Semantic Analysis of Amazon Reviews: Towards Understanding Consumer Perceptions

A Machine Learning Perspective

Abdullah Safder
Student Number: 22262455
abdullah.safder2@mail.dcu.ie

Yash Kishor Kadulkar
Student Number: 23266269
yash.kadulkar2@mail.dcu.ie

Janmesh Mahesh Boricha
Student Number: 23270924
janmesh.boricha2@mail.dcu.ie

Somtochukwu Igboanuzue
Student Number: 23267439
Somtochukwu.igboanuzue2@mail.dcu.ie

Ayesha Imran
Student Number: 23264020
ayesha.imran2@mail.dcu.ie

***Abstract*—This study introduces an advanced predictive model aimed at categorising unreviewed product feedback on e-commerce platforms like Amazon into positive or negative sentiments with exceptional precision. In other words, it aims to mine the unstructured reviews for meaningful patterns using a standard data mining technique, Knowledge Data Discovery (KDD), that helps predict their sentiment as positive or negative. By leveraging machine learning and sentiment analysis techniques, the model is tailored to tackle the vast and unstructured nature of customer reviews. Using a diverse dataset from Amazon, cutting-edge algorithms and sophisticated feature engineering like TF-IDF vectorizer are employed and optimised to extract relevant features and capture nuanced semantic information from the text. Standard evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess the models' effectiveness so that they can be compared widely. This research attempts to make significant advancements in e-commerce sentiment analysis, offering businesses valuable insights into consumer sentiments for informed decision-making in competitive markets.**

Keywords: product reviews, data mining, feature extraction, cluster analysis, TF-IDF vectorization

Source code: https://gitlab.computing.dcu.ie/kadulky2/sentimentanalysisamazon

I. Introduction

The digital age has witnessed an unprecedented surge in online commerce, catalysing a prolific growth in user-generated content, particularly product reviews. These reviews, often imbued with nuanced sentiments and candid opinions, constitute a treasure of consumer insights essential for driving market intelligence and informing strategic business decisions. As businesses increasingly recognize the pivotal role of customer feedback in shaping product development strategies, the demand for sophisticated tools capable of extracting and analysing sentiments expressed in textual data has never been greater.

In response to this increasing demand, sentiment analysis, a subfield of natural language processing (NLP), has emerged as a potent instrument for the complex analysis of sentiments embedded within online reviews. Leveraging advances in machine learning and deep learning techniques, sentiment analysis endeavours to decode the underlying emotions, attitudes, and opinions articulated by consumers in their reviews, thereby furnishing businesses with actionable insights into prevailing market sentiments.

In this paper, we embark on a comprehensive study focusing on sentiment analysis of Amazon product reviews—a domain with rich and diverse expressions of consumer sentiments. Our investigation is underpinned by a robust dataset comprising a staggering 1,800,000 training samples and 200,000 testing samples in each polarity sentiment category. This dataset encompasses three pivotal attributes: polarity, title, and text. The polarity index serves as the pivot of our analysis, categorizing reviews into distinct sentiment categories with labels 1 denoting negative sentiments and 2 signifying positive sentiments. Meanwhile, the 'title' column encapsulates summaries of the reviews, while the 'text' column unveils the heart of the reviews.

The choice of this dataset is based on its immense size and the balanced distribution of its target features, which makes it the perfect arena for conducting rigorous experiments and refining models. Utilizing the vast amount of information embedded within this comprehensive dataset, we aim to develop a predictive model with the remarkable ability to distinguish and categorize unmarked product reviews into clear sentiment categories with unparalleled accuracy and efficiency.

Our goal with this study is to develop a model that can accurately predict whether a product review is positive or negative. This model will be valuable for businesses operating in the competitive landscape of online commerce, helping them understand consumer preferences, connect with their target audience, and create products that meet customer needs and desires.

In the ensuing sections of this paper, we tend to study the intricacies of our research methodology, detailing the steps undertaken for data preprocessing, feature extraction, and model training. Subsequently, our study outlines the empirical findings and performance evaluation of our proposed predictive model. The research reveals the efficacy of the model in predicting online commerce trends and its potential implications for businesses operating in dynamic landscapes.

