Vol.15, Issue No 2, 2025

Enhancing Transparency in Sentiment Analysis: A Study of Explainable AI (XAI) Techniques in NLP

Aman Jha,
UG Student,
Department of CSE,
St. Martin's Engineering College,
amanjha1655@gmail.com

Ms. M. Maneesha,
Assistant Professor,
Department of CSE,
St. Martin's Engineering College,
Secunderabad, Telangana, India,
mmaneeshacse@smec.ac.in

Abstract: ***Sentiment analysis is a key Natural Language Processing (NLP) task that enables machines to interpret human emotions from text. However, deep learning models used in sentiment classification often function as "black boxes," making it difficult to understand their decision-making processes. Explainable AI (XAI) techniques aim to bridge this gap by providing insights into model predictions and enhancing transparency and trust. This paper explores the application of XAI techniques—SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model—Agnostic Explanations), and counterfactual explanations—to improve interpretability in sentiment analysis. As an example, we utilize deepseek-r1:1.5B, a lightweight yet powerful language model, along with the Twitter Entity Sentiment Analysis dataset to demonstrate how these methods uncover key factors influencing sentiment classification. We evaluate each XAI approach in terms of interpretability, computational efficiency, and real-world applicability. Our findings highlight the importance of explainability in NLP applications, particularly for AI-driven assistants, where transparency is essential for reliability. This study contributes to the growing need for interpretable AI, providing a practical framework for improving trust and accountability in automated sentiment analysis.***

Keywords: Explainable AI (XAI), Sentiment Analysis, SHAP, LIME, Counterfactual Explanations, Natural Language Processing (NLP), Model Interpretability

I. Introduction

Artificial intelligence (AI) has become an integral part of decision-making systems, with applications spanning healthcare, finance, social media analysis, and customer service. One critical area where AI is widely used is **sentiment analysis**, a Natural Language Processing (NLP) task that determines the emotional tone of textual data. Sentiment analysis plays a key role in **social media monitoring, brand reputation analysis, customer feedback processing, and automated assistants**. However, despite their impressive accuracy, modern deep learning models used in sentiment analysis often function as **black boxes**, offering little to no insight into how predictions are made. This lack of transparency raises concerns regarding **trust, accountability, and bias** in AI-driven decision-making systems [1], [2].

Explainable AI (XAI) addresses this challenge by providing methods that **make AI models more interpretable and transparent**. Techniques such as **SHAP (SHapley Additive Explanations) [2], LIME (Local Interpretable Model- Agnostic Explanations) [1], and counterfactual explanations [3]** help uncover how sentiment classification models derive their predictions. Understanding model behavior is particularly crucial for AI-driven assistants, where reliability and fairness are essential for building user trust.

This paper explores the application of **XAI techniques to sentiment analysis models**, focusing on their effectiveness in improving interpretability without compromising predictive performance. As an example, we use **deepseek-r1:1.5B**, a compact and efficient language model, along with the **Twitter Entity Sentiment Analysis** dataset [10]. The study evaluates

different XAI approaches based on **interpretability, computational efficiency, and real-world applicability**. Our goal is to highlight the strengths and limitations of each method, ultimately contributing to the development of more **transparent and trustworthy AI models** in NLP.

The remainder of this paper is structured as follows: **Section II** discusses related work in XAI for NLP, providing a review of existing techniques. **Section III** presents the methodology, detailing the dataset, model selection, and XAI techniques applied. **Section IV** showcases the implementation and results, comparing different explainability approaches. **Section V** discusses the insights gained, challenges encountered, and practical implications of XAI in sentiment analysis. Finally, **Section VI** concludes the paper and outlines directions for future research.

II. Related Work

The field of Explainable AI (XAI) has gained significant attention as AI models become more complex and widely deployed in real-world applications. In the context of **Natural Language Processing (NLP) and sentiment analysis**, researchers have explored various techniques to enhance model interpretability, ensuring transparency and trust in AI-driven decision-making [4].

A. Explainability in AI and NLP

Several studies have focused on the importance of **interpretability in deep learning models**, particularly in high-stakes domains such as healthcare, finance, and legal applications. Ribeiro et al. [1] introduced **LIME (Local Interpretable Model-Agnostic Explanations)** as a technique to approximate black-box models with simpler, interpretable models, allowing local explanations of predictions. Lundberg and Lee [2] developed **SHAP (SHapley Additive Explanations)**, based on cooperative game theory, to assign feature importance scores that provide both **global and local interpretability**.

