Sentiment Analysis of Customer Reviews

Jonathan Lew, Matthew Hong, Zac Allyn, Gale McDaris,
Pradyun Singh, Liew Jade Saefong, Kenneth Villafana, Jiashu Yu

Winter 2025

1 Introduction, Motivation, & Problem

Businesses cannot exist without customers. The foundation of any company's success is the result of happy customers. Businesses and customers co-exist in a mutually beneficial relationship. The giver and the receiver. Businesses provide necessary goods for their customers and customers who deem the quality satisfactory enough. It is a cycle that can benefit both sides; however, this cycle can potentially be broken. Sales from customers sustain business operations and enable potential expansion if there is great success. Happy customers result in revenue growth, while unhappy customers can find the same services elsewhere. Sometimes, it may be inconvenient for the customers, businesses that lose customers are more at risk as demand proves that a business offers value. If there is no demand, then there is no value, and therefore no customers.

In the growing age of technology, word-of-mouth has evolved into online reviews that can be accessed anywhere and anytime. Whether through food services or through retail sites, reviews can be found describing pros and cons, usability, past customer service experiences, and even the taste or temperature of the food they purchased. Easy accessibility to see and make reviews enables customers to highlight a business's success points and key necessary improvements.

Reviews can make or break a business. Popular applications such as Yelp or Google Reviews were created with the main purpose for users to look up reviews on anything that offers some sort of service. Upon viewing these reviews or stories of past experiences from previous customers, the user can gauge whether they will support that business. That being said, good reviews can influence a customer to walk through the door or can prevent them from ever entering. Past experiences are important and play a vital role in decision making; After all, how can future customers expect to not experience the same thing?

One bad experience can lead to a bad review and may just be enough to potentially ruin a business. However, a business resilience and professionalism rely on their responses and the actions they take to remedy past experiences. These reviews showcase a business's skill level, respectability, responsibility, and professionalism. Loyal customers help build credibility and trust. Businesses that have shown they have learned from past reviews show that they care and that they can improve for customers. This increases business reputation and influence new customers by exhibiting high levels of accountability.

Without customers, a business is just a concept without sustainability. That's why understanding customers' sentiments and feelings when they leave online reviews are vital to making sure businesses keep their brand and reputation alive and thriving. By understanding customers, businesses can find the context and root cause for such reviews so that they can fix the problem straight from the source.

Our project not only seeks to improve understanding of customers' sentiments, but to improve time and efficiency it takes for businesses to go through potentially hundreds or thousands of reviews that may be good or bad. Cultural factors, linguistic nuances, and differing contexts can make it extremely difficult to understand customer reviews. These reviews can consist of slang, humor, and new words or terms that can change the tone and actual meaning. This makes going through reviews time-consuming and hard on a busy user's cognition. With the aid of our sentiment analysis tool, we can offer vital support in sorting and understanding different kinds of reviews in a shortened amount of time. Through the use of BERT (Bidirectional Encoder Representations from Transformers), we can train models to understand these nuances and factors and create a tool that can go through reviews and categorize them.

2 Literature Review

Sentiment analysis is an approach to natural language processing that focuses on extracting/inferring subjective information from textual data to discern public opinion and emotions. The studies "Sentiment Analysis: A Comparative Study on Different Approaches" by Devika et al. (2016) and "Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review" by Mao et al. (2024) offer insights into the methodologies, applications, and challenges when conducting sentiment analysis.

Devika et al. (2016) categorize sentiment analysis into three primary levels: document level, sentence level, and aspect level. Sentiment analysis at the document level assesses whether the overall sentiment of an entire document is positive or negative, and assumes that each document is a single entity. At the sentence level, the goal of sentiment analysis is to decide whether a sentence contains a positive, neutral, or negative sentiment. Sentiment analysis at the aspect level is a finer analysis that focuses on individual terms and their sentiment polarity.

