analytics

MDPI

Article

Customer Sentiments in Product Reviews: A Comparative Study with GooglePaLM

Olamilekan Shobayo [1,*], Swethika Sasikumar [1], Sandhya Makkar [1] and Obinna Okoyeigbo [2]

[1] Department of Computing, Sheffield Hallam University, Sheffield S1 2NU, UK; swethika.sasikumar@student.shu.ac.uk (S.S.); s.makkar@shu.ac.uk (S.M.)
[2] Department of Engineering, Edge Hill University, Ormskirk L39 4QP, UK; obinna.okoyeigbo@edgehill.ac.uk
* Correspondence: o.shobayo@shu.ac.uk

Abstract: In this work, we evaluated the efficacy of Google's Pathways Language Model (GooglePaLM) in analyzing sentiments expressed in product reviews. Although conventional Natural Language Processing (NLP) techniques such as the rule-based Valence Aware Dictionary for Sentiment Reasoning (VADER) and the long sequence Bidirectional Encoder Representations from Transformers (BERT) model are effective, they frequently encounter difficulties when dealing with intricate linguistic features like sarcasm and contextual nuances commonly found in customer feedback. We performed a sentiment analysis on Amazon's fashion review datasets using the VADER, BERT, and GooglePaLM models, respectively, and compared the results based on evaluation metrics such as precision, recall, accuracy correct positive prediction, and correct negative prediction. We used the default values of the VADER and BERT models and slightly finetuned GooglePaLM with a Temperature of 0.0 and an N-value of 1. We observed that GooglePaLM performed better with correct positive and negative prediction values of 0.91 and 0.93, respectively, followed by BERT and VADER. We concluded that large language models surpass traditional rule-based systems for natural language processing tasks.

Keywords: sentiment analysis; natural language processing; GooglePaLM; product reviews; BERT; VADER; emotion detection; large language models

Citation: Shobayo, O.; Sasikumar, S.; Makkar, S.; Okoyeigbo, O. Customer Sentiments in Product Reviews: A Comparative Study with GooglePaLM. *Analytics* **2024**, *3*, 241–254. https://doi.org/10.3390/analytics3020014

Received: 29 April 2024
Revised: 12 June 2024
Accepted: 14 June 2024
Published: 18 June 2024

1. Introduction

In the digital era, online product reviews have become an indispensable tool for consumers making purchasing decisions. This is driven by the prevalence of e-commerce platforms like Amazon and Alibaba. These reviews, rich in user sentiment and opinions, significantly influence purchasing behavior through Word-of-Mouth (WOM) communication. They not only assess product quality but also sway decision-making, particularly for new items [1,2]. The burgeoning significance of product reviews across diverse industries underscores the critical role they play in shaping consumer perceptions and choices. For instance, positive online reviews can provide a safety net for online customers, acting as "social proof" for the willingness to purchase an item on any e-commerce platform. It fosters decision-making, authenticity, and provides unbiased perspectives from users' feedback on specific issues a prospective customer might have [3]. Online reviews have facilitated the purchase decision of any product or service for prospective customers, and this has been made more accessible by e-commerce platforms. The feedback from a review translates into positive improvement in products and services provided by organizations which, in turn, has led to high profit margins for organizations and a high propensity for customer purchases [4].

Natural Language Processing (NLP) has evolved significantly, starting from its inception in the 1950s with early endeavors like the Georgetown-IBM experiment, which laid the groundwork for machine translation [5]. Subsequent decades saw advancements in chatbots such as ELIZA and PARRY, incorporating statistical techniques like Hidden Markov

Models in the late 1980s [6]. The 2000s ushered in machine learning techniques, including neural networks and deep learning, revolutionizing NLP [7]. NLP has emerged as a pivotal tool for handling the vast volume of online reviews. NLP, a subset of artificial intelligence, empowers computers to understand, analyze, and generate human language [8]. Sentiment analysis, a crucial aspect of NLP, involves identifying emotions in textual content such as product reviews. Traditional methods relied on lexicons and rule-based systems such as Valence Aware Dictionary for Sentiment Reasoning (VADER) and TextBlobs, evolving into statistical methods like Naive Bayes and Support Vector Machines for classification [9]. However, challenges persist in the application of these methods, especially in the manual reconfiguration of text vectors when the rule of embedding changes, causing greater financial strain [10]. This has led to the development of deep learning techniques such as recurrent neural network (RNN), Long short-term memory (LSTM), Bidirectional long short-term memory (BiLSTM), and long sequence transformer models such as Bidirectional Encoder Representations from Transformers (BERT) and its variants for the automation of NLP tasks, such as sentiment analysis [11]. However, challenges such as recognizing sarcasm and cultural variations persist [12]. When viewing sentiments at the document level, these models have been known to struggle in analyzing sentiments correctly [13]. The development of newer large language (LLMs) context-aware sentiment models, which use transformers with self-attention mechanisms, has been used to address this issue by mapping sentiments to specific aspects mentioned in the online customer review implicitly. An example of such models developed is GPT and its variants, GooglePaLM and T5 [14]. The creation of Google's Pathways Language Model (GooglePaLM) marks a transformative phase in NLP. This model, which uses the self-attention transformer architecture, along with extensive training, has demonstrated exceptional proficiency in understanding intricate linguistic patterns [15].

Self-Attention Transformers

BERT was among the first set of models for NLP tasks based on transformer architecture. These models have been shown to provide excellent results in sentiment analysis when compared to neural network models like RNN and LSTM which use long-range dependencies [16]. Using pre-trained Large Language Models (LLMs) such as transformers for context-based sentiment analysis tasks is becoming increasingly popular. Researchers have leveraged large-scale pretraining for a substantial amount of textual data, made possible by big organizations such as Google's (Bard, PaLM) and Meta's Large Language Model Meta AI (LLaMA), providing storage and processing capabilities [17]. Transformers make use of self-attention, a mechanism that allows different words/vectors to compare with the subsequent words to develop dependencies within textual data up to 512 in length. (this can be fine-tuned to be more). They also process information using parallel servers so that words can be processed in parallel and stored, which improves efficiency when compared to neural network models that process data sequentially [18]. The self-attention transformer architecture is shown in Figure 1.

The self-attention mechanism of the transformer can be represented by the following:

$$\boldsymbol{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{Softmax}\left(\frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{d_k}}}\right).\boldsymbol{V} \quad (1)$$

The vectors Q, K, and V represent the Query, Key, and Value matrices of the input text. The similarity between words in a sentence is computed by multiplying Q and K^T, which is the scaled by a factor of $\sqrt{d_k}$. $\sqrt{d_k}$ represent the size of embeddings with a default of 512. The output of the attention layer is the element-wise product of the attention weights, and the value vector V after the SoftMax function has been applied [19].