In NLP, explainability techniques have been applied to tasks such as **text classification, sentiment analysis, and language modeling**. Arras et al. [4] demonstrated how **Layer-wise Relevance propagation (LRP)** can be used to interpret neural networks for text classification. Jain and Wallace [5] critically analyzed **attention mechanisms** in transformers, questioning whether attention weights alone provide meaningful explanations.

B. Counterfactual Explanations in Sentiment Analysis

Counterfactual explanations, as introduced by Wachter et al. [3], have emerged as an alternative to feature attribution methods like SHAP and LIME. These explanations focus on **identifying minimal changes in input text that would alter the model's prediction**, providing actionable insights for users. For instance, in sentiment analysis, a counterfactual explanation might highlight that a **single word change** (e.g., "terrible" to "amazing") could shift the sentiment classification. Recent studies have proposed counterfactual generation techniques for NLP, including **perturbation-based approaches** and **causal inference methods** [6].

C. Explainability Libraries and Tools

To facilitate the implementation of explainability techniques, several open-source libraries have been developed:

- **Alibi**: Provides LIME, Anchors, and counterfactual generators for NLP tasks [8].
- **InterpretML**: Supports Explainable Boosting Machines (EBMs) and SHAP for text models [9].
- **AI Explainability 360 (AIX360)**: Developed by IBM, offering a broad range of XAI methods for various AI applications [7].

These tools have made explainability more accessible for researchers and practitioners, enabling **easier integration of XAI techniques into NLP workflows**.

D. Gap in Existing Research

While LIME, SHAP, and counterfactual explanations have been widely studied, their effectiveness in **real-world sentiment analysis models** remains an open challenge. Most studies focus on **generic classification tasks**, leaving a gap in **comprehensive evaluations of explainability methods for NLP applications**. Additionally, research is still evolving in **evaluating**

explanation quality, computational efficiency, and usability in practical AI deployments [6].

This paper builds upon prior research by applying and comparing **SHAP, LIME, and counterfactual explanations** specifically in **sentiment analysis**, using **deepseek-r1:1.5B** as an example model. The study aims to assess the effectiveness of these techniques in providing meaningful explanations and improving model transparency.

III. Methodology

This section outlines the approach taken to evaluate explainability techniques in sentiment analysis. We describe the dataset used, the model selection process, and the implementation of explainability methods, including **SHAP, LIME, and counterfactual explanations**.

A. Dataset Selection

For this study, we use the **Twitter Entity Sentiment Analysis** dataset from Kaggle. This dataset consists of tweets labeled with sentiment categories (positive, negative, neutral), making it suitable for training and evaluating NLP models.

1) Data Preprocessing

To prepare the dataset for training, we perform the following preprocessing steps:

- **Text Cleaning**: Removing special characters, links, and unnecessary symbols.
- **Tokenization**: Converting text into word or subword tokens.
- **Label Encoding**: Mapping sentiment categories to numerical values.
- **Train-Test Split**: Dividing data into training (80%) and validation (20%) sets.

B. Model Selection

We use **deepseek-r1:1.5B**, a lightweight language model optimized for efficiency, as our sentiment classifier. The model selection is based on two possible approaches:

1) Fine-Tuning with AutoTrain

- The model is fine-tuned on the Twitter dataset using **Hugging Face AutoTrain** to optimize sentiment classification.
- Training parameters include batch size tuning, learning rate adjustments, and early stopping.

2) Direct Inference Using Pretrained Model

- Alternatively, the pretrained **deepseek-r1:1.5B** is used as a zero-shot or few-shot learner via **Ollama**, without additional training.
- The model classifies sentiment based on contextual understanding of text inputs.

The choice between these approaches depends on computational constraints and the need for domain-specific adaptation.

C. Explainability Techniques

To interpret model predictions, we apply the following explainability methods:

1) SHAP (SHapley Additive Explanations)

- Computes **feature importance scores** for each word in a tweet.
- Identifies the **most influential words** contributing to sentiment classification.
- Generates **SHAP value visualizations** to highlight positive and negative contributions.

2) LIME (Local Interpretable Model-agnostic Explanations)

- Creates **perturbed versions** of input text and trains a simpler interpretable model.

- Explains sentiment prediction at an **instance level** by showing word importance.
- Provides **bar charts and weights** for interpretability.

3) Counterfactual Explanations

- Identifies **minimal text changes** that would alter the sentiment classification.
- Example: "The product is **terrible**" (negative) → "The product is **great**" (positive).
- Helps users understand **how to change outcomes** based on textual adjustments.

