International Journal of AI, Big Data, Computational and Management Studies
Noble Scholar Research Group | Volume 2, Issue 2, PP.55-65, 2021
ISSN: 3050-9416 | https://doi.org/10.63282/3050-9416.IJAIBDCMS-V2I2P107

Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models

Ram Mohan Polam[1], Bhavana Kamarthapu[2], Ajay Babu Kakani[3], Sri Krishna Kireeti Nandiraju[4], Sandeep Kumar Chundru[5], Srikanth Reddy Vangala[6].
[1]University of Illinois at Springfield.
[2]Fairleigh Dickinson University.
[3]Wright State University.
[4]University of Illinois at Springfield.
[5]University of Central Missouri, Chundru.
[6]University of Bridgeport.

Abstract: Sentiment analysis is the practice of mining data, opinions, reviews, or statements using natural language processing (NLP) to predict the statement's emotion. Sentiment analysis involves categorizing content into three stages: "positive," "negative," and "neutral." It has an impact on a large number of individuals and companies globally. Sentiment analysis is an essential task in natural language processing, particularly in the e-commerce sector where understanding customer sentiment may significantly influence company decisions. Using the dataset of Amazon product evaluations, it examines the application of Bidirectional Encoder Representations from Transformers (BERT) in deep learning-based sentiment classification in this paper. Standard text normalization techniques, including tokenization, case folding, and stop word removal, are applied to the dataset. BERT is used to set a performance benchmark to compare with two baseline models. A TF-IDF vector served as the feature representation for both methods. However, an approach based on the SentiWordNet lexicon was employed for the lexicon-based method. The models are compared using evaluation metrics such as F1 score, recall, accuracy, and precision. According to the experimental results, the proposed BERT model outperforms traditional methods in terms of F1-score (88.97%), recall (89.67%), accuracy (89.84%), and precision (88.87%). According to the research, transformer-based models outperform large-scale review datasets in sentiment analysis tasks by efficiently learning contextual knowledge and categorization.

Keywords: Sentiment Analysis, Product Reviews, BERT Model, Text Preprocessing, Machine Learning Classification.

1. Introduction

Sentiment analysis, often known as opinion mining, is a branch of NLP that has attracted great attention in recent times and is used for identifying, extracting, and categorizing opinions or sentiments expressed in text [1]. Public perception, customer feedback, and market trends in the Business, Politics, Entertainment and other fields are largely dependent on it. With user-generated content via blogs, social media, and online stores Sentiment analysis, which is rapidly expanding in digital material, is becoming a crucial tool for businesses as it gives them insight into what consumers think of their goods and services [2]. Furthermore, n-grams, negation tags, and POS tags are commonly used to improve it. Negation tags and N-grams are meant to increase algorithm accuracy, whereas POS tags can prevent the ambiguity that comes with several POS of the same word.

In the domain of product reviews, Among the most important uses is sentiment analysis. Millions of customers review experiences and opinions on the e-commerce platforms that are available, like Amazon [3]. The sentiment information concealed in these reviews is often rich and nuanced, and if analyzed correctly, could provide useful information to manufacturers, sellers and other customers [4]. In contrast to numerical ratings, textual reviews cover subjective emotions, the pros and cons of a product, and user satisfaction in greater depth [5]. Although this richness in language makes use of language to express a sentiment very rich, it is also problematic to accurately extract and classify sentiments, particularly when that are talking about sarcasm, negation, or context-dependent phrases.

The magnitude and speed at which review data is accessible online have now categorized sentiment analysis research as a big data project. High-dimensional, unstructured text data requires data analysis methods and system architectures that are extremely flexible and scalable [6]. Standard approaches, which build a bag-of-words or use simplistic n-grams, cannot handle the complexities of such data [7]. Moreover, older systems often neglect natural language in that they only analyze a select aspect of it, for example, verbs or adjectives, without considering the whole context, which results in wrong classifications or over-simplification of sentiments [8]. The big data platforms and tools for sentiment analysis must be chosen most effectively in terms of handling text data in large quantities without losing semantic information.

The evolution in AI over the last couple of years, especially in the development of transformer architecture, and large language models, include BERT, Roberta, and GPT, has seen a paradigm shift in how sentiment classifications are accomplished [9]. These models can naturally learn syntactic structures, context, and the intensity of sentiment in long sequences of texts [10]. As compared with previous techniques, because LLMs take a fixed structure of a sentence into consideration and the precise number and position of each word in the structure, they are able to process the strict meaning of a sentence and characteristics of ambiguity, sarcasm, and multi-aspect sentiment expressions. When it comes to product reviews, LLMs present a great opportunity to classify sentiments based on a model that very much adheres to how people think [11]. This study assesses the efficacy of LLMs in large text using the Amazon Review Dataset sentiment analysis, comparing their performance with traditional ML models and demonstrating their superiority in both accuracy and contextual comprehension

This study uses the Amazon Review Dataset as a benchmark to examine the use of transformer-based LLMs for extensive sentiment analysis of product evaluations. The study illustrates the better accuracy, contextual comprehension, and scalability of LLMs in processing complex, high-volume text data by contrasting these sophisticated models using standard machine learning methods. The results show how these models could improve sentiment analysis and revolutionize real-world e-commerce analytics.

1.1. Motivation and Contribution of the Study

The main driving force for this study is the explosive growth of user-generated content (UGC) in the e-commerce sector, where it manifests as product reviews. Additionally, the customer insights found in these evaluations are a treasure trove that may influence brand perception and purchasing decisions. Sentiment classification of Product reviews helps businesses understand client satisfaction so they can make the right decisions. Such traditional sentiment analysis models come out with limitations in scalability, context understanding and feature extraction. To overcome these challenges, this study leverages advanced NLP techniques and ML models to perform large-scale sentiment analysis, providing an efficient, automated, and scalable solution for understanding customer opinions in the digital marketplace. The main contributions of this work are listed below:

- A robust methodology is introduced as part of the study to preprocess Amazon review data by handling missing values, lemmatizing, tokenizing, and eliminating stop words in order to get data ready for sentiment analysis.
- This uses TF-IDF to convert the textual data into numeric features that are useful to sentiment analysis, capturing key terms that indicate sentiment.
- This study utilizes BERT, an advanced approach for classifying sentiment, thus providing better accuracy in capturing the contextual relationship among the reviews.
- A detailed assessment framework is employed to assess the performance of the sentiment analysis model, providing a full knowledge of its efficacy through the use various metrics, such as F1-score, recall, accuracy, and precision.