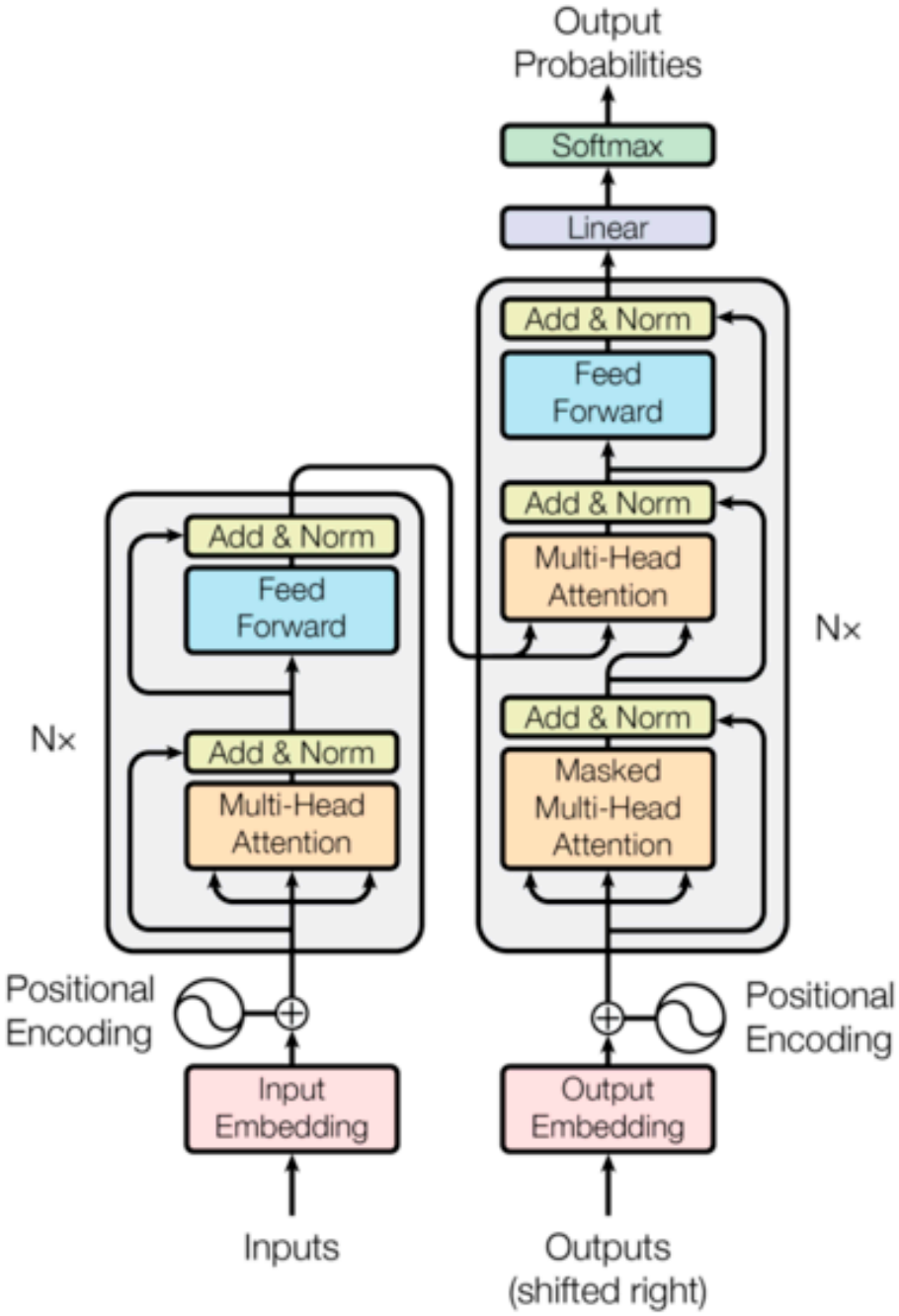

Figure 1. Transformer architecture [19].

This study presents a comparative analysis of different NLP tools, transcending traditional rule-based models such as VADER to the more recent transformer model such as BERT and LLMs such as GooglePaLM on the Amazon review dataset to observe how each model performs in relation to the ground truth, often including sarcastic comments made by individuals. The contributions of this research are as follows:

- A comparison of different NLP tools for sentiment analysis spanning different eras and technologies to deal with the contextual nuances of customer reviews.
- A proposed fine-tuned GooglePaLM large language model for sentiment analysis of online product reviews.

The rest of the paper is organized as follows. In Section 2, we analyze the related literature and show similar works to ours. Section 3 explains the methodology and the steps behind our study. Section 4 presents the findings from our comparative analysis. In

Section 5, we discussed our findings and in Section 6, we provided a conclusion based on our findings.

2. Related Works

Sentiment analysis, a crucial aspect of NLP, involves identifying emotions in textual content such as product reviews. Traditional methods relied on lexicons and rule-based systems, evolving into statistical methods like Naive Bayes and Support Vector Machines. However, challenges such as recognizing sarcasm and cultural variations persist [12]. The rest of this section discusses previous work conducted in relation to sentiment analysis of customer reviews of online products. It presents a review of the methodologies such as the traditional rule-based methods and the more recent large language transformer models.

2.1. Rule Based Approach

VADER, which is part of the Natural Language Toolkit (NLTK) library, has been effective in interpreting sentiments in social media by considering contextual clues [20]. Smith and Jones [21] performed a sentiment analysis utilizing the NLTK by exploring its applicability in assessing the emotional tone of product reviews. The study underscored NLTK's potential in deciphering sentiment nuances within textual data. Another study by Wang and Zang [22] employed the Stanford CoreNLP toolkit to analyze product sentiments, emphasizing its role in extracting sentiment features and identifying key aspects within reviews. Their research highlighted CoreNLP's effectiveness in uncovering nuanced sentiments and capturing the essence of customer opinions. A comparative study of sentiment analysis in e-commerce using NLTK and SpaCy was performed in [23], shedding light on the strengths and limitations of each toolkit in handling sentiment classification tasks. These findings provided valuable insights into the performance and suitability of NLP toolkits in analyzing customer sentiments within the e-commerce domain. This work was further improved by authors in [24] by incorporating Gensim into the study, offering a more comprehensive examination of sentiment analysis tools and their efficacy in capturing sentiment dynamics in e-commerce reviews. The authors in [25] developed a simple model to detect unconventional emotions such as skepticism in COVID-19-related textual data by mining users' opinions on Reddit using SpaCy. Emotion intensity was measured using NRC-EIL and the range of emotions used in the studies was defined by the Plutchik wheel of motion. They obtained the highest accuracy for the emotion of disgust.