D. Evaluation Criteria

The effectiveness of each XAI technique is evaluated based on

- **Interpretability**: How easily humans can understand explanations.
- **Computational Efficiency**: Execution time and resource consumption.
- **Actionability**: Whether explanations provide meaningful, actionable insights.

This methodology ensures a comprehensive assessment of **XAI in sentiment analysis**, bridging the gap between model performance and interpretability.

IV. Implementation & Results

This section details the implementation of sentiment analysis using **deepseek-r1:1.5B**, followed by the application of **SHAP, LIME, and counterfactual explanations**. The results of each explainability method are presented and compared based on their effectiveness in providing model transparency.

A. Sentiment Analysis Model Training & Inference

The sentiment classification model was implemented using two approaches:

1) Fine-Tuning deepseek-r1:1.5B using AutoTrain

The dataset was preprocessed as described in **Section III**, and the model was fine-tuned using **Hugging Face AutoTrain** with optimized parameters. After training, the model was evaluated on the test set, achieving a measurable level of accuracy and performance.

2) Zero-Shot Sentiment Analysis using deepseek-r1:1.5B via Ollama

Alternatively, the pretrained model was used for sentiment classification without additional training. The model inferred sentiment labels based on contextual understanding, and its performance was assessed by comparing predictions with labeled test data.

B. Explainability Methods & Results

Following sentiment classification, **XAI techniques** were applied to analyze model predictions.

1) SHAP (SHapley Additive Explanations) Results

SHAP was used to determine feature importance by analyzing how individual words contributed to sentiment predictions. The results indicated that sentiment-laden adjectives and modifiers had the most significant impact on model decisions. SHAP summary plots provided a global view of feature importance across multiple tweets, revealing consistent patterns in sentiment classification.

2) LIME (Local Interpretable Model-agnostic Explanations) Results

LIME was applied to generate instance-level explanations by perturbing input text and fitting an interpretable model. The results showed how individual words influenced sentiment classification on a case-by-case basis. LIME visualizations provided clear, interpretable weightings for words that contributed to positive or negative classifications, helping to explain specific predictions.

3) Counterfactual Explanations Results

Counterfactual explanations were generated to identify minimal text modifications required to change sentiment predictions. The results demonstrated how small alterations, such as changing sentiment-associated words, could flip the classification label. This method provided actionable insights for users by highlighting decision boundaries within the sentiment model.

C. Comparative Analysis of XAI Techniques

The effectiveness of **SHAP, LIME, and counterfactual explanations** was compared based on interpretability, computational efficiency, and practical application.

Method	Scope	Strengths	Weakness
SHAP	Global & Local	Detailed insights	Computationally heavy
LIME	Local	Fast & Simple	No global view
Counterf actuals	Local	Actionable results	Hard to generate

D. Summary of Findings

The comparative analysis of explainability techniques yielded the following insights:

- **SHAP provided the best global interpretability**, allowing an in-depth understanding of which features influenced sentiment classification across multiple samples. However, it required substantial computational resources.
- **LIME was effective for local interpretability**, generating fast and easy-to-understand explanations for individual predictions. However, it lacked insight into the model's overall decision-making behavior.
- **Counterfactual explanations were useful for understanding decision boundaries**, showing users how minimal text changes could alter sentiment classifications. This approach was particularly valuable in fairness and accountability applications.

These results demonstrate how **XAI techniques improve interpretability in sentiment analysis**, aiding in debugging, trust-building, and bias detection. By integrating explainability into sentiment analysis, AI-driven assistants can be made more transparent and reliable.

V. Discussion

The results from the explainability techniques applied to sentiment analysis reveal critical insights into how **SHAP, LIME, and counterfactual explanations** enhance transparency in machine learning models. This section discusses the strengths and limitations of these methods, their real-world applicability, and their role in improving trust in AI-driven sentiment analysis.

A. Strengths and Limitations of Explainability Techniques

Each explainability method presents unique advantages and challenges, making them suitable for different interpretability needs:

- **SHAP excels in global interpretability**, providing comprehensive insights into feature importance across multiple predictions [2]. However, its **computational cost** makes it impractical for large-scale real-time applications.
- **LIME offers quick, localized explanations**, making it a useful tool for analyzing individual predictions [1]. However, its **inability to provide global model behavior insights** limits its effectiveness in fully understanding the model.
- **Counterfactual explanations provide actionable insights**, enabling users to see what changes would alter a classification outcome [3]. However, **generating counterfactuals is computationally challenging**, requiring optimization techniques to maintain plausibility and realism.