II. Related Work

Sentiment analysis of Amazon product reviews has become a focal point in research. Recent studies have employed a spectrum of methodologies, ranging from traditional machine learning to advanced deep learning models, to decipher sentiments expressed in these reviews. This study provides a comprehensive overview of recent endeavours, critically evaluating the effectiveness, limitations, and potential advancements in sentiment analysis of Amazon product reviews.

[1] conducts a comparative analysis of sentence embedding techniques using machine learning algorithms on a Twitter dataset with a 3-class problem (positive, negative, neutral). GloVe embedding is employed, and various machine learning algorithms, including Logistic Regression, are evaluated, achieving an accuracy of 85.5% using Logistic Regression. The study contributes to the discourse on effective embedding techniques for sentiment analysis, showcasing the utility of GloVe embeddings in capturing sentiment nuances from short texts such as tweets. However, questions arise about the generalisability of the findings due to the focus on a specific dataset.

The Sentic GAT framework integrates context and common-sense knowledge for emotion recognition in conversation. Multi-head attention and sentiment-aware mechanisms are utilized, which helped achieve an accuracy ranging from 82% to 90%. [2] work emphasizes the importance of context in sentiment analysis, particularly in conversational settings. It builds upon previous studies by incorporating common-sense knowledge, showcasing advancements in capturing nuanced emotions. This study focused on the sentiment classification of Amazon product reviews using LSTM and Bidirectional LSTM models. In [3], GloVe embedding is utilized, revealing an accuracy of 80%, with a comparative analysis against Multinomial Naive Bayes. It sheds light on applying deep learning models, specifically LSTMs, in sentiment analysis of e-commerce data. While effective, using GloVe embeddings exhibits limitations in achieving higher accuracies. [4] research delves into sentiment analysis of Amazon product reviews, categorizing sentiments into positive, negative, and neutral (3-class problem). It employs Word2Vec and GloVe for word embedding, along with deep learning models like CNN, LSTM, and GRU. It provides insights into the effectiveness of various word embedding techniques and deep learning architectures. It builds upon previous research on sentiment analysis of e-commerce data, aiming to improve classification accuracy.

Building upon [5], this research introduced Convolutional Neural Networks (CNNs) for sentiment analysis of Amazon product reviews. It showcased state-of-the-art performance on benchmark datasets, emphasizing the hierarchical features learned from raw textual data. [6] contributed to the field of aspect-based sentiment analysis of Amazon reviews. It proposed a hierarchical attention network framework to extract aspect-specific sentiments, facilitating targeted product improvements and marketing strategies.

[7] research involved a domain adaptation approach based on adversarial training for sentiment analysis of Amazon reviews. It aims to minimise domain shift discrepancies while maximising sentiment classification accuracy. [8] research focused on multi-modal sentiment analysis, integrating textual and visual features from Amazon product reviews. The fusion-based approach showcased improved sentiment classification accuracy and interpretability.

A. Analysis

In recent years, the field of sentiment analysis of Amazon product reviews has witnessed a notable shift from traditional rule-based methodologies toward advanced deep learning models. This evolution signifies a significant stride towards capturing the intricate nuances of sentiments expressed by consumers in their reviews. Researchers have embraced this shift by exploring a diverse range of model architectures, from Long Short-Term Memory (LSTM) networks to Convolutional Neural Networks (CNNs) and attention mechanisms. This breadth of exploration reflects a concerted effort to extract deeper insights from the textual data inherent in Amazon product reviews, thus enriching the understanding of consumer sentiments. A key trend in recent studies, exemplified by papers such as [2], is the emphasis on context and common-sense knowledge integration. Researchers recognize the pivotal role of these elements in enhancing the models' ability to interpret sentiments accurately. This marks a progressive step towards developing more context-aware systems capable of capturing the subtle nuances embedded within consumer feedback.