The authors further explore various approaches to sentiment analysis. Machine learning approaches involve training algorithms on labeled datasets to classify sentiments. The discussed techniques used for a machine learning approach to sentiment analysis are Support Vector Machines (SVM), N-gram Analysis, and Maximum Entropy Models. SVMs utilize hyperplanes to separate data into different sentiment categories.

N-gram Analysis considers contiguous sequences of 'n' items (which may be words or characters) to capture both context and sentiment. Maximum Entropy Models employ the principle of making no assumptions beyond the constraints provided by the training data and they aim for uniform distribution unless otherwise specified.

There are also lexicon-based approaches, which rely on predefined dictionaries of words associated with positive or negative sentiments. The effectiveness of this method depends on the comprehensiveness of the lexicon and its adaptability to different contexts. The disadvantage of this approach is that each word that you want to include in your sentiment analysis must be predefined as positive or negative.

Mao et al. (2024) expand upon these methodologies by conducting a systematic literature review, highlighting the evolution of sentiment analysis techniques over time. They emphasize the integration of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which have demonstrated significant improvements in capturing contextual nuances and handling large-scale data.

Both studies underscore the diverse applications of sentiment analysis across various domains: analytics for marketing strategies and product development, efficient social media monitoring, political analysis, and patient feedback in healthcare industries.

Both papers discuss challenges that persist in the field. Because a word's meaning often depends on its context, it can be difficult to accurately judge its sentiment. Human sarcasm and irony are contradictory at face value, which makes it difficult to correctly match the meaning of the words to the tone. In addition to that, language is inherently dynamic and constantly changing, so sentiment analysis models need to undergo constant updates in order to stay current and accurate. Taking multiple languages into consideration also complicates a model, since different languages contain unique nuances and exist in varying cultural contexts. The studies by Devika et al. (2016) and Mao et al. (2024) outline a comprehensive overview of the methodologies, applications, and challenges in sentiment analysis. While traditional machine learning and lexicon-based approaches have laid the foundation, the integration of advanced deep learning techniques offers promising improvements. However, addressing challenges such as discerning context, identifying sarcasm, and addressing the evolving nature of language remains crucial for advancing the accuracy of sentiment analysis.

3 Dataset

The dataset used for the sentiment analysis is a set of 400,000 customer reviews from Amazon.com (1) . This dataset was specifically chosen as it was readily available on Kaggle and had a large number of reviews, sufficient to base a model on. Additionally, this dataset was a good choice for this project due to the various attributes it had. To begin with, it has a column called "Reviews" with 162492 unique text values which will be used as the text to train the sentiment analysis model. Additionally, it has a "Ratings" column where ratings are given from 1- 5 for different products (located in the "Product Name" column). The ratings from 1-5 allow us to group ratings 1 and 2 as negative, 3 as neutral, and 4 and 5 as positive. Thus, the model would be able to associate common words in each of the sentiment groups and those attributes are crucial for training. The dataset has additional columns such as "Brand Name", "Price", and "Review Votes" which is a column related to how many upvotes a certain review got. Since the dataset is based on real-world data, the data isn't split exactly into 33.33% neutral, 33.33% negative, and 33.33% positive sentiment. Instead, the sentiment breakdown of the data is 68.50% positive, 23.53% negative, and 7.97% neutral. Additionally, some rows that have missing values or special character and irregular capitalization as humans often make lexical errors when typing reviews.

4 Methodology

4.1 Models

In this project, we focused mainly on using BERT, a revolutionary bi-directional context processor. Unlike earlier models that only read text left-to-right, BERT analyzes each word by considering what comes before and after it. This is achieved through masked language modeling, where BERT is pre-trained to predict randomly masked words using their surrounding context. This creates a much richer understanding of languages.

When fine-tuned for sentiment analysis, BERT demonstrates remarkable performance. Its transformer architecture uses self-attention mechanisms to precisely weigh the importance of different words in determining sentiment. This approach achieves up to 95% accuracy in classifying reviews and requires only a fraction of the training data needed by traditional models. Most importantly, BERT excels at capturing subtle contextual cues that earlier models would miss.

