Sentiment Analysis Quality Report
Based on iHerb Data

Nov. 13, 2024

Introduction

The passionflower team has successfully completed web scraping on the iHerb platform, collecting 163,497 product reviews from 299 medicinal food products containing the ingredient passionflower. As the first stage of the task, our team conducted sentiment analysis using the VADER model to predict the sentiment of the review data. This report focuses on the next step: applying the BERT sentiment analysis model to predict sentiment and comparing its performance against VADER. The primary goal of this task is to ensure the sentiment results are accurate enough to support further analysis and insights.

Background and Model Selection

To identify a suitable sentiment analysis model for this project, our team conducted an in-depth comparison of various sentiment analysis approaches, including dictionary-based models (e.g., VADER), traditional machine learning models (e.g. Naïve Bayes), and deep learning models (e.g. BERT and RNN).

Model Name	Model Type	Accuracy (based on Yelp comment data)
NLTK VADER	Dictionary	0.73
Naïve Bayes	Machine Learning	0.67
HuggingFace BERT	Deep Learning	0.96
Flair RNN	Deep Learning	0.96
Flair DistilBERT	Deep Learning	0.97

Table 1: Summary of model performances on Yelp data

The research revealed that deep learning models consistently outperformed dictionary-based and traditional machine learning models in terms of accuracy and their ability to handle complex sentence structures (Fesenko, 2023).

Given these findings, our selected a fine-tuned BERT model trained on Amazon product reviews as the second model for sentiment prediction. This choice leverages BERT's contextual understanding and its proven performance in analyzing product review datasets similar to ours.

Methodology

For sentiment prediction, our team selected a fine-tuned RoBERTa model (which is a family of BERT models) trained on the Amazon reviews, available from HuggingFace. The model was fine-tuned for binary sentiment classification (positive and negative) by adding a classification head.

To evaluate and compare the performance of RoBERTa and VADER, we analyzed reviews with differing sentiment predictions. Specifically, reviews predicted as positive by one model but negative by the other were extracted for further inspection to understand the discrepancies.

Since the explicit sentiment were not available from web scraping, our team manually labeled a subset of 1,000 reviews to serve as a reliable ground truth for evaluating model performance. For instance, a review expressing sentiments such as "*the product helps me sleep*" or "*I can sleep well now*" was labeled as positive, and a review stating "*does not work*" or similar will be labeled as negative.

Using this labeled dataset, we evaluated and compared the accuracy and performance of the two models, identifying their respective strengths and weaknesses.

Results

Our team observed that approximately 11% of predictions between the two models were not coherent. Figure 1 presents the confusion matrix, summarizing the classification results from two models.

The VADER model correctly classified 33 reviews as negative but misclassified 382 as positive out of 415 negative reviews. For the 585 positive reviews, it identified 201 reviews as positive, resulting in an overall accuracy of 23.4%. In contrast, the BERT model successfully identified 382 of the 415 negative reviews and correctly identified 384 of the 585 positive reviews, achieving a higher accuracy of 76.6%.

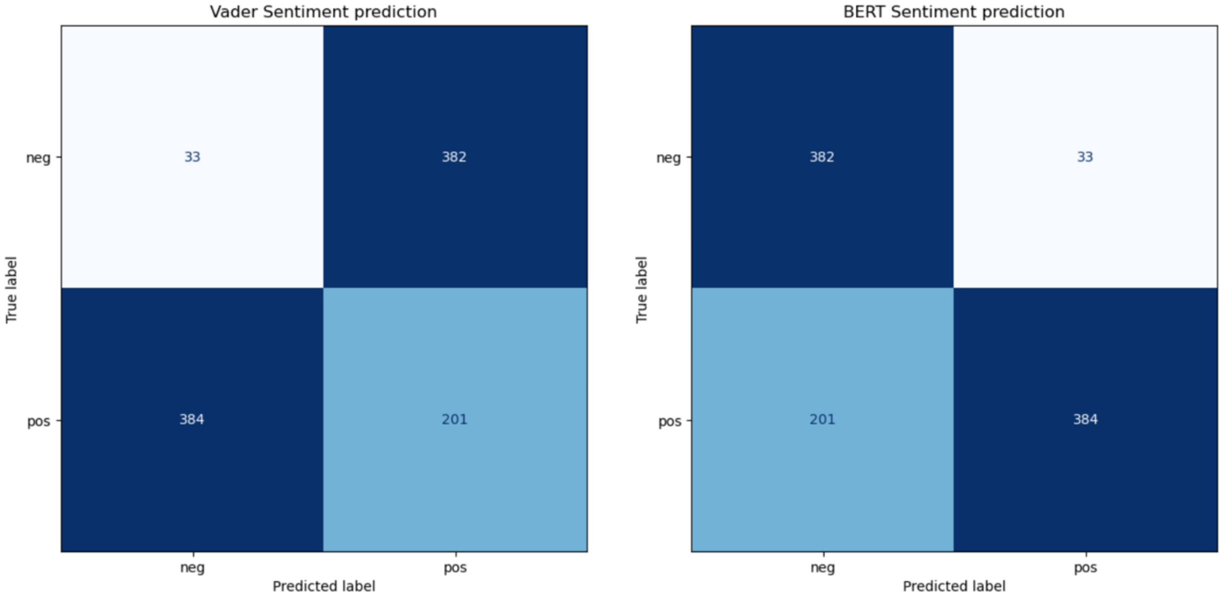

Figure 1: Confusion Matrices of the VADER and BERT model.

Model	Performance (Accuracy)
VADER	23.4%
BERT	76.6%

Table 2: Model performances on iHerb product reviews data

Discussion

It is worth noting that the VADER model performs better at understanding shot sentences, while the BERT model is more accurate in interpreting longer sentences. For example, the short phrase *"good stress complex"* is generally understood as referring to an effective stress-relief supplement and should be classified as positive sentiment. However, the BERT model incorrectly predicted it as negative sentiment, that is because the model is trained on general review data instead of medicinal food for stress relief. Conversely, for the longer sentence *"I bought it for an acquaintance with the problem of falling asleep and waking up early. It helped a lot!"*, which clearly indicates helping someone improve their sleep problems, the VADER model predicts it as negative sentiment, revealing its limitations in understanding complex sentence structures.

Conclusion

The above analysis demonstrates that deep learning models, such as BERT, generally perform better in sentiment classification tasks, particularly in understanding longer sentences.

However, the current challenge lies in the absence of a model specifically tailored to our research focus: the stress-relief effects of medicinal food. This limitation results in the model misinterpreting certain content. For instance, a product that helps someone fall asleep might be considered a negative sentiment in the context of movie reviews, but in our study, it should be classified as positive. Therefore, if a more suitable model becomes available, we are open to adopting it and improving the accuracy of our sentiment predictions.

References

Fesenko, P. (2023, October 5). *Best open-source models for sentiment analysis-part 1: Dictionary models*. Medium. https://medium.com/@pavlo.fesenko/best-open-source-models-for-sentiment-analysis-part-1-dictionary-models-ece79e617653

Fine-tuned RoBERTa for Sentiment Analysis on Reviews. AnkitAI/reviews-roberta-base-sentiment-analysis · Hugging Face. (n.d.). https://huggingface.co/AnkitAI/reviews-roberta-base-sentiment-analysis