An aspect-based sentiment analysis of product reviews using TextBlob, focusing on the toolkit's capability to analyze sentiments at a granular level, was conducted by the authors in [26]. Their research highlighted TextBlob's utility in dissecting product reviews into specific aspects and evaluating sentiment polarity within each aspect. Kim and Lee extended this exploration by employing Apache OpenNLP for sentiment classification of product reviews, illustrating the toolkit's proficiency in analyzing sentiment patterns and discerning contextual nuances within textual data. A comparative study of sentiment analysis tools, including NLTK, SpaCy, and Apache OpenNLP to evaluate their effectiveness in sentiment classification tasks, was carried out in [27]. Their research provided valuable insights into the comparative performance of these NLP toolkits, aiding researchers, and practitioners in selecting the most suitable toolkit for their specific sentiment analysis needs. Similarly, a study focused on an aspect-based sentiment analysis using Gensim showcased its ability to uncover sentiment dynamics across different aspects of product reviews presented in [28]. A sentiment analysis of online reviews using the TextBlob toolkit, emphasizing its simplicity and effectiveness in sentiment classification tasks, was performed in [29]. Their study underscored TextBlob's applicability in extracting sentiment features and gauging customer opinions from online reviews. A product review sentiment analysis using NLTK and Gensim was proposed in [30]. They provided insights into the synergistic application of multiple NLP toolkits in sentiment analysis tasks. This research highlighted the complementary strengths of NLTK and Gensim in capturing sentiment nuances within product reviews. Although the rule-based approach achieved

good performance, it had significant limitations in terms of maintenance of the model. Manual updates to the rules due to changing reviews will come at a significant cost; therefore, there is a need for the automatic extraction of sentiment features, especially for online product reviews [10].

2.2. BERT Model

BERTs originated from Google and employ bidirectional training to understand text contextually [31]. Lots of research has demonstrated BERTs' adaptability in sentiment analysis across various domains, including agriculture and e-commerce. A sentiment analysis of after-sales reviews of agricultural products was performed by the authors in [32]. They used an enhanced BERT model. They focused on the nuances found in customer reviews of agricultural products, such as non-standard expressions and sparse features. To measure the efficacy of their model, they obtained an F1 value of 89%, which was a better result compared to the original BERT model. The authors in [33] used a BERT with deep learning techniques such as BiGRU and Softmax on product reviews from 500 online customers from an e-commerce website. Their model was able to achieve a high accuracy of 95%, which outperformed models such as RNN and BiLSTM. The authors in [11] proposed the use of a fine-tuned BERT model to predict the sentiments of customers, based on reviews from Twitter, IMDB Movie Reviews, Yelp, and Amazon. They developed a dashboard that compared their proposed model with machine learning and neural network techniques such as LSTM, fastText, and BiLSTM. With a BERT, they obtained an accuracy of 90% which exceeded all of the other models compared. The use of BERT and its variants has also provided good metrics in analyzing sentiments. However, as it also stores sequences of vectorized text in its memory, this will limit the performance of the model when dealing with longer text, especially if the sentiment in the text contains contextual nuances implicitly embedded in the text.

2.3. Large Language Transformer Models

To deal with the nuances in the contextual meaning of the sentiments in customer online reviews, several large language models LLMs have been developed to provide better analyses. Novel approaches, including ChatGPT 3.5, have been explored for sentiment analysis, demonstrating its potential to understand customer emotions and emoticons [34]. The evolution of generative language models, notably the Generative Pre-trained Transformers (GPT) series, represents a significant advancement in NLP. GPT-3, with 175 billion parameters, stands out for its ability to generate text resembling human language [35]. InstructGPT further improves model refinement through human feedback, emphasizing the importance of user input [21]. In a similar work, the authors of [36] performed a comparison of two large language models, i.e., GPT-3 and LLaMA-2, for a product review sentiment analysis in predicting the star ratings of products provided by customers. They also included BERT and RoBERTa models in their comparison. From their experiments, LLMs performed better than the NLP, BERT, and RoBERTa models. GPT-3.5, however, gave the best performance with a predictive accuracy of 65%. This was closely followed by LLaMA-2 with a predictive accuracy of 62%. A performance evaluation of different NLP and LLMs using diverse datasets of online reviews was performed in [14]. They compared the strengths of PaLM and GPT-3.5-Turbo as the LLMs and ATAE-LSTM, flan-t5-large-absa, and DeBERTa as the NLP models. They used a wide range of product review datasets ranging from clothing to hotels. They obtained good accuracy with DeBERTa for tasks that do not require aspect-based sentiments. PaLM, however, did better for such tasks with an accuracy nearing the 90% mark, exceeding the GPT-3.5-Turbo model.

Despite advancements, challenges remain in understanding sarcasm and nuanced interpretations in sentiment analysis. In our research, we aim to propose GooglePaLM as a potential solution due to its strong language comprehension abilities and accessibility as a free tool. GooglePaLM was selected as the large language model based on its superior

performance, as suggested in [14,36] for empirical evaluation, and it will set the stage for future research in improving sentiment analysis of product reviews [37].

3. Materials and Methods

To conduct this research, we compared state-of-the-art specific sentiment analysis models such as VADER, BERT, and GooglePaLM for the sentiment analysis of reviews from fashion brands and predicted the assigned classes based on the scores provided in the dataset. We wanted to demonstrate the capabilities of GooglePaLM as a transformer model with a better reasoning ability and language spread [38], and how it outperforms other very popular sentiment models like BERT and VADER in performing sentiment analysis tasks. The proposed model is shown in Figure 2. The rest of this section will explain each process in detail.