4.2 Preprocessing

```
Original dataset size: 413840 reviews
Sampled 22000 reviews from the dataset
Using 'Reviews' as the review column
Using 'Rating' as the rating column
After removing missing values: 21993 reviews
Preprocessing reviews...

Sentiment distribution in the dataset:
positive: 15066 reviews (68.50%)
negative: 5175 reviews (23.53%)
neutral: 1752 reviews (7.97%)

Training set: 17594 reviews
Test set: 4399 reviews

Class distribution in training set:
positive: 68.50%
negative: 23.53%
neutral: 7.97%

Class distribution in test set:
positive: 68.52%
negative: 23.53%
neutral: 7.96%
```

Figure 1: Preprocessing Statistics

Clean data is imperative for a model's performance. Thus, preprocessing the data to make it suitable for BERT is a crucial step in creating a trained model. To begin with, a new column was added to the data table labeled "Sentiment" which is based on the ratings column and has a value of negative if rating is less than or equal to 3, neutral if the associated rating is 3, and positive if the rating is 4 or 5. Next, a column called "Sentiment Label" was created which converted those classifications into numbers (0 for negative, 1 for neutral, and 2 for positive). This allowed the sentiment to have a numerical value associated with it to make the data analysis and comparisons easier. Next, the reviews in the review column had to be preprocessed. Rows that had a missing review or rating column were dropped from the dataset. In order to ensure that the sentiment class breakdown (percentage of reviews with a positive, negative, or neutral sentiment), stratified sampling was applied to select 22,000 reviews to ensure the class breakdowns remained consistent. After sampling, the reviews were processed in order to remove any special character that were present as well as making all the text lowercase and normalizing any irregular whitespaces. This cleaned text was put into a new column called "review_clean". After all this pre-processing, the 22,000 reviews were stratified and split 80% training and 20% testing, having the same class breakdown as the original dataset. After the splitting, a BERT tokenizer was applied to the training and testing dataset as in order to use the bert-base-uncased model, each token has to map to an ID that is in BERT's existing vocabulary base.

The training process demonstrates that as the training loss decreases, accuracy increases, as shown by the numbers on the right. The model uses cross-entropy loss as it is a classification task, which helps guide its learning through forward passes and backpropagation. The decreasing loss values indicate that the model is learning effectively during forward passes (to calculate loss) and backpropagation (to update weights). After fine-tuning, the model achieved a final test accuracy of 88%. The training logs show that our fine-tuning process improved the model's accuracy from 85.94% to 96.42%, achieving a final test accuracy of 88.07%.

Following preprocessing, we proceeded to train the BERT-based sentiment analysis model. We utilized the bert-base-uncased model as our foundation and fine-tuned it specifically for classifying reviews into three sentiment categories (negative, neutral, and positive). The model architecture consisted of the pre-trained BERT layers followed by a classification head that outputs probabilities for each sentiment class. For training configuration, we selected hyperparameters based on both best practices for BERT fine-tuning and preliminary experiments. We used a batch size of 16 reviews, which provided a good balance between memory efficiency and training stability. The learning rate was set to 2e-5, a conservative value recommended for fine-tuning pre-trained transformers. We employed the AdamW optimizer, which enhances the standard Adam optimizer with decoupled weight decay regularization, helping to prevent overfitting. Our sequence length was capped at 128 tokens, sufficient for capturing the content of most reviews without excessive computational overhead. The training process was conducted on a MacBook Pro with an M4 chip over 4 epochs. Each epoch involved multiple steps: tokenized reviews were fed into the model, cross-entropy loss was calculated between predicted and actual sentiment labels, gradients were computed through backpropagation, and model weights were updated using the optimizer. At the end of each epoch, we evaluated the model's performance on the training data.

