Exploiting Large Language Models for Enhanced Review Classification Explanations through Interpretable and Multidimensional Analysis

Cristian Cosentino[1 [0000−0002−6368−373X]],
Merve Gündüz-Cüre[2 [0000−0002−7792−8056]],
Fabrizio Marozzo[1 [0000−0001−7887−1314]]✉, and
Şule Öztürk-Birim[2 [0000−0001−7544−8588]]

[1] University of Calabria, Rende, Italy
{ccosentino, fmarozzo}@dimes.unical.it
[2] Manisa Celal Bayar University, Manisa, Turkey
{merve.gunduz, sule.ozturk}@cbu.edu.tr

Abstract. In today's digital world, user-generated reviews play a pivotal role across diverse industries, providing invaluable insights into consumer experiences, preferences, and concerns. These reviews heavily influence the strategic decisions of businesses. Advanced machine learning techniques, including Large Language Models (LLMs) like BERT and GPT, have greatly facilitated the analysis of this vast amount of unstructured data, enabling the extraction of actionable insights. However, while achieving high classification accuracy is crucial, the demand for explainability has gained prominence. It is essential to comprehend the reasoning behind classification decisions to effectively utilize user-generated content analytics. This paper presents a methodology that leverages interpretable and multidimensional classification to generate explanations from user reviews. Compared to basic explanations readily available through systems like Chat-GPT, our methodology delves deeper into the classification of reviews across various dimensions (such as sentiment, emotion, and topics addressed) to produce more comprehensive explanations for user review classifications. Experimental results demonstrate the precision of our methodology in explaining why a particular review was classified in a specific manner.

Keywords: Large Language Models · Natural Language Processing · BERT · GPT · ChatGPT · Interpretable Models · Explainability

1 Introduction

In the digital age, the wealth of content available on the web, particularly user-generated reviews, has become an invaluable resource for businesses across various industries. These user reviews provide a wide range of insights, opinions and

experiences shared by consumers, offering detailed information on their satisfaction levels, preferences and pain points. From e-commerce platforms to hospitality services, businesses rely on user reviews to gauge the performance of their products or services, identify areas for improvement, and tailor their offerings to meet customer expectations. The widespread availability of online reviews underscores their significance as indicators of public sentiment and drivers of informed decision-making [20].

Accurate analysis and classification of user reviews are essential, driving the advancement of machine learning techniques, particularly Large Language Models (LLMs) like BERT and GPT [21]. These models excel at extracting insights from the vast amount of unstructured review data. However, their general pretraining, done on a wide range of text data from the Internet, may not fully capture the nuances of specific tasks or domains. Thus, fine-tuning on task-specific datasets becomes crucial, enhancing performance and adaptability for tasks such as sentiment analysis or topic modeling. This combination of pre-trained LLMs and fine-tuning is pivotal for robust user review analysis.

However, as classification techniques become more sophisticated and accurate, understanding the logic behind classification decisions becomes increasingly important and demanded. Techniques like LIME (Local Interpretable Model-agnostic Explanations) [23] and Integrated Gradients (IG) [25] have emerged as powerful tools, offering insights into classification decisions by highlighting influential features. Moreover, categorizing reviews based on dimensions such as analyzed topics or emotional expressions can enrich comprehension [6], facilitating the interpretation of user reviews and opinions to gain deeper insights into the various aspects expressed.

Once a review has been accurately classified and its underlying logic understood, leveraging ChatGPT to generate human-readable explanations becomes feasible [4]. ChatGPT can synthesize these insights into easily understandable narratives, elucidating the rationale behind classification decisions and offering context-rich explanations. Integrating these explanations into user feedback analytics systems provides deeper insights into customer sentiments, preferences, and experiences, facilitating more informed decision-making and targeted enhancements in products or services.

This paper introduces a methodology that employs interpretable and multidimensional classification to produce comprehensive explanations from user reviews. The methodology begins with fine-tuning pre-trained LLMs for sentiment analysis in user reviews. Subsequently, sentiment analysis results are interpreted and enhanced with multidimensional classification, such as emotion and topic analysis. Finally, human-readable explanations are generated using ChatGPT.