However, as the field advances, certain limitations have come to light. Many studies have been focused on specific datasets, prompting legitimate concerns regarding the generalisability of findings to broader e-commerce contexts. Moreover, the interpretability of deep learning models has emerged as a critical consideration. The opaque nature of these models underscores the pressing need for transparency, ensuring that decisions made by algorithms can be understood and justified. Addressing the challenges posed by noisy and unstructured data within Amazon reviews stands as a crucial frontier for future research. Developing more robust sentiment analysis systems hinges on effectively navigating these complexities, ensuring reliable and accurate insights into consumer sentiments. Transfer learning approaches, while promising, require further exploration, particularly in adapting sentiment models to handle domain-specific shifts within the Amazon product landscape. This avenue presents an opportunity for improved accuracy and adaptability of sentiment analysis systems.

The field is poised to delve deeper into multi-modal analysis, building upon frameworks such as [8]. By integrating diverse modalities like audio and user-generated content, researchers aim to glean richer insights into consumer sentiments, thus offering a more comprehensive view of product feedback. Another promising direction for future research lies in the realm of personalized recommendations. Tailoring sentiment models to cater to individual user preferences could revolutionize the e-commerce landscape, offering consumers more targeted and relevant product suggestions. Lastly, as sentiment analysis systems increasingly influence consumer decisions, addressing the ethical implications becomes imperative. Exploring the implications of sentiment-based algorithmic recommendations is crucial for future investigation. Ensuring transparency and accountability in these systems will be essential for maintaining consumer trust and ethical standards.

The trajectory of sentiment analysis in Amazon product reviews showcases a progressive march towards more sophisticated methodologies. Researchers are actively embracing advancements in model architectures, context-aware frameworks, and multi-modal analysis to enhance the accuracy and relevance of sentiment analysis systems. This study aims to address the shortcomings of the literature by optimising the feature space and reducing the computational costs by identifying the most feasible model that can correctly classify Amazon product reviews as positive or negative sentiment.

III. Methodology

A. Dataset Acquisition

The dataset that we are using has been taken from Kaggle – "Amazon Reviews". Each sentiment category has 200,000 samples for testing and 1.8 million samples for training in the dataset (a total of 3.6M samples). Three columns make up each sample: text, title, and polarity.

1. Polarity - This column is the target class, which denotes the sentiment of the review. The reviews are labelled 1 for negative and 2 for positive.

2. Title - The 'title' column contains the review title. It was extracted as a separate column.

3. Text - This column contains the review body. This column will act as an independent variable for this problem.

We picked this dataset because of its volume. The most important part we are interested in is the even distribution of the 2 classes.

B. Data Preprocessing

In this step, we aim to clean and standardise the text data. The preliminary steps include checking for missing information and deciding what to do with it. Moreover, we saw if there were any similar amounts of positive and negative sentiments in the dataset. Removing duplicated reviews was equally important, saving a few computation cycles. Preprocessing is an attempt to make the dataset suitable for sentiment analysis by removing noise and irrelevant information by performing these text preprocessing steps-

1) Make text lowercase: Convert all letters in the text to lowercase to ensure consistency and avoid treating the same word differently due to case variations (e.g., "Word" vs. "word").
2) Remove text written in brackets: Exclude any text enclosed within brackets, as it may contain additional information or metadata irrelevant to the sentiment expressed in the review.
3) Remove punctuation: Deleting all punctuation marks such as commas, periods, and exclamation points to combine words and enhance text clarity.
4) Remove words containing numbers: Eliminate words containing numerical characters as they may be meaningless to the sentiment analysis and could be noise.
5) Remove frequent words in both classes: Eliminate the words "book" and "one" as they appear very commonly in positive and negative reviews.
6) Remove stopwords: Strip out common words known as stopwords (e.g., "and," "the," "is") that occur frequently in the language but do not carry significant semantic meaning. Removing stopwords helps focus on content-carrying words.
7) Stemming and Lemmatization: Apply stemming and lemmatization to reduce words to their base or root forms. This step helps to further normalize the text data and reduce the dimensionality of the feature space by grouping words with similar meanings.

C. Exploratory Data Analysis

It is essential to inspect the dataset to get an overview of the results to be expected. Thus, we conducted an Exploratory Data Analysis. When dealing with text datasets, the features are not always obvious. Rigorous steps must be taken to select and extract the features from a long string of characters, and these were taken sequentially in the following manner:

1) **Word Frequency**: The foremost step in text data analysis is finding the most frequent words across all classes. This step helped us understand if a few words from the text data occurring frequently are enough to categorise the data as positive or negative. From Fig. 3.3.1, Fig. 3.3.2, and Fig. 3.3.3, we can observe that the words "book," "one," and "good" are present across all classes, which may further create a dilemma for the machine learning algorithms. There may be other words that do not impact the sentiment but frequently occur. Wordclouds often help us identify these kind of words.