5 Results

```
Epoch 1/4
Epoch 1: 100%|██████████████████████████████| 1100/1100 [21:45<00:00,  1.19s/it, loss=0.5450, acc=0.8594]
Epoch 1 - Loss: 0.3913, Accuracy: 0.8594, Time: 1305.90s

Epoch 2/4
Epoch 2: 100%|██████████████████████████████| 1100/1100 [21:36<00:00,  1.18s/it, loss=0.4316, acc=0.9071]
Epoch 2 - Loss: 0.2707, Accuracy: 0.9071, Time: 1296.26s

Epoch 3/4
Epoch 3: 100%|██████████████████████████████| 1100/1100 [21:28<00:00,  1.17s/it, loss=0.0360, acc=0.9403]
Epoch 3 - Loss: 0.1885, Accuracy: 0.9403, Time: 1288.21s

Epoch 4/4
Epoch 4: 100%|██████████████████████████████| 1100/1100 [21:33<00:00,  1.18s/it, loss=0.0035, acc=0.9642]
Epoch 4 - Loss: 0.1308, Accuracy: 0.9642, Time: 1293.08s

Training complete! Total time: 1h 26m 23s
Saved training stats plot to 'amazon_training_stats.png'

Evaluating model...
Evaluating: 100%|████████████████████████████████████████| 275/275 [01:11<00:00,  3.84it/s]

Accuracy: 0.8807

Confusion Matrix:
[[ 876  109   50]
 [  90  154  106]
 [  45  125 2844]]
```

Figure 2: Training Stats

The model demonstrated clear learning progression across the four epochs:

Epoch 1 showed promising initial results with an accuracy of 85.94% and a loss of 0.3913. This indicated that the pre-trained BERT model already provided strong linguistic understanding that transferred well to our sentiment classification task.

Epoch 2 brought substantial improvement, reaching 90.71% accuracy while reducing loss to 0.2707. The 4.77% accuracy gain suggested the model was effectively learning sentiment-specific features.

Epoch 3 continued the positive trend with 94.03% accuracy and 0.1885 loss. The consistent decrease in loss values confirmed that the model was successfully minimizing prediction errors.

By Epoch 4, the model achieved 96.42% training accuracy with a loss of 0.1308. This final epoch showed smaller improvements than previous ones, suggesting we were approaching convergence. The training process took a total of 1 hour, 26 minutes, and 23 seconds. Several optimization techniques contributed to this relatively efficient training time despite the computational demands of transformer models. We implemented gradient accumulation to reduce memory usage, employed mixed precision training to accelerate computations, and utilized efficient data loading configurations to minimize bottlenecks.

5.1 Model Performance

Classification Report:

	precision	recall	f1-score	support
negative	0.87	0.85	0.86	1035
neutral	0.40	0.44	0.42	350
positive	0.95	0.94	0.95	3014
accuracy			0.88	4399
macro avg	0.74	0.74	0.74	4399
weighted avg	0.88	0.88	0.88	4399

Figure 3: Classification Report

The model's overall classification performance is good, as seen in the test accuracy of 88.07% and significant gains from fine-tuning, where accuracy improved from 85.94% to 96.42% during training (Figure 2). This shows that the model effectively learned sentiment patterns using backpropagation and optimization. While the accuracy is high, a more detailed look at precision, recall, F1-score, and the confusion matrix reveals specific strengths and weaknesses in sentiment classification.

Positive sentiment classification is the model's strongest aspect, in which 94.4% (2844 out of 3014) of positive reviews were correctly classified. 4.1% (125 reviews) were mistakenly classified as neutral, and 1.5% (45) were mistakenly classified as negative. High precision (95%) for positive sentiment indicates that whenever the model classifies a review as positive, it is usually correct. In addition, the recall (94%) shows that the model captures nearly all actual positive reviews with low false negatives. This suggests that the model has learned unique positive sentiment patterns very well, possibly because there are ample positive reviews in the data set (68.5% of total data).

