Product and Business Performance Evaluation Based on Customer Reviews.

ABSTRACT

This research explores ASDA's business performance by analysing customer reviews from Trustpilot, employing sentiment analysis and machine learning models. The methodology involves systematic data collection, exploratory data analysis (EDA), and model development. Python and BeautifulSoup are employed for web scraping Trustpilot reviews. EDA encompasses various analyses, including sentiment analysis, topic modeling, and aspect-based analysis. Machine learning models are built, and their performance is evaluated using metrics like accuracy, F1 scores, ROC curve and Learning curve. Performance metrics, average review ratings, and sentiment analysis provide a comprehensive picture of customer sentiment. The exploration of what customers are saying reveals specific areas of commendation and concern. Latent Semantic Analysis (LSA) identifies key topics like delivery, customer service, and product issues. The dominance of negative sentiment underscores challenges in customer experiences. This research provides valuable insights into ASDA's business performance through an in-depth analysis of customer reviews. The Gradient Boosted Classifier emerged as the best classifier and model for sentiment prediction. The study contributes to the understanding of using customer reviews for business evaluation and lays a foundation for enhancing customer satisfaction through classification of future customers review for necessary action.

1. INTRODUCTION AND BACKGROUND

There has been the rise of online platforms like Trustpilot, Yelp dedicated to review products and businesses, additionally business also have incorporated reviews on their site. These reviews serve as repository of information, constituting a pivotal feedback mechanism for both consumers and businesses. This research aims to explore the business performance by utilizing collected customer reviews from Trustpilot on ASDA as a case study.

The rise and dominance of e-commerce has empowered consumers to share their opinions through reviews and ratings. This has made businesses to incorporate feedback in their

sites. These user-generated insights transcend mere anecdotal significance; they embody a trove of data that, when effectively analysed, can offer profound insights into product performance and the efficacy of business operations (1). Asda stands as a noteworthy example for sentiments and experiences of a diverse consumer base.

Customers reviews has been found to play a pivotal role in influencing purchasing decisions and shaping brand perception (2). Prior studies have illuminated the correlation between positive reviews and increased sales, as well as the potential impact of negative reviews on customer trust and loyalty.

1.1 Sentiment Analysis in Customer Review Analysis

Recent advancements in data science and AI have opened new avenues for extracting valuable insights from the vast pool of customer reviews. Liu's (3) work on sentiment analysis and opinion mining has been instrumental in understanding the nuances of customer sentiments expressed in textual data. Machine learning models, particularly those using natural language processing and deep learning, have shown effectiveness in capturing the complexities of sentiment in diverse textual content (4).

1.2 Machine learning in Reviews and Text Data

The application of predictive modelling from machine learning techniques, has been seen to correlate customer sentiments with business performance indicators. Hastie, Tibshirani, and Friedman's work on "The Elements of Statistical Learning" provide foundational insights into the use of machine learning for predictive modelling (5). Advancements in predictive modelling for sentiment classification are evident in research by Pang and Lee (4), emphasizing the effectiveness of machine learning approaches in capturing the nuances of sentiment within diverse textual content. As businesses increasingly recognize the significance of customer reviews, the development and refinement of predictive models play a pivotal role in extracting actionable insights from the vast landscape of customer reviews. Harish et al (17) showed how using a hybrid feature selection made up of machine learning based feature selection and lexicon-based feature selection gave a very good performance when used with classification model. They achieved this by using bag of words technique for the machine learning based feature selection then positive and negative word count for lexicon feature selection. Neelakandan and Paulraj (14) proved Gradient Boosted Decision Tree classified tweets better and faster when benchmarked with other models. They compared each more in terms of average sentiment score, accuracy, F-score, recall and precision. Shamantha *et al* (16) did feature selection based on each score word and evaluated the performance of each model used based on accuracy and precision.

1.3 Natural Language Processing (NLP) and Text Mining

Natural Language Processing (NLP) serves as a critical component in the analysis of text data. In their work "Speech and Language Processing" (6), Jurafsky and Martin delve into the complexities of Natural Language Processing (NLP), exploring subjects such as sentiment analysis, named entity recognition, and part-of-speech tagging. The integration of NLP techniques contributes significantly to the extraction of meaningful information from textual datasets.

Using methods such as sentiment analysis, machine learning models, and natural language processing, this study aims to unveil clear patterns and sentiments within the reviews. The research further explores developing predictive models that can predict the sentiment of customers reviews, offering a proactive approach to address challenges and capitalize on strengths.

2. METHODOLOGY, DATASET AND CODE IMPLEMENTATION

2.1 Data Collection

The data collection process for ASDA reviews on Trustpilot involved a systematic approach using Python and the BeautifulSoup library. Key steps included importing necessary libraries for web scraping and analysis, utilizing BeautifulSoup to extract information from TrustPilot pages, and systematically storing data in Pandas DataFrames. To overcome challenges like pagination and potential IP blocking, the process was divided into multiple stages with VPN employed for IP rotation. Three DataFrames were created from each stage and later concatenated to form a comprehensive dataset. The data underwent cleaning and conversion, with the final dataset saved to a CSV file. To focus the analysis, the dataset was filtered based on experience dates, covering reviews from January 1, 2017, to September 30, 2023. In total, 11,009 reviews were gathered for the analysis.

2.2 Exploratory Data Analysis

The primary goal of the Exploratory Data Analysis (EDA) on ASDA reviews from Trustpilot was to uncover patterns and trends in the dataset, thereby gaining valuable insights into how ASDA performed using customers sentiments and experience. The analysis encompassed various objectives, including examining word counts per review, visualizing the distribution of review ratings, exploring temporal patterns in reviews and ratings, identifying top words through lemmatization, conducting sentiment analysis using TextBlob and VADER Sentiment Analyzer, employing Latent Semantic Analysis (LSA) for

topic modelling, conducting aspect-based sentiment analysis through spaCy, generating WordCloud to visually represent descriptive terms for positive and negative sentiments, and evaluating performance metrics over distinct time frames. The EDA will provide a comprehensive understanding of customer feedback, sentiments, and trends of ASDA reviews.

