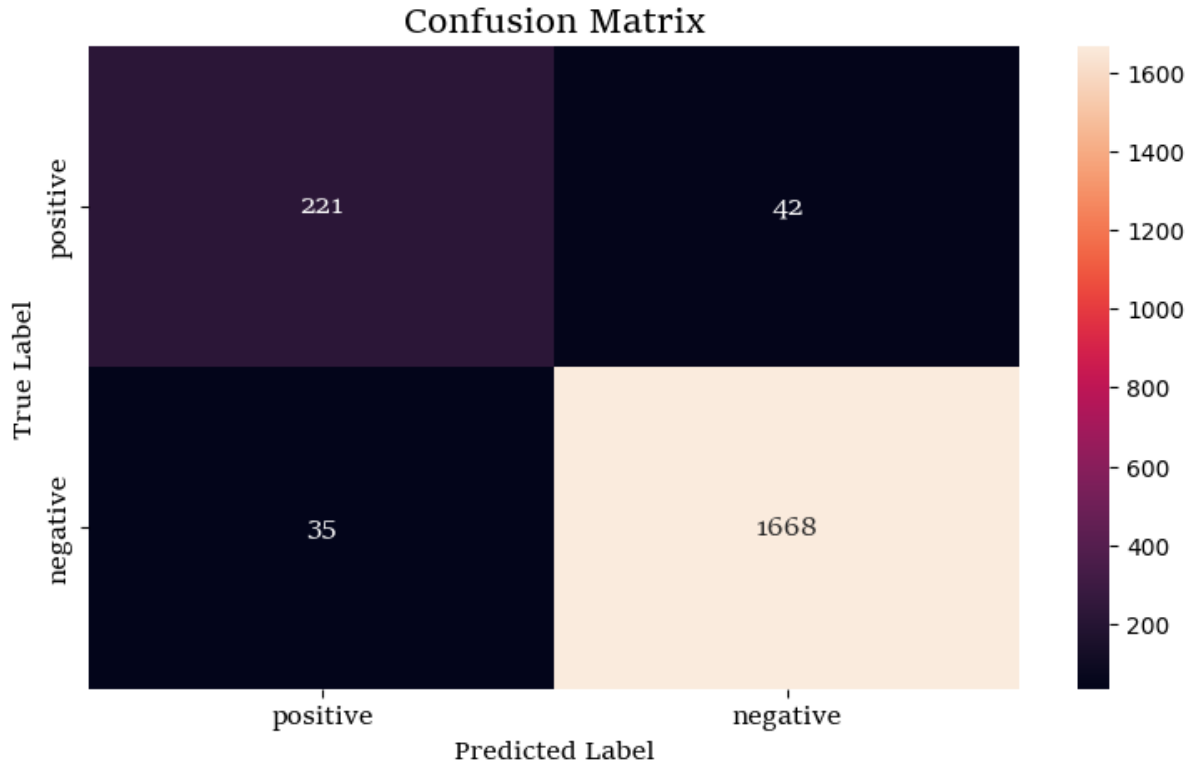
The BERT models trained on the eBay dataset both had similar metrics of 97%, 88%, 87%, and 87% for accuracy, precision, recall, and F1-score respectively, for the imbalanced and balanced datasets. For Amazon, again both balanced and imbalanced datasets achieved the same accuracy score of 96%, however, the model trained on the imbalanced data had a better precision score of 95%. Table 7 presents the results of the BERT models. As presented in Figure 9 for eBay, the BERT model on the imbalanced data returned the lowest number of FP of 26, a recall score of 75% for the positive class, while for the Amazon dataset, the BERT model on the balanced dataset returned the lowest FP of 42 and a recall of 84% for the positive class.