Negative sentiment classification is also good but not as precise as positive sentiment. The model was correct in classifying 84.6% (876 of 1035) of the negative reviews but incorrectly classified 10.5% (109 reviews) as neutral and 4.8% (50 reviews) as positive. The precision for negative sentiment is 87%, showing that most of the reviews the model classified as negative are negative, while the 85% recall suggests that the model gets most of the negative reviews but misses a few, assigning them to neutral. This suggests that milder negative reviews may not contain highly negative words and, therefore, be incorrectly labeled as neutral by the model.

Neutral classifying is the hardest task for the model, with the accurate classification of merely 44.0% (154 out of 350) of neutral reviews. 30.3% (106 reviews) were misclassified as positive, and 25.7% (90 reviews) were classified as negative. This low performance is also reflected in the low recall (44%) and precision (40%), with an F1-score of just 0.42, the lowest of all sentiment classes. The difficulty in classifying neutral reviews is likely explained by their linguistic closeness to positive and negative sentiment and their relative sparsity in the data set (just 7.97% of all reviews). The majority of the neutral reviews contain words that are either positive or negative, depending on the context, and therefore, they are harder to distinguish from the other two categories.

Figure 4: Confusion Matrix

The confusion matrix also highlights these patterns of classification. Neutral reviews recorded the greatest misclassification rate (56.0%), meaning a neutral review was more likely to be classified as either positive or negative than correctly classed. In comparison, 18.2% of the negative reviews were misclassified as having neutral sentiment. In contrast, positive reviews were least misclassified at a rate of 5.6%, upholding the faith of the model in classifying positive sentiment. This proves that when the model is uncertain, it will play it safe and classify a review as positive or negative but not neutral, assuming that neutral reviews lack strong sentiment indicators.

	negative	neutral	positive
negative	681	103	217
neutral	480	233	717
positive	65	53	2265

(a) Per-Class Performance

5.2 Key Challenges and Limitations

One of the major challenges in sentiment classification is class imbalance. The data is biased towards positive reviews (68.5%), with lesser counts of negative reviews (23.53%) and a much lesser fraction of neutral reviews (7.97%). This causes the model to become biased toward positive prediction, which is evident in the accuracy for positive sentiment being extremely high and the consistent misclassification of neutral reviews as positive. Because the model only sees neutral reviews less often during training, it has trouble identifying unique patterns for neutral sentiment.

Another significant challenge is neutral sentiment misclassification, as evident in this class's low precision and recall. The commonality of vocabulary among neutral, positive, and negative reviews creates ambiguity in classification. For example, a sentiment like "It works as expected" can be neutral but tagged positive, and "Nothing special, but not bad either" can be neutral but can be incorrectly tagged negative. The model cannot draw clean distinctions where there are no strong markers for sentiment.

Sarcasm and contextually complicated sentiments are also wonderful challenges. Sentimental models such as deep learning-based models like BERT rely strongly on word associations to determine sentiment. However, sarcasm and nuanced speech typically conflict with the meaning of literal words. For example, "Oh great, another product that broke in two days" contains the word "great," which is typically a positive sentiment word, but the sentiment overall is negative. The model cannot be excellent at correctly interpreting these situations without additional context-awareness mechanisms.

Additionally, short and long reviews impact classification accuracy. Short reviews provide less text for the model to analyze, which increases the chances of misclassification. For example, "Not bad" is a neutral phrase that could be misclassified as slightly positive. Longer reviews, however, provide more context and are easier to classify correctly. Since neutral reviews are probably shorter, this can make it more difficult for the model to identify them.

6 Conclusion

In conclusion, sentiment analysis offers a transformative tool for businesses aiming to understand and capitalize on customer feedback. By extracting nuanced emotional insights from vast amounts of review data, companies can tailor their marketing strategies, improve product offerings, and enhance overall customer service.

6.1 Future Improvements

Some of the following improvements could enhance performance. Firstly, addressing class imbalance by increasing the neutral reviews through data augmentation or resampling would balance the training process. Over-sampling of the neutral reviews or synthetic generation of neutral samples could make the model more adept at detecting patterns in this sentiment class.