Fig. 3.3.1: WordCloud from all reviews

Fig. 3.3.2: WordCloud from positive reviews

Fig. 3.3.3: WordCloud from negative reviews

Thus, we need to equalise the importance given to each word from all the reviews. TF-IDF vectorizer handles this problem by considering the frequency of the words across all reviews and in one particular review.

2) **Review Length Analysis**: Analyzing review length in sentiment analysis provides valuable insights into the expressiveness and depth of opinions. Longer reviews tend to contain more nuanced sentiments, offering a richer understanding of the subject's sentiment. Conversely, shorter reviews may convey more straightforward sentiments. By considering review length, we can tailor sentiment analysis models to account for the varying degrees of complexity and depth in reviews, leading to more accurate sentiment predictions. All reviews have lengths distributed around the mode (Fig 3.3.4). There is no significant difference between mode and average. Thus, we can retain maximum information by creating vectors using TF-IDF of the same length as the average review length.
3) **Derived metrics and Correlation**: The NLTK library from Python gives a polarity score to a text on a continuous scale of -1 to +1. We derived the polarity of each review with the NLTK library (which uses a dictionary-based technique to assign polarity scores) and the TextBlob library, which uses a similar approach to NLTK but had different results than NLTK. We also included subjectivity score as one other derived metric from the text. After analysing the correlation matrix, we observed that the polarity and subjectivity scores had no correlations with our target variable, i.e., the sentiment labels. This indicates that the labels are based on some context, and the dictionary-based approach is falling short. Thus, a dedicated model for Amazon product reviews is required to understand the context before classifying the reviews.

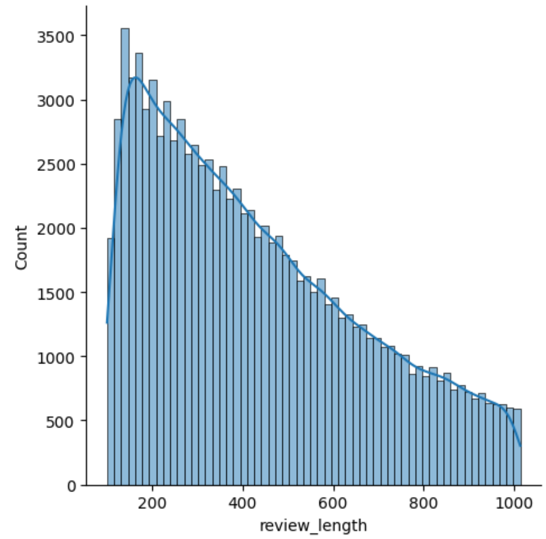

Fig. 3.3.4: Distribution of Review Lengths

4) **Cluster Analysis**: It is a technique to find if the data points (here, reviews in the form of vectors) assimilate in some manner. An iterative process of sampling (5000 random samples), modeling (KNN), and visualising suggested that the data did not naturally form two distinct groups (Fig. 3.3.5). The data points had no boundaries of separation; thus, no clusters formed. A deeper analysis was done by optimising the k-value of the KNN model. It was found that the distinction became clearer if the k-value was set greater than 2 (k=2). Fig 3.3.6 demonstrates the steady increase in the cross-validation score as the value of k increases.

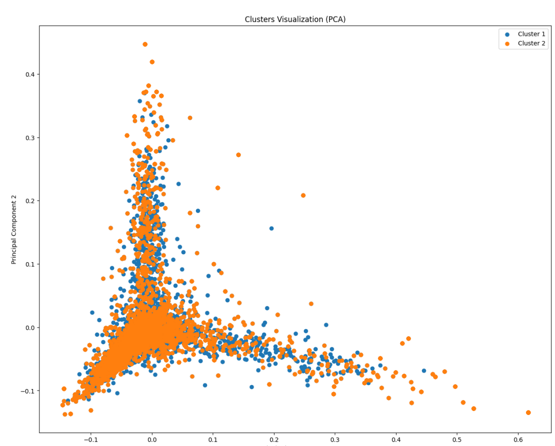

Fig. 3.3.5: 2-Cluster formation of the text reviews

We analysed the possibility of using many models on this dataset, including Logistic Regression, Support Vector Machine for Classification, K-nearest neighbour, Random Forest, Bidirectional LSTM, and many more.