Table 7: Results of the BERT models

Model	Company	Acc.	Precision	Recall	F1-Score
eBay	**Imbalanced**	97%	88%	87%	87%
	Balanced	97%	88%	87%	87%
Amazon	**Imbalanced**	96%	95%	89%	92%
	Balanced	96%	92%	91%	91%

Figure 9: BERT model confusion matrix on the imbalanced data for eBay (top) and BERT model confusion matrix on the balanced data for Amazon (bottom)





5.4 Discussion of Results

In total 36 experiments were conducted to classify sentiments into positive and negative classes. This was done for 2 groups of datasets covering reviews about eBay and Amazon. Across both datasets, the negative sentiment group was the majority class accounting for more than 75% of the total reviews. The SMOTE upsampling technique was used to address this class imbalance in the dataset, a technique that generates synthetic data using neighborhood information to balance the minority class with the majority class. The analysis compared the results of traditional ML models such as SVM and LR and DL models including Bi-LSTM and CNN. For the ML models, BOW and TF-IDF were used for feature extraction while for the DL models, Glove was used in the embedding layer for the CNN model as against the Keras embedding layer used for the other experiments.

Given the nature of the problem, whereby the goal was identifying all possible negative reviews as this would allow companies to gain insights into reasons why customers might be unhappy or dissatisfied with products and services, attention was paid to the recall score for the positive class, in this case, positive reviews, hence the aim was reducing the number of FP as much as possible. The thinking here is that when negative reviews are classified as positive, these relevant reviews which could otherwise have provided insights into what customers are disappointed or unhappy about will be missed and could potentially result in poor business performance, whereby these missed reviews then influence the decision making of potential new customers. Of all the models implemented for the eBay dataset, the BERT model trained on the imbalanced dataset had the lowest number of FP of 26 reviews, wherein the model classified 84% of the positive reviews properly. This model recorded an accuracy of 97% with balanced precision, recall, and F1-score of 88%, 87%, and 87% respectively. For the Amazon dataset, again the BERT model with balanced data recorded the lowest false positive results of 42 and this model had accuracy, precision, recall, and F1-scores of 96%, 92%, 91%, and 91% respectively.

Compared with the works of other researchers who have conducted studies on customer reviews about Amazon, the accuracy and recall scores of 96% and 91 % recorded by the BERT model on the balanced dataset are better than the results obtained by Haque et al. (2018) whose SVM model achieved a 93.6% accuracy for classifying reviews about phones and accessories. The results are also better than the results recorded by the BERT model developed by AlQahtani (2021) for their multiclass classification of reviews. However, the result in this study falls short of the results of their binary classification BERT model which recorded accuracy scores of 98.4%.

Furthermore, the best performing BERT model developed for this study had better accuracy scores than the model developed by Zhao and Sun (2022) for whom the BERT model used for their experiments scored an accuracy of 79.8%.

Three key points to highlight from the experiments conducted in this study include the importance of having balanced data when conducting SA experiments, the efficacy of the performance of BERT models in SA, and lastly, the accuracy metrics do not necessarily signify that a model will perform as expected when SA is concerned. The last point emphasizes why it is important to understand the dynamics of the dataset under consideration and the key outcomes for any SA project.

6 Conclusion

This study looked to explore the customer reviews provided by customers of Amazon and eBay and analyze sentiments shared by their customers. A total of 36 experiments were conducted using ML, DL, and transformer models including LR, SVM, Bi-LSTM, CNN, and BERT. In addition, feature extraction techniques of TF-IDF, BOW, and Glove were combined with the models. Given that the dataset was highly imbalanced, the SMOTE technique was used to balance the data, a process that involved generating synthetic data from the minority class and upsampling to match the majority class. Overall, of the five models used for the analysis, the BERT model demonstrated a good ability to limit the number of negative reviews incorrectly predicted as positive (FP). For the eBay dataset, the model recorded a recall score of 75% for the positive class, while this was 84% for the Amazon dataset. By using this metric, the goal is to address a situation whereby company stakeholders do not miss negative reviews which could be useful for them to improve the overall customer experience. The key limitation for this study is the lack of access to a more robust and larger dataset, wherein this would have allowed to implementation of other SA techniques such as Topic modelling which would have aided in understanding the common topics that customers were happy or unhappy about over time. For future works, for the DL models, other embedding techniques such as Word2Vec, Fast Text, and ELMO should be considered to assess their effectiveness versus the Glove technique. Furthermore, the performance of the Roberta model should be assessed as well to see how it performs versus the BERT model.