Second, adding context-sensitive linguistic properties might improve classification performance. Negation-based preprocessing methods (e.g., "not bad" vs. "bad") and sarcasm identification algorithms might be applied to improve the model's ability to read challenging cases. Fine-tuning the BERT model using additional domain-specific sentiment data might also improve its ability to identify subtle sentiment cues.

A further strategy is to alter decision thresholds for sentiment labeling. Currently, the model could label a review positive if the probability exceeds a standard threshold (for example, 0.5). Adjusting the threshold for neutral so that only high-intensity positive and negative signals lead to classification in those bins may reduce misclassifying neutral reviews.

Lastly, employing confidence-based reclassification would further improve performance. If the model predicts a sentiment label with low confidence, there might be a second decision-making process to reclassify borderline cases. This could involve processing additional linguistic features or an ensemble method where multiple models contribute to the final decision.

6.2 Final Thoughts

The application of models like BERT in our study demonstrates that advanced deep learning techniques can effectively distinguish between positive, negative, and neutral sentiments, even when faced with the intricacies of human language. In the end, we resulted with a final test accuracy of 88.07% and notable improvements in training performance, underscoring the potential of fine-tuning pretrained language models to capture subtle sentiment cues and inform decision-making processes in real-time.

However, several challenges remain inherent to the deployment of sentiment analysis systems. One major issue is class imbalance, as shown by the disproportionate representation of sentiment classes in our dataset; with mostly positive reviews dominating, the model often struggles to accurately identify neutral sentiments. The misclassification of neutral reviews which stem from their linguistic similarity to both positive and negative expressions, illustrates the difficulty in differentiating subtle shades of opinion. Additionally, the presence of sarcasm, contextual ambiguities, and colloquial language further complicates sentiment extraction, requiring continual updates and the integration of context-aware mechanisms.

Moving forward, several strategies can be adopted to enhance our model's performance and robustness. Addressing the class imbalance through data augmentation or oversampling of neutral reviews could provide the model with a more balanced perspective and improve its ability to discern subtle sentiment differences. Additionally, refining preprocessing methods to better handle linguistic complexity could further boost classification accuracy. Experimenting with a multitude of approaches or adjusting decision thresholds may also help mitigate the misclassification of borderline cases. Ultimately, continual refinement and adaptation of the model, in response to evolving language and customer behavior, will be key to leveraging sentiment analysis for actionable business insights. Overall, while sentiment analysis has proven to be a valuable asset for businesses, addressing these challenges is essential for enhancing the accuracy and reliability of sentiment-driven insights in dynamic, real-world environments.

