**Sentiment Analysis of Amazon Product Reviews for Consumer Insights**

Madhavi Kancham

DSCI 6004-1,2,3: Natural Language Processing

Vinaykumar Reddy Moku

DSCI 6004-1,2,3: Natural Language Processing

Nithish Kumar Mayavan

DSCI 6004-1,2,3: Natural Language Processing

**Abstract:** *This study investigates sentiment analysis on Amazon product reviews to extract insights into consumer behavior. A dataset of over 34,000 reviews from Kaggle was preprocessed and used to train several machine learning and deep learning models, including a feedforward neural network (FFNN), Long Short-Term Memory (LSTM), and transformer-based models such as DistilBERT and BERT. The models achieved high accuracy, with BERT reaching 96%, DistilBERT performing closely, and FFNN and LSTM also yielding strong results at 93.17% and 95%, respectively. Despite the success in positive sentiment classification, challenges arose with the classification of neutral and negative sentiments due to class imbalance. All models were successfully deployed via Gradio, offering an interactive interface for real-time sentiment analysis, demonstrating their practical application and performance in real-world scenarios.*

***Index Terms: Sentiment Analysis, Amazon Reviews, Feedforward Neural Network (FFNN), Long Short-Term Memory (LSTM), DistilBERT, BERT, Gradio, Deployment***

1. **Introduction**
   1. **Background and Motivation**

The proliferation of e-commerce, fueled by advancements in internet and network technologies, has shifted customer behavior significantly towards online retail platforms such as Amazon, Walmart, and Target. Before making purchase decisions, customers often rely on reviews shared by others, which provide rich, descriptive insights about products. These reviews serve a dual purpose: they not only guide potential buyers in comparing products and brands but also act as feedback for sellers. Sellers leverage these reviews to refine their sales strategies and enhance product quality, making consumer opinions a critical factor in the online retail ecosystem.

Amazon is a global e-commerce giant founded in 1994 by Jeff Bezos. Initially launched as an online bookstore, the platform has since evolved into a multi-faceted retail ecosystem that offers a wide range of products, including electronics, apparel, household goods, and more. Amazon’s success is built on its commitment to customer satisfaction, efficient logistics, and technological innovation. It provides features like personalized recommendations, fast shipping through Amazon Prime, and an extensive customer review system that helps users make informed purchasing decisions. The review system, in particular, is one of Amazon’s defining elements, as it enables consumers to share their experiences, rate products, and provide feedback that influences millions of shoppers worldwide.

Sentiment analysis, also known as opinion mining, is a key task within natural language processing (NLP) that has garnered substantial attention in recent years. Sentiments reflect emotions, expressions, thoughts, or judgments. Using sentiment analysis, businesses can analyze target audience sentiments toward specific entities. This form of text analysis identifies polarity—such as positive, negative, or neutral opinions—at various levels, including sentences, paragraphs, or entire texts.

Understanding customer emotions is invaluable for companies, as customers now freely share their thoughts and feelings through reviews, surveys, and social media. Advances in machine learning and automation have enabled the development of systems that automatically analyze this feedback, providing marketers with actionable insights. By listening closely to their customers, companies can tailor their products and services to meet evolving consumer needs, enhancing overall satisfaction and loyalty.

* 1. **Problem Statement**

While sentiment analysis has gained widespread adoption across various domains, achieving consistent and balanced performance across all sentiment classes—positive, neutral, and negative—remains a significant challenge. Many models excel in classifying positive sentiments but struggle with neutral and negative classes, often due to class imbalances, limited training data, or inherent model biases. This project aims to address these issues by leveraging a range of machine learning and deep learning models, carefully analyzing their strengths and weaknesses, and exploring methods to improve sentiment classification across all categories.