While each technique has its limitations, their combined use allows for a more complete understanding of sentiment classification decisions.

B. Practical Implications for AI Assistants

For AI-driven assistants to be **reliable, trustworthy, and fair**, they must be able to **justify** their decisions in a human-understandable manner. This study highlights how explainability techniques can:

- **Improve User Trust:** By showing how predictions are made, users can better understand and trust AI-driven sentiment analysis [6].
- **Enhance Debugging & Model Improvements:** Developers can use explainability methods to identify biases, errors, or unexpected model behavior [4].
- **Support Ethical AI & Compliance:** Regulatory frameworks, such as GDPR, require AI systems to provide explanations for their decisions. Implementing XAI techniques ensures compliance with transparency requirements [3].

By incorporating **SHAP, LIME, and counterfactuals**, AI assistants can provide more transparent, justifiable responses, leading to improved adoption and trust.

C. Alternative XAI Frameworks for NLP

While this study directly applied **SHAP, LIME, and counterfactual explanations**, several frameworks provide pre-built implementations of these and other XAI methods:

- **Alibi**: Offers implementations for LIME, SHAP, Anchors, and counterfactual explanations, specifically designed for NLP and tabular data [8].
- **InterpretML**: Provides an easy-to-use interface for SHAP-based feature importance analysis and Explainable Boosting Machines (EBMs) [9].
- **AI Explainability 360 (AIX360)**: Developed by IBM, this framework includes rule-based explainers, surrogate models, and additional transparency tools for machine learning models [7].

These frameworks facilitate the integration of explainability techniques into real-world applications, **reducing the complexity** of implementing XAI from scratch. Future studies could explore their effectiveness in sentiment analysis and compare them against custom-built implementations.

D. Challenges in Applying Explainability to NLP

Despite the benefits of explainability techniques, several challenges remain in applying them effectively to **sentiment analysis and NLP models**:

- **Complexity of Text Representations:** Unlike tabular data, NLP models rely on contextual embeddings, making it difficult to isolate the exact contribution of individual words [5].
- **High Computational Costs:** SHAP and counterfactual methods require significant computational resources, limiting their feasibility for real-time applications [2].
- **Trade-Off Between Interpretability and Accuracy:** Simplifying explanations may lead to **loss of information**, reducing the effectiveness of certain techniques [6].
- **Human-Centered Evaluation:** Existing methods focus on mathematical correctness, but evaluating whether explanations are truly useful and understandable remains an open research area [4].

Addressing these challenges is crucial for the **widespread adoption of XAI techniques in NLP applications**.

E. Future Directions

Future work should focus on enhancing the **efficiency and usability** of explainability techniques in NLP. Some key research directions include:

1. **Enhancing Real-Time Explainability**
    a. Many existing XAI techniques, especially SHAP and counterfactuals, are computationally intensive. Future work should explore more **efficient methods** for real-time sentiment analysis, balancing speed and interpretability [2].
2. **Hybrid Explainability Models**
    a. A promising direction is **combining multiple XAI techniques** to improve

explanation quality. For instance, integrating **SHAP with attention-based mechanisms** could enhance textual explanations in deep learning models [5].

3. **Human-Centric Evaluation**
    a. Explainability should be assessed not just for accuracy but also for **user comprehensibility and trust**. Future research should include **user studies** to evaluate how well different XAI methods align with human reasoning [6].
4. **Expanding to Other NLP Tasks**
    a. While this study focused on **sentiment analysis**, explainability techniques can be extended to tasks such as **text summarization, question answering, and bias detection** in AI models [4].
5. **Integration into AI Assistants**
    a. Given the growing use of AI-driven virtual assistants, **XAI techniques should be embedded into conversational AI systems** to make their responses more transparent and user-friendly [3].

By addressing these areas, future research can contribute to the development of **more interpretable, efficient, and user-aligned AI systems**, paving the way for trustworthy AI-driven decision-making.

F. Summary

This discussion highlights the **importance of XAI techniques in NLP**, emphasizing their role in improving model transparency, user trust, and ethical AI development. While **SHAP, LIME, and counterfactual explanations** offer valuable insights into sentiment analysis models, their practical challenges underscore the need for continued advancements in the field. **By leveraging existing XAI frameworks such as Alibi, InterpretML, and AIX360, future work can streamline implementation and enhance real-world applicability** [7], [8], [9].

