Ingénierie des Systèmes d'Information
Vol. 30, No. 2, February, 2024, pp. 505-516
Journal homepage: http://iieta.org/journals/isi

Evaluating Machine Learning Performance and Consumer Sentiments on E-Commerce Platforms: A Comprehensive Twitter Analysis of Amazon

Asmaa Sami Mirdan[1], Mohammed Rashad Baker[1*], Selim Buyrukoğlu[2]

[1] Software Department, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk 36013, Iraq
[2] Computer Engineering, Faculty of Engineering, Cankiri Karatekin University, Cankiri 18100, Turkey

Corresponding Author Email: mohammed.rashad@uokirkuk.edu.iq

https://doi.org/10.18280/isi.300222

Received: 26 December 2024
Revised: 16 January 2025
Accepted: 14 February 2025
Available online: 27 February 2025

Keywords:
unbalanced dataset, sentiment analysis, machine learning, classification, product review

ABSTRACT

The blogs, social networks, and review portals became the new form of marketing after consumers gained the right to express themselves through the internet on numerous topics, from production reviews to elements of pop culture. These include social media, which enhances customer interactions and proves to be the greatest source of business data required for analysis in efficient sales planning and customer relationship management. Data was accumulated about Amazon Inc. products and services involving utilizer tweets on Twitter from 31th May 2022 to 10th June 2022, with a total of 471,840 tweets. This research offers a comprehensive analysis of user sentiment sales prediction and customer retention for the small-scale business as well as the large-scale sales business using various machine learning (ML) models, including Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost) and Stochastic Gradient Descent (SGD). The outcomes reveal that the accuracy is higher in case of classification of the polarity of the product reviews by the ML models. In particular, it was identified that LR is the strongest algorithm with regards to this sort of assessment. In this regard, the proposed LR model produced consistently higher individual accuracies than its counterparts and, therefore, can be deemed the optimal solution for sentiment classification in product reviews. However, other models obtain the actual accuracy in some cases. These findings add to a growing literature of knowledge regarding SA and ML, specifically in their ability to deliver businesses with sound tools in analysing consumer sentiment, forecasting sales and ultimately improving CRM strategies. The findings of this research have a number of practical applications for a range of stakeholders in the business world and thus go beyond the academic context.

1. INTRODUCTION

The rise of online shopping has attracted many Internet users who prefer to purchase products and services online [1]. However, the vast array of products on e-commerce websites can sometimes overwhelm customers and make it challenging to find the right product. This high level of competition among global trading sites emphasizes the need for effective strategies to increase financial profit [2]. At present, people register their opinions on various reviews of related products, brands and services on the internet and exchange opinions with other people [3]. The opinions generated are valuable assets that can be used to inform important decisions. Checking opinions related to products and services can not only improve their quality, but also influence the purchasing decisions of users. Many people's purchases are guided by learning about products and services in the virtual space and based on comments provided by other users.

Many large companies use customer reviews from online shopping sites and other web pages to develop and improve their business, including customer relationship management, increasing customer satisfaction, customer retention, and sales [4]. They also use online reviews to build their reputation and brand awareness. Despite the availability of free text comments online, businesses have invested significantly in studies and consultants to understand consumer perspectives on their products and services. Companies can support their offerings and take steps towards organizational success by utilising survey techniques and comment analysis. Applications of machine learning (ML) and natural language processing (NLP) have been widely researched for analyzing user behavior in e-commerce platforms [5, 6]. Using these technologies, businesses can gain insight into customer preferences, buying patterns and other valuable information to optimize marketing strategies, improve customer satisfaction and increase revenue.

Currently, with the increase in people using websites and social media to express opinions, the ability to perform sentiment analysis (SA) to predict attitudes has also grown [7]. SA involves classifying opinions into positive, negative, or neutral opinions through NLP. However, understanding consumer behavior is complex and not an easy job [8]. In this sense, researchers have applied various data mining approaches and ML models to historical online customer data

to predict the likelihood of future behavior [9]. Although SA has many applications, it faces NLP-related technical challenges. Issues in NLP limit the efficiency and accuracy of SA. Consequently, data mining techniques have become the focus of SA [10]. Additionally, social media has become a great source for getting users' opinions about products, services, and various topics. Blogs and websites are a real-time tools for gathering product feedback. However, the overabundance of blogs in the cloud has generated a huge amount of information in various forms such as opinions, comments and reviews. Therefore, there is an urgent need to find a method to extract meaningful information from big data, classify it into different categories and predict the behavior or emotions of the end user.

Various studies have explored SA and e-commerce, but few have focused specifically on analyzing user sentiments about e-commerce on Twitter. Bao et al. [11] explored the impact of preprocessing methods on Twitter sentiment classification, which is relevant to the topic; however, they did not specifically analyze e-commerce-related tweets. Awiagah et al. [12] discussed factors influencing e-commerce adoption in Ghana, but their study did not focus on SA or Twitter data.

In addition, several studies have developed SA techniques using ML algorithms like Support Vector Machines (SVMs) [13], or provided surveys of SA methodologies [14]. While these studies provide valuable insights into SA techniques, they did not concentrate on analyzing Twitter data or e-commerce SA specifically. Singla et al. [15] used ML for SA of customer product reviews, which is highly relevant to the e-commerce domain. Similarly, Dridi et al. [16] investigated the role of semantics in SA of social media, providing insights into analyzing user-generated content. However, neither study focused on Twitter data or e-commerce SA on social media platforms. Demircan et al. [17] developed a Turkish SA model using various classifiers on e-commerce product reviews, which is highly relevant to the topic. Noor and Islam [18] performed SA on women's e-commerce reviews from Amazon.com, which also aligned well with the e-commerce domain. While these studies analyzed e-commerce product reviews, they did not specifically analyze user-generated content on social media platforms like Twitter. Zhao et al. [19] proposed an algorithm for SA of online product reviews, and Rezaei et al. [20] extracted user ratings for products from Flipkart.com, both focusing on product reviews. However, these studies did not analyze Twitter data or e-commerce SA on social media platforms.

Savci and Das [21] discussed SA in the context of e-commerce, social media, and forums, providing a comprehensive overview of SA techniques and their performance across different languages. Their study is highly relevant to the topic, as it covers e-commerce, social media, and SA across multiple languages, which could provide valuable insights for the given topic of predicting user sentiments about e-commerce on Twitter. Lin et al. [22] used several deep learning (DL) models and related neural network models to analyze Weibo online-review short texts to perform SA. They compared proposed models with the vector representation generated by Word2Vec's CBOW model, they found that BERT's word vectors can obtain better SA results. In this sense, SA on Twitter data related to Amazon products and services, using ML models to predict user sentiments and compare the performance of different models can be considered important. This would provide insights into public perception of Amazon's offerings on social media platforms.

Thus, the primary aim of this study is to use different ML models, including LR, XGBoost, AdaBoost and SGD with different scales and using SMOTE technique for upsampling the minority class to validate the efficiency of these models in predicting sentiments. The models will be applied to a unique dataset collected from Twitter, with the goal of gaining insights into public sentiment towards Amazon. This study also contributes to the field by applying SA to Twitter data within the e-commerce domain, specifically focusing on Amazon's products and services. By analyzing public sentiment towards a major e-commerce player on an influential social media platform, this study addresses identified gaps in the existing literature. This could potentially provide businesses with robust tools for understanding customer sentiment, predicting sales, and enhancing customer retention strategies.

2. METHODOLOGY

In this section, we have detailed the steps of the proposed method, starting with the general flowchart of our study, which is shown in Figure 1.