* 1. **Contributions**
* Developed and implemented multiple models for sentiment classification, including a Feedforward Neural Network (FNN), Long Short-Term Memory (LSTM), DistilBERT, and BERT. Each model was optimized and tailored to process and analyze Amazon product reviews effectively.
* Evaluated the performance of these models using robust metrics, including accuracy, precision, recall, and F1-score, to assess their effectiveness and identify areas of improvement.
* Discovered and analyzed the gaps in classifying negative and neutral sentiments, which are often underrepresented or misclassified. The findings offer valuable insights and recommendations for future work in improving sentiment analysis systems, particularly in handling class imbalances and improving model interpretability.
* Successfully deployed the sentiment analysis models to a production-like environment, enabling real-time predictions and practical application. Leveraged pre-trained models such as DistilBERT and BERT to ensure scalability and computational efficiency, facilitating the integration of sentiment analysis capabilities into existing system

By addressing these aspects, the project not only provides a deeper understanding of sentiment analysis challenges but also contributes to the development and deployment of more reliable and balanced models for real-world applications.

1. **Related Work**

Extensive research has been conducted in recent years to analyze and classify customer reviews effectively. Notably, Guner et al. [1] from KTH Royal Institute of Technology, Stockholm, analyzed 60,000 randomly selected product reviews from Amazon. Their dataset, obtained from Kaggle, comprised over 4 million reviews. The study evaluated the performance of three algorithms: Multinomial Naïve Bayes (MNB), Linear Support Vector Machine (LSVM), and Long Short-Term Memory networks (LSTM). Using metrics like accuracy, area under the curve (AUC), precision, recall, and F1-score, they concluded that the LSTM model performed best, achieving a precision greater than 0.90 and an AUC of 0.96 for binary sentiment classification (positive and negative).

Fang and Zhan [2] collected and analyzed over 5.1 million Amazon product reviews across four categories: beauty, books, electronics, and home. They used Naïve Bayes, Support Vector Machines, and Random Forest classifiers for sentiment analysis. Their study emphasized evaluating sentiment polarity and concluded that Random Forest produced more reliable results. Additionally, they observed that for larger datasets, SVM outperformed Naïve Bayes in sentiment classification.

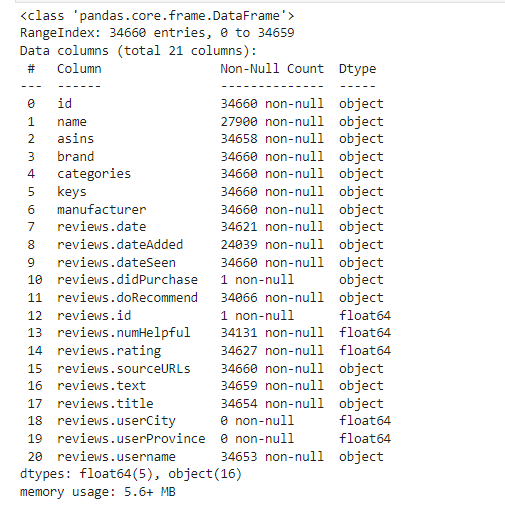
Tan et al. [3] applied both traditional machine learning algorithms, including Naïve Bayes, SVM, and k-nearest neighbor, as well as deep learning models like LSTM and recurrent neural networks, to an Amazon reviews dataset. Their dataset comprised 34,627 reviews split into 21,000 training and 13,627 test samples. Their findings revealed that LSTM achieved the highest test accuracy of 71.5%, although they attributed the lower accuracy to data imbalance in their dataset.

Rain [4] conducted sentiment classification on a dataset containing 50,000 Amazon book reviews, distinguishing between positive and negative sentiments. Using Naïve Bayes and decision-list classifiers, along with feature extraction methods like bag-of-words and bigrams, Rain concluded that Naïve Bayes outperformed decision-list classifiers. Furthermore, bag-of-words proved to be the most effective feature extraction technique for sentiment labeling.