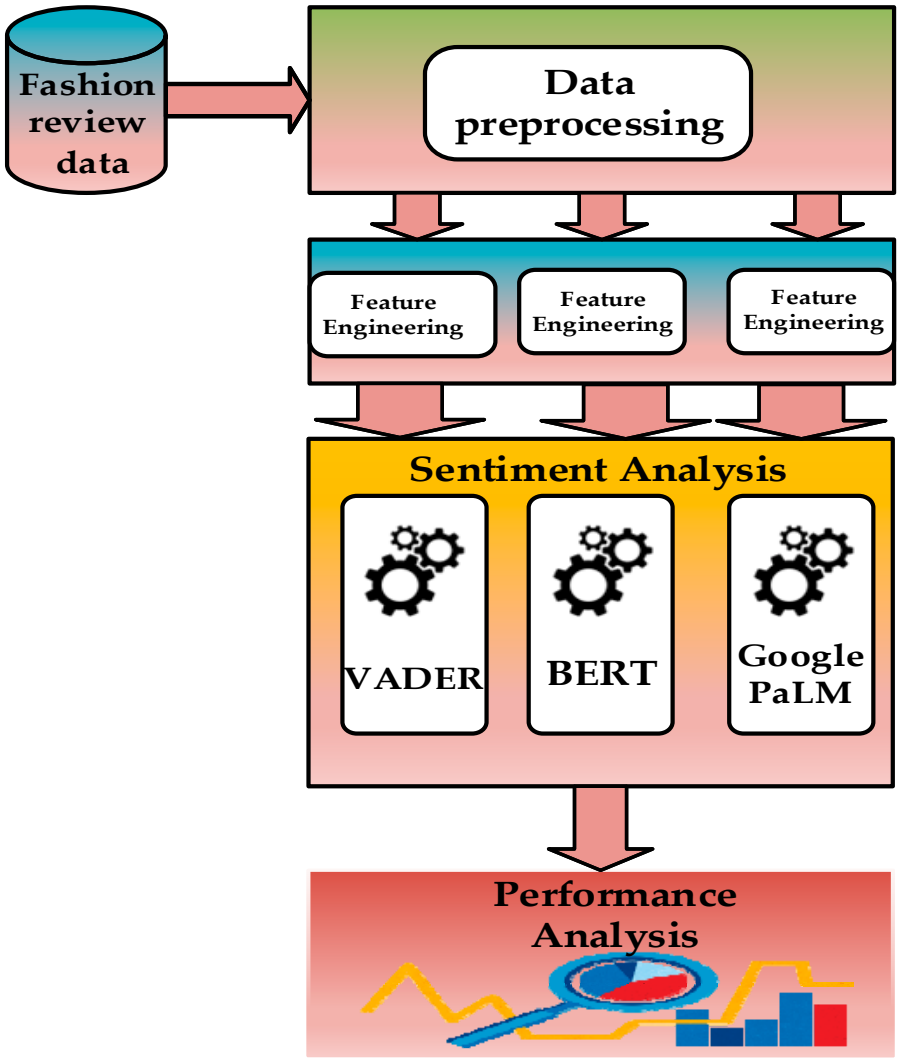

Figure 2. Experimental design.

3.1. DataSet

The dataset used was the Amazon Fashion Dataset [39], which comprises customer reviews for a range of fashion products. It includes 883,636 data points, and 10 variables are listed below.

- reviewerID—the ID of the reviewer, e.g., A2SUAM1J3GNN3B;
- asin—the ID of the product, e.g., 0000013714;
- reviewerName—the name of the reviewer;

- vote—helpful votes of the review;
- style—a dictionary of the product metadata, e.g., "Format" is "Hardcover";
- reviewText—the text of the review;
- overall—the rating of the product;
- summary—a summary of the review;
- unixReviewTime—the time of the review (unix time);
- reviewTime—the time of the review (raw);
- image—images that users post after they have received the product.

3.2. Data Pre-Processing

The dataset had a small number of missing values in the reviewText column. The number of missing values, 1233, was relatively insignificant compared to the total count of 883,636. Consequently, these missing rows were eliminated from the dataset. A new column named" sentiment" was generated based on the ratings column. This was performed manually, and it was used to represent the ground truth for classification. Ratings below 3 were labeled as negative, those above 3 as positive, and a neutral label was assigned to ratings of 3. The distribution of the sentiment classes is shown in Figure 3.

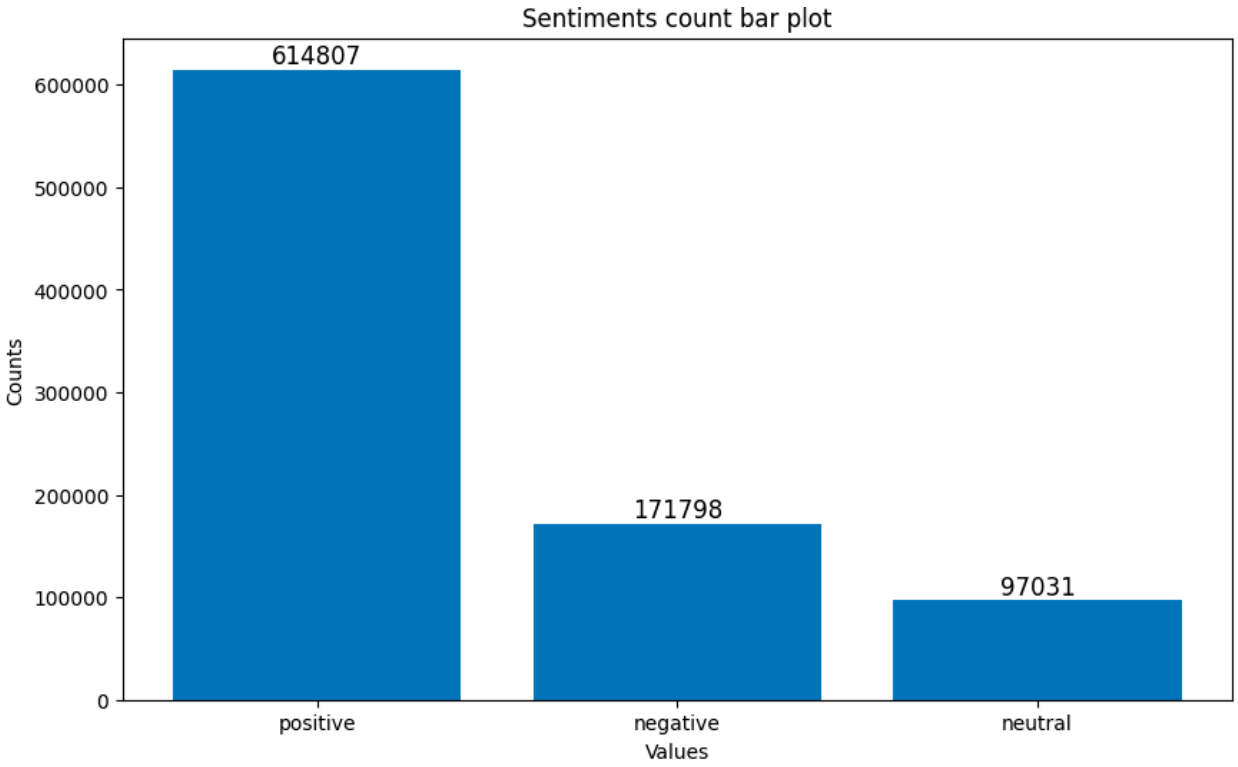

Figure 3. Counts of different sentiment classes.