1.2. Justification and Novelty

The use of sentiment analysis on the Amazon review dataset is justified because understanding consumer sentiment has become increasingly important in large-scale platforms, and it serves as an important aspect for businesses to understand in order to improve their products and services. In order to arrive at effective sentiment classification, the research preprocesses the data, using tokenization, lemmatization, and also removing stop words, and then applies feature extraction with TF-IDF. Novelty is introduced by the use of the BERT model, a cutting-edge transformer-based architecture, that harnesses its bidirectionality to absorb contextual meaning of reviews for better sentiment understanding. Additionally, the study does not stop at the traditional classification methods, but also uses the metrics of precision, recall, accuracy and F1 score to measure the effectiveness of the model. This combination of advanced techniques and comprehensive evaluation offers both practical and methodological contributions, improving sentiment analysis capabilities for real-world applications.

1.3. Structure of the paper

The structure of this paper is as follows: Section II reviews related work in sentiment analysis and product review classification. Section III details the proposed methodology and experimental setup. Section IV presents the results and analysis, while Section V concludes the study and outlines future directions. Artificial Intelligence (AI) has fundamentally transformed the landscape of cybersecurity, offering advanced capabilities that significantly enhance threat detection and response. By leveraging machine learning and predictive analytics, AI systems can analyze vast amounts of data to identify patterns and anomalies indicative of potential cyber threats. This capability is crucial for managing the complexity and volume of modern cyber threats, providing organizations with a powerful tool to detect and mitigate risks more effectively. Machine learning algorithms, a subset of AI, are particularly valuable in cybersecurity. These algorithms are designed to learn from historical data and identify patterns that may signal malicious activity.

Artificial Intelligence (AI) has fundamentally transformed the landscape of cybersecurity, offering advanced capabilities that significantly enhance threat detection and response. By leveraging machine learning and predictive analytics, AI systems can analyze vast amounts of data to identify patterns and anomalies indicative of potential cyber threats. This capability is crucial for managing the complexity and volume of modern cyber threats, providing organizations with a powerful tool to detect and mitigate risks more effectively. Machine learning algorithms, a subset of AI, are particularly valuable in cybersecurity. These algorithms are designed to learn from historical data and identify patterns that may signal malicious

activity. Artificial Intelligence (AI) has fundamentally transformed the landscape of cybersecurity, offering advanced capabilities that significantly enhance threat detection and response. By leveraging machine learning and predictive analytics, AI systems can analyze vast amounts of data to identify patterns and anomalies indicative of potential cyber threats.

This capability is crucial for managing the complexity and volume of modern cyber threats, providing organizations with a powerful tool to detect and mitigate risks more effectively. Machine learning algorithms, a subset of AI, are particularly valuable in cybersecurity. These algorithms are designed to learn from historical data and identify patterns that may signal malicious activity. Artificial Intelligence (AI) has fundamentally transformed the landscape of cybersecurity, offering advanced capabilities that significantly enhance threat detection and response. By leveraging machine learning and predictive analytics, AI systems can analyze vast amounts of data to identify patterns and anomalies indicative of potential cyber threats. This capability is crucial for managing the complexity and volume of modern cyber threats, providing organizations with a powerful tool to detect and mitigate risks more effectively. Machine learning algorithms, a subset of AI, are particularly valuable in cybersecurity. These algorithms are designed to learn from historical data and identify patterns that may signal malicious activity.

2. Literature Review

The paper's part gives a summary of the corpus of work on sentiment classification in product reviews. Most of the reviewed works focus on classification techniques and approaches for analyzing large-scale textual data. Some of the key studies are as follows:

Long, Zhou and Ou (2019) social media text sentiment analysis, because social multimedia material is growing at an exponential rate. SVM, NB, hybrid models, and other standard ML techniques have had difficulty classifying social media material due to natural language ambiguities and indirect attitudes. This article aims to investigate the sentiment analysis of Chinese text on social media by combining Bidirectional Long-Short Term Memory (BiLSTM) networks with a Multi-head Attention (MHAT) mechanism [12].

Yadav and Bhojane (2019) sentiment analysis for the assessment of Hindi multidomain. Pre-classified words are used in NN Prediction to classify the data. Data classification is accomplished with SentiWordNet (HSWN), which uses NN prediction with pre-classified phrases as labelled data. In conclusion, it reports accuracy in all methods. They have gathered datasets for their manual and random reviews in a variety of domains, including health, business, current events, tourism, films, technology, and products. They comprise mixed Hindi terms with an accuracy of 71.5%, such as "brave," "careful," "mineral," etc [13].

Goularas and Kamis (2019) researchers have come to favor DL approaches and sentiment analysis in Twitter data, which help solve a variety of issues at the same time. CNN, which are particularly effective in image processing, and RNN, which are successfully used in NLP activities, are the two particular varieties of neural networks that are employed. In order to compare different word embedding systems, such as Word2Vec and the global vectors for word representation (GloVe) models sentiment analysis, an evaluation process done under a single testing framework using the same dataset and computing environment is also used to look at the performance, advantages, and disadvantages of the aforementioned techniques [14].

Kant et al. (2018) A useful application of NLP to real-world data is sentiment categorization. They show that fine-tuning in conjunction with large-scale unsupervised language modelling provides a workable solution to this problem on challenging datasets, such as those with domain-specific context and label class imbalance. Unsupervised language modelling and fine-tuning is a simple methodology that yields excellent sentiment classification results in the actual world. By using 40GB of text (Amazon reviews) to train an attention-based Transformer network and fine-tuning on the training set, their model attains an accuracy of 0.69 [15].