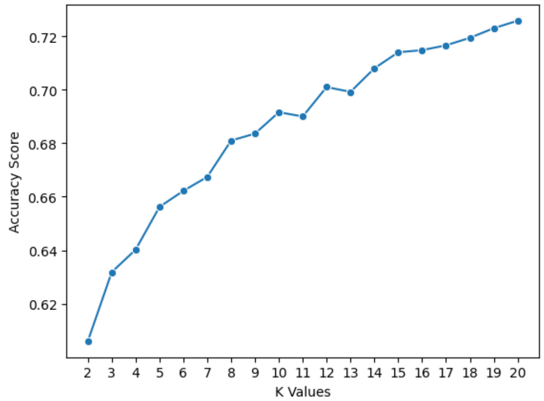

Fig. 3.3.6: Cross-validation score against k-value of the KNN model

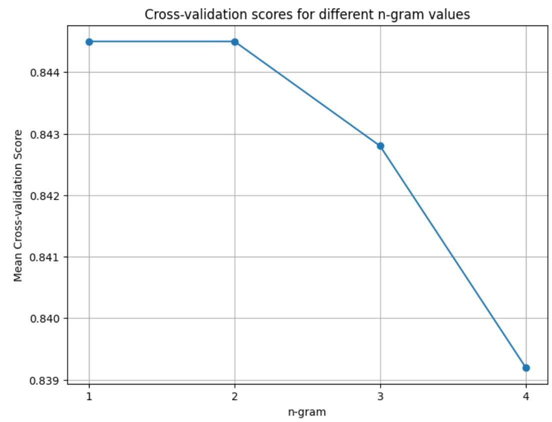

Fig. 3.4.1: Optimising n-gram values

D. Feature Engineering

The crucial part of producing information from raw data is selecting and extracting the most prominent features. Feature Engineering mainly considers the following factors which include-

1) **Term Frequency**: Being able to identify the number of unique words and their significance in a text review is crucial because it helps us distinguish between positive and negative reviews. For instance, if a review contains the word "amazing" repeatedly, we can assume that the review is positive. However, the frequency of occurrence of a word may not always be an accurate indicator of its importance. Some words that appear less frequently may have a more significant impact on the review's sentiment. For example, the word "extravagant" may not be commonly used in positive reviews, but if it appears in a review, it can significantly influence the overall sentiment of the text. To address this problem, we use the TF-IDF (Term Frequency-Inverse Document Frequency) technique for feature extraction. This method considers the frequency of a word in a review and the frequency across multiple reviews.
2) **N-gram Analysis**: It is a common practice in Natural Language Processing (NLP) to consider the text in the form of segments called 'grams.' When considering 2 words in a segment, we call it 2-gram or bigram (here, n=2). Similarly, a trigram consists of segments with 3 words (here n=3). This helps us understand the frequently occurring text patterns in the reviews, such as phrases, proverbs, or slang. To perform experiments on the n-grams, we chose 10,000 training samples at random. The n-grams were analysed from 1 to 4. Every time we ran the experiment, the accuracy declined from 1-gram to 2-gram and decreased further from 2-gram to 3-gram (Fig. 3.4.1). Thus, the optimum value for n-grams was concluded to be 1. This indicates that the model did not find a significant group of words enough to distinguish between a positive and a negative review.
3) **Vectorization**: In order for machine learning models to analyse text data, it needs to be converted into a numerical format. TF-IDF vectorizer is a tool that can convert text data into numerical format, allowing the models to interpret it. When creating these vectors, the number of maximum features to consider is an important factor. In this analysis, a range from 2,000 to 30,000 was considered with a step size of 1,000. It was discovered that a feature space of 2,000 was too limited for the model to accurately predict sentiment, as shown in Figure 3.4.2. The cross-validation score peaked at 20,000 features, indicating that a feature space of 20,000 was optimal for distinguishing between positive and negative reviews.

