

Interpreting User Opinions: A Multidimensional Approach Leveraging Explainable AI and Generative Models

Summary of the contribution

1. What is the main claim of the paper? Why is this an important contribution to the machine learning literature?

*This paper introduces a novel methodology for **multidimensional and explainable classification of user opinions**, addressing the growing need for transparency and interpretability in automated opinion analysis. While prior approaches typically focus on single dimensions—such as sentiment—our method combines **fine-tuned large language models**, **explainable AI (XAI) techniques**, and **generative models** to produce detailed, human-readable explanations across multiple latent dimensions (e.g., sentiment, topic, emotion, and domain-specific aspects).*

*This contribution is important to the machine learning literature for several reasons. First, it enhances trust and usability in real-world applications where understanding why a classification was made is critical (e.g., mental health, political discourse). Second, by introducing **BERTopicXAI**, our framework bridges the gap between unsupervised topic modeling and interpretability, offering visual and token-level explanations at the topic level. Third, the integration of GPT-based generation with XAI-derived signals supports **transparent, context-rich reasoning**, improving explainability beyond typical black-box summaries. Finally, a comprehensive evaluation—including human expert assessments, LLM-as-a-Judge, and semantic metrics—demonstrates that our approach significantly outperforms existing baselines in both interpretability and explanatory quality [11, 8, 10].*

2. What is the evidence you provide to support your claim? Be precise.

*To support our claim, we conducted a **comprehensive evaluation** across four annotated datasets—**Booking, Amazon, Political, and Depression**—spanning domains such as hospitality, consumer electronics, political discourse, and mental health. Each opinion was classified across multiple dimensions: **sentiment, topic, emotion**, and **dataset-specific attributes** (e.g., political stance or mental state).*

*We compared three prompting strategies: **GPT-base** (raw opinion only), **GPT-interm** (adds classification labels), and **GPT-adv** (adds both labels and word-importance explanations from **LIME** [29] and our proposed **BERTopicXAI**). Explanations were generated using **GPT-4o via API** [36]. Our evaluation included:*

- **Textual and Semantic Metrics:** We assessed **readability, coherence, complexity**, and **semantic similarity** using Dice, TF-IDF, ROUGE-L, and S-BERT scores [35]. **GPT-adv outperformed** the other variants, producing more fluent, aligned, and well-structured explanations.

- **LLM-as-a-Judge:** Using the **LLM-as-a-Judge paradigm** [37, 38], GPT-4 evaluated explanations based on **detail coverage, clarity, dimensional mapping, and analytical insight**. GPT-adv consistently scored highest.
- **Domain Expert Evaluation:** With input from **40 domain experts**, who reviewed explanation triplets in random order, **GPT-adv was most frequently preferred** for its clarity and accurate association between text and classification dimensions.

In summary, the combined **quantitative, automated, and expert evaluations** demonstrate that our methodology—integrating **fine-tuned LLMs, explainability cues, and generative models**—delivers significantly more **informative, interpretable, and contextually rich explanations** than standard generative approaches.

3. What papers by other authors make the most closely related contributions, and how is your paper related to them?

Among the most closely related works are studies on **opinion mining, explainable AI, and the use of large language models (LLMs)** for interpretability. Our paper builds upon and extends several key contributions in these areas:

- **Multidimensional opinion analysis** as explored in [24] and [27], where opinions are examined across multiple facets such as sentiment and emotion. Our work expands this by introducing **topic-level analysis** and additional dataset-specific dimensions (e.g., mental health, political stance), providing a more nuanced and comprehensive understanding of user opinions.
- In terms of **explainability**, we are inspired by XAI methods such as **LIME** [29] and **Integrated Gradients** [28], widely used to reveal model decision rationale at the token level. We apply these techniques specifically across multiple classification dimensions and integrate them into a **generative explanation process** using GPT.
- We also build upon **BERTopic** [30], a state-of-the-art framework for unsupervised topic modeling. However, BERTopic does not offer token-level interpretability. To address this, we introduce **BERTopicXAI**, our own extension that enables **visual and token-level explanations** for topic assignments, bridging the gap between topic modeling and model transparency.
- Lastly, our use of the **LLM-as-a-Judge paradigm** [37, 38] aligns with recent efforts to **automate evaluation** using generative models themselves as meta-evaluators, providing a reliable and scalable benchmark for assessing explanation quality.

Overall, our contribution lies in **unifying fine-tuned LLM classification, XAI-driven interpretation, and GPT-based generative explanation** within a single framework—advancing interpretability in opinion mining beyond current approaches in both **depth and dimensionality**.

4. Have you published parts of your paper before, for instance in a conference? If so, give details of your previous paper(s) and a precise statement detailing how your paper provides a significant contribution beyond the previous paper(s).

*This article is a **substantial extension** of our prior work:*

C. Cosentino, M. Gunduz-Cure, F. Marozzo, S. Ozturk-Birim, "**Exploiting Large Language Models for Enhanced Review Classification Explanations Through Interpretable and Multidimensional Analysis**". 27th International Conference on Discovery Science (DS2024), pp. 3-18, 2024.