2.3 Model Data Pre-processing

This stage involves machine learning focus EDA, Data processing and feature engineering. A comprehensive data preparation and sentiment labels that will be appropriate for building a model was done here. Data preparation and pre-processing stage employs text cleaning techniques such as lowercase conversion, removal of punctuation and stop words, lemmatization, and stemming. Lemmatization was done to make the words readable, and Stemming was done to reduce long reviews for computational speed (7). Also, various analysis, including the calculation of stop word percentages, positive word percentages, and evaluation of polarity and subjectivity using TextBlob and VADER Sentiment Analyzer, are performed. Getting the percentage of positive words was aided with positive word lexicon compiled by Shekhar Gulati[1]. The lemmatized words were used calculate the Polarity and Subjectivity of the reviews using TextBlob. Data preparation for the machine learning had positive and negative for sentiment labels. To determine which review features best correlate with sentiment, I analysed the median values for different attributes across positive and negative reviews by visualising the distribution of percentage of stop words, percentage of positive words, and stemmed word counts within the review text. The training data's target variable (sentiment) was categorised into negative and positive sentiments to address the data imbalance, with a predominance of negative sentiments (**Figure 1**). Also, this was done to reduce the complexity of the model and to provide a clear distinction between negative and positive sentiment. The pre-processed data is then saved for subsequent utilization.

[1] Shekhar Gulati: https://github.com/shekhargulati/sentiment-analysis-python/blob/master/opinion-lexicon-English/positive-words.txt

Figure 1: Distribution of Count of Sentiment for training Data

The model processing stage involves loading the pre-processed data, categorizing reviews into positive and negative sentiments, and converting tokenized reviews into TF-IDF word vectors. This generally involve converting the text review data to what machine learning can understand and a number-based feature that has relationship with the label we want to predict. The use of CountVectorizer and TF-IDF to convert the review data into a format suitable for machine learning models, where each document is represented by a vector indicating the count of each word in the review. Both vectorizers were tested for accuracy and the ideal number of token pair(ngrams) was also determined to ensure the ideal number to use. When converting review data into a format suitable for machine learning models using CountVectorizer and TF-IDF, the choice of "ngrams" involves determining the length of the sequences of words to consider. This is an exploration of different combinations of unigrams, bigrams to find the most effective representation of the review data for the machine learning models. The goal is to identify the ngram size that best captures the relevant information and patterns in the text data. A logistic regression model is trained for sentiment analysis, and its performance is assessed using a confusion matrix and accuracy score. Confusion matrix was employed for result interpretation. The TF-IDF vectorizer is fine-tuned through adjustments in n-grams, minimum document frequency (min_df), and maximum document frequency (max_df), with optimal hyperparameter

values identified. Setting a minimum document frequency helps filter out terms that are too rare and might be considered noise in the dataset. It can be used to focus on more common and potentially more informative terms while maximum document frequency helps filter out terms that are too common and might not provide much discriminatory power (8).

Machine learning model training involves utilizing TF-IDF vectorization on stemmed sentences for numerical representation. The logistic regression model is then trained and evaluated on a test set, considering variations in n-grams and minimum document frequency to ascertain optimal values. Logistic Regression is particularly well-suited for binary classification tasks. To ensure data hygiene, we store the processed data in a separate location.

2.4 Model Building

In this phase, diverse models are employed to refine the initial predictions made by the base logistic regression. Four classification models, namely Logistic Regression, Random Forest, Support Vector Classification, and Gradient Boosted Classifier, undergo testing on the dataset. The model building process involves the creation of a pipeline incorporating standard scaling and classification. GridSearch is employed to identify the optimal set of hyperparameters for each classifier (9). Standard scaling, a technique ensuring uniform feature scales, is utilized to prevent certain features from overshadowing others during training. The classifiers, algorithms predicting input categories, are integral to the model's output predictions. GridSearch is employed to fine-tune hyperparameters, crucial settings determined before training begins. The methodology includes generating a confusion matrix and assessing key performance metrics such as Accuracy, Precision, Recall, Specificity, and F1 Score. These metrics will collectively underscore the overall effectiveness of the best model for the data.

3. RESULTS

The primary objective of this research is to evaluate the business performance of ASDA by analysing customer reviews on Trustpilot. Diverse performance metrics will be employed to provide a comprehensive assessment of the business performance based on customers reviews. Additionally, a machine learning model will be trained to predict the sentiments expressed in these reviews, adding a predictive dimension to the analysis. This approach aims to offer a fine and data-driven understanding of ASDA's performance, combining quantitative metrics with advanced predictive capabilities for a more insightful

evaluation. The period under for analysis is 2017 to 2023 (Before lockdown 2020 during lockdown and after lockdown).

3.1 Performance Metrics

Average Review Ratings

In business reviews, Trustpilot serves as a platform where customers not only share their experiences but also assign a performance score to businesses using a star rating system. Notably, the average rating, indicated by the number of stars, was identified as 1.459, reflecting the collective sentiment of customers towards the ASDA on the platform. It is noteworthy that the year 2020, witnessed a substantial surge in the number of reviews as shown in **figure 2b**, coinciding with a period marked by covid lockdowns where physical contacts was limited a potential correlation with increased online orders.

Interestingly, in **figure 2a** shows the average ratings reached its peak in 2020, suggesting a potential correlation with heightened customer satisfaction during the lockdown period. Post-2020, there was a discernible decline in the average reviews, marking a shift in the trend.

These fluctuations in the average rating and review counts unveil valuable insights into the evolving dynamics of customer sentiment, potentially influenced by external factors such as the global lockdown and subsequent adaptations in consumer behaviour after the lockdown.