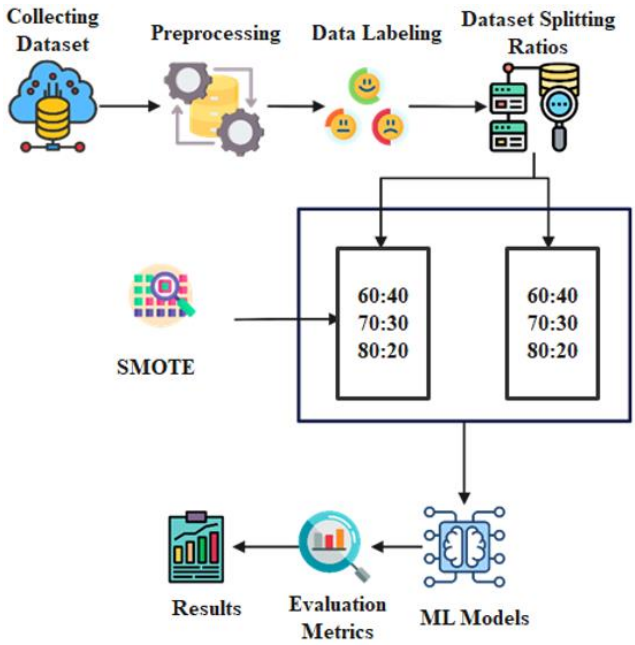

Figure 1. The proposed research framework

2.1 Collection of dataset

The dataset used in this paper is a collection of 471,840 tweets obtained using the Twitter API, which represents a source of data for SA and other NLP tasks. These include the text of the tweet, details of the user, the time that the tweet and any replies, likes, retweets, etc., was created and other associated metadata. For this purpose, tweets were collected from May 31, 2022 to June 10, 2022 and the study included tweets which contained either #amazon or #amazonprime to keep the research specific to Amazon related products and services. Tweets in other languages were not considered to reduce the level of translation required during analysis as only English tweets were considered. Table 1 provides an overview of the data some of the tweets considered in the analysis are presented.

Table 1. A snapshot of dataset used in this study

Datetime	Tweets
2022-06-10 07:24:16+00:00	pi is not dependent on hype like bitcoin, pi is creating value! In the future, pi will definitely surpass Amazon #Picoin https://t.co/3Uu7jBAo52
2022-06-10 07:24:15+00:00	https://t.co/nSdlrS99l7: memory foam pillow king https://t.co/n5svg1DzzJ via @amazon Tempur-Pedic - 15440325P TEMPUR-Cloud Breeze Dual #CoolingPillow,#King\nTempur-Pedic-15440325P TEMPUR-Cloud Breeze Dual Cooling Pillow,King #Amazon #follow4follow #EasyIFB #FastIFB #ShopAndShare https://t.co/bMvsTfCYrz
2022-06-10 07:24:14+00:00	Unicorn Flash Laser Pencil Case for Girls, Students Cute Pen Pencil Pouch Hol... https://t.co/eCzo6j24oP via @amazon
2022-06-10 07:24:08+00:00	IT WILL BE AVAILABLE IN SEPTEMBER ON AMAZON !! https://t.co/NxLP2TUO9v
2022-06-10 07:24:03+00:00	Someone actually dropped link to The Boys Season 3 Episode 4 but it’s not on Amazon prime yet lmao what
2022-06-10 07:24:03+00:00	@gamerlyfe502 @wtfkxz @PattyNest Lol bro thanks, works. I have Amazon prime but I ain’t tryna stay up all night for this 💀
2022-06-10 07:24:03+00:00	There are many ways to make money out of a regular job.\n\nI know how to sell on Amazon.\n\nLet me show you how I've been doing it for the last 6 years here: https://t.co/9o3enYLUT3
2022-06-10 07:24:02+00:00	Like It👍 from Folk in Amazon\n\nThe Wreck of the Edmund Fitzgerald\n\nhttps://t.co/cNKSHkohSp
2022-06-10 07:24:02+00:00	Amazon Prime fucking sucks I waited til 3am just for them to shit the bed
2022-06-10 07:24:02+00:00	Amazon\nMatching Love\nhttps://t.co/A5ham6Y2L0

2.2 Data preprocessing

All social media tweet data Twitter often contains a large number of words and characters that are ineffective for data analysis. For example, there are data tweets such as “@safemoonjustv?, hilari and educ. The data found useless words or characters such as “@” and “?”. Data cleaning techniques paired with regex can find superfluous characters and remove them from the core data to enhance the dataset’s quality [23]. Unstructured text often contains a large amount of noise, especially if techniques such as web or page scraping are used. HTML and HTTP tags are usually components that do not add much value to the understanding and analysis of the text [24]. Therefore, we removed the unnecessary tags and kept the textual information in all the documents.

In the preprocessing stage, all elements particular to Twitter were dealt with. For example, Emojis were dropped from the dataset as they did not have any value to the sentiment analysis goals of the study. Any @mention was also removed to filter out mentions of particular users, which would skew the set towards a particular orientation. Hashtags were preserved but transformed to enhance the ease with which they could be analysed; the ‘#’ symbol was excluded and hashtags were separated into individual words where possible (e.g. #AmazonPrime was converted to Amazon Prime). These steps made it possible to clean the data set and prepare it for sentiment analysis without losing such context as hashtags.

Every text document has special terms that represent special institutions, and have a more informative aspect and a unique framework. These entities are called named entities. This term specifically refers to objects in the real world such as people, places, organizations, etc., which often have specific names. Because these entities do not provide us with meaningful information for SA, we removed them from this section. In addition, We removed all sets of punctuation marks including [~{|}’_[\]@?<=>;:/.-,+*()’&%$# “!]. Because these signs are seen in all sentences, removing them will lead to positive results. Next vocabulary separation, also referred to as Tokenization was applied. This step involves breaking down larger blocks of text into smaller, more manageable units of language called tokens. Tokens represent the smallest linguistic units and can be anything from words and numbers to punctuation marks.

Stop words are actually words that are commonly used. Words that are meaningless or have no special meaning, especially when semantic features are extracted from the text. These items are usually very frequent in the text, and usually these words include adjectives, conjunctions, additions and such. Some examples of stop words include and, the, an, a, and the like. During NLP, there is no tendency for these types of words to occupy space or take up valuable processing time. For this reason, these words can be easily removed in this section.

Lemmatization helps in standardizing words to their canonical form, which is linguistically correct. For example, the words "running," "runs," and "ran" would all be converted to their base form "run." This process is crucial in SA, as it allows for the grouping and analysis of similar sentiments expressed through variations of a word. While it doesn't directly handle slang or non-standard terminology, it helps to unify different forms of standard words, thereby enhancing the accuracy of the SA. Besides, stemming also was applied to reduce a word to its stem or root form, which may not necessarily be a valid word on its own. For instance, stemming could reduce the word influences to the simpler form "influence". Search engines and other text analysis tools often use stemming to improve the efficiency and relevancy of their results [25].

2.3 VADER for data labeling

VADER is a lexicon and rule-based SA tool that performed exceptionally well on social media SA according to study [26]. A distinctive feature of VADER is that it eschews polarity from the document scoring process, instead providing positive, negative, neutral, and compound scores. An advantage of VADER is that it does not require training data, thus enabling us to apply it to previously unseen data [27].

In order to maintain the reliability and coherence of sentiment classification we used certain thresholds derived from the compound score given by VADER. Those sentiments with the compound scores > 0.05 were categorized as positive.

The sentiment was considered negative if the compound score <-0.05. Results between -0.05 and 0.05 were defined as having a mild sentiment which means the information conveyed bears a neutral or no specific positive/negative feeling. These thresholds offered timely baselines to help categorize sentiment on differences between tweets and social media comments. Table 2 is showing the score of each sentiments used in our study.