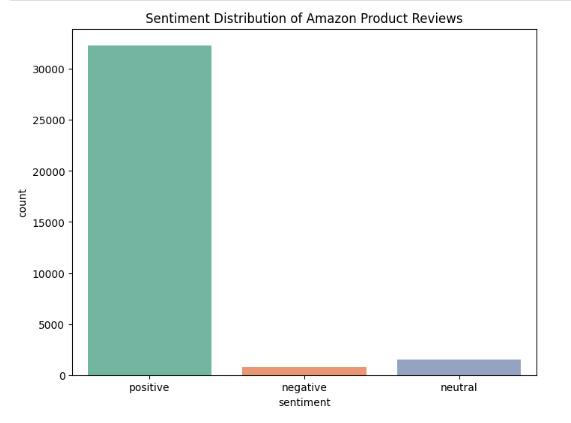
Shrestha and Nasoz [5] explored sentiment analysis of Amazon reviews using recurrent neural networks (RNNs) with gated recurrent units (GRU). They utilized a dataset of approximately 3.5 million Amazon product reviews and developed a novel approach combining paragraph vectors (inspired by word vectors) and GRU-derived product embeddings. These embeddings were used to train a support vector machine (SVM) classifier. The classifier achieved 81.29% accuracy with review embeddings alone, while the inclusion of product embeddings slightly increased accuracy to 81.82%. The authors suggested that this technique could also be adapted for user profiling tasks.

1. **Dataset and Preprocessing**
   1. **Dataset Description**

The dataset, sourced from Kaggle, comprises over 34,000 reviews with 21 variables, including review text, ratings, and metadata. The sentiments are labeled as positive, neutral, or negative, offering a comprehensive view of consumer feedback.



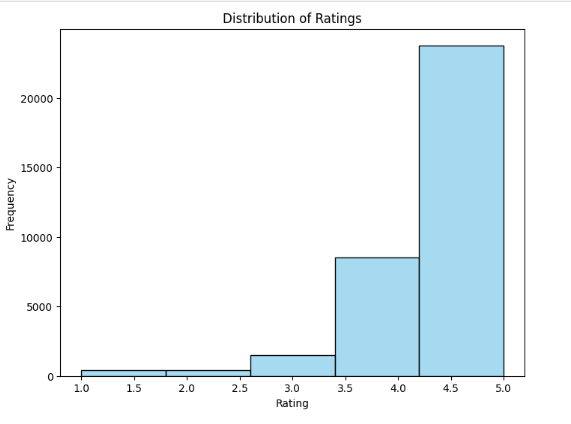
* 1. **Exploratory Data Analysis (EDA)**
* **Sentiment Distribution**: Positive reviews dominate, with fewer neutral and negative sentiments.



* **Most Frequent Words**: Positive sentiments feature words like *"love"* and *"use,"* while negative reviews include *"poor"* and *"broken."*

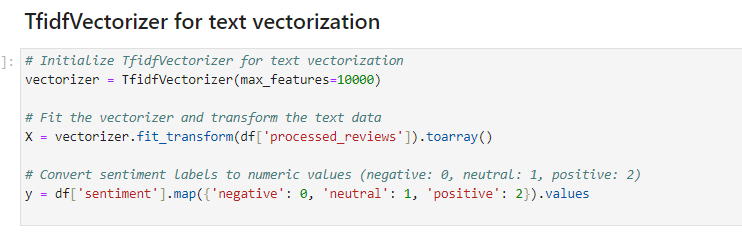


* **Ratings Distribution**: Reviews are skewed towards higher ratings (4 and 5 stars).



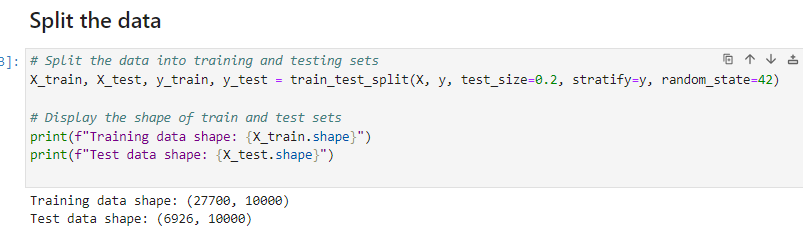
* 1. **Preprocessing**
  2. **Text Cleaning**, **Tokenization, Vectorization**, and **Encoding**

The preprocessing phase of the analysis involved three main steps to prepare the data for sentiment classification. First, text cleaning was performed by removing punctuation, stopwords, and special characters, while also converting all text to lowercase for uniformity. Second, tokenization and vectorization were applied, where the reviews were transformed into numerical feature representations using a TfidfVectorizer with a vocabulary size of 10,000 terms, effectively capturing important textual patterns [6]. Sentiment labels were encoded into numeric values, mapping negative, neutral, and positive sentiments to 0, 1, and 2, respectively, ensuring compatibility with machine learning models.