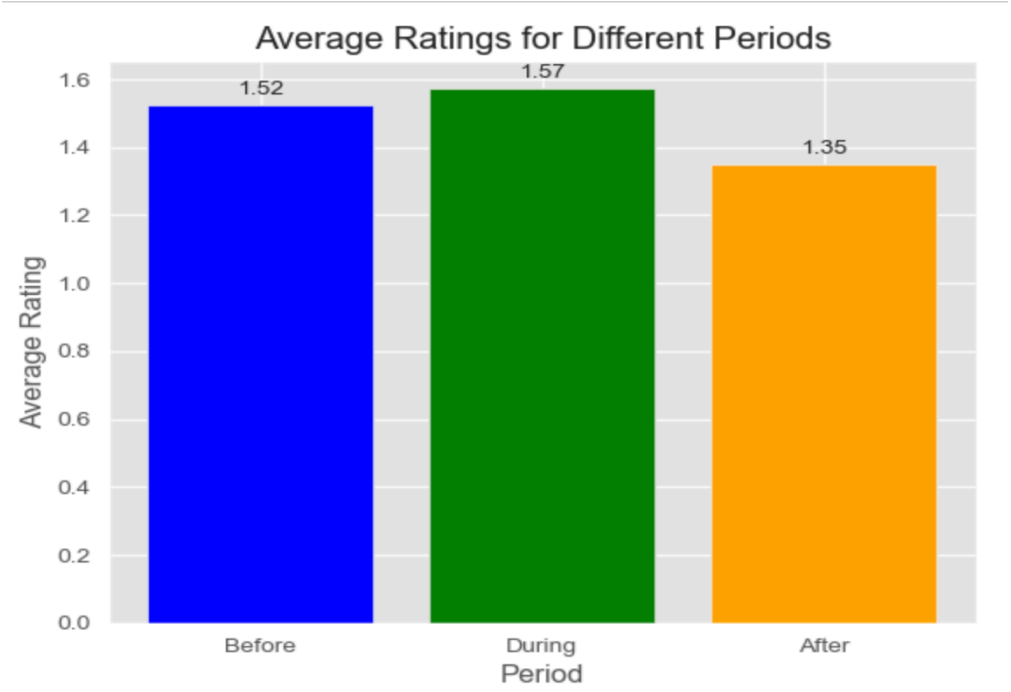

Figure 2a: Average ratings for different periods

Figure 2b: Trend of Reviews for different years

What Customers are Saying?

Upon analysis, customers were observed to talk much on the level of "customer service" they received (**figure 3a**). This underscores the significance of delivering exceptional service, whether in-store or online. Specific aspects such as "click and collect," interactions with drivers, delivery experiences, and refund processes were also prominently discussed, indicating a notable focus on online shopping and instances where customers expressed dissatisfaction, leading to refund requests.

Further breakdown of the top words based positive and negative reviews in **figure 3b** revealed positive reviews were characterized by commendatory words such as "amazing," "helpful," "excellent," and "friendly," potentially reflecting the positive impact of the services rendered. Conversely, negative reviews featured words like "worst," "rude," "shocking," "awful," and "poor," suggesting potential concerns related to the conditions of products and the quality of services provided. This insights into the customer sentiment, shedding light on areas for improvement and enhancement in both product offerings and service delivery.

Figure 3a: What customers are saying.

Figure 3b: What customers are filtered on sentiments.

The topics extracted from the Latent Semantic Analysis (LSA) of ASDA reviews provide insights into various themes and concerns expressed by customers as shown in **table 1**. These can be broken down as follows.

- Delivery and Ordering Process (Topics 1, 5, 7, 10): Customers frequently discussed aspects related to the delivery process, including orders, delivery drivers, and online shopping experiences. Topic 7 suggests that the click-and-collect service, a popular feature in modern retail, was a notable part of customer discussions.

- Customer Service and Staff (Topics 3, 6, 8, 9): The presence of topics related to staff and customer service (Topics 3 and 6) indicates that customers often shared their experiences with the service quality and interactions with staff members. Topic 8 highlights positive sentiments related to great service and prices, reflecting positive customer experiences.

- Product Issues (Topics 4, 12, 14): Topics 4 and 14 suggest concerns about missing items in stock and issues related to ordered products. This may indicate challenges in maintaining inventory or fulfilment issues.

- App and Online Shopping (Topics 11, 13): Customers discussed their experiences with the ASDA apps, including features like rewards on the reward app (Topic 11) and potential frustrations, as hinted by the term "waste" in Topic 13.

- Refund and Waiting Time (Topics 15): The last topic, centred around refund and waiting time, points to concerns customers may have regarding the speed and efficiency of refund processes.

It's essential to note that Topic 2 appears to contain terms that may not directly contribute to a coherent theme, potentially indicating noise or unclear patterns in the data. Further investigation may be needed for this topic.

Topics and Top Words	
Topic 1	delivery asda order
Topic 2	nan cartwheeling sniffed
Topic 3	staff store asda
Topic 4	items missing stock
Topic 5	delivery driver home
Topic 6	service customer missing
Topic 7	order collect click
Topic 8	great service prices
Topic 9	collect click service
Topic 10	shopping online delivered
Topic 11	rewards app asda
Topic 12	food date bought
Topic 13	time app waste
Topic 14	stock dont ordered
Topic 15	refund time waiting

Table 1: Common words or themes in identified topics.

Sentiment Analysis

Examining the sentiment distribution in the reviews **figure 4** revealed a predominant negative sentiment, constituting over 60% of the 10,790 reviews. This consistent pattern was observed across the three distinct periods under review in the analysis. A deeper dive into the reviews made clear the dominance of negative sentiment when considering the percentage of positive opinions (**figure 5**) in the reviews, underscoring the limited occurrence of positive sentiments in the overall review.