Table 2. Sentiment scores

Sentiment	Score
Positive	372,653
Negative	55,650
Neutral	43,537

The dataset used in the study consists of three sentiments as depicted in Table 2. However, for the purpose of this study, we only considered binary sentiment classification whereby positive and negative samples were retained while the neutral samples were completely removed to enhance the analysis with the ultimate aim of determining the sentiment polarity. Lack of neutral sentiments means that neutral or balanced opinions which might blur the focus and add confusion to the interpretation of models' performance. To achieve that, we limited the scope of our analysis to binary classification in order to increase the accuracy of the SA and get a better insight into the model capacity to separate different types of emotional messages. This approach also follows typical trends in SA researches that mostly target simple positive or negative sentiment orientations.

2.4 SMOTE

Considering that the number of positive and neutral data in this collection is much more than the negative data, one of the important challenges in this section is the imbalance of the dataset in this study due to the existing challenges. In the data set, we use the Recall criterion to evaluate the model. SMOTE (Synthetic Minority Oversampling Technique) is a popular resampling method used to balance class distributions in datasets with a skewed class imbalance [28]. SMOTE addresses class imbalance by generating new synthetic minority class samples rather than merely duplicating existing samples, as in basic oversampling. The algorithm selects a minority class sample and computes its k-nearest neighbors from the minority class. New samples are interpolated between the selected sample and its neighbors.

2.5 Dataset splitting ratios

In this study, the training-test split ratios used were 60:40, 70:30 and 80:20 in order to determine the efficiency of the ML models. These splits were made with the purpose of having sufficient data to train as well as to have another set of data to test the outcome. The 60/40 is a way of testing models at a lower level of training the models The 70/30 split is generally favored because it is a globally acceptable level of split. The 80/20 split optimizes the size of the training dataset in response to which the different models can learn more examples yet still reserved enough samples to objectively evaluate their performance. This multi-ratio approach allows for determining all the possible strengths and weaknesses of the models, incorporating the specifics of the distribution of the given data.

2.6 ML models

This study considers several well-known ML techniques for the SA problem. We dissect the operation of four of the most commonly applied classifiers for SA: LR, XGBoost, AdaBoost, and SGD. For the models employed in this study, we used the various default hyperparameters as offered by the libraries of each model. No attempts were made to fine-tune or optimize the hyperparameters of the models, and the results are reported on the models in their default state.

2.6.1 LR

LR is a basic linear classification model that uses a weighted summation of the input features to fashion a prediction [29]. LR requires a text input and classifies the sentiment as positive or negative depending on the presence of some words and phrases and the number of times these appear [30]. It is simple to use and understands its results, but it has problems with comprehending sophisticated language. The IMDB database was pre-processed to remove stop words and punctuations based on the use of LR for sentiment classification. Nevertheless, it gives a relatively small yield and is usually inferior to more sophisticated techniques [31]. However, it is generally outperformed by more advanced methods.

2.6.2 XGBoost

XGBoost is an implementation of gradient-boosted decision trees designed for speed and performance [32]. For SA, XGBoost builds an ensemble of decision trees where each tree learns from the errors of the previous one to make increasingly accurate predictions. XGBoost performs well on SA tasks across many domains and datasets XGBoost captures non-linear relationships and interactions between words better than linear models like LR [33]. It is also efficient in training and tuning, making it popular for sentiment modeling.

2.6.3 AdaBoost

AdaBoost, is another ensemble method that combines multiple weak learners, typically DT [34]. For SA, AdaBoost iteratively trains classifiers on modified versions of the data that emphasize previously misclassified examples [35]. This focus on hard examples improves accuracy. AdaBoost has outperformed LR and Naive Bayes models for sentiment classification [36].

2.6.4 SGD

SGD optimizes models by taking small steps along estimated gradients, leading to fast model updates [37]. For SA, SGD enables efficient training of linear classifiers like LR on large datasets [38]. It is easy to implement and appropriate for text data where examples are high-dimensional and sparse. However, convergence can be slow and tuning is required to control step sizes. SGD remains a popular optimization technique but is often used within more advanced neural network architectures.

2.7 Evaluation metrics

Various criteria have been introduced to evaluate the performance of the text classification system [39]. The accuracy criterion expresses the number of correct predictions made by the classifier divided by the total number of predictions made by the same classifier. In general, accuracy refers to how well the model predicts the output. These metrics are defined as follows:

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Pre. = \frac{TP}{TP + FP} \quad (2)$$

$$Rec. = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Pr. \times Re.}{Pr. + Re.} \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (5)$$

TP, FP, TN, and FN stand for true positive, false positive, true negative, and false positive, respectively.

3. RESULTS AND DISCUSSION

The dataset was initially raw and unannotated, and we needed to assign appropriate labels to each instance to indicate the target variable or outcome we were trying to predict. The labeling process involved human annotators who were provided with clear guidelines and context to assign labels based on the specific task. Depending on the nature of the study, the labeling could be binary (e.g., positive/negative, yes/no) or multiclass (e.g., low/medium/high).

The final labeled dataset became the foundation for our supervised ML models. Using VADER, we assigned a sentiment score to each text instance in the dataset. The scores ranged from -1 (most negative) to +1 (most positive), with 0 indicating a neutral sentiment. VADER utilizes a lexicon and rule-based approach, which makes it particularly suitable for social media texts and short informal texts.

Figure 2 presents a word cloud visualization of words frequently associated with negative sentiment tweets. The dominant terms include "ref", "bad", "worst", "horrible", "terrible", "devil" and "pain" which indicate strong negative emotions. On other terms like "money", "cost" and "price" suggest dissatisfaction regarding financial elements. The prevalence of words like "not", "don't", "can't" and "wouldn't" imply negation and refusal. Thus, the word cloud provides a summarized snapshot of key terms in tweets expressing negative opinions and attitudes.

Figure 2. Word cloud sentiment in negative

Figure 3 shows a wordcloud for terms associated with positive sentiment. Dominant words include "love", "like", "good", "best", "great", "happy", "amazing" which convey strong positive emotions. The terms "thank", "help" and "support" imply gratitude and appreciation. Words such as "friend", "family" and "child" suggest positive affiliations and relationships. The prevalence of present tense verbs indicates optimism and enthusiasm. Thus, this word cloud summarizes key terms in tweets with favorable and upbeat sentiments.

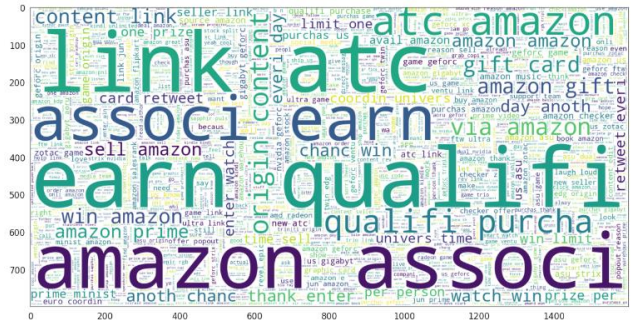

Figure 3. Word cloud sentiment in positive

Comparing the two graphs, the negative sentiment terms appear more diffuse while the positive words are more concentrated around a few main concepts like "love", "like" and "good". The positive terms also seem more personal and relationship-oriented while the negative words relate more to general emotions. The word clouds provide an efficient visualization of the lexical differences between tweets of opposing sentiment. Analyzing the frequent terms provides useful insights into how sentiment is expressed in the text data.

3.1 ML model application without SMOTE

In this section, we tested several ML models on the raw data without adjusting the distribution of the variables. The aim was to evaluate the performance of these models despite the imbalanced class composition. We employed commonly used classifiers such as LR, XGBoost, AdaBoost, and SGD. Evaluation metrics included accuracy, precision, recall, F1-score, cross-validation, and the area under the Receiver Operating Characteristic (ROC) curve. These metrics are outlined in Table 3.