To evaluate the effectiveness of our approach, we conducted an in-depth evaluation using hotel review datasets that included various opinions, ranging from negative to positive. Using ChatGPT, we generated explanations for each review analyzed. We explored scenarios in which ChatGPT only received the review and its classification as input, as well as scenarios where additional information, such as LIME or IG interpretations, representations of the topics described, and the

emotions expressed, was provided. Comparative analysis demonstrated that explanations generated using this advanced approach outperformed those produced by the baseline model. They showed greater accuracy and informativeness, as confirmed by quantitative measures such as linguistic scores and semantic similarity scores, as well as qualitative assessments conducted by both automatic tools and human experts.

The remainder of the paper is structured as follows. Section 2 offers a concise review of related works. Section 3 describes the proposed methodology. Sections 4 and 5 discuss the results and compare the different outcomes achieved. Finally, Section 6 concludes the paper.

2 Related work

In the age of AI driven by advanced Large Language Models (LLMs), data analysis has undergone a revolution, offering efficient processes for extracting insights. LLMs, powered by natural language processing (NLP) and machine learning, comprehend user queries and produce textual reports with relevant information. These models, stemming from the Transformer architecture, are categorized into decoder-based (e.g., GPT) and encoder-based (e.g., BERT) models. The former excels in causal language modeling, while the latter generates semantic representations in a latent space.

These LLMs act as interactive guides, leading users through data analysis and presenting results in an understandable manner. Across various domains like education, e-commerce, healthcare, and entertainment, LLMs are utilized primarily for information retrieval tasks and generation of informative reports. *Educational* LLMs assist with class schedules and materials, while *healthcare* LLMs aid in the pre-diagnosis of both physical [3] and mental [4] illnesses. *E-commerce* LLMs provide customer support and product information, enhancing shopping experiences [14]. In *media and communication*, LLMs have shown superior performance compared to crowd-workers in text annotation tasks [10]. In the corporate context, GPT models have been used to transform structured tabular data into coherent natural language descriptions and summaries [26]. In *finance*, GPT models are utilized for financial reports, summaries, and sentiment analysis [5]. In *information technology*, ChatGPT simplifies log analysis, improving organization and comprehensibility [19].

In terms of understanding the results of deep learning models, a significant obstacle lies in providing clear explanations for their predictions, especially in crucial areas such as clinical and legal fields. This challenge has led to the development of eXplainable AI (XAI) techniques, which include both *post-hoc* and *self-explanatory* techniques[12], designed to address this problem. Post-hoc techniques aim to explain predictions from pre-trained black-box models. Currently, the most popular approaches are *model-agnostic*, meaning they can be applied to any underlying black-box model, with no assumption on their internal working and structure. Among them, LIME (Local Interpretable Model-agnostic Explanation) [23] and SHAP (SHapley Additive exPlanations) [17] determine a weight

assignment as a proxy for feature importance by following a regression and game theory approach, respectively. Similarly, MAPLE (Model Agnostic Supervised Local Explanations) [22] provides explanations by combining local linear models and Random Forest-based ensembles. Integrated Gradients [25] is a feature attribution approach and also used as an XAI method to explain image and language processing [18]. Integrated gradients method evaluates the importance of features by averaging the gradient of the model output, which is interpolated along a straight-line trajectory in the input data space.

Differently, self-explanatory techniques are trained to provide explanations alongside predictions. However, these methods commonly encounter challenges related to flexibility and integration with other deep learning models [15]. AI techniques, crucial for understanding deep learning models, extend to fields like topic modeling, enhancing interpretability. Leveraging methods such as Topic-Word Attention (TWA), XAI uncovers themes in textual data, aiding in understanding user sentiments and review analysis. Integration of XAI with topic modeling offers clearer insights into the decision-making process of machine learning models, bridging gaps in comprehension across various domains [16].