Ejaz et al. (2017) process of gathering people's thoughts, feelings, and experiences via reviews, blogs, and other sources is known as product review. Consists of three well-known ML algorithms: DT learner with document vector, RF learner with n-gram, and RF learner with word vector. It is known as a dictionary-based method for opinion mining with n-grams. Amazon's Product Review information has been used to forecast both positive and negative opinions. The experimental result demonstrates that the lexicon-based strategy works better than conventional ML techniques, with a 79% model accuracy [16].

Shivaprasad and Shetty (2017) Sentiment analysis is A kind of research that understands and draws conclusions from reviews. Text analytics, polarity classification, computational linguistics, and natural language processing are all used in the analytical process. Every algorithm has several applications. The taxonomy of several sentiment analysis techniques is presented in this work. Additionally, this study demonstrates that SVM outperforms NB and maximal entropy approaches with an accuracy of 80.45% [17]. Table I provides a comparative summary of the reviewed literature, outlining the main findings, identified limitations, and suggested future research directions for each study.

Table 1: Comparative Analysis of Large Language Model Approaches for Sentiment Classification in Product Reviews

Author	Dataset	Methodology	Findings	Limitations	Future Work
Long, Zhou, and Ou (2019)	Chinese social media text	BiLSTM with Multi-head Attention (MHAT)	Improved handling of ambiguous and indirect sentiments in Chinese text	Limited to Chinese language; generalizability to other languages not tested	Extend to multilingual sentiment analysis and domain-specific datasets
Yadav and Bhojane (2019)	Manually collected Hindi reviews (multi-domain)	SentiWordNet + Neural Networks using pre-classified sentences	Achieved 71.5% accuracy; handled mixed Hindi-English sentiment terms	Language-specific and domain-specific dataset; limited scalability	Expand to larger, more diverse datasets; improve accuracy using deep learning
Goularas and Kamis (2019)	Twitter data	CNN, RNN with Word2Vec and Glove embeddings	Compared embeddings; demonstrated effectiveness of DL in NLP sentiment tasks	No domain adaptation; performance dependent on embedding selection	Integrate hybrid embeddings and real-time sentiment analysis
Kant et al. (2018)	Amazon Reviews (40GB)	Transformer-based unsupervised language modeling + fine-tuning	Achieved 0.69 accuracy; demonstrated effectiveness of large-scale pretrained models	Label imbalance and domain-specific nuances not fully addressed	Enhance label handling; apply model to other review platforms
Ejaz et al. (2017)	Amazon Product Reviews	Lexicon-based + ML (Random Forest, Decision Tree, n-grams)	Lexicon-based approach outperformed other ML models with 79% accuracy	Limited to binary classification; no context-aware modeling	Explore context-aware DL models and hybrid lexicon-ML approaches
Shivaprasad and Shetty (2017)	Not explicitly mentioned	SVM, Naïve Bayes, Max Entropy for sentiment polarity classification	SVM achieved highest accuracy of 80.45% compared to other classical models	No deep learning models considered; limited exploration of embedding techniques	Extend to DL methods; evaluate on varied datasets and more complex tasks

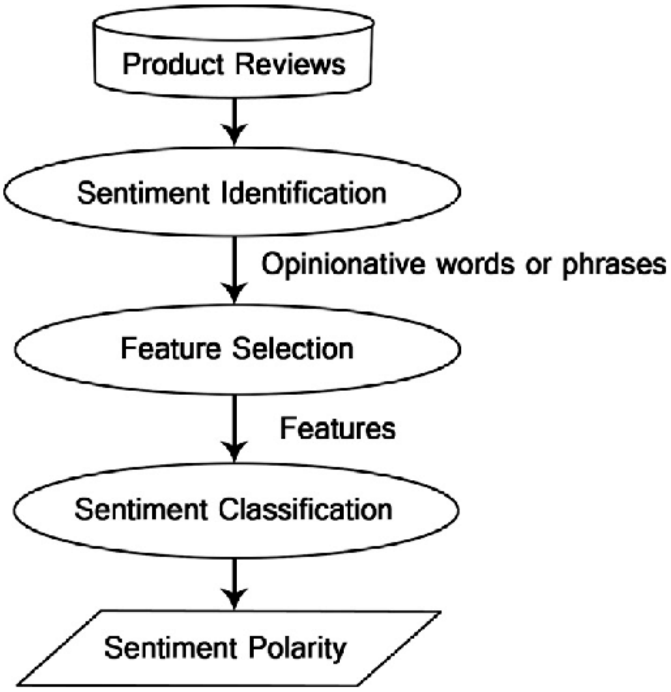

Figure 1: Flowchart of Sentiment Analysis for Amazon Review Dataset

3. Methodology

The first step in the sentiment analysis approach for the Amazon review dataset is data pretreatment, which involves handling missing and duplicate values, converting text to lowercase, and applying further text cleaning steps such as stop word removal, tokenization, and lemmatization to ensure consistency and prepare the data for analysis illustrate ion Figure 1. The cleaned text is then converted into numerical attributes using TF-IDF, which captures the keywords that express mood. The BERT model is utilized for classification, and the dataset is split between training subsets. This model is trained using the training data and predicts sentiment labels (positive, negative, or neutral) based on the textual properties. The results are summarized for interpretation, and metrics like accuracy, precision, recall, and F1-score are used to evaluate the model's performance and give a comprehensive assessment of its effectiveness. The flow diagram describing the suggested technique is shown in Figure 1.

