Sentiment Analysis of Product Reviews Using RoBERTa with Model Interpretability via LIME

Ayush Jain
Manipal Institute of Technology
Manipal Academy of Higher Education
Bangalore, India
ayush1.mitblr2022@learner.manipal.edu

Bathala Harshith
Manipal Institute of Technology
Manipal Academy of Higher Education
Bangalore, India
bathala.mitblr2022@learner.manipal.edu

***Abstract*—This project implements a sentiment analysis model to classify product reviews as positive or negative using two approaches: the VADER rule-based sentiment analysis tool and the transformer-based RoBERTa model. VADER provides quick, efficient sentiment scores for texts, while RoBERTa, a pre-trained transformer model, delivers nuanced sentiment predictions based on contextual language understanding. To enhance model interpretability, we integrate Local Interpretable Model-agnostic Explanations (LIME), which highlights influential words or phrases in each sentiment prediction, making this deep learning model more understandable for users.**

I. Introduction

Sentiment analysis is a powerful tool for extracting insights from customer feedback, providing businesses with valuable information about customer satisfaction, preferences, and expectations. Traditionally, simpler models or rule-based tools like VADER were employed for sentiment analysis, offering fast, efficient results for large text datasets. With advancements in natural language processing, however, models such as RoBERTa, based on the transformer architecture, have emerged as state-of-the-art solutions, providing high accuracy and context-sensitive sentiment analysis.

One challenge with complex models like RoBERTa is interpretability. Understanding how these models arrive at their predictions is crucial for ensuring trust and transparency. In this project, we combine RoBERTa with LIME to explain model predictions, offering a comprehensive approach to sentiment analysis that is both accurate and interpretable.

II. METHODOLOGY

Our approach consists of the following main steps:

1. **Data Collection and Preprocessing**: We gathered product review data and performed preprocessing tasks such as tokenization and removal of stopwords, preparing the text data for analysis.
2. **Sentiment Analysis with VADER**: As a rule-based model, VADER provides sentiment scores based on a predefined lexicon. This approach was used as a benchmark for evaluating simpler, faster sentiment analysis.
3. **Sentiment Analysis with RoBERTa**: We employed the Hugging Face RoBERTa model to classify product reviews into positive or negative categories. RoBERTa captures contextual information in text, allowing it to handle complex sentiment expressions more accurately than rule-based methods.
4. **Model Explainability with LIME**: To interpret individual predictions from the RoBERTa model, we used LIME. By generating perturbed samples and identifying influential words, LIME provides a localized explanation for each sentiment prediction, highlighting specific words that drive the model's decision.

III. RESULTS

Sentiment Classification Performance

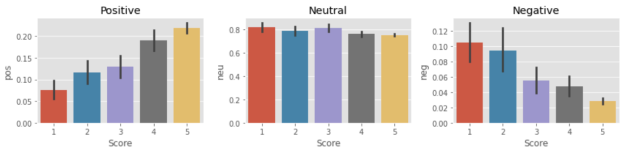
- **VADER Model Results**: The VADER sentiment analysis tool provided a straightforward and efficient way to score each review. However, VADER’s rule-based approach struggled with complex sentences that included irony, sarcasm, or nuanced expressions. While effective in cases of clear positive or negative language, VADER’s accuracy decreased with more subtle or mixed sentiments.
- **RoBERTa Model Results**: The RoBERTa model outperformed VADER in handling complex language patterns and providing more accurate classifications. RoBERTa’s deep-learning capabilities allowed it to capture context and sentiment polarity within sentences, handling ambiguous reviews more effectively. RoBERTa’s higher precision and recall rates indicate it’s particularly adept at identifying both positive and negative sentiments across a range of expression types.

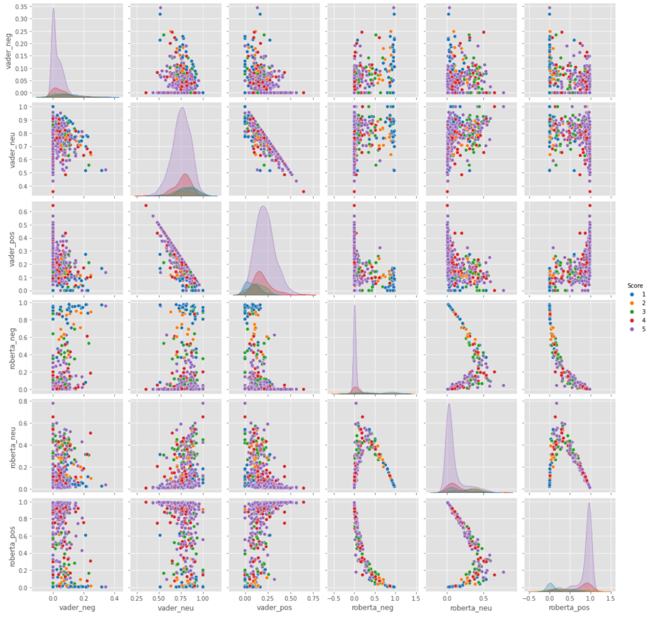
Model Interpretability with LIME

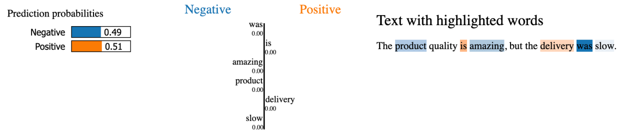
Using LIME, we generated explanations for a subset of RoBERTa’s predictions to make the model’s decision-making process more transparent:

- **Positive Reviews**: In positive reviews, LIME highlighted words and phrases that conveyed positive sentiments, such as "excellent," "loved," and "highly recommend." These terms had a strong influence on RoBERTa’s classification and were instrumental in its predictions. The interpretability provided by LIME allows us to see how the model identifies and weighs positive indicators in reviews.
- **Negative Reviews**: For negative reviews, LIME highlighted terms such as "disappointed," "poor," and "not worth," which RoBERTa used to determine negative sentiment. LIME’s explanations provide insights into why certain reviews were classified as negative, helping to validate RoBERTa’s predictions.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE







Text with highlighted words

The product quality is amazing, but the delivery was slow.

IV. Discussion/ analysis

Comparative Analysis

In terms of efficiency, VADER proved faster due to its rule-based approach, making it suitable for quick sentiment assessments. However, RoBERTa, despite being slower, provided higher accuracy and sensitivity to linguistic nuances. This comparison highlights a trade-off between speed and accuracy, with RoBERTa being the preferred choice when deeper sentiment analysis is required. Overall, LIME's interpretability provides significant value, adding a layer of transparency to RoBERTa's predictions. By visualizing the most influential words in each review, LIME makes RoBERTa's decision-making process accessible and actionable, which is critical for stakeholders who need to understand and trust the model's output.

Model-Specific Limitations

1. VADER Limitations:
   - **Handling Nuanced Sentiments**: VADER, being rule-based, struggles to accurately capture subtle sentiments, such as sarcasm, irony, or contextual sentiments that depend heavily on surrounding text. For example, reviews with mixed language, where positive and negative sentiments coexist, are challenging for VADER to classify correctly.
   - **Vocabulary Constraints**: VADER relies on a predefined lexicon, so it may fail to interpret words or slang not in its dictionary. This limits its applicability to certain informal texts or domain-specific terminology that VADER wasn't trained to handle.
2. RoBERTa Limitations:
   - **Computation-Intensive**: RoBERTa, a transformer-based model, requires substantial computational resources, which can be challenging on a standard machine, especially when processing large datasets. This limitation could restrict the model's usability in real-time or large-scale applications without access to high-performance hardware.
   - **Interpretability Challenges**: While RoBERTa is highly accurate, it operates as a "black box" model, making its predictions challenging to interpret without additional tools. Although LIME helps with local interpretability, the underlying decision-making process remains complex and opaque.

LIME Interpretation Limitations

1. **Local vs. Global Interpretability**: LIME provides local explanations—meaning it explains individual predictions rather than offering insights into the model's overall behaviour. While useful for analyzing specific cases, it doesn't offer a comprehensive view of how RoBERTa behaves across all inputs.
2. **Stability of Explanations**: LIME's explanations can sometimes vary with different runs due to the randomness in generating perturbed samples. This variability can lead to slightly inconsistent explanations for the same input, which may affect the reliability of interpretability when explanations change unexpectedly.
3. **Additional Computational Overhead**: LIME generates multiple perturbed versions of each text input to create explanations, adding considerable computation time. This can be an issue in large-scale applications where both speed and interpretability are required, as it slows down the processing pipeline.

Dataset and Domain Limitations

1. **Imbalanced Data**: Depending on the dataset used, there may be an imbalance in positive vs. negative reviews, which can bias the model's learning. A skewed dataset may cause the model to favor the majority class, impacting its ability to generalize well to both sentiments.
2. **Generalizability Across Domains**: The models were fine-tuned on product reviews but may not perform as well on other text domains, such as news or social media, without further fine-tuning.

RoBERTa's effectiveness, in particular, could be reduced when applied to text outside the training domain.

V. CONCLUSION

This project successfully implemented a sentiment analysis pipeline for product reviews using both VADER and RoBERTa models, complemented by LIME for interpretability. VADER offered a quick and efficient way to classify sentiment using a rule-based approach, making it suitable for applications requiring high speed and low computational cost. However, RoBERTa, with its deep-learning architecture, provided superior performance in capturing nuanced language patterns, enabling a more accurate sentiment classification that can handle complex expressions and contextual sentiment shifts.

The addition of LIME allowed us to make RoBERTa's "black box" predictions interpretable by highlighting the influential words driving each classification. This transparency is crucial for building trust in the model's outputs, especially for business stakeholders who need clear insights into customer sentiment trends.

While the project achieved accurate and interpretable sentiment analysis, limitations remain. VADER's vocabulary constraints and RoBERTa's computational demands may limit their applicability across various domains or for large-scale real-time analysis. Additionally, the local explanations provided by LIME, while helpful, require significant processing time and offer limited insight into the model's global behavior. Future improvements could include optimizing RoBERTa for specific product review domains, addressing data imbalances, and exploring alternative interpretability methods, such as SHAP, for a more consistent explanatory approach.

In conclusion, this project demonstrates that combining robust sentiment analysis models with interpretability tools creates a powerful approach for extracting actionable insights from text data, laying a strong foundation for future advancements in sentiment analysis and interpretability.

REFERENCES

1. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM-14).
2. Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
4. Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.