Our research differs from state-of-the-art work in that it seeks to leverage interpretable and multidimensional classification techniques to provide comprehensive explanations from user reviews. Specifically, we harness the capabilities of LLMs, employing BERT models for multi-dimensional classification in user reviews and ChatGPT for generating human-readable explanations. Additionally, we propose methods for evaluating the quality of these generated explanations and conducting comparative analyses between them.

3 Proposed methodology

This section presents our proposed methodology for achieving explainable classification of user reviews across various domains, including but not limited to products on Amazon, hotels on Booking, restaurants on Tripadvisor, and websites and services on Trustpilot. Our objective is to elucidate and comprehend customer sentiments expressed in these reviews, aiming to provide detailed explanations regarding the judgments conveyed by users, including their positive and negative scores.

The methodology comprises three distinct phases: (*i*) fine-tuning pre-trained LLMs for sentiment analysis of user reviews; (*ii*) interpreting sentiment analysis results and enriching with multi-dimensional classification; and (*iii*) generating human-readable explanations with ChatGPT. In the following, we provide a detailed description of the main steps of our approach, whose execution flow is depicted in Figure 1.

The initial phase of our methodology — *fine-tuning pre-trained LLMs for sentiment analysis of user reviews* — involves selecting the target of the reviews (such as hotels, restaurants, products, or services) and acquiring related training datasets classified by sentiment (positive and negative). We then proceed to refine the pre-trained models, utilizing for example BERT models like RoBERTa,

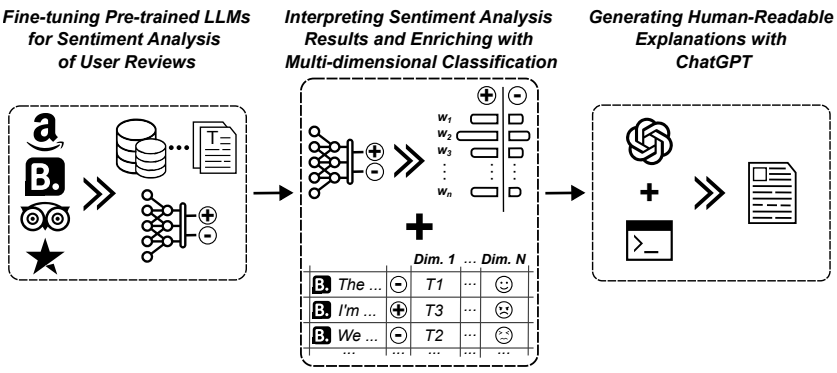


Fig. 1. Execution flow of the proposed methodology.

DistilBERT, and ALBERT. Through this fine-tuning process, we aim to enhance the models' precision in sentiment recognition, enabling them to better understand and classify the nuanced sentiments expressed in user reviews. Evaluation of the models' performance on a test set allows us to compare their accuracy, validating their effectiveness in accurately discerning sentiments across diverse user-generated content.

In the second phase of our methodology, termed *interpreting sentiment analysis results and enriching with multi-dimensional classification*, the focus shifts to explaining the model classification decisions. This is crucial for understanding the rationale behind the predictions and enhancing transparency. Different explainability techniques can be utilized, such as leveraging attention mechanisms in LLMs to highlight influential words or employing post-hoc methods like Integrated Gradients and LIME for explanations. These techniques produce interpretable outputs, including visualizations and detailed feature attributions, which can be used to gain insights into the factors influencing sentiment classification decisions and inform decision-making processes in various domains. Regarding the enrichment with multidimensional classification, reviews can be categorized across various dimensions, such as the addressed topic or the expressed emotion. This approach provides additional insights into the user opinion across different aspects, facilitating a more comprehensive understanding of the sentiment expressed in their review.