In Table 3, the best results were obtained by LR model for all measurements, the accuracy, precision, recall, and F1-score were 0.96 each, MCC 0.84, AUC 0.89. Notably, its training time was relatively short at 3.30 seconds. SGD followed with accuracy, precision, recall, an F1-score of 0.92, an MCC of 0.61, an AUC of 0.71, and a notably brief training time of 0.77 seconds. XGBoost demonstrated commendable performance with accuracy, precision, and recall, an F1-score of 0.94, an MCC of 0.75, and an AUC of 0.82; however, its training time was considerably longer at 52.96 seconds. The lowest performance was observed in AdaBoost with the following values: accuracy 0.91, precision 0.91, recall 0.91, F1-score 0.91, MCC 0.57, and AUC 0.72 as well as significantly longer training time equal to 2132.73 seconds. These analyses indicate that in a 60/40 split without SMOTE, a higher accuracy with reasonable computation time is attainable with LR as the best algorithm, while SGD is a viable option if the computation time is restricted.

In Table 4, LR achieved an accuracy, precision, recall, and F1-score of 0.97, an MCC of 0.86, and an AUC of 0.90, with a training time of 6.06 seconds. SGD also remained stable and had an accuracy of 0.92, precision of 0.92, recall of 0.92, and

F1-score of 0.92, MCC of 0.62, AUC of 0.72, and a training time of 0.69 seconds. XGBoost showed good performance with accuracy, precision, recall, and F1-score of 0.95, an MCC of 0.75, and an AUC of 0.82, albeit with an increased training time of 68.30 seconds. AdaBoost remained the least effective, with accuracy, precision, recall, and F1-score of 0.91, an MCC of 0.57, an AUC of 0.72, and a significantly extended training time of 2554.49 seconds. Table 4 also implies the same conclusion which we made above that LR is indeed a good contender for this task, and even if the training sample size is larger, while SGD can also be used, but if it is more efficient to use.

Table 5, LR led the performance metrics, with accuracy, precision, recall, and F1-score of 0.97, an MCC of 0.86, an AUC of 0.90, and a training time of 5.22 seconds. SGD followed closely, with accuracy, precision, recall, an F1-score of 0.92, an MCC of 0.61, an AUC of 0.71, and a short training time of 0.75 seconds. XGBoost demonstrated good performance with accuracy, precision, recall, and an F1-score of 0.94, an MCC of 0.75, and an AUC of 0.82, with a training time of 60.32 seconds. AdaBoost performance continued with previous splits having the worst performance with accuracy of 0.91 and the highest standard deviation of 0.05 and 0.02 in precision and recall respectively, F1-score of 0.91, MCC of 0.57, AUC of 0.72 and longest training time of 3100.65 seconds. Across all three data split ratios without the application of SMOTE, the results consistently indicate that LR provides the most favorable performance profile for this sentiment analysis task. It always delivers the best performance metrics and reasonably shorter training times. When it comes to performance across the splits, SGD is second only to AdaBoost, and characterized by exceptionally short training times, which points to its applicability in the situations when computational time is of the essence. Nonetheless, XGBoost has shown fairly promising performance improvement with the major disadvantage of significantly longer training time compared to that of LR and SGD. AdaBoost also performs the worst of all the four models studied here and has the longest training time, suggesting that this model is less suitable for this particular sentiment analysis task than the other models when SMOTE is not used. The trends presented for all the splits are similar which indicate the robustness of the conclusions made here: LR and SGD are effective when it comes to achieving the best balance between accuracy and efficiency when SMOTE is not used.

The ROC curves for ML models on an SA task without SMOTE using a 60/40 split are shown in Figure 4. The LR gives the maximum accuracy, having an AUC of 0.890, which means the curve is close to the top left corner, hence having high TPR and low FPR at various thresholds signifying a high TPR and a low FPR across various thresholds. XGBoost comes in second with an AUC of 0.825, hence showing a fairly good but low discrimination capacity between the two classes.

Table 3. Result analysis of SA by splitting the dataset into 60:40 (without using SMOTE)

ML Models	Acc.	Pre.	Rec.	F1-Score	MCC	AUC	Training Time (Seconds)	Confusion Matrix
LR	0.96	0.96	0.96	0.96	0.84	0.89	3.30	[[17477, 4776], [815, 148254]]
XGBoost	0.94	0.94	0.94	0.94	0.75	0.82	52.96	[[14656, 7597], [1302, 147767]]
AdaBoost	0.91	0.91	0.91	0.91	0.57	0.72	2132.73	[[10268, 11985], [2283, 146786]]
SGD	0.92	0.92	0.92	0.92	0.61	0.71	0.77	[[9819, 12434], [467, 148602]]

Table 4. Result analysis of SA by splitting the dataset into 70:30 (without using SMOTE)

ML Models	Acc.	Prec.	Rec.	F1-Score	MCC	AUC	Training Time (Seconds)	Confusion Matrix
LR	0.97	0.97	0.97	0.97	0.86	0.90	6.06	[[13399, 3198], [657, 111237]]
XGBoost	0.95	0.94	0.95	0.95	0.75	0.82	68.30	[[11074, 5523], [1019, 110875]]
AdaBoost	0.91	0.91	0.91	0.91	0.57	0.72	2554.49	[[7693, 8904], [1758, 110136]]
SGD	0.92	0.92	0.92	0.92	0.62	0.72	0.69	[[7407, 9190], [359, 111535]]

Table 5. Result analysis of SA by splitting dataset to 80:20 (without using SMOTE)

ML Models	Acc.	Prec.	Rec.	F1-Score	MCC	AUC	Training Time (Seconds)	Confusion Matrix
LR	0.97	0.97	0.97	0.97	0.86	0.90	5.22	[[8998, 2058], [461, 74144]]
XGBoost	0.94	0.94	0.94	0.94	0.75	0.82	60.32	[[7305, 3751], [669, 73936]]
AdaBoost	0.91	0.91	0.91	0.91	0.57	0.72	3100.65	[[5101, 5955], [1174, 73431]]
SGD	0.92	0.92	0.92	0.92	0.61	0.71	0.75	[[4880, 6176], [246, 74359]]

AdaBoost and SGD have the lowest AUC values of 0.723 and 0.719, respectively; their curves are the least close to the top-left corner, which means that they have the least ability to differentiate between positive and negative sentiments. This figure solidifies the evidence that LR provides the highest classification accuracy in this context than the other methods, by XGBoost, AdaBoost, and SGD, as reviewed in Table 3.

Figure 5 shows the ROC curves of SA task and the ML models without the SMOTE technique where 70% of the dataset was used for training and 30% for testing. The curve shows that LR yields the best result with an AUC of 0.901 to reach the top-left corner representing high true positive rate and low false positive rate with varying thresholds. The next is XGBoost with the AUC of 0.829, which means that it has a somewhat lower capacity for classification between classes compared to LR. The worst results are given by AdaBoost and SGD with AUC of 0.724 and 0.722, respectively; their curve is less close to the top-left corner, showing lower ability to separate positive and negative sentiments. This visualization also supports the conclusion that LR has the highest classification accuracy in this case, XGBoost, followed by AdaBoost, and SGD are not as good as the first two, which is consistent with the results of Table 4 for the 70/30 split without SMOTE.

Figure 6 displays the ROC curves of the ML models in an SA task without using SMOTE, with an 80:20 ratio of data

split. The figure shows that LR is again able to sustain its high accuracy of 0.904 AUC, as the curve climbs steeply towards the top-left area, implying high true positive rate and low FPR across the threshold. The top-left corner, indicating a high TPR and a low FPR across various thresholds. XGBoost comes next with an AUC of 0.826, which shows a decent, but lower performance of discriminating between classes than that of LR. AdaBoost and SGD are the worst classifiers with the AUC of 0.723 and 0.719 respectively; their graphs are least close to the upper left quadrant showing poor ability of classifying positive sentiment from negative. This visualization also stands to support this conclusion where it is clear that LR has the best classification, followed by XGBoost whereas AdaBoost and SGD have poor performance, as was observed in the same case, but with out of sample data of 80/20 without SMOTE. The trends observed in Figure 6 are similar to those in Figures 4 and 5 with similar conclusions that the relative performances of the models are consistent across different data splits when SMOTE is not used.