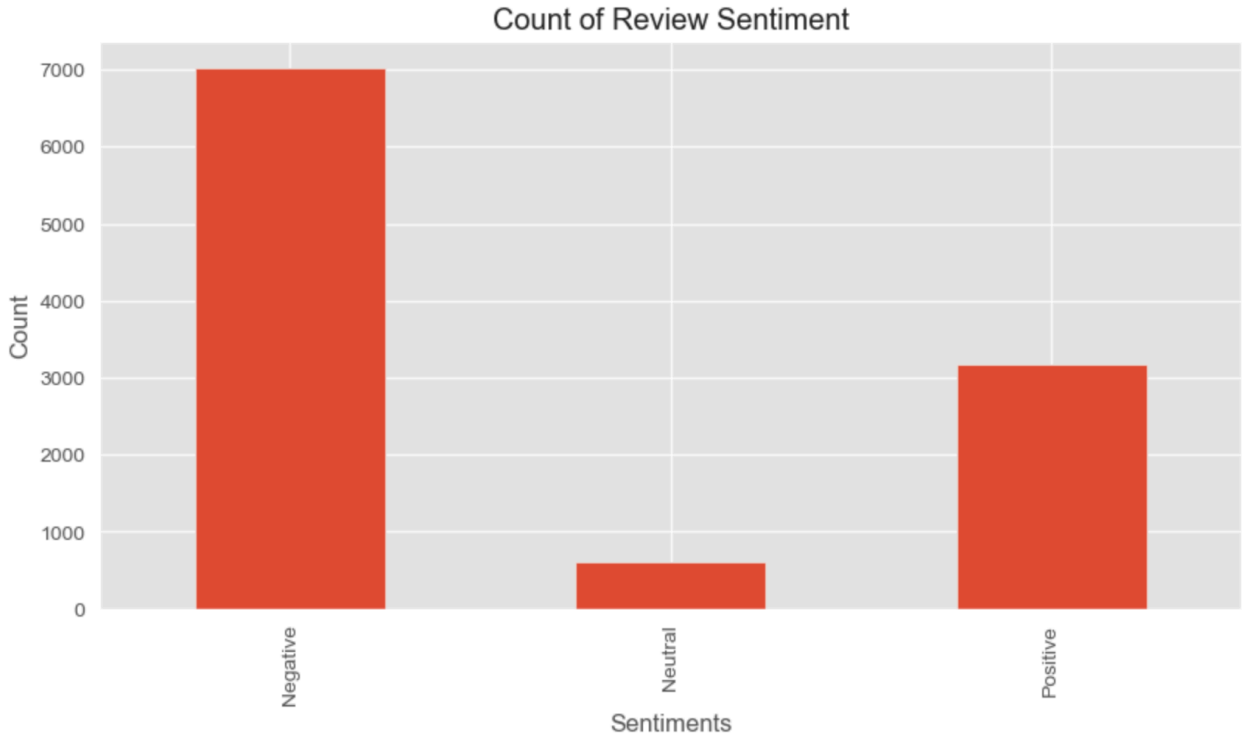

Figure 4: Count of Review Sentiment

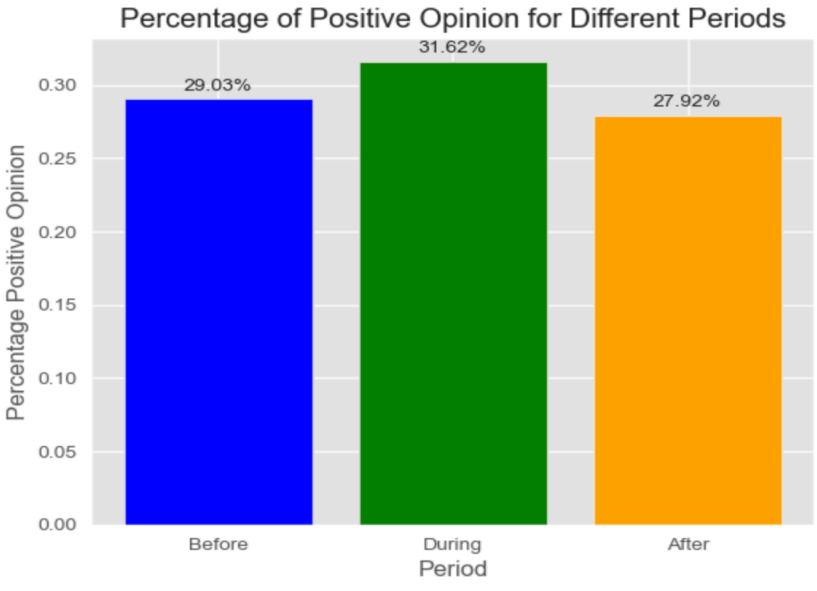

Figure 5: Percentage of Positive Opinion

Machine Learning Model

Feature Selection

Stop word percent, positive word percent and stemmed word count in the reviews were analysed to get the best features. Each bar in the histograms shows how many reviews in each category fall into a particular range of the Stop word percent, percentage of positive words and stemmed word count. Of these features, the percentage of positive words in each reviews provided the clearest differentiation between positive and negative sentiment. As shown in **Table 2** and the associated histogram (**Figure 6)**, reviews labelled as positive sentiment had a higher median percentage of explicit positive words (0.1296%) compared to negative reviews (0.0968%) this is 33.8% increase. This makes positive word percentage a more determining feature for the sentiments. Using the median (the blue vertical line in the histograms) helps provide a more robust measure of the central tendency of each feature, making the analysis less susceptible to the impact of outliers. This indicates that reviews containing more positive word tend to have an overall positive sentiment while that with lower positive word percent has negative sentiment. The feature analysis highlights that the relative frequency of positive vocabulary within customer reviews exhibits the strongest correlation with positive review labelling. This factor appears more differentiate the sentiments than stop word percentage and stemmed word count. The percentage of positive words could therefore serve as a feature when building a model for categorizing sentiment for new customer feedback.

	FEATURES		
Sentiments	**Stop Word Percentage**	**Positive Word Percentage**	**Stemmed Word Count**
Positive	0.469	0.1296	43.0
Negative	0.465	0.0968	42.0
% difference between categories	0.86%	33.8%	2.38%

Table 2: Feature Analysis of Positive and Negative Sentiments

Table 2 presents the feature analysis of sentiment categories, including stop word percentage, positive word percentage, and stemmed word count. The last row provides the percentage difference between positive and negative sentiments for each feature to show how the feature differentiate the categories.