In the final phase of our methodology, focused on *generating human-readable explanations with ChatGPT*, we utilize natural language generation to synthesize comprehensive explanations. By integrating insights from sentiment analysis results and augmenting with multidimensional classification, ChatGPT produces explanations that are clear and consider the specific details of each review. Leveraging ChatGPT ability to generate coherent and contextually relevant text, we ensure that our explanations are insightful and informative. This process effectively communicates the rationale behind our sentiment classification decisions, enhancing transparency and facilitating informed decision-making.

4 Experimental results

Our study involves the analysis of labeled review datasets, where each review is tagged with a sentiment label indicating positive or negative feedback. In particular, we analyzed reviews written by users about a hotel that offer valuable information on various aspects of their experience, including quality of service, comfort, cleanliness and general satisfaction. The datasets considered and used come from Booking [1] and Tripadvisor [2], two of the main hotel booking platforms in the world. In particular, we focused on the Booking dataset, which includes over 515,000 customer reviews and ratings of 1,493 hotels across Europe, integrated with geographic location information for deeper analysis and contextual understanding.

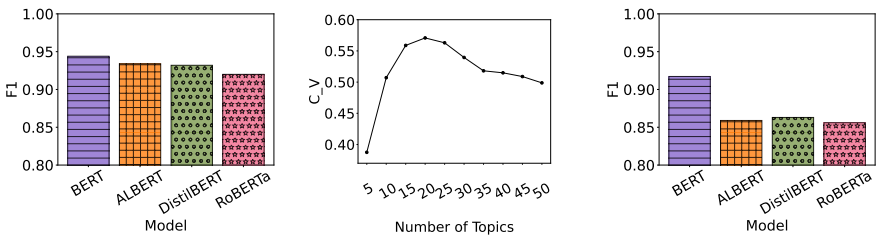
In the following sections, we comprehensively discuss the experimental results we achieved. Section 4.1 details the fine-tuning process of BERT models for sentiment analysis, as well as for topic and emotion extraction. In Section 4.2, we utilize the best performing BERT model to generate explanations related to sentiment classification using Integrated Gradients and LIME. Finally, Section 4.3 outlines how all the obtained information from the reviews is fed into ChatGPT to transform the explanations generated by IG and LIME into human-understandable text.

4.1 Leveraging BERT Models for Sentiment, Topic, and Emotion Extraction

This section investigates the application of Large Language Models for sentiment analysis, topic modeling, and emotion recognition in review data. We leverage BERT-based models for these classification tasks due to their proficiency in natural language understanding [9].

Regarding the classification of positive or negative sentiment, we tried several BERT models, DistilBERT, RoBERTa and ALBERT. We perform model fine-tuning using a training dataset comprising 42,500 reviews, a validation set with 5,000 reviews and then evaluate these models on a separate test set containing 2,500 reviews. Our evaluation criteria included standard metrics for evaluating the accuracy of a model such as accuracy, precision, recall, and F1 score. Figure 2(a) shows the F1 value. BERT, although slightly, appears to be the best model among those considered, with an F1 score of 0.94, followed closely by ALBERT and DistilBERT with 0.93, and RoBERTa at 0.92.

For topic extraction, we adopted BERTopic, based on the recommendation in [11], where it outperformed other techniques in terms of both consistency and diversity of topics. To determine the optimal number of topics, we assessed various metrics, including coherence. Coherence serves as a performance indicator for a topic model, with the number of topics requiring a balance between having a large number, which may result in overly specific categories, and a smaller number, which might blend meaningful subcategories [7]. Specifically, in Figure 2(b), coherence values are illustrated across different numbers of topics. As depicted, around 20-25 topics yield the highest coherence values. We selected 25



(a) Sentiment (F1) (b) Topic similarity (c) Emotion (F1)

Fig. 2. Evaluation of the scores obtained by BERT for the extraction of sentiment and emotion (F1), and by BERTopic for topic detection.

as it offers specific topics that aptly capture the content of the reviews with more granularity compared to fewer topics. For instance, topics related to pillows and beds, TV services, and alarm systems are encompassed in the model with 25 topics but not in the one with 20.