The overall steps of the flowchart for sentiment classification on product review are provide in below:

3.1. Data Collection

A publicly accessible dataset of Amazon reviews was used for the sentiment classification research, which was obtained via the Kaggle platform. This dataset consists of user reviews for five different product categories: watches, cameras, groceries, furniture, and mobile gadgets. Despite having over 100 million reviews in its entirety, a balanced selection of 80,000 reviews from each category was chosen, yielding 400,000 recordings in total. Numerous specific attributes are included in each entry, such as the marketplace, the number of helpful votes, the total number of votes, the verified buy status, the company ID, the customer ID, the review ID, the product ID, the parent product, the product title, the category, and the star rating. These general characteristics provide a solid basis for carrying out in-depth sentiment analysis and assessing the model's effectiveness.

3.2. Data Analysis and Visualization

Data analysis and visualization are key in sentiment classification for product reviews, helping to identify sentiment trends. EDA involves examining sentiment distribution using pie charts to show proportions of positive, neutral, and negative reviews, and bar charts to compare sentiment counts across different product categories. These visualizations assist in understanding the overall sentiment and guide the selection of appropriate ML models.

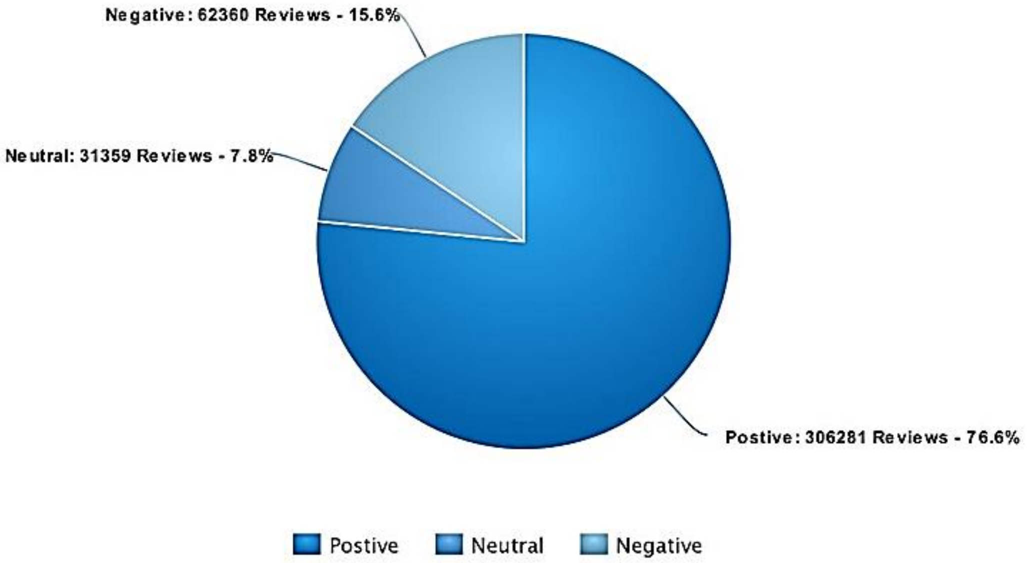

Figure 2: Sentiment Classification of Star Rating Distribution of Reviews

Figure 2 illustrates the sentiment distribution of product reviews. The majority of reviews are positive (76.6%, 306281 reviews), followed by negative (15.6%, 62360 reviews), and then neutral reviews (7.8%, 31359 reviews). This distribution highlights a strong positive sentiment towards the reviewed products.

The bar chart showing the distribution of review counts for various product categories is shown in Figure 3: Watches, groceries, cameras, mobile electronics, and furniture. Mobile Electronics exhibits the highest number of reviews, significantly exceeding the counts for other categories. This distribution of review data across product categories forms the basis for text analysis aimed at sentiment classification, where the varying volumes of feedback can influence model training and evaluation for understanding customer opinions within each category.

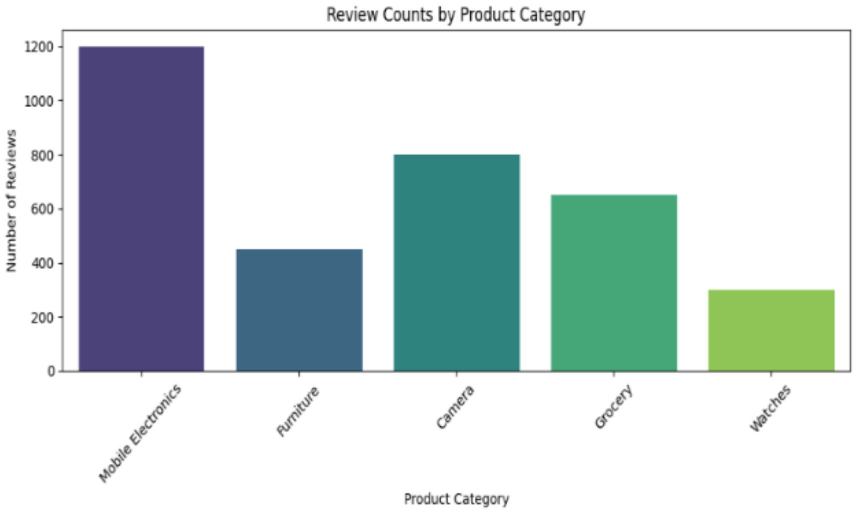

Figure 3: Bar Chart of Review Count by Product Category

3.3. Data Preprocessing

The quality of the data used to generate proposed models is directly impacted by data preprocessing, making it an essential step in the multimodal sensing workflow. In the data pretreatment procedures, duplicate and missing values were handled. The text was normalised by eliminating stop word tags and changing it to lowercase in order to cut down on noise. Tokenization, lemmatization and Feature extraction was done transforming the text into numerical representations with TF-IDF. The following pre-processing steps are listed below:

- **Handling Duplicate Value:** Handling duplicate values involves identifying and removing identical entries from the dataset to guard against bias in analysis and guarantee data integrity. This step ensures that each record is unique, improving The quality of data utilized in modelling.
- **Handling Missing Value:** In order to address the issue of missing values in the dataset, they will use Python's fillna() method to fill in the null values in the "review body" and "star rating" features. Regarding the "star rating" function, the Interpolate method is applied to handle missing values more effectively.
- **Lowercase Conversion:** All review words are converted to lowercase. For example, "great" and "amazing" are substituted by "great" and "amazing." Through case-insensitive word treatment, lowercasing decreases the dimensionality of the data and helps standardize the language.