To balance the data, downsampling of all the classes was performed by random samples of data based on each class to generate more balanced classes with 97,031 samples for each class. We downsampled as we did not want to introduce too much synthetic data in the minority classes. Also, pre-trained transformer models such as the GooglePaLM and GPT are known to provide modest accuracies when used in sentiment analysis tasks, even in small data sample sizes. This is due to the large number of trainable parameters (540 billion) that have been saved during model training [14,40]. The dataset was split in the ratio of 70:20:10 to represent training, testing, and validation. The frequency of words in the positive and negative sentiment cloud is shown in Figures 4 and 5.

Figure 4. Word cloud for positive sentiments.

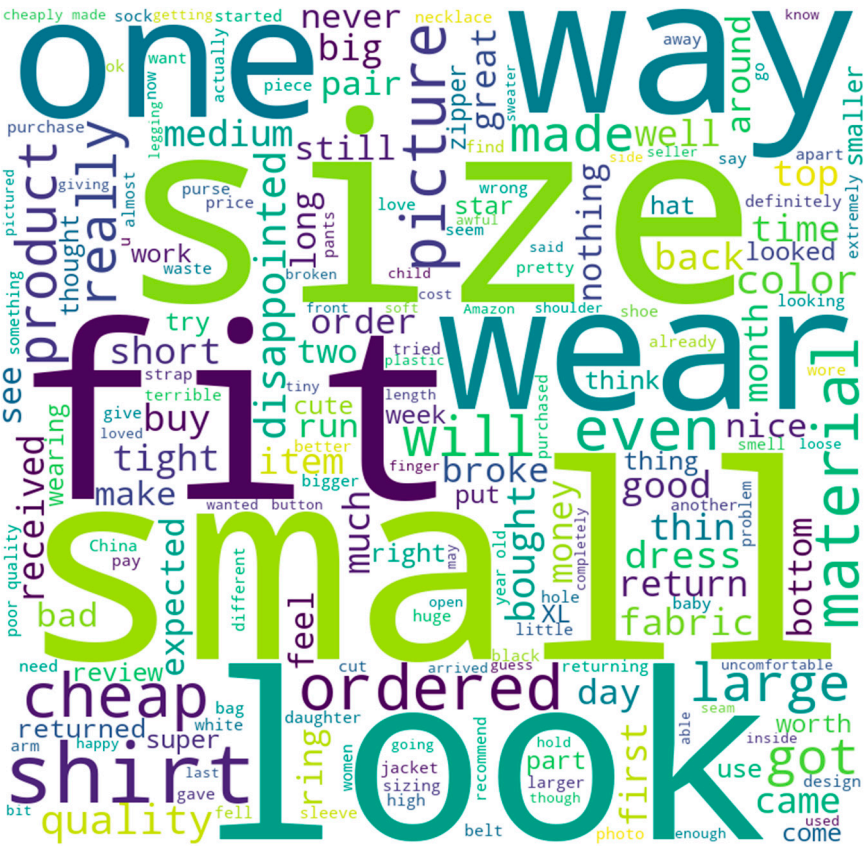

Figure 5. Word cloud of the negative sentiments.

3.3. Feature Engineering and Model Tuning

In this research, we analyzed the sentiments of the preprocessed Amazon fashion review dataset with three NLP toolkits. The data were then prepared for each of the models used in the study. Some of the parameters used for each model are shown in Table 1.

Table 1. Models and parameters.

Model	Tokenizer	Model Architecture	Temperature	N-Value
BERT	BertTokenizerFast	BertForSequenceClassification	Default	Default
Google PaLM	AutoTokenizer	models/text-bison-001	0.0	1
VADER	Word Tokenization	VADER	N/A	N/A

The values for BERT's parameters, including Temperature, are typically set to default values provided by the library or framework implementing the BERT model. These default values are chosen based on empirical observations and considerations of model performance across various tasks. For instance, in many implementations of BERT, the default values for these parameters are set to reasonable values that generally work well for a wide range of applications and datasets. As for the Google PaLM model, a Temperature value of 0.0 was used. This choice effectively eliminated randomness during text generation, resulting in deterministic outputs. The decision to set the Temperature to 0.0 reflects a preference for deterministic behavior in text generation tasks, which may be desirable in certain applications where consistency and predictability are crucial. Other parameters with default values used for the BERT model were Top P, Top K, and Max Output Tokens, which was not applicable to the GooglePaLM model. The N-value, which refers to the number of chat completions generated by models for each input prompt was selected as 1 for GooglePaLM and the default value was used for BERT. It determines the quantity of alternative responses or completions provided by the model for a given prompt, allowing users to explore different possible outcomes or interpretations. A tokenizer was used to break down the text into individual words, phrases, or other meaningful elements called tokens. In the context of natural language processing (NLP) models like BERT, the tokenizer converts input text into tokens that the model can understand and process. The model architecture refers to the specific design and structure of a machine learning or deep learning model. It includes the arrangement and configuration of layers, nodes, and connections within the model. For example, in BERT, the model architecture consists of multiple transformer layers that enable the model to capture contextual relationships in text. VADER is a rule-based sentiment analysis tool that operates on pre-tokenized text inputs. It does not involve a text generation process where the selection of tokens or the length of the output is determined by configurable parameters.

Instead, VADER analyzes the sentiment of pre-existing text passages without generating new content, making these parameters unnecessary for its functionality [14,20,41]. The evaluation techniques used for the performance analysis models include metrics such as precision, recall, accuracy, and F1 score [42].

4. Results

The confusion matrix of the sentiment classes for the model was obtained as shown in Figures 6–8.