Regarding emotion identification, we considered six emotions: sadness, anger, love, surprise, fear, and joy. We conducted experiments on a textual dataset annotated with emotions [24], employing various fine-tuned BERT models. The dataset consisted of 10,000 training examples, with 3,500 instances each allocated for validation and testing. In this case, the F1 values, as depicted in Figure 2(c), are lower than those for sentiment analysis, given the complexity of identifying six classes rather than just two. Once again, but more markedly, the BERT model outperforms DistilBERT, RoBERTa and ALBERT by 0.05 F1 points.

To describe the dataset and better understand positively and negatively ranked reviews, we provide two examples for each class in Table 1. These examples highlight the distinguishing features of positive and negative reviews. Positive reviews emphasize pleasant experiences such as enjoying warm cookies, friendly staff attitudes, ideal location, and delightful breakfast. Conversely, negative reviews mention issues such as high prices, drainage problems, and water temperature issues in the shower. These examples offer insights into the sentiment distribution within the dataset and aid in understanding the sentiments expressed by the reviewers.

4.2 Interpreting sentiment analysis results with IG and LIME

In the second step of the methodology, the focus shifts to explaining the classification decisions made by the trained model. This is critical to understanding the logic behind the model predictions and improving transparency in the analysis process. Various explainability techniques can be employed for this purpose. One approach is to exploit the attention mechanisms inherent to LLMs, which highlight the most salient words or phrases in the input text that influenced the classification result. Alternatively, post-hoc explanation techniques such as Integrated Gradients (IG), LIME (local interpretable model-independent expla-

Review	Sentiment	Topic	Emotion
We loved the warm cookies and the staff were so friendly welcoming and helpful. We really enjoyed our stay	Positive	[‘cookie’, ‘arrival’, ‘warm’, ‘chocolate’, ‘check’, ‘touch’, ...]”	Trust
Location was ideal. Breakfast was a dream. We enjoyed our trip and stay at hotel	Positive	[‘hotel’, ‘location’, ‘staff’, ‘close’, ‘station’, ‘room’, ...]	Surprise
Price paid similar last year for 2 nights	Negative	[’price’, ’expensive’, ’value’, ’money’, ’little’, ’prices’, ...]	Sadness
Drainage in the shower was not adequate and it took a while for all the water to drain. Also the water temperature setting went from warm water in the cold	Negative	[’shower’, ’pool’, ’water’, ’bathroom’, ’bath’, ’swimming’, ...]	Anger

Table 1. Sample reviews showcasing sentiment analysis, identified topics, and expressed emotions.

nations) or SHAP can be applied to provide explanations for model predictions. These explanations can take the form of visualizations or textual summaries, highlighting the key features that contribute to each classification decision. In this study, we apply IG and LIME approaches to create explanations for sentiment classification. We chose these methods because they provide instance-based explanations, meaning an explanation can be generated for each selected review. In contrast, SHAP provides global explanations based on all reviews or groups of reviews, making it less suitable for single review analysis. Given our focus on review-based explanations, LIME and IG are selected as post-hoc approaches for explainable artificial intelligence (XAI) in sentiment classification.

Pred. label	*Score*	*Word importance*
Positive (0.96)	**2.81**	the spa was first rate, beds were very confortable, staff friendly
Positive (0.96)	**3.33**	We loved the warm cookies and the staff were so friendly welcoming and helpful. We really enjoyed our stay
Negative (0.97)	**-1.32**	Not enough tea bags only 1 provided
Negative (0.99)	**-0.34**	The receptionist of the check in was not friendly as we expected. The comunication with him was difficult and he doesn't made effort to really help us. They need to improve it

Fig. 3. Examples of predicted sentiments interpreted with LIME. Positive sentiment words displayed in varying shades of green, while negative words are depicted in shades of red.