By refining explainability methods and integrating them into AI-driven assistants, we can build **more reliable, transparent, and user-friendly AI systems**.

VI. Conclusion & Future Work

A. Conclusion

Explainable AI (XAI) plays a crucial role in improving transparency, trust, and accountability in AI-driven sentiment analysis. This study explored three key explainability techniques—**SHAP, LIME, and counterfactual explanations**—to interpret the decision-making process of a sentiment classification model. By applying these techniques to **deepseek-r1:1.5B** on the **Twitter Entity Sentiment Analysis** dataset, we demonstrated how each method provides unique insights into model predictions.

The findings from this study highlight that:

- **SHAP is effective for global feature importance analysis** but requires substantial computational resources.
- **LIME provides fast, localized explanations** but lacks a broader model-level interpretability.
- **Counterfactual explanations offer actionable insights** by suggesting minimal text changes needed to alter predictions, though they are computationally expensive to generate.

These results emphasize that **a combination of explainability techniques is necessary to achieve comprehensive transparency in NLP applications**. By integrating XAI into sentiment analysis, AI models can become more interpretable, reliable, and aligned with user expectations.

B. Future Work

While this study provides a comparative analysis of explainability techniques, several areas remain open for future research and improvement:

1. **Enhancing Real-Time Explainability**
    a. Many existing XAI techniques, especially SHAP and counterfactuals, are computationally intensive. Future work should explore more **efficient**

methods for real-time sentiment analysis, balancing speed and interpretability.

2. **Hybrid Explainability Models**
    a. A promising direction is **combining multiple XAI techniques** to improve explanation quality. For instance, integrating **SHAP with attention-based mechanisms** could enhance textual explanations in deep learning models.
3. **Human-Centric Evaluation**
    a. Explainability should be assessed not just for accuracy but also for **user comprehensibility and trust**. Future research should include **user studies** to evaluate how well different XAI methods align with human reasoning.
4. **Expanding to Other NLP Tasks**
    a. While this study focused on **sentiment analysis**, explainability techniques can be extended to tasks such as **text summarization, question answering, and bias detection** in AI models.
5. **Integration into AI Assistants**
    a. Given the growing use of AI-driven virtual assistants, **XAI techniques should be embedded into conversational AI systems** to make their responses more transparent and user-friendly.

By addressing these areas, future research can contribute to the development of **more interpretable, efficient, and user-aligned AI systems**, paving the way for trustworthy AI-driven decision-making.

VII. References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, **"Why should I trust you? Explaining the predictions of any classifier,"** in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD'16)*, San Francisco, CA, USA, 2016, pp. 1135–1144.

[2] S. M. Lundberg and S.-I. Lee, **"A unified approach to interpreting model predictions,"** in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS'17)*, Long Beach, CA, USA, 2017, pp. 4765–4774.

[3] S. Wachter, B. Mittelstadt, and C. Russell, **"Counterfactual explanations without opening the black box: Automated decisions and the GDPR,"** *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2017.

[4] L. Arras, F. Horn, G. Montavon, K.-R. Muller, and W. Samek, **"What is relevant in a text document? An interpretable machine learning approach,"** *PLOS ONE*, vol. 12, no. 8, p. e0181142, Aug. 2017.

[5] S. Jain and B. C. Wallace, **"Attention is not explanation,"** in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist. (NAACL'19)*, Minneapolis, MN, USA, 2019, pp. 3543–3556.

[6] O. Biran and C. Cotton, **"Explanation and justification in machine learning: A survey,"** in *Proc. IJCAI-2017 Workshop Explainable AI (XAI'17)*, Melbourne, Australia, 2017.

[7] IBM Research, **"AI Explainability 360: An open-source toolkit for interpretability,"** 2020. [Online]. Available: https://aix360.mybluemix.net

[8] SeldonIO, **"Alibi: Algorithms for explainable AI,"** 2020. [Online]. Available: https://github.com/SeldonIO/alibi

[9] Microsoft Research, **"InterpretML: Machine learning interpretability package,"** 2021. [Online]. Available: https://github.com/interpretml/interpret

[10] Kaggle, **"Twitter entity sentiment analysis dataset,"** 2024. [Online]. Available: https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis

[11] DeepSeek AI, **"DeepSeek LLM: Open-source large language model,"** 2024. [Online]. Available: https://github.com/deepseek-ai