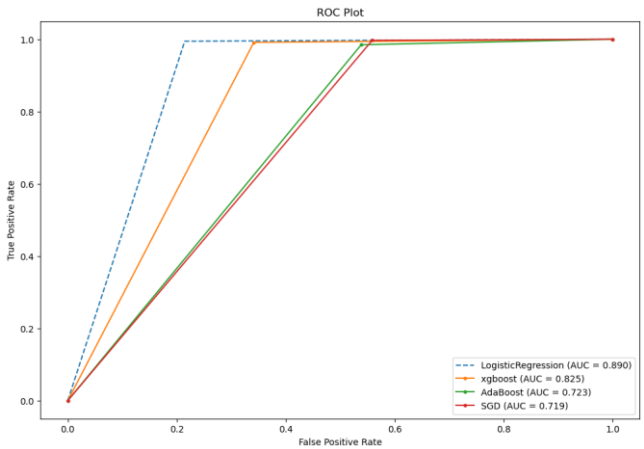

Figure 4. The ROC plot results for obtained ML models after applying 60:40 before applying SMOTE

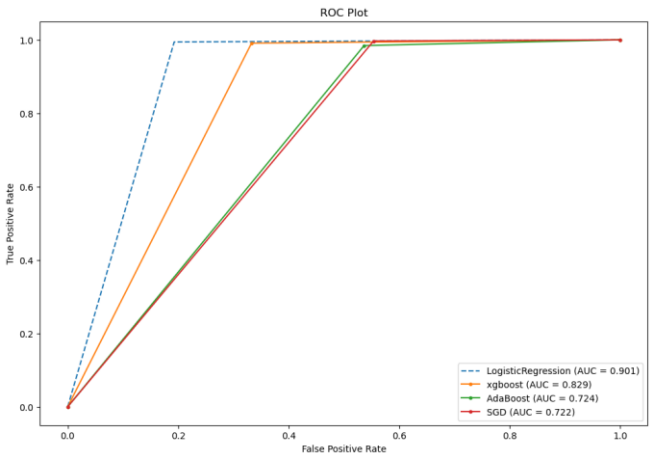

Figure 5. The ROC plot results for obtained ML models after applying 70:30 before applying SMOTE

Therefore, the above Figures 4, 5, and 6 show that when applied to an SA task without SMOTE, LR performs the best among the other ML models for all the three data splits. The ROC curve of LR rises sharply towards the top-left corner, indicating that LR ultimately has the highest AUC values of 0.890, 0.901 and 0.904. XGBoost is the next best model with an AUC of 0.825, 0.829 and 0.826 for the three folds, AdaBoost and SGD perform the worst with an AUC of approximately 0.72. These ROC plots support the quantitative results shown in Tables 3, 4 and 5 by emphasizing that LR has a higher capacity to distinguish between positive and negative sentiments in this case than XGBoost, AdaBoost, and SGD. These consistent trends to all three figures irrespective of the different data split ratios indicate that these performance differences are quite stable when SMOTE is not used.

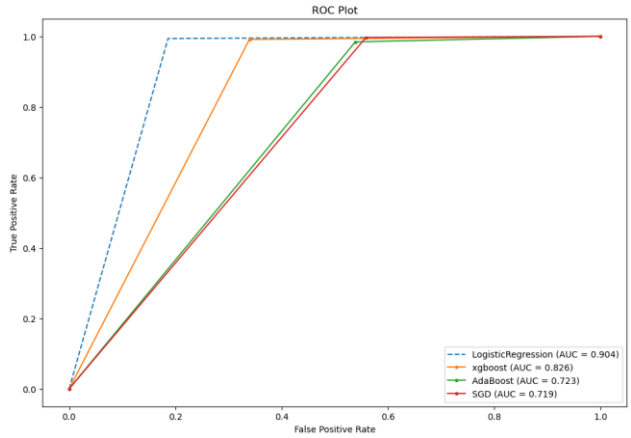

Figure 6. The ROC plot results for obtained ML models after applying 80:20 before applying SMOTE

3.2 ML model application with SMOTE

We utilized the Synthetic Minority Over-sampling Technique (SMOTE) to rectify the class imbalance in our dataset. SMOTE creates artificial samples for the underrepresented group by extrapolating between nearby real-world instances. Table 6, showing that that the dataset was highly imbalanced before applying SMOTE. The original class distribution was as follows: Class positive (majority class) had 372,653 samples while the Class negative (minority class) had 55,650 samples. After applying SMOTE the dataset was balanced that is the number of instances of the class positive was 372,653 and that of class negative was also 372,653.

Table 6. Class distribution before and after using SMOTE

Sentiment Class	Before SMOTE	After SMOTE
Negative	55,650	372,653
Positive	372,653	372,653

This section explains the implementation of SMOTE and its impact on student enrollment. Subsequently, we applied the same family of ML models to the SMOTE-balanced dataset. The objective of this study was to ascertain whether balancing the dataset could improve model performance. We present the final results, including metrics such as accuracy, precision, recall, F1-score, Matthew's Correlation Coefficient (MCC), and Area under the ROC Curve (AUC), in Table 7.

Table 7 shows the ML models performance of models trained and tested using SMOTE with a 60/40 data split. Data imbalance was addressed using SMOTE. LR was outstanding with an accuracy of 0.98, precision of 0.98, recall of 0.98, F1-score of 0.98, MCC of 0.98 and AUC of 0.99. This implies that the system has almost perfect capability of classifying instances while having a fairly good balance of precision and recall. The high MCC value of 0.96 also indicates strong accuracy of the predictability of the classification that was

assigned to the given items. LR also took a very short time to train, just 1.59 seconds. The accuracy, precision, recall, and F1-score of XGBoost was satisfying at 0.94. However, it was slower in its training time which took 75.02 seconds on average. AdaBoost was slower and less accurate with 0.82 and a training time of 2575.30 seconds, thus, it is not as fit for this SA task and has a high computational cost. SGD had an accuracy of 0.95, almost as good as LR, and took only 1.64 seconds to train. Looking at the confusion matrices, the least number of misclassifications were obtained by the LR model followed by the SGD model, then the XGBoost model and the highest number of misclassifications were obtained by the AdaBoost model.

In Table 8, the outcomes based on the 70/30 data split patterns When it comes to the results yields the same pattern as viewed. LR remained at the peak of its efficiency with the scores of 0.98, while its training time rose to 5.69 seconds. The confusion matrix for LR reveals that the model has nearly equal numbers of false positives and false negatives. SGD also performed well with values around 0.95 and 1.44 seconds for the training time, and a confusion matrix with an approximate split of errors. All the other metrics were somewhat around 0.94 for XGBoost while the training time was surprisingly higher and took around 106.08 seconds. In its confusion matrix, it seems to have a slightly higher probability of false negatives. AdaBoost was the slowest and had metrics close to 0.82, while training time climbed to 3645.19 seconds. This table supports the previous observation that LR and SGD are two of the best candidates for this task even if the training data set is much bigger.

Table 9 presents the findings in the condition where the number is divided in 80/20. LR came out on top with the most uniform results, 0.98, and a training time of 3,220 seconds. Its confusion matrix stayed neutral. The SGD model was further analyzed and showed values close to 0.95 on the metrics, a remarkably low of 1.12 seconds for the training time, and relatively equal values in the confusion matrix. XGBoost had similar performance with the metrics being around 0.94 seconds as was the training time, which was 113.43 seconds and the confusion matrix showed that it had a slightly higher tendency to misclassify as negative. AdaBoost Metrics were still fairly close to previous splits, with 0.82 and a training time of 4128.12 seconds, its confusion matrix had more false positives and false negatives than other classifiers.