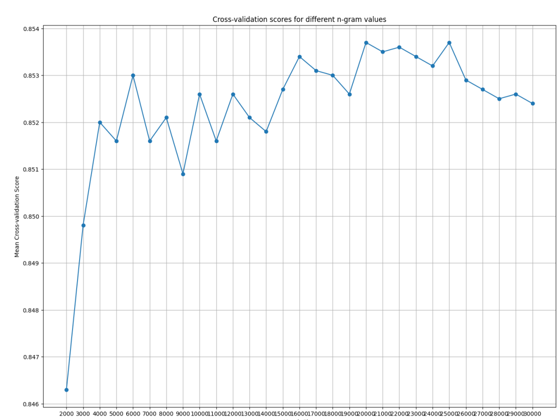

Fig. 3.4.2: Optimising n-gram values

E. Model Training

The study focused on 5 different machine learning algorithms for conducting experiments. These algorithms are Logistic Regression, Support Vector Machine, K-nearest neighbors, Random Forest Classifier, and Bidirectional LSTM. Each model has its own unique classification algorithm and was chosen for its ability to accurately distinguish between two target classes.

Logistic Regression, a probabilistic technique based on a sigmoid function, was selected. Support vector machine, Support Vector Machine was chosen for its ability to identify a decision boundary for distinguishing between positive and negative reviews. The K-Nearest Neighbors Classifier was used to evaluate the natural clustering of reviews into two groups, one consisting of positive reviews and the other of negative reviews. Random Forest Classifiers were selected because they can handle complex text data and avoid overfitting. They offer some interpretability and can classify reviews as positive or negative, even when the sentiment is nuanced.

Bidirectional LSTMs were chosen for their ability to analyze text in both directions. This allows them to capture the context of words and how they influence sentiment throughout a sentence. This is important for understanding sarcasm, negation, or phrases where word order is essential for sentiment. Bidirectional LSTMs can achieve higher accuracy in sentiment analysis tasks compared to simpler models. Each of the 5 models presents its own interpretation of the text data and their results, which are covered in the following section.

IV. Results

A. Logistic Regression

The Logistic Regression model can be primarily used to classify a two-class problem, such as positive and negative reviews. Regression has been widely used as a baseline in the literature [1], [3]. The results have been competitive, but the studies present a comparatively complex model that can capture the complexity of unstructured data in text reviews. Our study defies the research in this domain thus far. Logistic Regression performed the best in various scenarios, for instance, when the machine learning algorithms in this study were trained on a large sample (100,000 training samples and 10,000 testing samples), as well as when trained on the complete dataset altogether (3.6M training samples and 200,000 testing samples). After conducting 5-fold cross-validation and tuning the parameters of the vectorizer, we observed an accuracy of 89.0% and 87.53% on the sample dataset and the complete dataset, respectively, with consistent precision, recall, and F1 score using Logistic Regression. Thus, Logistic Regression outperformed all the models used in this study with minimum computational cost and time. This experiment further suggests that linear models are equally efficient as complex models in understanding the relationship among the features of a text review in the case of the classification of Amazon product reviews.

B. Support Vector Machine (SVM)

An SVM model performed well on two datasets of varying sizes (100,000 train/10,000 test and 3.6M train/200,000 test). The model achieved high accuracy (over 86%) in both scenarios, with balanced precision, recall, and F1 scores. This suggests that the model can effectively handle a variety of data quantities. The model's consistency across sizes demonstrates its robustness and ability to generalize effectively to new data. This is critical for real-world applications in which data sizes vary. Beyond its consistent performance, the model's strengths include handling high-dimensional data (20,000 features in this case) and identifying complex decision boundaries. This, combined with the balanced precision and recall scores, makes SVM a solid choice for classification tasks. However, there is room for improvement. SVM's performance can be improved further by tuning hyperparameters, selecting the optimal kernel function, and considering specific dataset characteristics. Finally, this analysis demonstrates the power and generalisability of SVM as a classification model. Its ability to perform well on datasets of varying sizes makes it a viable candidate for real-world use. With further optimization, its performance is likely to be even better.