3.4. Removal Stop Word

In text mining, stop words are phrase elements that don't apply to every field. Everybody has removed all stop words, punctuation, and HTML elements found in their corpus of reviews. This preparation stage enhances computing performance and lowers data noise.

3.5. Tokenization

The method of tokenization involves dividing a text sequence can be divided into distinct components called tokens, which might comprise individual words, phrases, or even entire sentences [18]. These tokens can then be used as inputs for a number of additional processes, such as parsing and text mining. It assists models in concentrating on the significance of discrete elements instead of digesting the text as a whole.

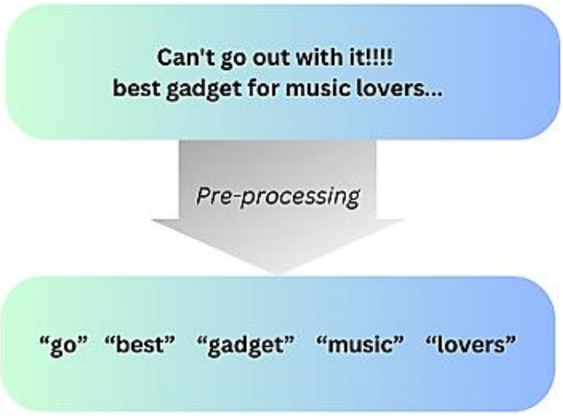
Can't go out with it!!!!
best gadget for music lovers...
Pre-processing
"go" "best" "gadget" "music" "lovers"

Figure 4: Tokenization of Sentiment Word

Figure 4, look of the text before to and following preprocessing Tokenizers also give papers their dependability.

3.6. Lemmatization

A lexicon normalization technique called lemmatization breaks words down to their most basic form according to their context and meaning. Morphological analysis is used to highlight word construction, while vocabulary is used to show dictionary value. This keeps the normalised sentence's context and meaning intact. To make the text simpler, non-word elements like symbols, numerals, and punctuation are eliminated in the last stage. In order to provide low-dimensional space features for ML models, Lexicon normalization transforms high-dimensional information. Normalizing words that appear more than once is necessary to prevent textual noise.

3.7. Feature Extraction with TF-IDF

An essential step in the process, feature extraction involves numerically representing unprocessed text. Many ML techniques use TF-IDF as their foundational component. When looking for and determining Two roughly related metrics for determining a word's relevance to a document are TF-IDF, or TF-IDF [19]. The most often used term weighting technique, the TF-IDF algorithm, is used to evaluate a word's importance in a text quantitatively.

A word's relevance for the document or documents is determined by its score frequency using TF-IDF, which is based on the word's frequency. Equations (1-3) include the formulas needed to determine the TF-IDF score step-by-step:

$$tf(w,d) = \log(1 + fw,d)$$

$$idf(w,D) = \log(Nf(w,d))$$

$$tfidf(w,d,D) = tfw,d * idf(w,D)$$

The TF-IDF (w, d, D) depicts the TF-IDF score for term t in document d with respect to document collection D.

3.8. Data Splitting

In order to conduct the analysis of sentiment categorization in product evaluations, the dataset underwent preprocessing and was separated into testing and training sets. Seventy percent of the data in the training set was used to train a model. The remaining 30% was set aside just to assess the model's performance.

3.9. Classification with BERT Model

An architecture for DL called BERT can be employed for downstream NLP jobs. The design is based on a transformer-derived layered encoder layer. Both pre-training and fine-tuning are essential parts of BERT. Pretraining involves training BERT in an unlabeled big corpus employing two tasks that are not under supervision: NSP and MLM for predicting the following statement. This results in a pre-trained model. To begin fine-tuning, it loads the model with the pretrained parameters. Then, for tasks like classification, it uses labelled data to fine-tune each parameter individually.

The input tokens are represented by H = 768-shaped vectors in the pre-trained model's black box. One or more phrases can make up a sequence, they can be joined by a [SEP] token or started with a [CLS] token [20]. An output layer is used to model and fine-tune all parameters for the classification task. In actuality, the entire sequence should only be represented by the output from the [CLS] token. Therefore, Figure 5 displays the whole fine-tuning BERT architecture for the classification job. Equation (4) shows the likelihood of label c, which is predicted by adding a simple SoftMax classifier on top of the model. where the task-specific parameter matrix is denoted by W. A log-probability of correctly labelling by maximizing the parameters from both BERT and W together.

$$p(c|h) = softmax(Wh)$$

A stand for the number of attention heads, and H for the number of hidden embedding sizes. A smaller model architecture may be employed with less compute resources and requires fewer parameters to train.

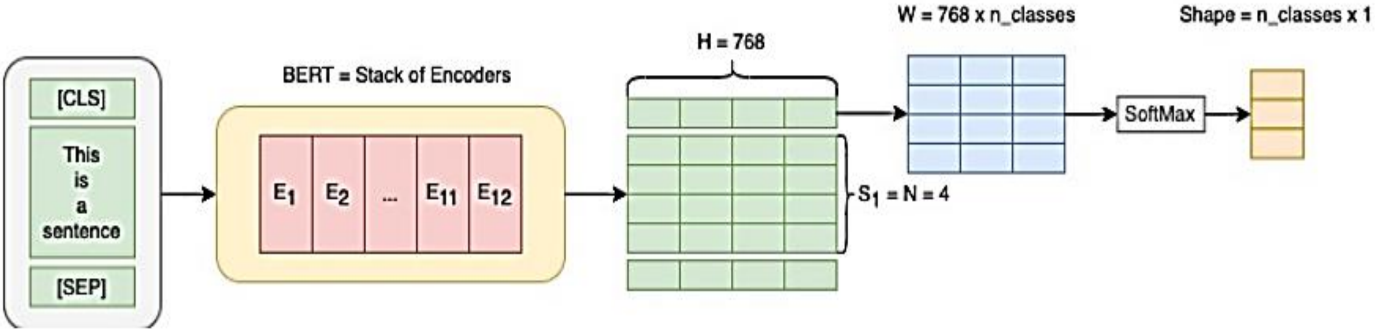

Figure 5: Fine tuning BERT Architecture for Classification Task

The ideal hyperparameter settings vary depending on the job; for example, setting the maximum epoch number to 4 and saving the best model for testing on the validation set shown in Figure 5.