7 Sentiment Analysis Tool

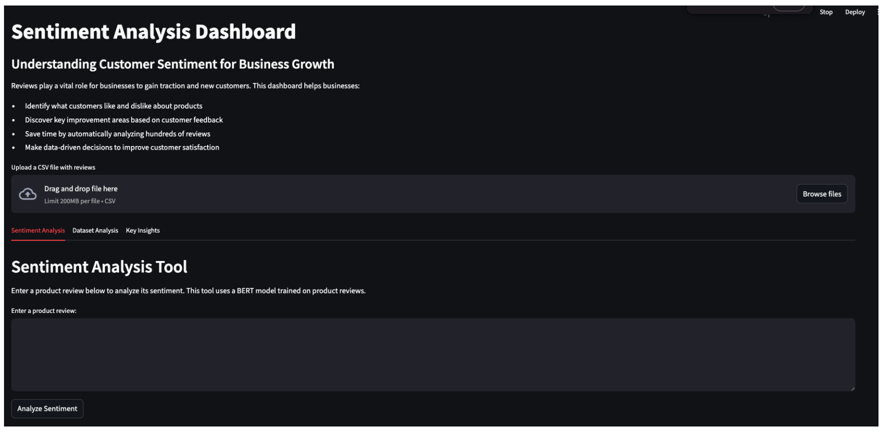

Figure 6: Dashboard of Sentiment Analysis Tool

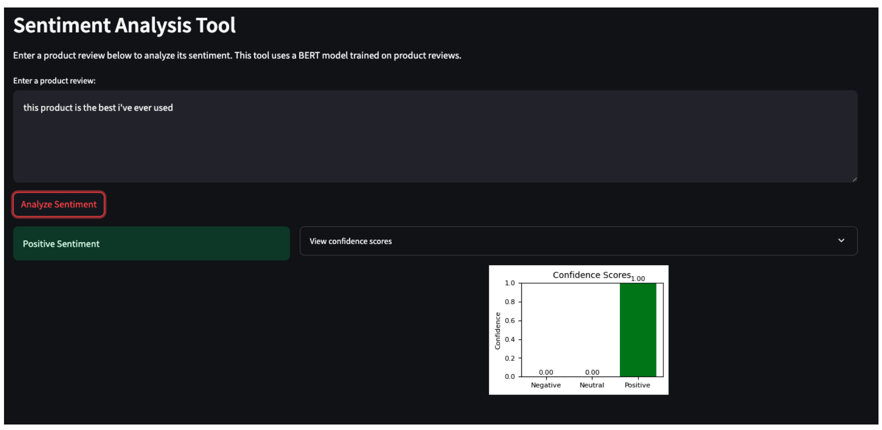

Figure 7: Positive Review Input on the Sentiment Analysis Tool

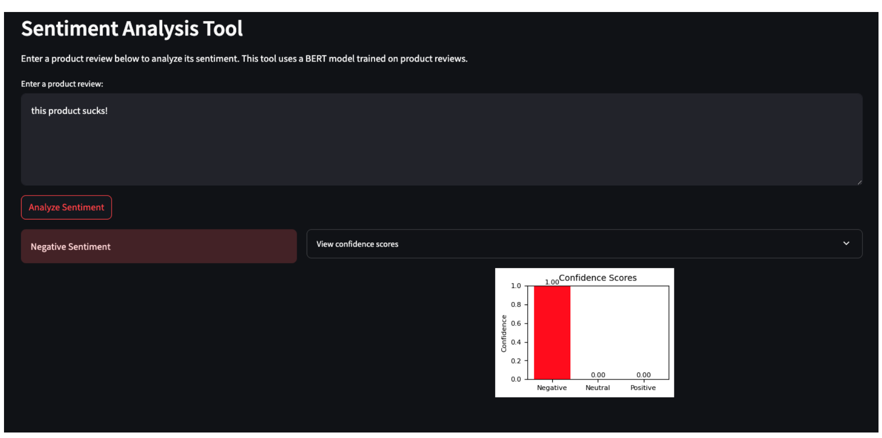

Figure 8: Negative Review Input on the Sentiment Analysis Tool

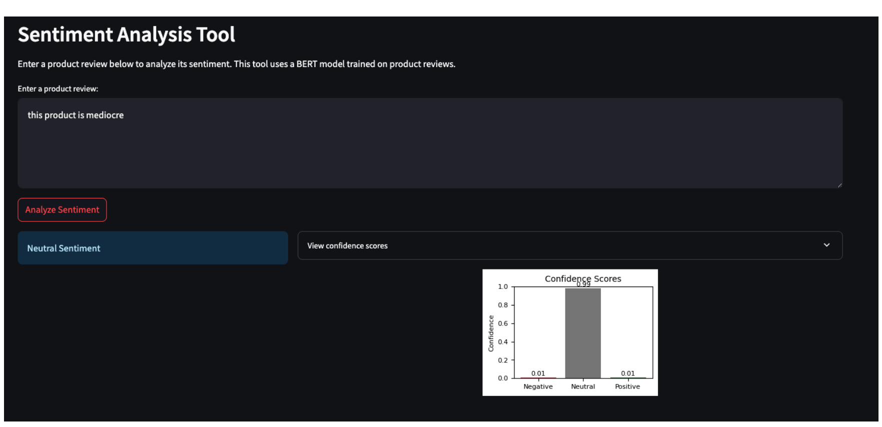

Figure 9: Neutral Review Input on the Sentiment Analysis Tool

8 Project Roadmap

Week 1: February 3, 2025

The research problem is to develop an accurate sentiment analysis model that categorizes Amazon customer reviews into positive, neutral, and negative sentiments using a large-scale dataset. The challenge is to effectively handle real-world data with noise and imbalanced classes while leveraging state-of-the-art techniques such as BERT. Research questions include: What text preprocessing methods best enhance model performance? How can we improve classification accuracy when class distributions are uneven? This phase sets the overall project scope, outlines expected outcomes (e.g., achieving high accuracy and robustness), and defines key performance indicators.

Week 2: February 10, 2025

The background study involves a systematic review of existing literature on sentiment analysis and BERT-based text classification. This review will cover seminal research papers and recent case studies that address similar challenges in analyzing customer reviews. The goal is to identify best practices, common pitfalls, and gaps in current methodologies. Insights from this review will inform the design of our preprocessing strategies, model architecture, and evaluation techniques, ensuring that our approach is grounded in established research while addressing identified limitations.

Week 3: February 17, 2025

The dataset exploration and understanding phase focuses on a dataset of 400,000 Amazon reviews. The dataset includes columns like Reviews, Ratings, Product Name, Brand Name, Price, and Review Votes. Key tasks include cleaning the data by dropping rows with missing reviews or ratings, removing special characters, normalizing text (such as converting to lowercase and correcting irregular whitespaces), and creating a new Sentiment column based on ratings. This column is further converted into numerical values (sentiment labels). Stratified sampling will be used to extract 22,000 reviews while maintaining the original sentiment distribution (approximately 68.50% positive, 23.53% negative, and 7.97% neutral). Exploratory data analysis (EDA) will be performed to generate summary statistics and visualizations that uncover underlying patterns in the data.

Week 4: February 24, 2025