Table 7. Result analysis of SA by splitting the dataset into 60:40 (using SMOTE)

ML Models	Acc.	Prec.	Rec.	F1-Score	MCC	AUC	Training Time (Seconds)	Confusion Matrix
LR	0.98	0.98	0.98	0.98	0.96	0.98	1.59	[[146943, 2152], [3504, 145524]]
XGBoost	0.94	0.94	0.94	0.94	0.88	0.94	75.02	[[136651, 12444], [4518, 144510]]
AdaBoost	0.82	0.83	0.82	0.82	0.66	0.82	2575.30	[[109057, 40038], [12182, 136846]]
SGD	0.95	0.95	0.95	0.95	0.91	0.95	1.64	[[142826, 6269], [6302, 142726]]

Table 8. Result analysis of SA by splitting the dataset into 70:30 (using SMOTE)

ML Models	Acc.	Prec.	Rec.	F1-Score	MCC	AUC	Training Time (Seconds)	Confusion Matrix
LR	0.98	0.98	0.98	0.98	0.96	0.98	5.69	[[110205, 1710], [2578, 109099]]
XGBoost	0.94	0.94	0.94	0.94	0.88	0.94	106.08	[[102623, 9292], [3417, 108260]]
AdaBoost	0.82	0.83	0.82	0.82	0.66	0.82	3645.19	[[81898, 30017], [8885, 102792]]
SGD	0.95	0.95	0.95	0.95	0.95	0.910	1.44	[[107228, 4687], [4673, 107004]]

Table 9. Result analysis of SA by splitting the dataset into 80:20 (using SMOTE)

ML Models	Acc.	Prec.	Rec.	F1-Score	MCC	AUC	Training Time (Seconds)	Confusion Matrix
LR	0.98	0.98	0.98	0.98	0.96	0.98	3.22	[[73597, 1107], [1679, 72679]]
XGBoost	0.94	0.94	0.94	0.94	0.88	0.94	113.43	[[68541, 6163], [2195, 72163]]
AdaBoost	0.82	0.83	0.82	0.82	0.66	0.82	4128.12	[[54849, 19855], [6070, 68288]]
SGD	0.95	0.95	0.95	0.95	0.91	0.95	1.12	[[71553, 3151], [3093, 71265]]

Therefore, the LR model had higher accuracy, precision, recall, F1-score, MCC, and AUC across all the three data split ratios; and the training time was low, though slightly fluctuating. SGD was again amongst the most accurate with the best and almost equally best training time which makes it one of the best. Although, XGBoost yielded good results, it was observed that the training time of this model increased with the expansion of training data. AdaBoost was the most unstable and was always the worst performing algorithm and had long training times proving that it is not suitable for this particular sentiment analysis task. The stable performance of both LR and SGD in different data split setting further strengthen the argument about their suitability for this task though the selection between the two may depend on certain preference with regards to accuracy over computational cost. Thus, these results can be of great significance for the selection of the models in the sentiment analysis, with a particular focus on the efficiency of LR and SGD methods.

Figure 7 is the ROC plot that presents the results of applied ML models with the dataset split at 60/40 after using SMOTE for oversampling. The graph shows the TPR against the FPR for each model at different thresholds.

The LR model shows AUC of 0.981. This curve closely hugs the top-left quadrant of the plot, meaning that the model has a high TPR and a very low FPR at nearly all threshold levels. The XGBoost shows good performance AUC 0.943, although it is slightly lower than LR, with the curve lies slightly away from the origin of the graph. AdaBoost with an AUC of 0.825 is the lowest performing model among the four proposed models because its curve is located far from the top left position. Lastly, the red line is for SGD which has the AUC of 0.958. This performance is very close to LR which means that the model has a similar capability of classifying between the two classes.

The ROC curves of the following ML models are presented in Figure 8 based on 70/30 data split and application of

SMOTE. The ROC curve is a graphical summary of a model's diagnostic accuracy showing the TPR against the FPR at different threshold values. The LR has the highest AUC of 0.981 proving that it has high discrimination ability. This curve coincides with the top left corner of the plot indicating a high TPR and low FPR at various thresholds. SGD model also shows good performance with an AUC of 0.958, but the curve is slightly behind LR, especially low FPR values. representation of a model's diagnostic ability, plotting the TPR against the FPR at various threshold settings. The LR exhibits the highest AUC of 0.981, indicating excellent discriminatory capability. Its curve closely follows the top-left corner of the plot, signifying a high TPR and a low FPR across different thresholds. SGD model also demonstrates strong performance with an AUC of 0.958, its curve slightly below that of LR, especially at lower FPR values. XGBoost model has an AUC of 0.943 which shows its curve is below the LR and SGD curve but still very close to the top left corner. AdaBoost model presents the lowest AUC equal to 0.826 among all four models. Its curve is located significantly lower and to the right of the top-left corner, suggesting a lower TPR and/or higher FPR at different thresholds, as the tabular results showed.

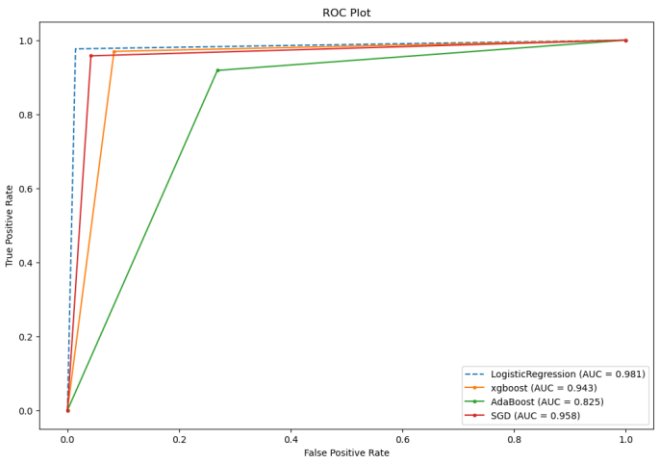

Figure 7. The ROC plot results for obtained ML models after applying 60:40 after applying SMOTE

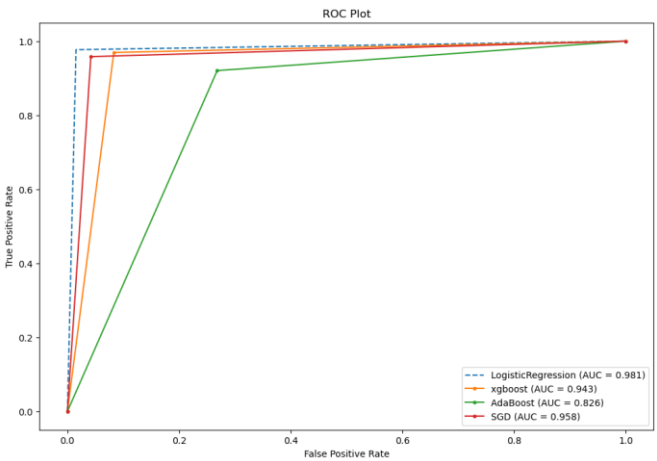

Figure 8. The ROC plot results for obtained ML models after applying 70:30 after applying SMOTE

Figure 9 shows that the AUC value of the LR model is 0.981 and it is the highest, which proves that this model has very high discriminatory power. Its curve also stays near the top-left corner of the plot and shows high TPR and low FPR at various thresholds. The performance of SGD model also remains competitive with an AUC of 0.958, just a tad behind LR. XGBoost model achieves an AUC of 0.944, which is slightly higher than in the 70/30 split with AUC of 0.943. Thus, the improvement is achieved with the help of the expanded training dataset, which has a lower top left position of its curve than that of both LR and SGD but is slightly lower in performance. AdaBoost model still has the lowest AUC score of 0.826, which confirms the scores derived from the 70/30 split and the tabular analysis.