3.10. Performance Metrics

Evaluating the results of ML algorithms in sentiment classification on product reviews, well-known evaluation measures are employed to evaluate each classifier's effectiveness. The selection of assessment measures is predicated on their pertinence to the sentiment analysis task. Commonly applied evaluation metrics for sentiment classification include F1-score, confusion matrix, recall, accuracy, and precision. These metrics, which provide values like TP, TN, FP, and FN, are used to assess each classifier's performance. Here, a positive class denotes a feeling that is positive, while a negative class denotes a sentiment that is these negative words have the following definitions.

- **TP:** (True Positive) indicates the quantity of accurately categorized data.
- **TN:** (True Negative) is the quantity of inaccurately categorized data.
- **FP:** (False Positive) shows the quantity of accurate data that was incorrectly categorized.
- **FN:** (False Negative) reflects the quantity of inaccurate data that is categorized as accurate.

Accuracy: The accuracy of the classifier shows how often the correct prediction will be made. The ratio of correct forecasts to all predictions serves as the accuracy metric. It is expressed in Equation (5):

$$Accuracy = \frac{\text{TN + TP}}{\text{TP + TN + FP + FN}} \qquad (5)$$

Precision: A measure of accuracy called precision indicates how many of all positive forecasts turn out to be accurate. It is calculated as the overall number of anticipated positives as a percentage of all categorized positives. An effective model should have a high degree of accuracy. According to Equation (6), precision equals:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (6)$$

Recall: The number of accurately anticipated courses with a good outcome or proportion of accurately identified and positively categorized groups to all positively classed classes is known as the recall. The recall rate of a good model should be high recall is defined in Equation (7):

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (7)$$

F1-Score: A high F1-score is indicative of strong recall and accuracy as the score includes data on these two characteristics. It is defined as follows in Equation (8):

$$\text{F1} = \frac{2*(\text{precision}*\text{recall})}{\text{precision}+\text{recall}} \qquad (8)$$

These matrices are also used for comparison for model performance.

4. Result And Discussion

A high-performance computing setup was used for experimental analysis for effectively managing the Amazon review dataset and measurement of model performance. It utilized an NVIDIA RTX 4090 GPU with 24 GB of VRAM, an Intel Core i9-13900K CPU (1.0 GHz), and 64 GB of DDR5. Windows 11 Pro was installed. The study used BERT as the classification model on the dataset of Amazon product reviews for sentiment analysis. Accuracy, precision, recall, and F1-score are popular classification metrics that were used to evaluate the model's performance. These metrics provide information about the model's ability to categorize sentiment categories. Analysis reveals that BERT is indeed capable of solving the problem of text classification in the context of the product review domain.

Table 2: Proposed BERT model Performance on Amazon review dataset

Matrix	BERT
Accuracy	89.84%
precision	88.87%
Recall	89.67%
F1-score	88.97%

Table II below shows the results of the performance evaluation of the proposed BERT model using the Amazon review dataset. BERT has extraordinarily high-performance measures, including accuracy of 89.84%, precision of 88.87%, recall of 89.67%, and F1 score of 88.97%. In this instance, these metrics demonstrate the model's capacity to classify with the highest accuracy while minimizing false positives and false negatives. These findings show that the BERT model works consistently and dependably for sentiment categorization on product reviews across all assessment measures.

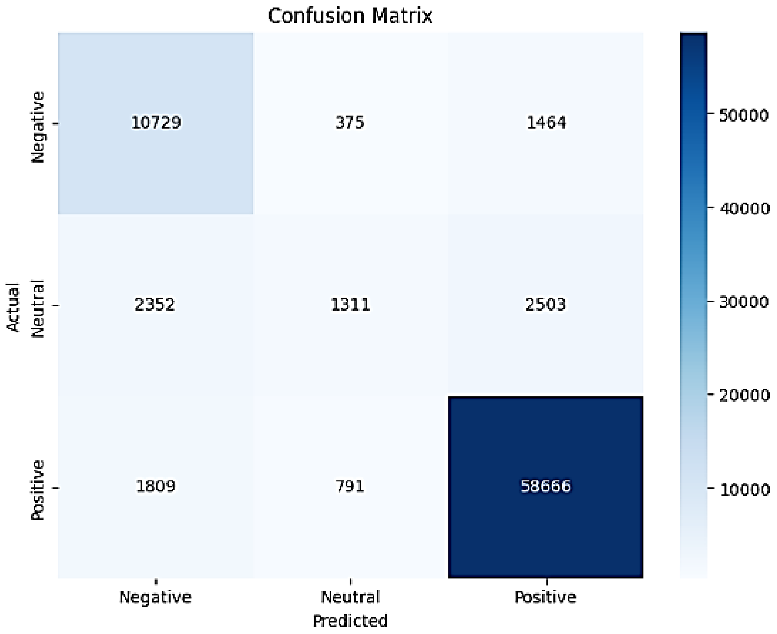

Figure 6: Confusion Matrix of BERT Model

Figure 6 represents the Confusion matrix for sentiment classification using BERT. The model correctly classified 58,666 positive, 10,729 negative, and 1,311 neutral reviews. Misclassifications occurred mainly in the neutral class, often confused with positive and negative sentiments. This indicates strong performance on positive and negative classes, with room for improvement in neutral sentiment detection.