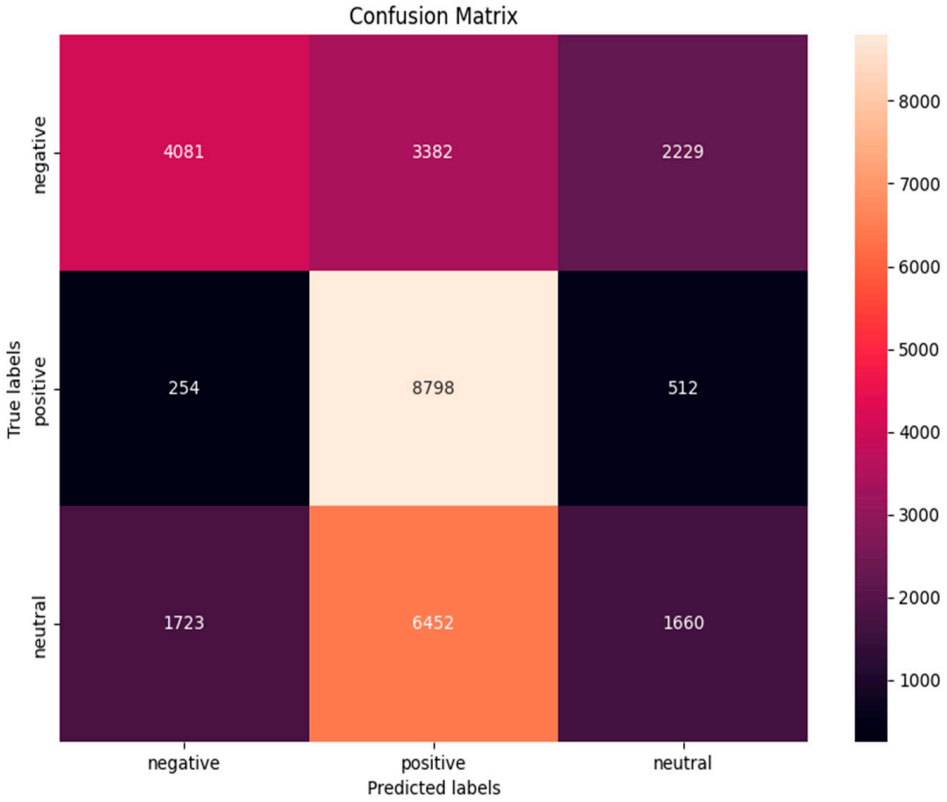

Figure 6. Confusion matrix for VADER.

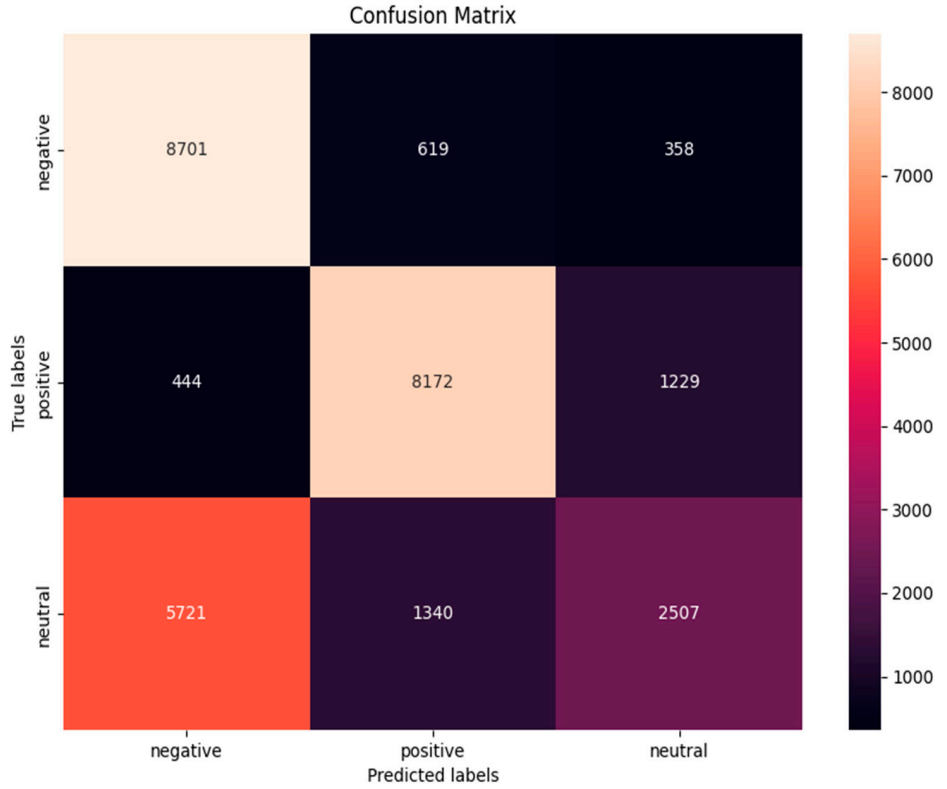

Figure 7. Confusion matrix for BERT.

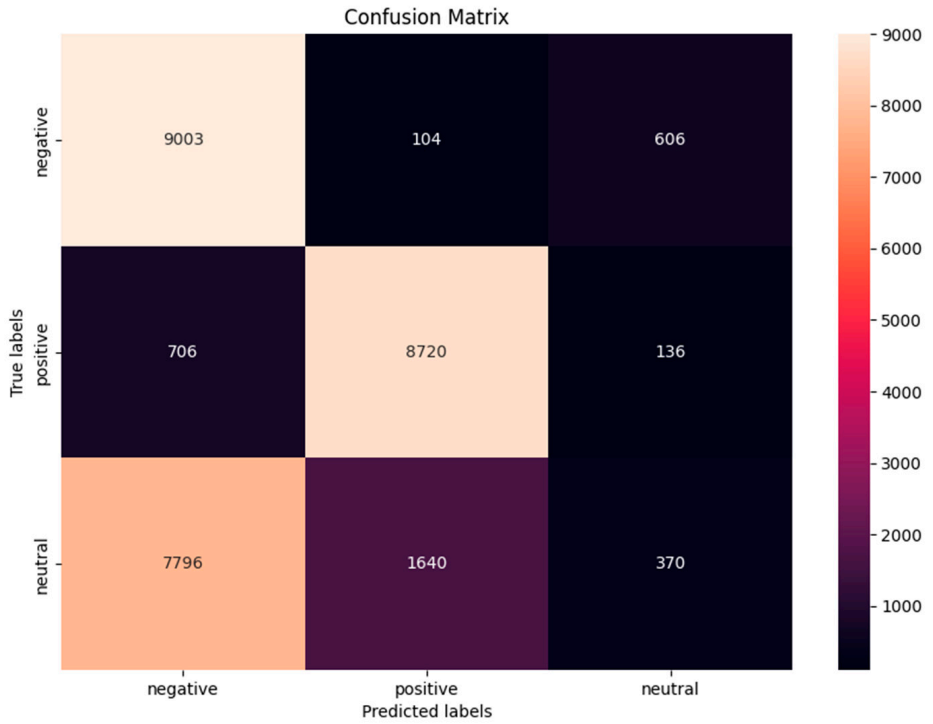

Figure 8. Confusion matrix for GooglePaLM.