The current manuscript builds upon the earlier version by introducing several key enhancements:

- **Broader Scope of Opinion Sources:** While the previous version focused exclusively on hotel reviews for sentiment classification, this study incorporates diverse datasets—including product reviews, political discourse, and mental health posts—broadening the applicability of our methodology across multiple domains [25, 26, 27, 28].
- **Multidimensional Classification:** We move beyond sentiment-only classification to include topics, emotions, and dataset-specific aspects (e.g., political support and psychological distress), offering a more complete view of user opinions [20, 21, 22, 23].
- **Structured and Aligned Explanation Outputs:** A novel feature of this work is the generation of structured textual explanations that explicitly align linguistic elements with their corresponding classification dimensions, thereby improving interpretability and analytical clarity.
- **Explainability Extension with BERTopicXAI:** A major addition is the design and implementation of BERTopicXAI, a novel method that introduces token-level visual explanations for topic modeling—addressing the interpretability gap in unsupervised topic detection [19].
- **Advanced Generative Explanation via GPT:** We integrate GPT-based natural language generation into our pipeline to synthesize detailed, human-readable commentaries based on classification outcomes and interpretability signals, which was not covered in the conference version [36, 37].
- **Expanded Evaluation Framework:** The evaluation has been significantly extended with textual and semantic metrics, LLM-as-a-Judge analysis, and human expert reviews, enabling a more rigorous assessment of explanation quality across all prompting strategies [38, 39, 40].

*In addition to these technical contributions, we have **revised and expanded all major sections**, including related work, methodology, experimental design, and analysis. The result is a **substantially enriched and generalized framework** for interpretable opinion classification that significantly advances beyond our earlier work.*

References:

[8] Cantini, R., Cosentino, C., Marozzo, F.: *Multi-dimensional Classification on Social Media Data for Detailed Reporting with Large Language Models*. In: IFIP Int. Conference on Artificial Intelligence Applications and Innovations, pp. 100–114 (2024). Springer.

- [10] Belcastro, L., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: *Detecting Mental Disorder on Social Media: A ChatGPT-Augmented Explainable Approach*. arXiv preprint arXiv:2401.17477 (2024).
- [11] Cosentino, C., Gündüz-Cüre, M., Marozzo, F., Öztürk-Birim, Ş.: *Exploiting Large Language Models for Enhanced Review Classification Explanations Through Interpretable and Multidimensional Analysis*. In: Int. Conference on Discovery Science, pp. 3–18 (2025). Springer.
- [19] Guidotti, R., et al.: *A Survey of Methods for Explaining Black Box Models*. ACM Computing Surveys 51(5), 1–42 (2018).
- [20] Lundberg, S.M., Lee, S.-I.: *A Unified Approach to Interpreting Model Predictions*. NeurIPS 30 (2017).
- [21] Plumb, G., Molitor, D., Talwalkar, A.S.: *Model Agnostic Supervised Local Explanations*. NeurIPS 31 (2018).
- [22] Makino, M. et al.: *The Impact of Integration Step on Integrated Gradients*. In: EACL Student Research Workshop, pp. 279–289 (2024).
- [23] Kumar, P., Raman, B.: *A BERT-Based Dual-Channel Explainable Text Emotion Recognition System*. Neural Networks 150, 392–407 (2022).
- [24] Morichetta, A., Casas, P., Mellia, M.: *Explain-it: Towards Explainable AI for Unsupervised Network Traffic Analysis* (2020).
- [25] Laato, S. et al.: *How to Explain AI Systems to End Users: A Systematic Literature Review and Research Agenda*. Internet Research 32, 1–31 (2022).
- [26] Zhang, H., Shafiq, M.O.: *Survey of Transformers and Towards Ensemble Learning Using Transformers for NLP*. Journal of Big Data 11(1), 25 (2024).
- [27] Grootendorst, M.: *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure* (2022).
- [28] Chen, Y., Peng, Z., Kim, S.-H., Choi, C.W.: *What We Can Do and Cannot Do with Topic Modeling: A Systematic Review*. Communication Methods and Measures 17(2), 111–130 (2023).
- [29] Mohammad, S.M., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: *SemEval-2018 Task 1: Affect in Tweets*.
- [30] Plutchik, R.: *A General Psychoevolutionary Theory of Emotion*. Emotion: Theory, Research, and Experience 1 (1980).
- [35] Egger, R., Yu, J.: *A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts*. Frontiers in Sociology 7 (2022).
- [36] Davis, J. et al.: *The Temperature Feature of ChatGPT: Modifying Creativity for Clinical Research*. JMIR Human Factors 11(1) (2024).
- [37] Zheng, L. et al.: *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. NeurIPS 36, 46595–46623 (2023).
- [38] Poświata, R., Perełkiewicz, M.: *OPI@LT-EDI-ACL2022: Detecting Signs of Depression from Social Media Text Using RoBERTa Pre-Trained Language Models*. In: Second Workshop on Language Technology for Equality, Diversity and Inclusion, pp. 276–282 (2022).
- [39] Belcastro, L., Branda, F., Cantini, R., Marozzo, F., Talia, D., Trunfio, P.: *Analyzing Voter Behavior on Social Media During the 2020 US Presidential Election Campaign*. Social Network Analysis and Mining 12(1), 83 (2022).
- [40] Hansen, L. et al.: *Textdescriptives: A Python Package for Calculating a Large Variety of Metrics from Text*. Journal of Open Source Software 8(84), 5153 (2023).