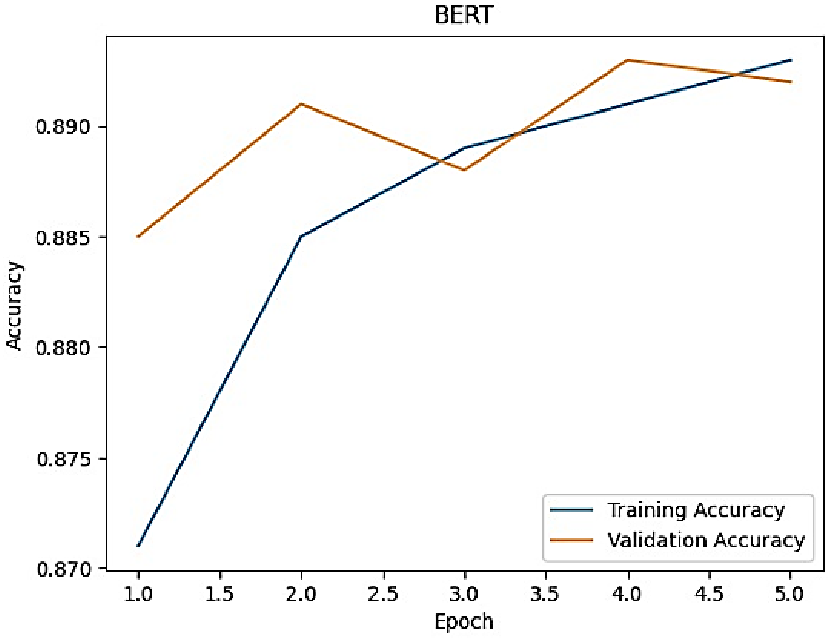

Figure 7: Accuracy Curve of BERT Model

Figure 7 shows the BERT model's training and validation accuracy curves throughout five epochs. The graphic shows how accuracy has changed over time on both training and validation datasets. Initially, both accuracies increase with each epoch, indicating model learning. The validation accuracy slightly surpasses training accuracy until the third epoch, after which Training accuracy is continuously increasing. The precision of the validation peaks at the fourth epoch and shows a minor decline thereafter, potentially signaling the onset of overfitting. These findings show how well the BERT model captures contextual patterns for product review sentiment categorization tasks.

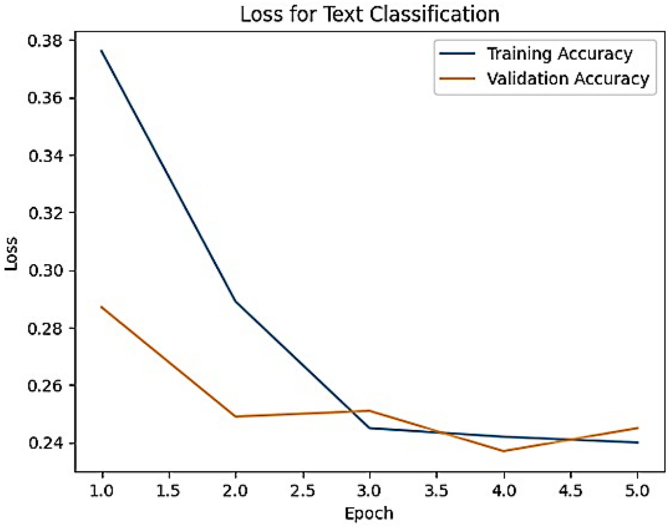

Figure 8: Loss curve of BERT Model

The loss in training and validation curves for text categorization using BERT are shown in Figure 8. At first, there is a significant decrease in training and validation loss, suggesting early learning is successful. By the third epoch, the losses converge closely, showing improved model generalization. From epochs 3 to 5, the losses continue to decline slightly, with a minor increase in validation loss at epoch 5, which could suggest the beginning of overfitting. Overall, the plot demonstrates a stable training process with a good balance between training and validation performance.

Table 3: Comparative analysis for Sentiment classification on product review Existing models' performance

Matrix	**BERT**	**Logistic Regression**[21]	**SentiWordNet Lexicon**[22]
Accuracy	89.84%	81.90%	80%
precision	88.87%	87%	88%
Recall	89.67%	88%	88%
F1-score	88.97%	88%	88%

The comparative results of sentiment classification performance are summarized in the Table III, highlighting the effectiveness of the proposed BERT model against two baseline models: LR and the SentiWordNet Lexicon approach. The BERT model outperformed the baseline models for every assessment measure. It gave the highest accuracy rate of 89.84%, which exceeded that of LR (81.9%) and SentiWordNet (80%). BERT scored 88.87% in terms of precision, a tad better than LR 87% and SentiWordNet 88%. Compared to baselines, BERT achieved 89.67% for recall, which is higher than both of them (88%). Finally, while F1 is not the most important metric, the BERT model had an F1 score of 88.97%, beating the 88% final F1 scores of the baseline methods. The results indicate that this BERT-based approach is superior in terms of its performance and robustness over sentiment classification tasks.

As a result, the proposed BERT model has advantages over other models for sentiment classification, such as bidirectional encoding that enables the model to comprehend the contextual meaning and thus better interpret the sentiment. It shows considerable improvement regarding F1 score, memory, accuracy, and accuracy compared to currently used conventional models, indicating its possible robustness. Additionally, BERT lessens the requirement for intensive feature engineering, as it automatically extracts meaningful features from raw text. It can adapt well to various and subtle review content due to its strong generalization capability, and its scalability supports its application to handle big data in real-world scenarios.

5. Conclusion And Future Work

Product review sentiment classification deals with categorizing the expressed opinion of a customer as positive, negative or neutral. Companies might more accurately gauge customer happiness and improve their products and services. by utilizing a strong foundation for sentiment categorization of Amazon product reviews, utilizing cutting-edge ML models. Extensive experiments on the proposed BERT-based sentiment analysis model show its superiority compared to traditional ML and lexicon-based approaches in terms of capturing nuanced sentiment in large-scale product reviews. BERT shows good results with its understanding of context and sentiment within the text, achieving 89.84% accuracy, 88.87% precision, 89.67% recall, and an F1 score of 88.97%.