C. K-Nearest Neighbours (KNN) Classifier

The K-Nearest Neighbors (KNN) model with n=2 demonstrated varying performance across dataset sizes. On the smaller dataset (Train=100,000, Test=10,000), KNN struggled to produce satisfactory results, with an accuracy of 51.8% and low precision, recall, and F1-score. However, when applied to a larger dataset (Train=3,600,000, Test=200,000), KNN's accuracy increased significantly, reaching 63.98% with improved precision, recall, and F1-score. These findings highlight the sensitivity of KNN performance to the k value and dataset characteristics. Specifically, using k=2 proved suboptimal, particularly on the smaller dataset, indicating the model's difficulty in effectively classifying data points with this configuration. As a result, experimenting with different k values and performing feature engineering may improve KNN's performance. Compared to models such as Logistic Regression, Support Vector Machine, Random Forest, and Bidirectional LSTM, the KNN with n=2 had the lowest accuracy and F1-score across both dataset sizes. This emphasizes the significance of parameter selection and optimization in achieving competitive results. In conclusion, the KNN model with n=2 performed poorly, particularly on the smaller dataset, where accurate classification proved difficult. Further experimentation with different k values and model optimization may be required to improve KNN's effectiveness and align it with the other models in the analysis.

D. Random Forest Classifier

The Random Forest Classifier demonstrated its effectiveness with accuracy rates of 85.66% on the smaller dataset and 85.13% on the larger dataset, particularly in handling text data complexity and preventing overfitting. The architecture of the Random Forest Classifier, which builds several decision trees and combines their predictions, is essential to improving prediction accuracy and noise resistance. This ensemble approach is a dependable option for sentiment analysis since it successfully captures the many patterns and complexity in the sentiment expressed in Amazon product evaluations.

E. Bidirectional Long-Short Term Memory (LSTM)

Long-Short Term Memory (LSTM) is a modified application of Recurrent Neural Networks (RNN), known for their

Dataset size	Vectorizer	Model	Accuracy (%)	Precision	Recall	F1-score
Train=100000, Test=10000	TF-IDF (features=20000, ngram_range=(1, 3)	**Logistic Regression**	**88.9**	**0.89**	**0.89**	**0.89**
		Support Vector Machine	88.88	0.88	0.88	0.88
		KNN (n=2)	51.8	0.58	0.51	0.39
		Random Forest Classifier	85.66	0.86	0.86	0.86
		Bidirectional LSTM	84.3	0.84	0.84	0.84
Train=3600000, Test=200000	TF-IDF (features=20000, ngram_range=(1, 1)	**Logistic Regression**	**87.52**	**0.88**	**0.88**	**0.88**
		Support Vector Machine	86.75	0.87	0.87	0.87
		KNN (n=2)	63.98	0.64	0.64	0.64
		Random Forest Classifier	85.13	0.85	0.85	0.85
		Bidirectional LSTM	84.03	0.84	0.84	0.84

TABLE I: Performance of the Machine Learning Algorithms under different circumstances

ability to understand sequences and temporal data, were a vital part of this study as a method to capture the complex relationships among the features created from the text data. An embedding layer converts input sequences into dense numerical vectors, facilitating semantic understanding. With a vocabulary limit of 20,000 words, the tokeniser translates text data into numerical sequences, which are essential for model input. Post-LSTM layers, batch normalization, and standardized activations enhance training efficiency and stability. Dropout regularisation mitigates overfitting by randomly disconnecting neurons during training. Dense layers may be added after the LSTM layers for classification or regression tasks. For binary classification, the output layer often has one neuron with a sigmoid activation function.

Throughout the training process spanning ten epochs, the model consistently improved accuracy on the training data, culminating in a commendable accuracy of 95.73% by the fourth epoch. However, the model's performance on the validation dataset exhibited marginal fluctuations, with the validation accuracy peaking at 87.72% in the third epoch and marginally decreasing to 86.49% in the fourth epoch. This disparity between the training and validation accuracies in later epochs suggests the onset of potential overfitting. However, it also presents a promising opportunity to explore regularisation techniques or model adjustments further to improve generalisation capabilities.