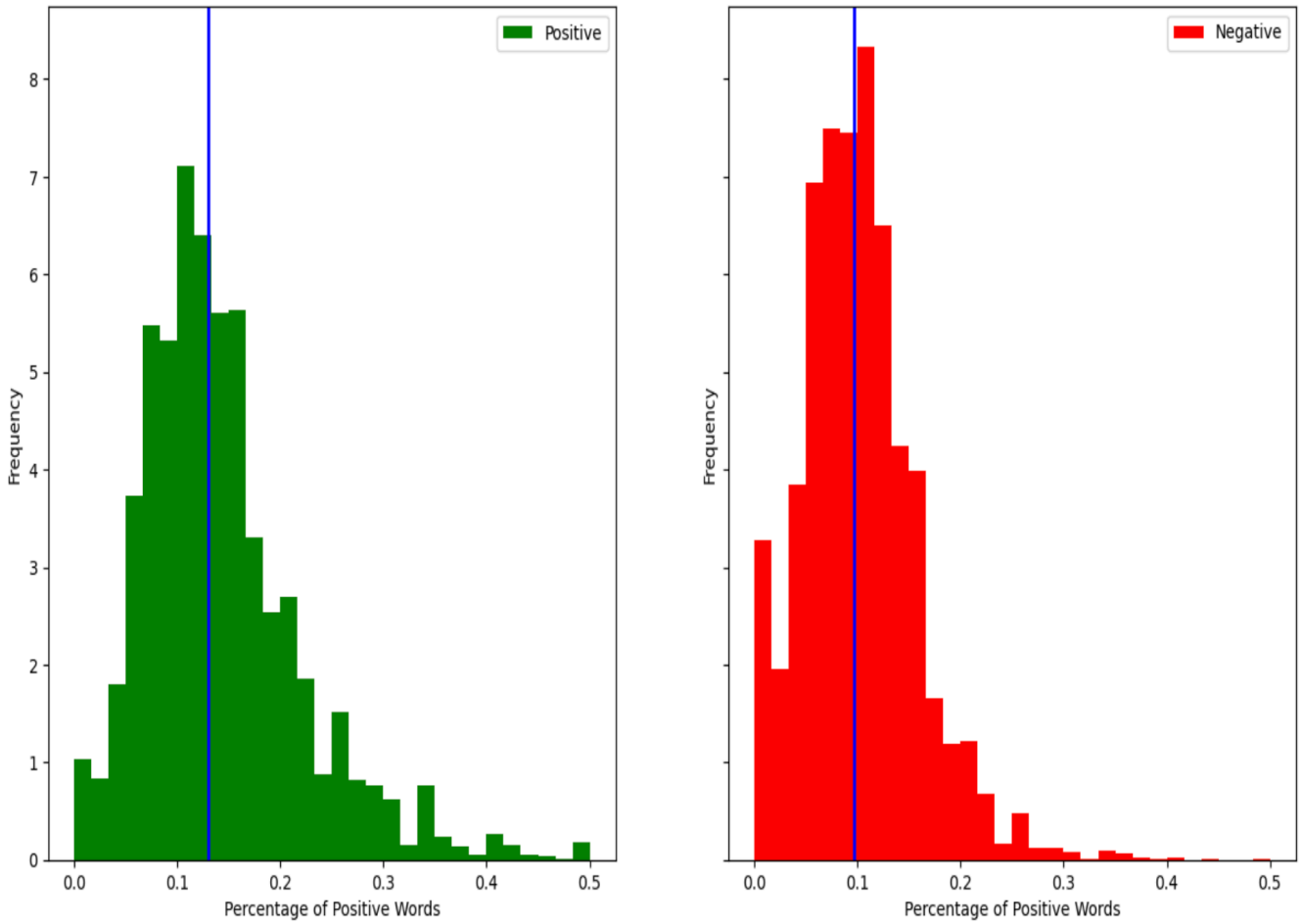

Figure 6: Distribution of Positive word percent Across Negative and Positive Sentiment Reviews

CountVectorizer and TfidfVectorizer were employed to transform the textual reviews into numerical representations. Both had accuracies of 86% and 84% respectively (**table 3**) when tested on a base logistic regression model, the similarity in performance is notable. I opted for TF-IDF (Term Frequency-Inverse Document Frequency) is preferred as it places emphasis on the significance of words within each individual document relative to their occurrence across the entire set of reviews. This distinction is crucial because reviews are diverse, being written by different individuals expressing their unique thoughts and perspectives.

	CountVectorizer	**TF-IDF**
Accuracy Score	0.8576	0.8397
ACC Macro	0.8576	0.8397
F1 Macro	0.8527	0.8333
FPR Macro	0.1520	0.1719
Kappa	0.7064	0.6683
NPV Macro	0.8636	0.8480

	CountVectorizer	TF-IDF
Overall ACC	0.8576	0.8397
PPV Macro	0.8636	0.8480
SOA1 (Landis & Koch)	Substantial	Substantial
TPR Macro	0.8480	0.8281
Zero-one Loss	430	484

Table 3: A side-by-side comparison of the metrics for both CountVectorizer and TF-IDF approaches

The optimal value for the minimum word frequency or number of occurrences (min_df) in the TFIDF vectorizer, resulting in the highest accuracy score of the model was 34 as shown in **figure 7**. Additionally, the ideal number of n-grams for the best accuracy was determined to be 1 (**figure 8**).

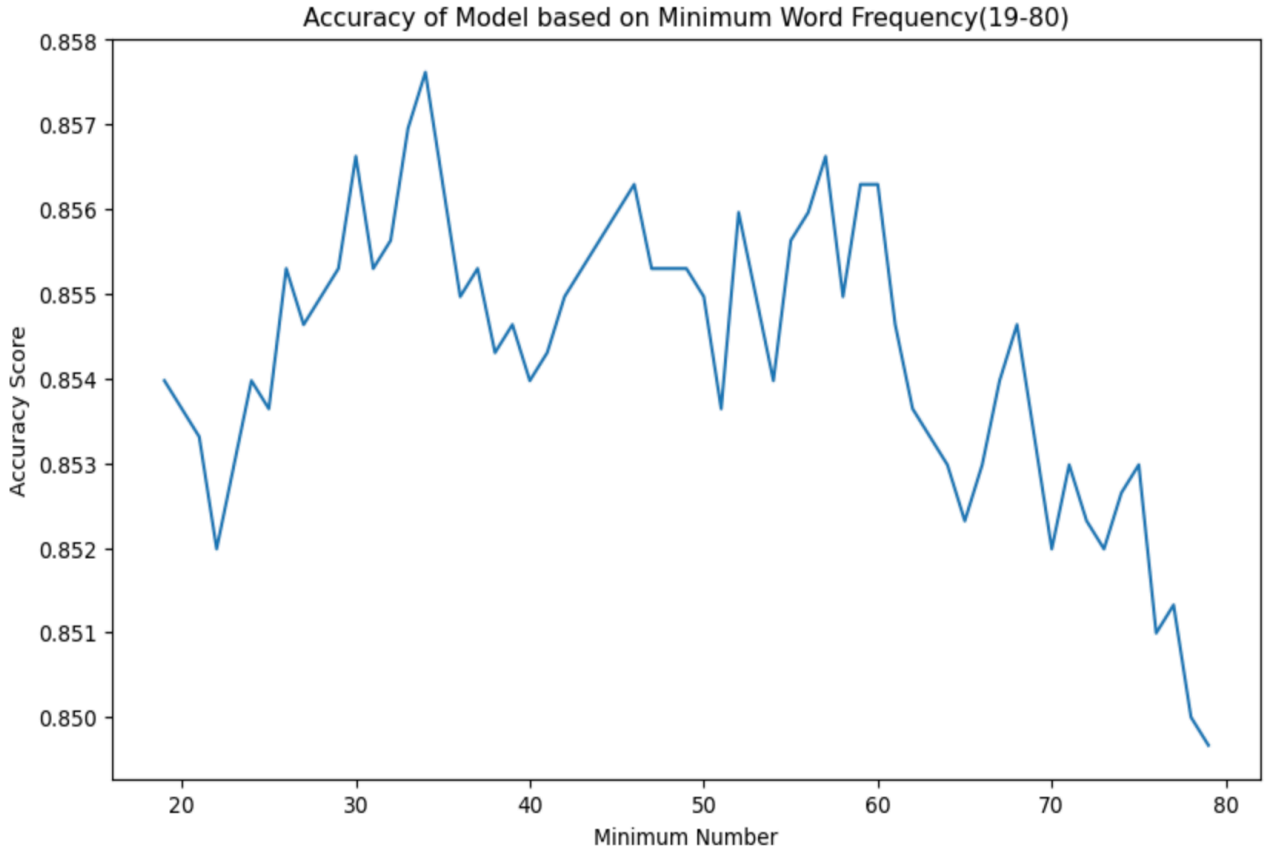

Figure 7: Accuracy of model based on minimum word frequency(19-80).