The ROC plots in the Figures 7, 8 and 9 clearly indicate that the two best models for this sentiment analysis task are LR and SGD. LR always provides the highest AUC values, which signifies almost perfect classification ability, and SGD is accompanied by high AUC values and extremely short training time for all the splits of data. Despite the relatively good performance XGboost found to exhibit, its much longer time to train makes it less suitable. AdaBoost remains the weakest performer with the lowest AUC, and the longest training time and cannot be used in this application. The observed performance trends in terms of AUC across different splits of the data support these findings and the ROC plots provide visualization of these trends confirming the superiority of both LR and SGD compared to other methods on this particular type of sentiment analysis task, while the choice between these two methods would be made based on whether slightly better accuracy of the LR model is valued over the substantially faster convergence of the SGD method.

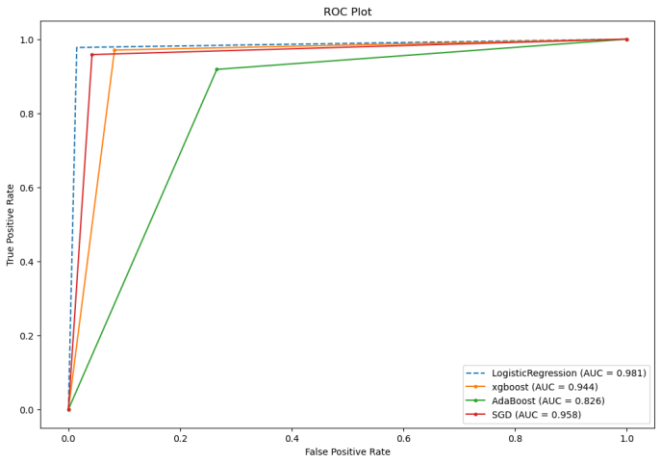

Figure 9. The ROC plot results for obtained ML models after applying 80:20 after applying SMOTE

In this study, we evaluated the impact of different dataset splitting ratios (60:40, 70:30, and 80:20) on the performance of four ML models, both with and without the application of SMOTE, to address the class imbalance. The results, when using SMOTE, showed that the performance metrics across the different ratios were relatively high and consistent for LR and SGD. When not using SMOTE, the performance was generally consistent across all three splits, with LR showing the best performance. The performance differences between the splits were not substantial when using or not using SMOTE, although the models performed generally better when using SMOTE. These results indicate that the number of training data does play a role to some extent, however, the models, especially the LR and SGD models, are fairly insensitive to changes within the range explored here. These findings do not

definitively prove that the studies do highlight the fact that the 70:30 or 80:20 split is universally superior, especially in selecting the right sampling ratio and how to handle a class imbalance in cases where the datasets are imbalanced.

4. CONCLUSIONS AND RECOMMENDATION

In this study, we performed an automatic tagging and SA method on raw Twitter data related to e-commerce activity using VADER for sentiment polarity detection, along with four ML algorithms. Before data analysis, the data was cleaned through processes like folding, data deletion, rewording, stop word removal, and stemming. Once cleaned, the data could be automatically tagged and classified with the four ML algorithms. The combined workflow of VADER sentiment polarity detection and ML algorithms effectively analyzed raw Twitter data across three scenarios. From our experiments, it is evident that LR consistently outperformed the other algorithms across most evaluation metrics. It demonstrated the highest accuracy, precision, recall, F1-score, MCC, and AUC, making it the optimal choice for this particular task. XGBoost and SGD also exhibited competitive performance and could be viable alternatives, particularly when interpretability is less critical, and a higher degree of model complexity is acceptable. However, it is important to note that the performance of these algorithms may vary depending on the specific characteristics of the dataset and the nature of the task. Therefore, we recommend conducting further experiments on different datasets and tasks to validate the generalizability of the results.

Another important result of this study is the superiority of the LR model in all compared performance indicators. This can be explained by features of the sentiment analysis data and the advantages of the LR algorithm. The data, as an output of product reviews, seems to be, to some extent, linearly separable where the presence or absence of particular words indicates a positive or negative attitude. As a linear classifier, LR is very effective in detecting such linear correlations. Additionally, compared with other models including XGBoost and SGD, LR has less tendency to overfit due to a simple model structure when dealing with high dimensional text data. Hence, the default settings of LR were used, although it is important to note that the algorithm sometimes incorporated an implicit regularization, which helps improve its performance. Thus, for this particular task, straightforward linear models, which were implemented in LR, together with their simplicity and inherent regularization, were sufficient to provide the best results compared to more sophisticated models that might be accurate for data containing non-linear patterns.

Based on the evaluation metrics, LR achieved the best results overall, boasting the highest accuracy, precision, recall, F1-score, MCC, and AUC. XGBoost and SGD also demonstrated reasonable performance on this dataset, while AdaBoost fell behind in terms of performance. It's important to note that these findings need to be validated through additional experiments on different data to determine their generalizability across tasks. Factors such as dataset characteristics and problem type can significantly impact relative model performance. Therefore, while LR appears optimal for this case, other algorithms like XGBoost and SGD may prove superior given different criteria and data. This underscores the need for more research to compare the models' applicability fully. In the future, we will strive to apply various ways to make our model more efficient, as well as employ other aspect-based analytic methodologies on negative review data sets to identify problematic product attributes. As a result, the business may improve the quality of its products and boost its sales rate.

Additionally, in the future work will also look at the possibility of studying the effects of the length of the texts on the models as well. This will entail a finer division of the set into subgroups, for example by the length of the text and assessing how different models behave on each of them with a special interest in the extremely short and the extremely long texts in the context of this data set. An analysis of this nature could help identify strengths and weaknesses of specific models when dealing with various kinds of textual data and also help establish how the performance of the models changes with increased information content of the text. Such a finer-grained view of model behavior could in turn help design better strategies for sentiment analysis in environments such as Twitter where text length variability is a given fact.

The models used for SA are rapidly advancing, but there are still many unsolved problems in the field of opinion research. Based on this, in this part and according to the axes of this research, titles with the aim of continuing research and developing knowledge in the field of research. Every opinion text with a positive polarity can be considered a valuable resource which always suggests products or services to other users and customers. Based on this, companies, organizations, as well as e-commerce websites can benefit from this comment text with the help of recommender systems, products and services to other customers and users who choose the category. Products and services have the same taste as the opinions of the people providing them.

REFERENCES

[1] O'cass, A., Fenech, T. (2003). Web retailing adoption: Exploring the nature of internet users Web retailing behaviour. Journal of Retailing and Consumer Services, 10(2): 81-94. https://doi.org/10.1016/S0969-6989(02)00004-8

[2] London, T., Hart, S.L. (2004). Reinventing strategies for emerging markets: Beyond the transnational model. Journal of International Business Studies, 35: 350-370. https://doi.org/10.1057/palgrave.jibs.8400099

[3] Liao, S.H., Widowati, R., Hsieh, Y.C. (2021). Investigating online social media users' behaviors for social commerce recommendations. Technology in Society, 66: 101655. https://doi.org/10.1016/j.techsoc.2021.101655

[4] Zeng, Y.E., Wen, H.J., Yen, D.C. (2003). Customer relationship management (CRM) in business-to-business (B2B) e-commerce. Information Management & Computer Security, 11(1): 39-44. https://doi.org/10.1108/09685220310463722

[5] Baker, M.R., Mahmood, Z.N., Shaker, E.H. (2022). Ensemble learning with supervised machine learning models to predict credit card fraud transactions. Revue d'Intelligence Artificielle, 36(4): 509-518. https://doi.org/10.18280/ria.360401

[6] Alamoodi, A.H., Zaidan, B.B., Zaidan, A.A., Albahri, O.S., Mohammed, K.I., Malik, R.Q., Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review.