Furthermore, reducing loss values across epochs for training and validation datasets signifies the model's progressive convergence towards optimal parameter values, aligning more closely with ground truth labels.

V. Conclusion & Future Work

In this work, to completely examine their performance, the models were trained across various training and testing dataset sizes using statistical measurements. This research was able to present the assessment of various machine learning models, such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest Classifier, and Bidirectional LSTM, and has provided valuable insights into their respective outcomes in accurately classifying product reviews as positive or negative sentiments. it becomes evident that the combination of advanced algorithms and feature engineering has set the path for significant progress in understanding and analyzing consumer sentiments. The Logistic Regression model's exceptional performance illustrates the potential of linear models in sentiment analysis, highlighting the importance of model optimization over the inherent complexity of the models. In the future, there will likely be more research and innovation in the field of sentiment analysis in e-commerce, especially on sites like Amazon. Subsequent investigations may focus on incorporating multi-modal data sources, such as videos and photos, to enhance the sentiment analysis framework.

References

[1] Poornima, A. and Priya, K.S., 2020, March. A comparative sentiment analysis of sentence embedding using machine learning techniques. In 2020 6th international conference on advanced computing and communication systems (ICACCS) (pp. 493-496). IEEE.

[2] Tu, G., Wen, J., Liu, C., Jiang, D. and Cambria, E., 2022. Context-and sentiment-aware networks for emotion recognition in conversation. IEEE Transactions on Artificial Intelligence, 3(5), pp.699-708

[3] Kumar, R.P., Elakkiya, R. and Rithani, M., 2023, October. Automated Sentiment Classification of Amazon Product Reviews using LSTM and Bidirectional LSTM. In 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT) (pp. 1-6). IEEE.

[4] Gope, J.C., Tabassum, T., Mabrur, M.M., Yu, K. and Arifuzzaman, M., 2022, February. Sentiment analysis of Amazon product reviews using machine learning and deep learning models. In 2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE) (pp. 1-6). IEEE.

[5] Chen, Y., 2015. Convolutional neural network for sentence classification (Master's thesis, University of Waterloo).

[6] Suhariyanto, S., Sarno, R., Fatihah, C. and Faisal, E., 2020. Aspect based sentiment analysis: a systematic literature review. Journal of Applied Intelligent System, 5(1), pp.8-22.

[7] Areshey, Ali, and Hassan Mathkour. 2023. "Transfer Learning for Sentiment Classification Using Bidirectional Encoder Representations from Transformers (BERT) Model" Sensors 23, no. 11: 5232.

[8] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. and Hussain, A., 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, pp.424-444.

[9] J. Sadhasivam and R. B. Kalivaradhan, "Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm," International journal of mathematical, engineering and management sciences, vol. 4, no. 2, pp. 508–520, 2019.

[10] S. A. Aljuhani and N. Saleh, "A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones," International journal of advanced computer science & applications, vol. 10, no. 6, 2019.

[11] M. K. Shaik Vadla, M. A. Suresh, and V. K. Viswanathan, "Enhancing Product Design through AI-Driven Sentiment Analysis of Amazon Reviews Using BERT," Algorithms, vol. 17, no. 2, pp. 59-, 2024.
[12] S. J and K. U, "Sentiment analysis of amazon user reviews using a hybrid approach," Measurement. Sensors, vol. 27, pp. 100790-, 2023.
[13] T. a/p Sinnasamy and N. N. A. Sjaif, "Sentiment Analysis using Term based Method for Customers' Reviews in Amazon Product," International journal of advanced computer science & applications, vol. 13, no. 7, 2022.
[14] N. Lestari Putri, B. Warsito, and B. Surarso, "Impact of Feature Engineering for Improved Sentiment Analysis in Amazon Product Reviews Using K-Nearest Neighbor," E3S web of conferences, vol. 448, pp. 2030-, 2023.
[15] C. Son, J. Kim, and Y. Kim, "Developing scenario-based technology roadmap in the big data era: an utilisation of fuzzy cognitive map and text mining techniques," Technology analysis & strategic management, vol. 32, no. 3, pp. 272–291, 2020.