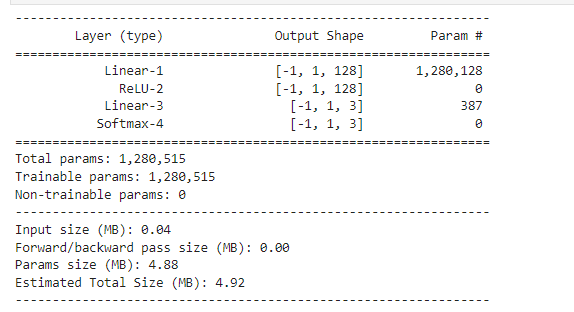
* 1. **Data Splitting**

The dataset was split into 80% training and 20% testing sets using stratified sampling to maintain the proportional distribution of sentiment classes across both subsets. This ensured balanced representation of sentiments for effective model training and evaluation. After the split, the training data comprised 27,700 samples with 10,000 features, while the test data contained 6,926 samples with the same feature dimensions.



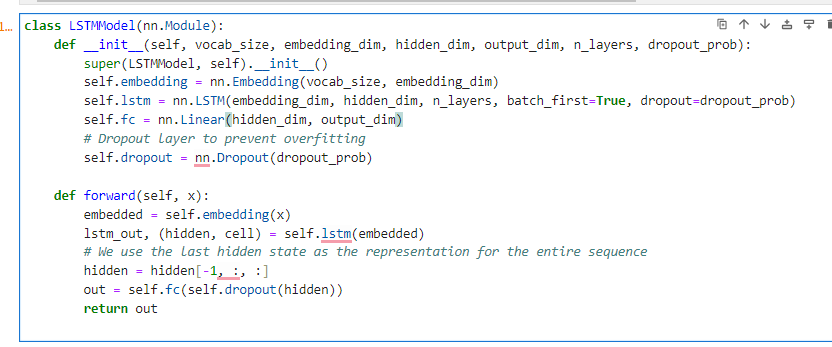
1. **Methodology**
   1. **Feedforward Neural Network (FFNN)**

The architecture of the SentimentNN model consists of an input layer that takes in 10,000 features derived from TF-IDF vectorization. The model includes a hidden layer with 128 units and ReLU activation to introduce non-linearity. The output layer has 3 units, corresponding to the three sentiment classes (negative, neutral, and positive), with a softmax activation function to output probability distributions across these classes. The model is trained using CrossEntropyLoss as the loss function for multi-class classification, and the Adam optimizer is employed with a learning rate of 0.001 to efficiently update model parameters during training.



* 1. **Long Short-Term Memory (LSTM)**

The architecture of the LSTM-based sentiment analysis model consists of an embedding layer that converts input words into dense vectors of fixed size, followed by an LSTM layer with multiple stacked units to capture the sequential dependencies within the text. The LSTM processes the embedded input and generates hidden states that represent the contextual information of the sequence. The final hidden state is passed through a dropout layer to reduce overfitting and is then fed into a fully connected layer to produce the sentiment classification output, with three output units representing the three sentiment classes: negative, neutral, and positive. The model incorporates dropout regularization to prevent overfitting and is designed to handle large vocabularies and long text sequences.



* 1. **Transformer Models**

1. **DistilBERT**

A distilled version of BERT, optimized for faster inference and reduced memory usage while retaining high accuracy.

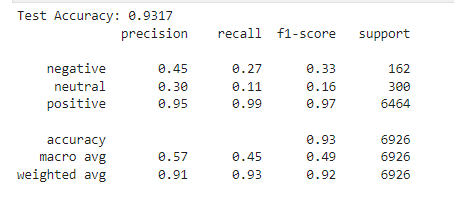
* Pretrained weights were fine-tuned on the Amazon reviews dataset.

1. **BERT**

The BERT architecture incorporates 12 transformer layers with self-attention mechanisms to contextualize input text.

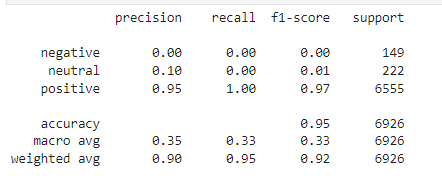
* Fine-tuned using BertForSequenceClassification for three-class sentiment classification.