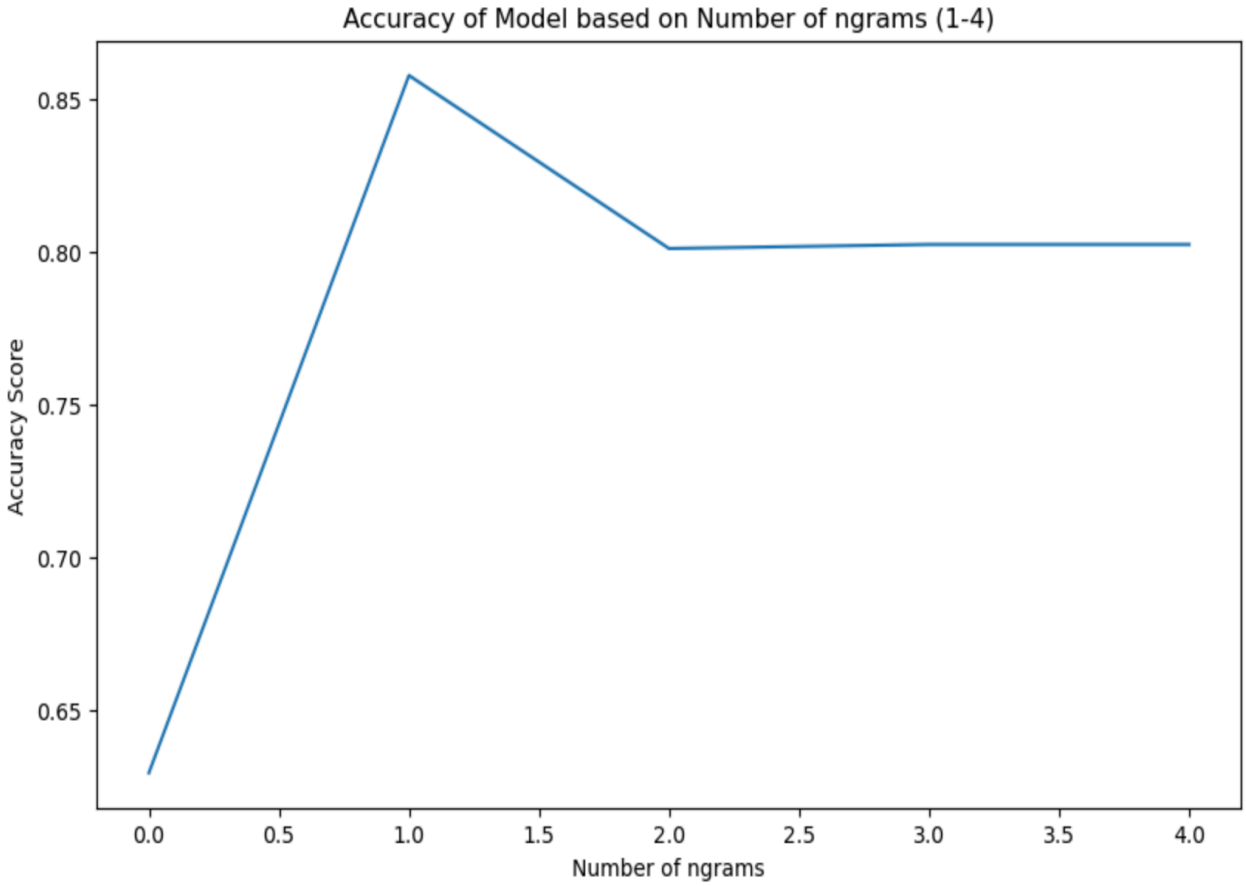

Figure 8: Accuracy of model based on number of ngrams (1-4).

Models Result and Performance Evaluation

The review data was tested on different classification model; Logistic regression, random forest, Support Vector Classification, and Gradient Boosted Classifier model and the result is present in **table 4**.

Model	Accuracy	F1_Macro	F1_Micro	F1_Weighted
Logistic Regression	0.864	0.861	0.864	0.864
Random Forest	0.807	0.798	0.807	0.804
Support Vector Classification	0.861	0.858	0.861	0.861
Gradient Boosted Classifier	0.879	0.876	0.879	0.879

Table 4: Performance Metrics of Different Classification Models

Logistic Regression achieved the highest accuracy among the models at 86.4%, with strong F1 scores across macro, micro, and weighted categories. Random Forest displayed a lower accuracy of 80.7%, with slightly lower F1 scores in comparison to Logistic Regression. Support Vector Classification (SVC) showed competitive performance with an accuracy of 86.1% and strong F1 scores. Gradient Boosted Classifier outperformed other models with the highest accuracy at 87.9% and strong F1 scores across different metrics.

Gradient Boosted Classifier was chosen as it performed highest. Summary statistic (**table 5** and **table 6**) of the confusion matrix further highlights the model accuracy.

Overall Statistic	
Metric	**Value**
ACC Macro	0.87934
F1 Macro	0.87608
FPR Macro	0.1265
Kappa	0.75232
NPV Macro	0.87987
Overall ACC	0.87934
PPV Macro	0.87987
SOA1(Landis & Koch)	Substantial
TPR Macro	0.8735
Zero-one Loss	243

Table 5: Performance Metrics of the Gradient Booster Classifier Model

Class Statistic		
Classes	**Negative**	**Positive**
ACC(Accuracy)	0.87934	0.87934
AUC (Area under the ROC curve)	0.8735	0.8735
AUCI (AUC value interpretation)	Very Good	Very Good
F1	0.8962	0.85596
FN (False negative)	96	147
FP (False positive)	147	96
FPR	0.16916	0.08384
N (Condition negative)	869	1145
P (Condition positive)	1145	869
POP(Population)	2014	2014
PPV(Precision)	0.87709	0.88264
TN (True negative)	722	1049
TON (Test outcome negative)	818	1196
TOP (Test outcome positive)	1196	818
TP (True positive)	1049	722
TPR (Sensitivity)	0.91616	0.83084

Table 6: Performance Metrics Comparison for Negative and Positive Classes of the Gradient Boasted Classifier Model.