BERT yields better performance than baseline models, such as LR with TF-IDF and the Sent WordNet lexicon, because BERT uses deep contextualization of words, thus serving as a useful tool in sentiment classification for e-commerce usage. In light of these findings, it appears that BERT can provide important benefits to automated review analysis and decision-making systems, and it may make for good business assets in helping businesses glean more insight into what customers like and do not like about the businesses. Moving forward, it could try tuning the model for better detection of neutral sentiment and also try out other models like Roberta or GPT-based architectures to improve the performance. Moreover, this approach can be expanded on other types of text data, such as reviews from other platforms or customer opinions on services. In addition, the use of domain-specific lexicons and hybrid models could increase the precision of the sentiment classification task.

References

[1] D. M. E.-D. M. Hussein, “A Survey on Sentiment Analysis Challenges,” J. King Saud Univ. - Eng. Sci., vol. 30, no. 4, pp. 330–338, Oct. 2018, doi: 10.1016/j.jksues.2016.04.002.

[2] H. Zou, X. Tang, B. Xie, and B. Liu, “Sentiment Classification Using Machine Learning Techniques with Syntax Features,” in 2015 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, Dec. 2015, pp. 175–179. doi: 10.1109/CSCI.2015.44.

[3] Q. Pan, X. Zheng, and G. Chen, “A Mix-model based Deep Learning for Text Sentiment Analysis,” in 2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB), IEEE, Nov. 2018, pp. 1–6. doi: 10.1109/ICCBB.2018.8756420.

[4] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[5] N. Sahoo, C. Dellarocas, and S. Srinivasan, “The impact of online product reviews on product returns,” Inf. Syst. Res., 2018, doi: 10.1287/isre.2017.0736.

[6] X. Fang and J. Zhan, “Sentiment analysis using product review data,” J. Big Data, 2015, doi: 10.1186/s40537-015-0015-2.

[7] A. H. Anju, “Extreme Gradient Boosting using Squared Logistics Loss function,” Int. J. Sci. Dev. Res., vol. 2, no. 8, pp. 54–61, 2017.

[8] L. Muliawaty, K. Alamsyah, U. Salamah, and D. S. Maylawati, “The concept of big data in bureaucratic service using sentiment analysis,” Int. J. Sociotechnology Knowl. Dev., 2019, doi: 10.4018/IJSKD.2019070101.

[9] V. Kolluri, “A Comprehensive Analysis on Explainable and Ethical Machine: Demystifying Advances in Artificial Intelligence,” TIJER - Int. Res. Journals, vol. 2, no. 7, pp. 2349–9249, 2015.

[10] S. Karimi and F. S. Shahrabadi, “Sentiment Analysis Using BERT (Pre-Training Language Representations) and Deep Learning on Persian Texts,” Technol. Deep Learn., 2019.

[11] A. S. Rathor, A. Agarwal, and P. Dimri, “Comparative Study of Machine Learning Approaches for Amazon Reviews,” in Procedia Computer Science, 2018. doi: 10.1016/j.procs.2018.05.119.

[12] F. Long, K. Zhou, and W. Ou, “Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention,” IEEE Access, vol. 7, pp. 141960–141969, 2019, doi: 10.1109/ACCESS.2019.2942614.

[13] M. Yadav and V. Bhojane, “Semi-Supervised Mix-Hindi Sentiment Analysis using Neural Network,” in 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, Jan. 2019, pp. 309–314. doi: 10.1109/CONFLUENCE.2019.8776943.

[14] D. Goularas and S. Kamis, “Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data,” in 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), 2019, pp. 12–17. doi: 10.1109/Deep-ML.2019.00011.

[15] N. Kant, R. Puri, N. Yakovenko, and B. Catanzaro, “Practical Text Classification With Large Pre-Trained Language Models,” 2018.

[16] A. Ejaz, Z. Turabee, M. Rahim, and S. Khoja, “Opinion mining approaches on Amazon product reviews: A comparative study,” in 2017 International Conference on Information and Communication Technologies, ICICT 2017, 2017. doi: 10.1109/ICICT.2017.8320185.

[17] T. K. Shivaprasad and J. Shetty, “Sentiment analysis of product reviews: A review,” in 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017, pp. 298–301. doi: 10.1109/ICICCT.2017.7975207.

[18] V. S and J. R, “Text Mining: open Source Tokenization Tools – An Analysis,” Adv. Comput. Intell. An Int. J., vol. 3, no. 1, pp. 37–47, Jan. 2016, doi: 10.5121/acii.2016.3104.

[19] S. Pei, L. Wang, T. Shen, and Z. Ning, “DA-BERT: Enhancing part-of-speech tagging of aspect sentiment analysis using BERT,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019. doi: 10.1007/978-3-030-29611-7_7.

[20] Z. Miftahutdinov, I. Alimova, and E. Tutubalina, “KFU NLP Team at SMM4H 2019 Tasks: Want to Extract Adverse Drugs Reactions from Tweets? BERT to The Rescue,” 2019. doi: 10.18653/v1/w19-3207.

[21] T. U. Haque, N. N. Saber, and F. M. Shah, “Sentiment analysis on large scale Amazon product reviews,” in 2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018, 2018. doi: 10.1109/ICIRD.2018.8376299.

[22] A. Veluchamy, H. Nguyen, M. L. Diop, and R. Iqbal, “Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches,” SMU Data Sci. Rev., vol. 1, no. 4, pp. 1–22, 2018.

[23] Kalla, D., & Samiuddin, V. (2020). Chatbot for medical treatment using NLTK Lib. IOSR J. Comput. Eng, 22, 12.

[24] Kuraku, S., & Kalla, D. (2020). Emotet malware a banking credentials stealer. Iosr J. Comput. Eng, 22, 31-41.