1. **Results and Discussion**
   1. **Feedforward Neural Network**

The Feedforward Neural Network achieved an overall accuracy of 93.17%, demonstrating strong performance for certain sentiment classes. Specifically, it showed high precision and recall for the positive class, but struggled with the neutral and negative classes. The neutral class had a precision of 30% and recall of 11%, while the negative class achieved a precision of 45% and recall of 27%. This suggests that the model faces difficulties in accurately classifying nuanced sentiments, which can be attributed to its limited depth, potentially limiting its ability to capture complex patterns in the data.

* 1. **LSTM Model**

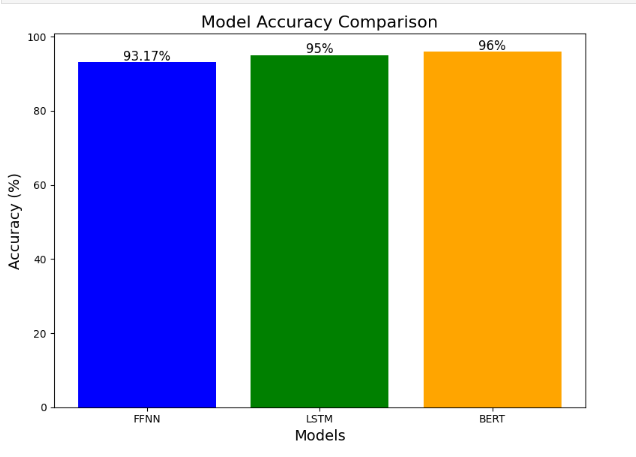
The LSTM model achieved an overall accuracy of 95%, showcasing impressive performance for the positive sentiment class with a precision of 95% and recall of 100%. However, it encountered significant challenges with the neutral and negative classes, similar to the Feedforward Neural Network, likely due to class imbalance. For the negative class, precision and recall were both 0%, while the neutral class had a precision of 10% and recall of 0%. Despite these issues, the model performed exceptionally well on the positive class, with an F1-score of 0.97. The macro and weighted averages reveal that, although the model excels in positive sentiment classification, it struggles with the minority classes, affecting overall performance on the full dataset.



* 1. **Transformer Models**

The Transformer models, including DistilBERT and BERT, demonstrated notable improvements over the Feedforward Neural Network (FFNN) and LSTM, particularly in terms of inference speed and generalization. DistilBERT, a lighter version of BERT, offered faster inference times while maintaining strong performance, though it still faced similar class imbalance issues, favoring positive sentiments. BERT achieved an overall accuracy of 96%, showcasing its robustness in understanding context and excelling in positive sentiment classification. However, BERT's performance came at the cost of being computationally expensive, and it struggled with predicting minority classes, reflecting ongoing challenges with class imbalance despite its superior accuracy.

**5.4. Models Analysis**

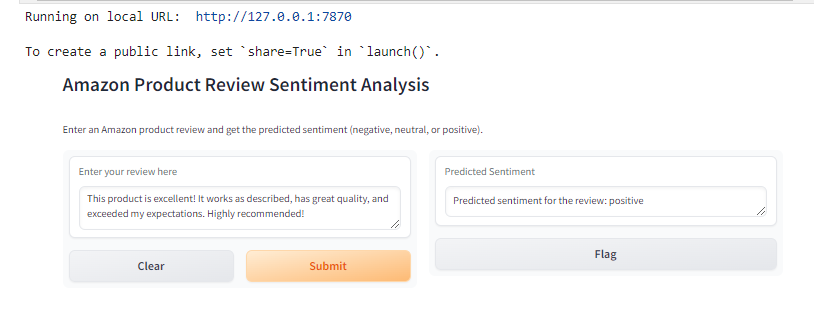


The performance of the models varies significantly, with the Transformer-based models outperforming the Feedforward Neural Network (FFNN) and LSTM in overall accuracy. The FFNN achieved an accuracy of 93.17%, but struggled with nuanced sentiment classification, particularly for neutral and negative sentiments. The LSTM model showed an improvement with an accuracy of 95%, excelling in positive sentiment classification, although it faced challenges with class imbalance, particularly for neutral and negative sentiments. The Transformer models, especially BERT, delivered the best results with an accuracy of 96%, offering robust contextual understanding and high performance on positive sentiments. However, both Transformer models exhibited similar issues with class imbalance, where positive sentiments were favored, while struggling with minority class predictions, particularly in neutral and negative categories.