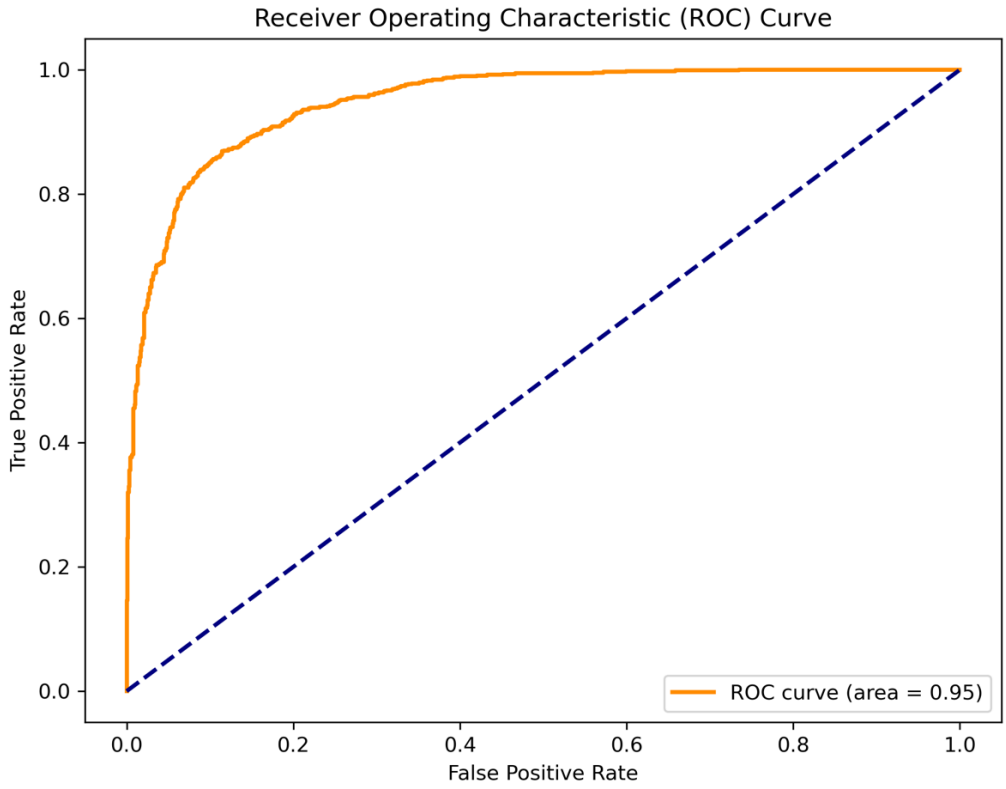

Figure 9a: ROC curve for the GBC model.

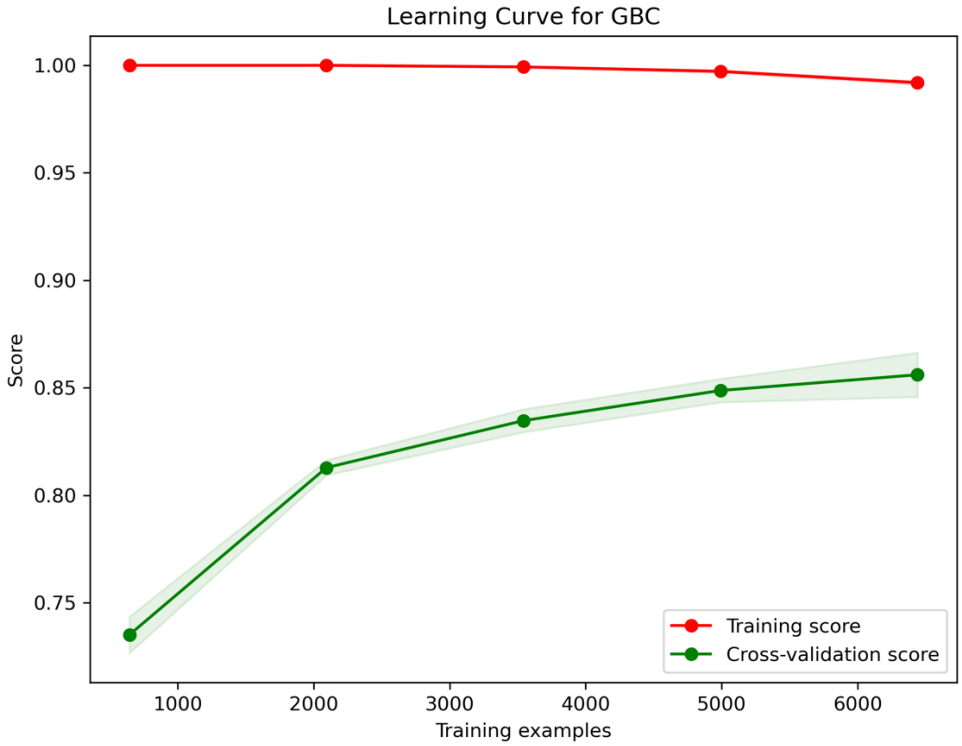

Figure 9b: Learning curve for the GBC model.

These metrics collectively provide a comprehensive evaluation of the model's performance on each class and overall. The GBC model demonstrates strong predictive capabilities, achieving high accuracy and F1 scores across both classes. The false positive and false negative rates are relatively low, indicating good performance in classifying both Negative and Positive instances. Also, ROC cure in **figure 9a** shows the model is very good at distinguishing between positive and negative reviews. The learning curve in **figure 9b** shows the curves are converging towards a similar accuracy value. Both curves also reach high accuracy levels in

4. DISCUSSION

In year 2020, there was a notable peak in the average rating, which can be attributed to the surge in online orders during the global lockdown, as argued by (10). This alignment suggests a correlation between increased customer reviews and the increased prevalence of online transactions during the lockdown period. The identified topics, including "Delivery and Ordering Process," "Customer Service and Staff," and "Product Issues," concise with prevalent themes observed in retail customer feedback, as documented (11). This indicates consistency with established patterns and concerns commonly expressed by customers in the retail and ecommerce.

The prevalence of negative sentiment, accounting for over 60% across all periods, suggests challenges in customer experiences. Upon examining the average star ratings of similar stores to ASDA on Trustpilot, it was noted that these stores also received low reviews. This observation led to the conclusion that customers often turn to Trustpilot primarily to voice complaints about their experiences with these stores. The consistent pattern underscores the persistence of negative sentiments and highlights specific areas that warrant improvement.