For the VADER model, of the overall negative opinions, 3017 were accurately classified, but a substantial quantity of 4498 were wrongly classified as positive, and 2177 as neutral. This resulted in moderate accuracy with a precision score of 0.51 and a recall of 0.47. While excelling in detecting positive sentiments, VADER struggled with negative sentiments, misclassifying them as positive in 46.41% of cases. For the positive sentiment classes, the model demonstrated robustness in categorizing positive feelings, accurately detecting 8921 occurrences. Nevertheless, 175 instances were erroneously categorized as negative and 468 as neutral. It also struggled in correctly classifying the neutral sentiments as shown in Figure 6. For the BERT model, among all the instances classified as negative, 8701 were correctly identified as true negatives, while 619 were incorrectly identified as positive, and 358 as neutral. The model accurately identified 8172 cases as positive attitudes (true positives), but incorrectly classified 444 instances as negative and 1229 instances as neutral. Classifying neutral sentiments proved to be more arduous, as 2507 were accurately classified, while 5721 were misclassified as negative and 1340 were misclassified as positive, as shown in Figure 7. For GooglePaLM, of the total negative sentiments, 9003 were correctly identified, while 104 were misclassified as positive and 606 as neutral. The model improved in terms of positive sentiments as it correctly classified 8720 instances, misclassified 706 as negative, and 136 as neutral. The classification of neutral sentiments posed a significant challenge for the model, with 7796 instances being misclassified as negative and 1640 as positive, while only 370 were correctly identified. This is also shown in the confusion matrix chart in Figure 8. The summary of the evaluation metrics for the models is shown in Table 2.

Table 2. Performance metrics of sentiment analysis models.

Model	Evaluation Metrics					
	Precision	**Recall**	**Accuracy**	**F1-Score**	**Correct Positive**	**Correct Negative**
Google PaLM	0.28	0.31	0.62	0.27	0.91	0.93
BERT	0.67	0.66	0.66	0.63	0.83	0.89
VADER	0.50	0.47	0.46	0.41	0.93	0.31

5. Discussion

GooglePaLM excelled in accurately categorizing positive and negative sentiments but faced difficulties with neutral attitudes, indicating areas for improvement. Notably, the model demonstrated a high degree of accuracy in correctly classifying positive and negative sentiments, with correct positive and negative classification at 91.17% and 92.68%, respectively. This suggests that while the model faces challenges with overall sentiment classification, its performance is notably stronger when distinguishing between clearly positive and negative sentiments. The percentages of positive sentiments classified as negative and vice versa were 7.38% and 1.07%, respectively, for the PaLM model. These figures further highlight the model's proficiency in distinguishing between positive and negative sentiments, despite the lower overall precision and recall scores. The BERT model, which comes close to the GooglePaLM model in terms of the evaluation metrics, demonstrated a balanced performance but has challenges in interpreting neutral sentiments, just as the GooglePaLM model does. The BERT model showed significant efficacy in accurately categorizing sentiments, with an 83.01% accuracy in categorizing positive sentiments, an even greater 89.90% accuracy in classifying negative sentiments, and a misclassification rate of 4.51% for positive sentiments erroneously labeled as negative and 6.40% for negative sentiments wrongly classified as positive. While exhibiting high precision, the BERT model struggled with nuanced expressions of sentiment, necessitating further tuning to enhance accuracy across the diverse text input. Finally, the VADER model, which has a high reputation for detecting positive sentiments, showed promise in scenarios where identifying positive feedback is paramount. The model exhibited a notable level of accuracy in accurately categorizing positive attitudes, with a correct positive rate of 93.28%. Nevertheless, the accurate negative percentage was markedly smaller, measuring at 31.13%. This discrepancy suggests that the model has superior performance in detecting positive sentiments compared to negative attitudes. The proportion of good sentiments misclassified as negative was relatively small, at 1.83%, which further demonstrates the model's ability to accurately identify positive sentiments. In contrast, the model demonstrated a greater inclination to incorrectly identify negative thoughts as positive, with an occurrence rate of 46.41%. This result further shows the ineffectiveness of the rule-based models in detecting emotions in text when used for sentiment analysis.

6. Conclusions

This research aimed to evaluate the effectiveness of advanced language-generative models, particularly GooglePaLM, in sentiment analysis concerning product reviews, comparing them to traditional natural language processing (NLP) techniques. The comparative analysis unveiled insights into the performance of the different models used in this study. While conventional models like VADER showed proficiency in detecting positive sentiments, they struggled with accurately categorizing negative and neutral tones. GooglePaLM exhibited exceptional precision in discerning positive and negative thoughts but faced challenges with neutral sentiments, as with all other models. The BERT model demonstrated moderate efficacy across sentiment categories but highlighted the need for further refinement. In terms of implications and contributions, this research provides empirical evidence that advanced models like GooglePaLM surpass classic NLP techniques in specific areas of sentiment analysis. It offers a pragmatic structure and approach that can be readily applied or adapted in subsequent sentiment analysis investigations. Despite its significant insights, the study acknowledges limitations, including its focus on a single language and the exclusion of emojis and multimedia content in sentiment analysis. These limitations present opportunities for future research to explore multilingual sentiment analysis, incorporate non-textual sentiment indicators, and investigate sophisticated fine-tuning methods for language models. Also, future studies will look to explore GooglePaLM 2, which is the predecessor to GooglePaLM with more learnable parameters.