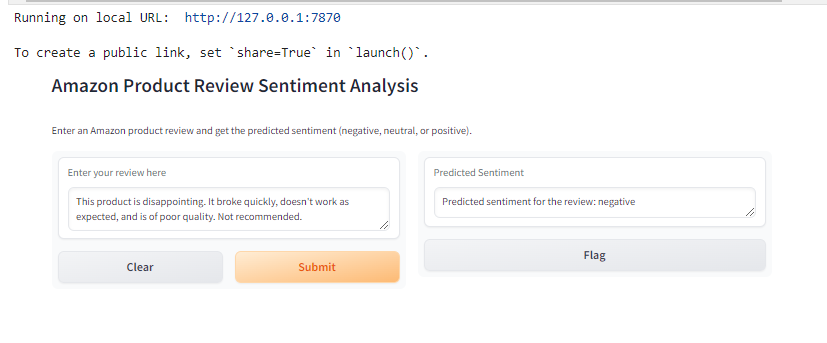
1. **Real World Deployment**

For the successful deployment of the sentiment analysis models, we leveraged **Gradio**, a powerful Python library that facilitates the creation of interactive web interfaces. Gradio was used to develop an interface that allows users to interact with the different sentiment analysis models implemented in this project: Feedforward Neural Network, LSTM, DistilBERT, and BERT.

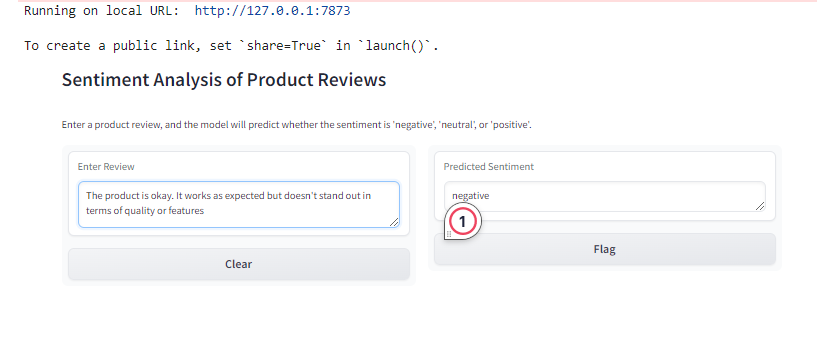
1. **Feedforward Neural Network**

****

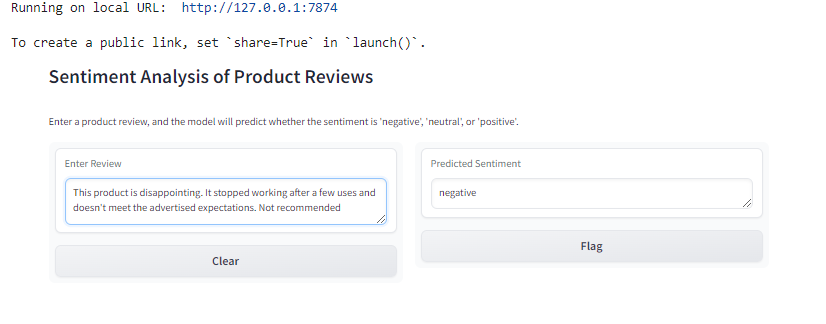
1. **LSTM**

****

1. **DistilBERT,**

****

1. **BERT.**

****

The interface allows users to input product reviews as text and see the predicted sentiment (positive, neutral, or negative) generated by each model. The Gradio interface was designed to present the predictions from all models in parallel, giving users the ability to compare their results in real-time. This provided an intuitive and accessible way for non-technical users to explore the performance of the various sentiment classification algorithms.