The TF-IDF approach was chosen over CountVectorizer, despite nearly identical accuracies, due to TF-IDF's emphasis on the significance of word choice in individual reviews and data size. Optimal values for min_df and n-grams were identified, offering valuable insights into feature selection for sentiment analysis. A study by (12) also compared the accuracies of CountVectorizer and Term Frequency-Inverse Document Frequency (TF-IDF) techniques, confirming the superiority of TF-IDF, particularly in handling large datasets.

Gradient Boosted Classifier outperformed other models, aligning with (13) where ensemble models were described as a "state-of-the-art solution for many machine learning challenges. This result is further supported in (14) work on classifying sentiment of twitter data where Gradient Boosted Decision Tree classifier (GBDT) gave an efficient result. The model achieved high accuracy, precision, and recall, indicating robust predictive capabilities. This also agrees with (17) where high accuracy was achieved when they used hybrid feature selection with classification model.

This study faces limitations from several factors, starting with the data source. The data lacks specific details about the products or stores that customer visited or purchased. This could potentially restrict the depth of insights into sentiments specific to products or stores. To partially mitigate this limitation, topic modelling was employed to provide a broad overview of the subjects discussed by customers.

In the evaluation of the machine learning model, the learning curve revealed promising performance. However, a discernible gap between the training and cross-validation curves was observed. Although these curves appeared to converge, there was an indication that the model could benefit from additional training data. Future iterations of this study could involve using the entirety of the Trustpilot data on ASDA review, rather than restricting the analysis to specific eras or years under review.

During the data collection phase, Trustpilot site-imposed restrictions on my IP address due to an excessive number of requests. This issue was successfully addressed by implementing a Virtual Private Network (VPN) for the connection. While this solution effectively resolved the problem, it underscores the importance of considering potential technical constraints in large-scale web scraping projects.

5. CONCLUSION

This research has provided valuable insights into ASDA's business performance through a comprehensive analysis of customer reviews on Trustpilot. The exploration of various performance metrics, sentiment analysis, and machine learning models has contributed to a great understanding of customer sentiment. The Gradient Boosted Classifier emerged as the most efficient in predicting sentiment. The robust performance of the model, as evidenced by high accuracy, precision, and recall, adds confidence to its applicability in sentiment analysis.

The contribution and significance of this research is using customer reviews for business evaluation and performance and ultimately provide a significant foundation for improving customer satisfaction. The models were able to identify the feedback type (Positive or Negative) and these models can be used on future reviews to track changes over time allowing businesses to adapt and evolve in response to shifting consumer preferences and market dynamics.

Recommendations for future work in this research could explore the impact of external events on customer sentiment and using review data that delve deeper into specific product or service aspects. Additionally, incorporating deep learning algorithms like Recurrent Neural Networks with Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) which in Kansara et al (15) work, proved deep learning techniques performed better when it was benched marked with traditional machine learning techniques. Also, the integration of customer demographic data could further enhance the predictive capabilities of sentiment analysis models.

References

(1) Anderson EW, Sullivan MW. The Antecedents and Consequences of Customer Satisfaction for Firms. *Marketing Science*. 1993;12(2):125-43.

(2) Chevalier JA, Mayzlin D. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*. 2006;43(3):345-54.

(3) Liu B. *Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies*. 2012;5(1):1-167.

(4) Pang B, Lee L. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 2008;2:1-135. https://doi.org/10.1561/1500000011

(5) Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second. Springer; New York: 2009.

(6) Jurafsky D, Martin JH. *Speech and Language Processing*. Pearson; 2019.

(7) Pramana R, Debora, JJ Subroto, AAS Gunawan, Anderies. Systematic Literature Review of Stemming and Lemmatization Performance for Sentence Similarity. *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*. Yogyakarta, Indonesia; 2022. pp. 1-6. https://doi.org/10.1109/ICITDA55840.2022.9971451.

(8) Behera SK, Dash R. Performance of ELM Using Max-Min Document Frequency-Based Feature Selection in Multilabeled Text Classification. In: Mishra D, Buyya R, Mohapatra P, Patnaik S, eds. *Intelligent and Cloud Computing*. Smart Innovation, Systems and Technologies, vol 194. Springer, Singapore; 2021. https://doi.org/10.1007/978-981-15-5971-6_46

(9) Shekar BH, Dagnew G. Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). Gangtok, India; 2019. pp. 1-8. https://doi.org/10.1109/ICACCP.2019.8882943

(10) Guthrie C, Fosso-Wamba S, Arnaud JB. Online consumer resilience during a pandemic: An exploratory study of e-commerce behavior before, during and after a COVID-19 lockdown. *Journal of Retailing and Consumer Services*. 2021;61. https://doi.org/10.1016/j.jretconser.2021.102570.

(11) Lim J, Park M, Anitsal S, Anitsal MM, Anitsal I. Retail Customer Sentiment Analysis: Customers' Reviews of Top Ten US Retailers' Performance. Global *Journal of Management & Marketing (GJMM)*. 2019;3(1).

(12) Raza GM, Butt ZS, Latif S, Wahid A. Sentiment analysis on COVID tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models. 2021 International Conference on *Digital Futures and Transformative Technologies* (ICoDT2). 2021 May 20. IEEE. pp. 1-6.

(13) Sagi O, Rokach L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2018;8(4):e1249.

(14) Neelakandan S, Paulraj D. A gradient boosted decision tree-based sentiment classification of twitter data. *International Journal of Wavelets, Multiresolution and Information Processing*. 2020;18(04):2050027.

(15) Kansara D, Sawant V. Comparison of Traditional Machine Learning and Deep Learning Approaches for Sentiment Analysis. In: Vasudevan H, Michalas A, Shekokar N, Narvekar M, editors. Advanced Computing Technologies and Applications. Algorithms for Intelligent Systems. Springer; Singapore; 2020. https://doi.org/10.1007/978-981-15-3242-9_35

(16) Shamantha RB, Shetty SM, Rai P. Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance. In: *Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*; 2019; Singapore. IEEE; 2019. p. 21-25. https://doi.org/10.1109/CCOMS.2019.8821650.

(17) Harish BS, Kumar K, Darshan HK. Sentiment analysis on IMDb movie reviews using hybrid feature extraction method. *International Journal of Interactive Multimedia and Artificial Intelligence.* 2019. http://doi.org/10.9781/ijimai.2018.12.005