Expert Systems with Applications, 167: 114155. https://doi.org/10.1016/j.eswa.2020.114155
[7] Jihad, K.H., Baker, M.R., Farhat, M., Frikha, M. (2022). Machine learning-based social media text analysis: impact of the rising fuel prices on electric vehicles. In International Conference on Hybrid Intelligent Systems, Switzerland, pp. 625-635. https://doi.org/10.1007/978-3-031-27409-1_57
[8] Umar, S.U., Rashid, T.A., Ahmed, A.M., Hassan, B.A., Baker, M.R. (2024). Modified bat algorithm: A newly proposed approach for solving complex and real-world problems. Soft Computing, 28(13): 7983-7998. https://doi.org/10.1007/s00500-024-09761-5
[9] Cirqueira, D., Hofer, M., Nedbal, D., Helfert, M., Bezbradica, M. (2019). Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda. In International Workshop on New Frontiers in Mining Complex Patterns, Springer International Publishing, pp. 119-136. https://doi.org/10.1007/978-3-030-48861-1_8
[10] Khan, M.T., Durrani, M., Ali, A., Inayat, I., Khalid, S., Khan, K.H. (2016). Sentiment analysis and the complex natural language. Complex Adaptive Systems Modeling, 4: 1-19. https://doi.org/10.1186/s40294-016-0016-9
[11] Bao, Y., Quan, C., Wang, L., Ren, F. (2014). The role of pre-processing in twitter sentiment analysis. In Intelligent Computing Methodologies: 10th International Conference, ICIC 2014, Taiyuan, China, pp. 615-624. https://doi.org/10.1007/978-3-319-09339-0_62
[12] Awiagah, R., Kang, J., Lim, J.I., (2016). Factors affecting e-commerce adoption among SMEs in Ghana. Information Development, 32(4): 815-836. https://doi.org/10.1177/0266666915571427
[13] Devi, D.N., Kumar, C.K., Prasad, S. (2016). A feature based approach for sentiment analysis by using support vector machine. In 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, pp. 3-8. https://doi.org/10.1109/IACC.2016.11
[14] Abirami, A.M., Gayathri, V. (2017). A survey on sentiment analysis methods and approach. In 2016 Eighth International Conference on Advanced Computing (ICoAC), Chennai, India, pp. 72-76. https://doi.org/10.1109/ICoAC.2017.7951748
[15] Singla, Z., Randhawa, S., Jain, S. (2017). Sentiment analysis of customer product reviews using machine learning. In 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, India, pp. 1-5. https://doi.org/10.1109/I2C2.2017.8321910
[16] Dridi, A., Reforgiato Recupero, D. (2019). Leveraging semantics for sentiment polarity detection in social media. International Journal of Machine Learning and Cybernetics, 10: 2045-2055. https://doi.org/10.1007/s13042-017-0727-z
[17] Demircan, M., Seller, A., Abut, F., Akay, M.F. (2021). Developing Turkish sentiment analysis models using machine learning and e-commerce data. International Journal of Cognitive Computing in Engineering, 2: 202-207. https://doi.org/10.1016/j.ijcce.2021.11.003
[18] Noor, A., Islam, M. (2019). Sentiment analysis for women's e-commerce reviews using machine learning algorithms. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, pp. 1-6. https://doi.org/10.1109/ICCCNT45670.2019.8944436
[19] Zhao, H., Liu, Z., Yao, X., Yang, Q. (2021). A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. Information Processing & Management, 58(5): 102656. https://doi.org/10.1016/j.ipm.2021.102656
[20] Rezaei, S., Kahani, M., Behkamal, B., Jalayer, A. (2022). Early multi-class ensemble-based fake news detection using content features. Social Network Analysis and Mining, 13(1): 16. https://doi.org/10.1007/s13278-022-01019-y
[21] Savci, P., Das, B. (2023). Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. Journal of King Saud University-Computer and Information Sciences, 35(3): 227-237. https://doi.org/10.1016/j.jksuci.2023.02.017
[22] Lin, W., Zhang, Q., Wu, Y.J., Chen, T.C. (2023). Running a sustainable social media business: The use of deep learning methods in online-comment short texts. Sustainability, 15(11): 9093. https://doi.org/10.3390/su15119093
[23] Gozudeli, Y., Karacan, H., Yildiz, O., Baker, M., Minnet, A., Kalender, M., Akcayol, M., (2015). A new method based on Tree simplification and schema matching for automatic web result extraction and matching. Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, pp. 1-5.
[24] Gozudeli, Y., Yildiz, O., Karacan, H., Baker, M.R., Minnet, A., Kalender, M., Akcayol, M.A. (2014). Extraction of automatic search result records using content density algorithm based on node similarity. In the International Conference on Data Mining, Internet Computing, and Big Data (BigData2014), Asia Pacific University of Technology and Innovation (APU), Kuala Lumpur, Malaysia, pp. 69-75.
[25] Rajput, G.K., Kundu, S., Kumar, A. (2021). The impact of feature extraction on multi-source sentiment analysis. In 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, pp. 510-515. https://doi.org/10.1109/SMART52563.2021.9676201
[26] Alamoodi, A.H., Baker, M.R., Albahri, O.S., Zaidan, B.B., Zaidan, A.A., Wong, W.K., Baqer, M.J. (2022). Public sentiment analysis and topic modeling regarding COVID-19's three waves of total lockdown: A case study on movement control order in Malaysia. KSII Transactions on Internet & Information Systems, 16(7): 2169-2190. https://doi.org/10.3837/tiis.2022.07.003
[27] Baker, M.R., Taher, Y.N., Jihad, K.H. (2023). Prediction of people sentiments on twitter using machine learning classifiers during Russian aggression in Ukraine. Jordanian Journal of Computers & Information Technology, 9(3): 187-204. https://doi.org/10.5455/jjcit.71-1676205770
[28] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953
[29] Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X. (2013). Applied logistic regression. John Wiley & Sons.
[30] Baker, M.R., Utku, A. (2023). Unraveling user perceptions and biases: A comparative study of ML and

DL models for exploring twitter sentiments towards ChatGPT. Journal of Engineering Research, 11: 23. https://doi.org/10.1016/j.jer.2023.11.023

[31] Araque, O., Corcuera-Platas, I., Sánchez-Rada, J.F., Iglesias, C.A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications, 77: 236-246. https://doi.org/10.1016/j.eswa.2017.02.002

[32] Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, pp. 785-794. https://doi.org/10.1145/2939672.2939785

[33] Kumar, V., Kedam, N., Sharma, K.V., Khedher, K.M., Alluqmani, A.E. (2023). A comparison of machine learning models for predicting rainfall in urban metropolitan cities. Sustainability, 15(18): 13724. https://doi.org/10.3390/su151813724

[34] Freund, Y., Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1): 119-139. https://doi.org/10.1006/jcss.1997.1504

[35] Esuli, A., Sebastiani, F. (2009). Training data cleaning for text classification. In Conference on the Theory of Information Retrieval, Berlin, Heidelberg, pp. 29-41. https://doi.org/10.1007/978-3-642-04417-5_4

[36] Bangyal, W.H., Qasim, R., Rehman, N.U., Ahmad, Z., Dar, H., Rukhsar, L., Ahmad, J. (2021). Detection of fake news text classification on COVID-19 using deep learning approaches. Computational and Mathematical Methods in Medicine, 2021(1): 5514220. https://doi.org/10.1155/2021/5514220

[37] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 19th International Conference on Computational StatisticsParis France, 2010 Keynote, Invited and Contributed Papers, Paris, France, pp. 177-186. https://doi.org/10.1007/978-3-7908-2604-3_16

[38] Aziz, E.F., Baker, M.R. (2024). Enhancing multi-class password strength prediction through machine learning and ensemble techniques. International Journal of Safety & Security Engineering, 14(5): 1635-1645. https://doi.org/10.18280/ijsse.140530

[39] Baker, M.R., Mohammed, E.Z., Jihad, K.H. (2022). Prediction of colon cancer related tweets using deep learning models. In International Conference on Intelligent Systems Design and Applications, Switzerland, Springer, pp. 522-532. https://doi.org/10.1007/978-3-031-27440-4_50