1. **Conclusion and Future Work**
   1. **Conclusion**

In this project, sentiment analysis on Amazon product reviews was successfully explored using four different models: Feedforward Neural Network (FFNN), Long Short-Term Memory (LSTM), DistilBERT, and BERT. The models performed admirably, achieving accuracies ranging from 93.17% for FFNN to 96% for BERT. Each model demonstrated significant potential in classifying sentiments, with transformer-based models, particularly DistilBERT and BERT, excelling in contextual understanding and outperforming traditional architectures. Despite challenges in classifying neutral and negative sentiments due to class imbalance, the models were successfully deployed through Gradio, providing an accessible and intuitive interface for users to interact with. The successful deployment of all models showcases their practical application and effectiveness in real-world scenarios.

* 1. **Future work**

Future work can focus on addressing class imbalance by exploring techniques like oversampling, undersampling, or SMOTE to improve the classification of neutral and negative sentiments. Additionally, ensemble models could be implemented to combine predictions from multiple models for enhanced performance. Fine-tuning transformer models like BERT on domain-specific datasets, such as e-commerce reviews, could further improve generalization. Furthermore, optimizing and deploying lightweight models for real-time applications would make them more practical in production environments. Finally, the sentiment analysis techniques could be adapted for user profiling, helping to create more personalized customer experiences.

**References**

1. L. Guner, E. Coyne, and J. Smit, “Sentiment analysis of Amazon.com reviews,” Mar. 2019. Available: <https://www.researchgate.net/publication/332622380_Sentiment_analysis_for_Amazoncom_reviews>.
2. X. Fang and J. Zhan, “Sentiment analysis using product review data,” *Journal of Big Data*, vol. 2, no. 1, pp. 16, Jun. 2015. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2.
3. W. Tan, X. Wang, and X. Xu, “Sentiment analysis for Amazon reviews,” Available: <http://cs229.stanford.edu/proj2018/report/122.pdf>.
4. C. Rain, “Analysis in Amazon reviews using probabilistic machine learning,” 2012. Available: https://www.semanticscholar.org/paper/Analysis-in-Amazon-Reviews-Using-Probabilistic-Rain/f0afe9ea9d286248336ee9dc4e954aecde3475bb.
5. N. Shrestha and F. Nasoz, “Deep learning sentiment analysis of Amazon.com reviews and ratings,” *International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI)*, vol. 8, no. 1, pp. 1-10, Feb. 2019. Available: <https://arxiv.org/abs/1904.04096>.
6. “sklearn.metrics.precision\_recall\_fscore\_support,” Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision\_recall\_fscore\_support.html.
7. P. D. Turney, “Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews,” in *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, Dec. 2002, pp. 417-424. Available: <https://arxiv.org/abs/cs/0212032>.
8. M. I. Zul, F. Yulia, and D. Nurmalasari, “Social media sentiment analysis using K-means and Naïve Bayes algorithm,” in *Proc. 2nd International Conference on Electrical Engineering and Informatics (Icon EEI)*, Oct. 2018.
9. Amazon Review Data (2024) Available: <http://jmcauley.ucsd.edu/data/amazon/>.
10. Available: http://datameetsmedia.com/staging/3908/bag-of-words-tf-idf-explained.
11. S. Banerjee, “A dynamic business model for IT industries,” *Test Engineering and Management*, vol. 81, pp. 1–6, Jan.-Feb. 2019. Available: http://testmagzine.biz/index.php/testmagzine/article/view/3/2.
12. R. K. Krishnam, “A study on tools and techniques for business models,” *Test Engineering and Management*, vol. 81, pp. 7–12, Jan.-Feb. 2019. Available: http://testmagzine.biz/index.php/testmagzine/article/view/4/3.
13. J. Lee, “Study of migration and mobility in the age of disruption with socio-economic changes,” *Test Engineering and Management*, vol. 81, pp. 1–4, May-Jun. 2019. Available: http://testmagzine.biz/index.php/testmagzine/article/view/7/6.
14. L. Alasa, “The role of Internet of Things in healthcare system with security and sensor networks,” *Test Engineering and Management*, vol. 81, pp. 5–8, May-Jun. 2019. Available: http://testmagzine.biz/index.php/testmagzine/article/view/8/7.