Sentiment analysis of Amazon product reviews

Sharofiddin Juraev

September 10, 2023

1 Introduction

In e-commerce, understanding customer sentiment towards products post-purchase is crucial. This understanding can enhance customer satisfaction, offer sellers a clearer view of their product's reception. An additional challenge in e-commerce is the systematic organization and presentation of product reviews. Prioritizing misleading reviews can have financial implications and reduce customer trust.

1.1 The dataset

The dataset focuses on Amazon product ratings and reviews. Collected through web scraping using Selenium, a significant portion of these reviews pertains to electronics. After a rigorous data cleaning process, the dataset comprises over 82,000 balanced positive and negative reviews.

1.2 Why NLP models should be opted

Amazon product reviews are inherently unstructured and varied, making their sentiment analysis a challenging task. Natural Language Processing (NLP) models offer an advantage due to their ability to process and understand complex textual patterns and contexts. Moreover, NLP models can handle the intrinsic unorganized nature of these reviews, discerning subtle linguistic nuances that might be overlooked by traditional methods. Their adaptability ensures that domain-specific jargons and terminologies are effectively addressed. Consequently, for the sentiment analysis of such intricate datasets as Amazon reviews, NLP models present a methodologically sound approach.

2 Comparison of LSTM with GloVe Embeddings vs BERT Model for Sentiment Analysis on Amazon Product Reviews

Two state-of-the-art approaches for sentiment analysis on Amazon product reviews are evaluated: a Bidirectional LSTM model utilizing pre-trained GloVe word embeddings, and a BERT-based model. This analysis compares their performances and investigates reasons behind their respective efficacies.

2.1 LSTM with GloVe Embeddings Model:

- **Architecture**:
  - The model takes reviews as inputs, tokenized to sequences of length 1000.
  - Uses pre-trained GloVe embeddings to represent words, which provides semantic understanding based on prior training on large datasets.
  - A Bidirectional LSTM layer captures sequential patterns within the text.
  - Dropout is implemented to mitigate overfitting.
  - Flattening and Dense layers finalize the architecture, leading to an output for binary classification.
- **Performance**:
  - Accuracy: 90.65%

- F1 Score: 90.68%
- ROC AUC Score: 96.80%
- Cross-Entropy Loss: 0.2298

2.2 BERT Model:

- **Architecture**:
  - Utilizes the BERT model for sequence classification.
  - First 9 layers of BERT are frozen, ensuring faster training and leveraging pre-trained features.
  - A custom classifier with dropout and L2 regularization is added to the BERT model, fine-tuning for the specific sentiment analysis task.
  - Early stopping is utilized during training to prevent overfitting.
- **Performance**:
  - Accuracy: 90.17%
  - F1 Score: 90.15%
  - ROC AUC Score: 96.66%
  - Cross-Entropy Loss: 0.2553

2.3 Comparative Analysis:

- **Performance**: Both models achieve over 90% accuracy, but the LSTM with GloVe embeddings slightly outperforms the BERT model in most metrics. The difference, however, is marginal.
- **Efficiency & Complexity**:
  - The LSTM model with GloVe embeddings might be simpler and faster to train given the lighter architecture and use of pre-trained embeddings.
  - BERT, while robust and versatile, requires more computational power and memory due to its deeper architecture.
- **Generalizability**:
  - GloVe embeddings are based on general contexts from Twitter, potentially capturing a broad sense of semantics.
  - BERT, being pre-trained on vast text data, might be more adaptable to diverse contexts but requires fine-tuning for specific tasks.

2.4 Why These Architectures Suit the Sentiment Analysis Task?

1. **Semantic Understanding**: Both GloVe embeddings and BERT provide deep semantic understanding of text. This is crucial for sentiment analysis, where the difference between positive and negative can hinge on nuances.
2. **Context Capture**: Bidirectional LSTMs and BERT both understand the context by considering words/tokens from both directions. Amazon reviews can be long and context-heavy, making this feature essential.
3. **Regularization and Overfitting**: Both models incorporate techniques (dropout and L2 regularization) to counteract overfitting. Given the diverse nature of Amazon reviews, this ensures the model generalizes well.
4. **Fine-tuning**: While the LSTM model is built and trained from scratch for the task, BERT is fine-tuned. Fine-tuning a pre-trained BERT model can lead to better results faster, as the model already has a broad understanding of language.

2.5 Conclusions

Both models offer strong performance for sentiment analysis on Amazon product reviews. The LSTM with GloVe embeddings has a slight edge in terms of performance metrics, possibly due to its architecture specifically designed for this task and the rich semantics captured by GloVe embeddings. We have chosen intentionally GloVe embeddings based on Twitter context. However, the difference in results is minimal. As our computational power is limited, the LSTM with GloVe embeddings are preferred. Conversely, for diverse contexts and a wider range of NLP tasks, fine-tuning a BERT model could be advantageous.