7 References

AI Multiple (2023) *Top 4 Use Cases of Sentiment Analysis in Marketing in 2023*. Available online: https://research.aimultiple.com/sentiment-analysis-marketing/#:~:text=For%20instance%2C%20KFC%2C%20Pizza%20Hut,can%20benefit%20from%20sentiment%20analysis. [Accessed 7/12/2023].

Aljuhani, S. A. & Alghamdi, N. S. (2019) A comparison of sentiment analysis methods on Amazon reviews of Mobile Phones. *International Journal of Advanced Computer Science and Applications*, 10(6).

AlQahtani, A. S. (2021) Product sentiment analysis for amazon reviews. *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, 13.

Cavazos, R. (2019) *The economic cost of bad actors on the internet*. Available online: https://f.hubspotusercontent00.net/hubfs/5228455/Research/Fake%20Online%20Reviews%202021.pdf [Accessed 07/12/2023].

Channelsight (2023) *Selling on Amazon vs eBay*. Available online: https://www.channelsight.com/blog/selling-on-amazon-vs-ebay#:~:text=Recent%20data%20shows%20that%20Amazon,the%20global%20eCommerce%20market%20respectively. [Accessed 7/12/2023].

Hameed, Z. & Garcia-Zapirain, B. (2020) Sentiment classification using a single-layered BiLSTM model. *Ieee Access*, 8, 73992-74001.

Haque, T. U., Saber, N. N. & Shah, F. M. (2018) Sentiment analysis on large scale Amazon product reviews, *2018 IEEE international conference on innovative research and development (ICIRD)*. IEEE.

Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W. & Adisaputra, M. W. (2022) Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier. *Procedia Computer Science*, 197, 660-667.

Huang, H., Asemi, A. & Mustafa, M. B. (2023) Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions. *IEEE Access*.

Insider Intelligence (2021) *Global Ecommerce Update 2021: Worldwide Ecommerce Will Approach $5 Trillion This Year*. Available online: https://www.insiderintelligence.com/content/global-ecommerce-update-2021 [Accessed 7/12/2023].

Jeffrey Pennington, Richard Socher & Manning, C. D. (2014) *GloVe: Global Vectors for Word Representation*. Available online: https://nlp.stanford.edu/projects/glove/ [Accessed 30/11/2023].

Jiang, H. (2003) Sentiment analysis on imbalanced airline data. *Proceedings of Jiang2016AnalysisOI*.

Khatri, A. (2020) Sarcasm detection in tweets with BERT and GloVe embeddings. *arXiv preprint arXiv:2006.11512*.

Kumar, A., Chatterjee, J. M. & Díaz, V. G. (2020) A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing. *International Journal of Electrical and Computer Engineering*, 10(1), 486.

Singh, U., Saraswat, A., Azad, H. K., Abhishek, K. & Shitharth, S. (2022) Towards improving e-commerce customer review analysis for sentiment detection. *Scientific Reports*, 12(1), 21983.

Statista (2023) *Retail e-commerce sales worldwide from 2014 to 2026*. Available online: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/#:~:text=In%202021%2C%20retail%20e%2Dcommerce,8.1%20trillion%20dollars%20by%202026. [Accessed 7/12/2023].

Tammina, S. & Annareddy, S. (2020) *Sentiment Analysis on Customer Reviews using Convolutional Neural Network*.

Think Tank European Parliament (2015) *Online consumer reviews: The case of misleading or fake reviews*. Available online: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2015)571301 [Accessed 7/12/2023].

Trustpilot (2020) *The critical role of reviews in Internet trust*. Canvas8. Available online: https://cdn2.hubspot.net/hubfs/2749863/2019-trustpilot/The%20Critical%20Role%20of%20Reviews%20in%20Internet%20Trust%20(UK)%20-%20final.pdf [Accessed 7/12/2023].

Tusar, M. T. H. K. & Islam, M. T. (2021) A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data, *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*. IEEE.

Zhang, X., Chen, F. & Huang, R. (2018) A combination of RNN and CNN for attention-based relation classification. *Procedia computer science*, 131, 911-917.

Zhao, X. & Sun, Y. (2022) Amazon Fine Food Reviews with BERT Model. *Procedia Computer Science*, 208, 401-406.