Author Contributions: Conceptualization, S.S. and O.S.; Data curation, S.S.; Formal analysis, S.S., O.S., S.M. and O.O.; Investigation, S.S., S.M., O.O. and O.S.; Methodology, S.S. and O.S.; Supervision, O.S. and S.M.; Visualization, S.S., S.M., and O.S.; Writing—original draft, S.S.; Writing—review and editing, O.S., S.M., O.O. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this work and the code is available publicly via this link: https://github.com/Swethijith/Customer-sentiments-in-product-review (accessed on 19 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hanaysha, J.R. An examination of the factors affecting consumer's purchase decision in the Malaysian retail market. *PSU Res. Rev.* **2018**, *2*, 7–23. [CrossRef]
2. Ozcan, K.; Ramaswamy, V. Word-of-mouth as dialogic discourse: A critical review, synthesis, new perspective, and research agenda. *Adv. Consum. Res.* **2004**, *7*, 528–532.
3. Kang, M.; Sun, B.; Liang, T.; Mao, H. A study on the influence of online reviews of new products on consumers' purchase decisions: An empirical study on JD. com. *Front. Psychol.* **2022**, *13*, 983060. [CrossRef]
4. Devedi, P.; Sujatha, R.; Pathak, R. A study on parameters of online reviews content that influence consumers buying behaviour-an Indian perspective. *J. Bus. Retail. Manag. Res.* **2017**, *11*, 12–21. [CrossRef]
5. Hutchins, W.J. The Georgetown-IBM experiment demonstrated in January 1954. In Proceedings of the Conference of the Association for Machine Translation in the Americas, Washington, DC, USA, 28 September–2 October 2004; pp. 102–114.
6. Mor, B.; Garhwal, S.; Kumar, A. A systematic review of hidden Markov models and their applications. *Arch. Comput. Methods Eng.* **2021**, *28*, 1429–1448. [CrossRef]
7. Cronin, R.M.; Fabbri, D.; Denny, J.C.; Rosenbloom, S.T.; Jackson, G.P. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int. J. Med. Inform.* **2017**, *105*, 110–120. [CrossRef]
8. Wang, B.; Xiong, S.; Huang, Y.; Li, X. Review rating prediction based on user context and product context. *Appl. Sci.* **2018**, *8*, 1849. [CrossRef]
9. Ansari, A.A. Evolution of sentiment analysis: Methodologies and paradigms. In *Trends of Data Science and Applications: Theory and Practices*; Springer: Singapore, 2021; pp. 147–174.
10. Yang, L.; Li, Y.; Wang, J.; Sherratt, R.S. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access* **2020**, *8*, 23522–23530. [CrossRef]
11. Durairaj, A.K.; Chinnalagu, A. Transformer based contextual model for sentiment analysis of customer reviews: A fine-tuned bert. *Int. J. Adv. Comput. Sci. Appl.* **2021**, *12*, 474–480. [CrossRef]
12. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731–5780. [CrossRef]
13. Jiang, Q.; Chen, L.; Xu, R.; Ao, X.; Yang, M. A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 6280–6285.
14. Mughal, N.; Mujtaba, G.; Kumar, A.; Daudpota, S.M. Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis. *IEEE Access* **2024**, *12*, 60943–60959. [CrossRef]
15. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modelling with pathways. *J. Mach. Learn. Res.* **2023**, *24*, 1–113.
16. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *AI Open* **2022**, *3*, 111–132. [CrossRef]
17. Chavez, M.R.; Butler, T.S.; Rekawek, P.; Heo, H.; Kinzler, W.L. Chat Generative Pre-trained Transformer: Why we should embrace this technology. *Am. J. Obstet. Gynecol.* **2023**, *228*, 706–711. [CrossRef] [PubMed]
18. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
20. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225. [CrossRef]

21. Smith, J.; Jones, A. Sentiment analysis of product reviews using Natural Language Toolkit (NLTK). *J. Appl. Linguist.* **2018**, *25*, 112–125.
22. Wang, Y.; Zhang, L. Product sentiment analysis using Stanford CoreNLP. *Int. J. Comput. Linguist.* **2019**, *15*, 220–235.
23. Chen, H.; Liu, M. Sentiment analysis in e-commerce: A comparative study of NLTK and SpaCy. *J. Nat. Lang. Process.* **2020**, *32*, 321–335.
24. Patel, R.; Gupta, S. Comparative analysis of sentiment analysis using NLTK, SpaCy, and Gensim in the e-commerce domain. *J. Comput. Intell.* **2021**, *8*, 45–56.
25. Basile, V.; Cauteruccio, F.; Terracina, G. How Dramatic Events Can Affect Emotionality in Social Posting: The Impact of COVID-19 on Reddit. *Future Internet* **2021**, *13*, 29. [CrossRef]
26. Zhang, Q.; Li, W. Aspect-based sentiment analysis of product reviews using TextBlob. *J. Inf. Sci.* **2019**, *36*, 180–195.
27. Kim, S.; Lee, J. Sentiment classification of product reviews using Apache OpenNLP. *J. Nat. Lang. Process. Tech.* **2018**, *21*, 305–320.
28. Gupta, R.; Sharma, V. Comparative study of sentiment analysis tools: NLTK, SpaCy, and Apache OpenNLP. *Int. J. Comput. Intell. Appl.* **2020**, *17*, 1450012.
29. Chen, Y.; Wang, Z. Aspect-based sentiment analysis of customer reviews using Gensim. *Expert Syst. Appl.* **2019**, *45*, 256–270.
30. Park, S.; Kim, M. Sentiment analysis of online reviews using the TextBlob toolkit. *J. Inf. Technol. Res.* **2020**, *27*, 78–92.
31. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
32. Cao, Y.; Sun, Z.; Li, L.; Mo, W. A study of sentiment analysis algorithms for agricultural product reviews based on improved bert model. *Symmetry* **2022**, *14*, 1604. [CrossRef]
33. Liu, Y.; Lu, J.; Yang, J.; Mao, F. Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax. *Math. Biosci. Eng.* **2020**, *17*, 7819–7837. [CrossRef]
34. Jagdale, R.S.; Shirsat, V.S.; Deshmukh, S.N. Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing: Proceeding CISC*; Springer: Singapore, 2019; Volume 2017, pp. 639–647.
35. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
36. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. *Nat. Lang. Process. J.* **2024**, *6*, 100056. [CrossRef]
37. Liu, X.; Zhang, Q. Product review sentiment analysis using NLTK and Gensim. *J. Nat. Lang. Process. Tech.* **2018**, *20*, 412–425.
38. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. Palm 2 technical report. *arXiv* **2023**, arXiv:2305.10403.
39. Ni, J.; Li, J.; McAuley, J. Justifying recommendations using distantly labelled reviews and fine-grained aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 188–197.
40. Majdik, Z.P.; Graham, S.S.; Shiva Edward, J.C.; Rodriguez, S.N.; Karnes, M.S.; Jensen, J.T.; Barbour, J.B.; Rousseau, J.F. Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study. *JMIR AI* **2024**, *3*, e52095. [CrossRef] [PubMed]
41. Gregory, P.A.; Bert, A.G.; Paterson, E.L.; Barry, S.C.; Tsykin, A.; Farshid, G.; Vadas, M.A.; Khew-Goodall, Y.; Goodall, G.J. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nat. Cell Biol.* **2008**, *10*, 593–601. [CrossRef]
42. Shobayo, O.; Zachariah, O.; Odusami, M.O.; Ogunleye, B. Prediction of stroke disease with demographic and behavioural data using random forest algorithm. *Analytics* **2023**, *2*, 604–617. [CrossRef]