The goal of this phase is to develop an accurate prediction model using a BERT-based approach. The process starts with preparing the training pipeline where the cleaned text data (review_clean) and numerical sentiment labels (sentiment_label) are used. The text is tokenized using the BERT tokenizer to map tokens to the BERT vocabulary. The model is then fine-tuned through forward passes and backpropagation while monitoring cross-entropy loss and accuracy improvements. Detailed training logs will document how the model's performance evolves, aiming for robust classification that accurately predicts sentiment even when faced with the inherent noise of real-world text data.

Week 5: March 3, 2025

The evaluation phase involves a rigorous assessment of the trained model's performance. Using a reserved test set from the 80/20 stratified split, the model will be evaluated on metrics such as accuracy, precision, recall, F1 score, and a confusion matrix. Visualizations, including a heatmap of the confusion matrix and plots showing the distribution of actual versus predicted sentiments, will be generated. The evaluation process will also merge predictions with metadata like product IDs and review dates to analyze sentiment trends over time. This comprehensive testing is designed to identify strengths and areas for further improvement in the model.

Week 6: March 10, 2025

The final phase is the development of a web-based front-end that makes the sentiment analysis model accessible to end users. This involves designing a user interface where users can input new review text, trigger the model, and view prediction results along with confidence scores and visual insights. The interface will be integrated with the backend model to ensure smooth and responsive operation. User acceptance testing will be performed to verify that the application is intuitive and meets the project's functional requirements, thereby finalizing the system for practical deployment.

References

[1] Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, *36*(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

[2] Devika, M., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: a comparative study on different approaches. *Procedia Computer Science*, *87*, 44–49. https://doi.org/10.1016/j.procs.2016.05.124

[3] Rahulvks. (2017, February). Customer Review Analysis. *Kaggle.com*; Kaggle. https://www.kaggle.com/code/rahulvks/customer-review-analysis/input

[4] Horev, R. (2018, November 10). BERT Explained: State of the art language model for NLP. *Medium*; TDS Archive. https://medium.com/data-science/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270