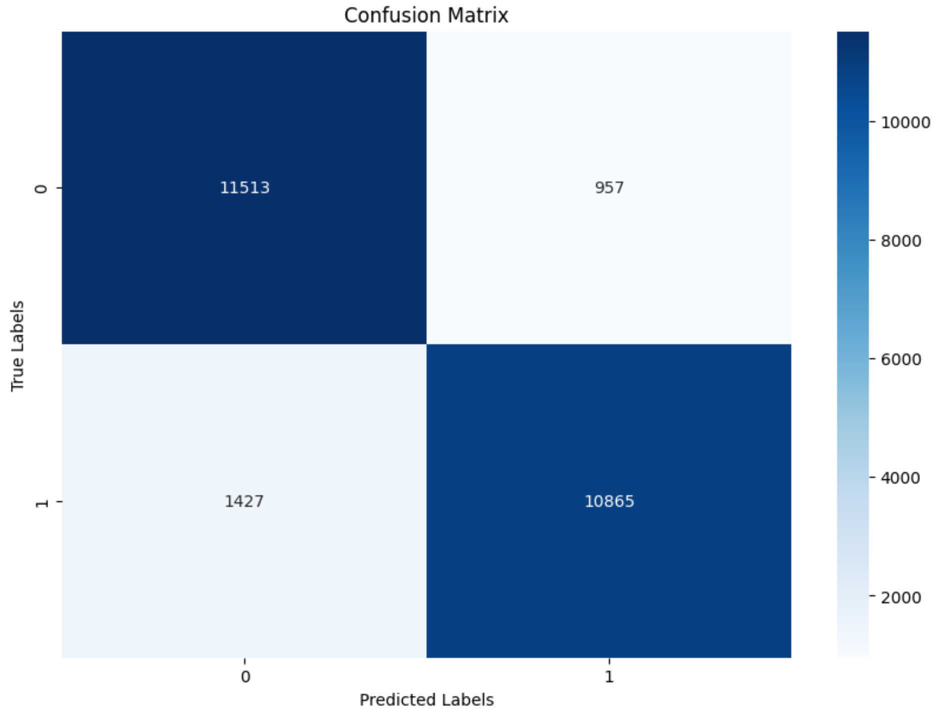

Figure 1: Confusion matrix for LSTM model.

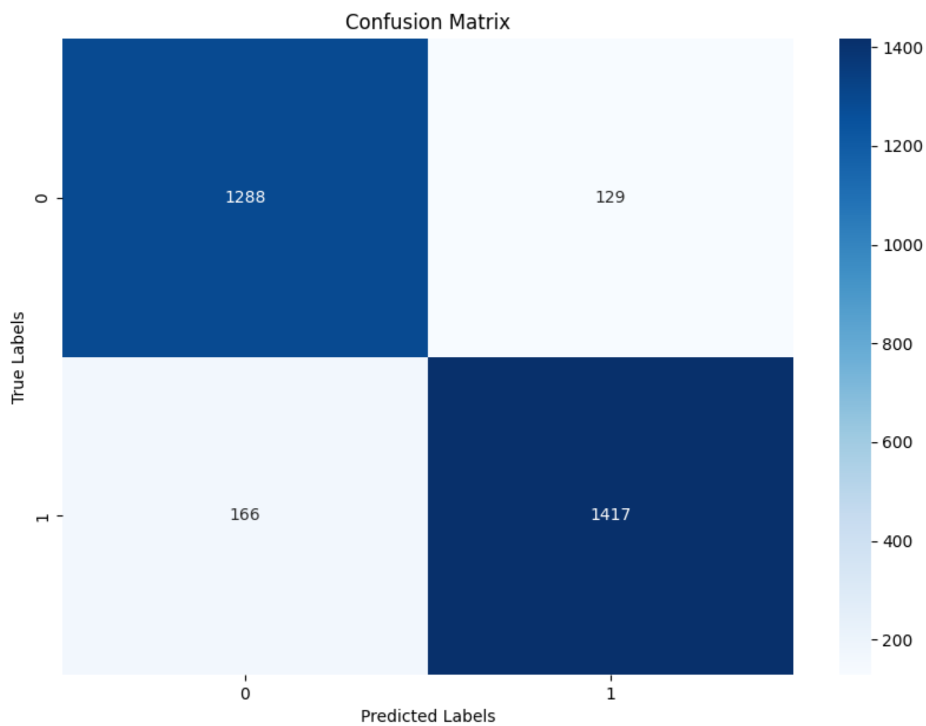

Figure 2: Confusion matrix for BERT model (Note that the model only run on 10,000 reviews as we had limited computing power).