Figure 3 shows examples of sentiment predicted by LIME. Terms associated with a positive sentiment are represented with various shades of green, while negative terms are represented with shades of red. The text provides a series of sentences, each labeled with a sentiment and an associated score. Positive sentences are characterized by words such as ‘*comfortable*’ and ‘*helpful*’, while negative sentences contain terms such as ‘*difficult*’ and ‘*not friendly*’. Overall, the reviews reflect a mixed experience, with some positive points, such as comfortable beds and friendly staff, but also some criticisms, such as a lack of availability of tea bags and an unfriendly reception by the reception desk. Overall, the figure

provides a clear and intuitive illustration of the sentiment predictions associated with each sentence, highlighting the key words that influence their ranking.

4.3 Generating human-readable explanations with ChatGPT

The goal of this section is to outline the methodology for generating clear and comprehensive explanations of sentiment analysis results using ChatGPT. For this purpose, we use the GPT-3.5-turbo version with a temperature setting of 0 for more focused (less creative) tasks in the generated explanations [8]. The following prompt provides a structured guide for using ChatGPT to generate human-readable commentary on the explanation of a model classification. It begins by presenting essential elements: the input review x, the sentiment achieved with BERT s, the interpretation i comprising word-importance pairs obtained from IG or LIME, topic representations t, and the emotional tone of the review e. The task is clearly outlined: generating a concise commentary on the review x and the expressed sentiment s, integrating topic representations t and emotional tone e. The commentary must adhere strictly to the information provided in i, avoiding the introduction of additional information or personal deductions. Overall, the prompt guides the generation of insightful commentary while ensuring relevance and conciseness.

Review (`$x`): {*input review*}, **Sentiment (`$s`)**: {*BERT output class*}, **Interpretation (`$i`)**: {*(word, importance) pairs*}, **Topic (`$t`)**: {*topic representation*}, **Emotion (`$e`)**: {*emotion extracted*}

Task: You are given a review `$x` written by a customer of a hotel, along with its sentiment `$s`, which can be either "`NEGATIVE`" or "`POSITIVE`". Additionally, an interpretation `$i` is provided as a list of (word, importance) pairs, indicating the significance of each word in determining the sentiment classification. The topic addressed in the review `$t` is also given as a string list, representing the relevant words for the topic to which the review belongs. Furthermore, the emotion `$e` of the review is provided, indicating the predominant emotion expressed. Explain why `$x` was classified as `$s` using `$i`, `$t`, and `$e`, highlighting words in `$x` within quotes. The explanation should be about 100 words, avoiding any additional information not provided in `$i`, and without introducing your own comments or deductions.

To thoroughly explore how additional information aids ChatGPT to create explanations for the opinions expressed in reviews, we have developed multiple suggestions. In the *base* prompt (*ChatGPT-base*), solely the review (x) and the sentiment (s) are utilized. Moving to an *intermediate* level, we incorporate the review (x) with its interpretation (i), with distinct prompts for IG (*ChatGPT-interm*-IG) and LIME (*ChatGPT-interm*-LIME). Finally, in the *advanced* prompt, we integrate the review (x), sentiment s, interpretation (i), topic (t), and emotion (e), enhancing the depth of understanding. Again, separate prompts are designated for IG (*ChatGPT-adv*-IG) and LIME (*ChatGPT-adv*-LIME). This systematic approach enables a comprehensive examination of the impact of varied information inputs on the ChatGPT explanation generation process.

5 Performance comparison

In this section, we describe the evaluation of our methodology performance compared to a standard approach. As previously outlined, our aim is to assess how the extra data provided to ChatGPT (interpretation and classification of topics and emotions) aids in crafting precise explanations for the sentiments expressed in reviews. This approach allows us to delve deeply into the subject, uncovering the subtle effects of various information inputs on ChatGPT ability to generate explanations.

Below, we present a positive review example with the sentiment, interpretation, topic and emotion, alongside their commentaries generated by ChatGPT using the three prompts introduced before: *ChatGPT-base*, *ChatGPT-interm*, and *ChatGPT-adv*. We opted for LIME as the model for data interpretation. Below we report one positive review example and how it was described with *ChatGPT-base*, *ChatGPT-interm* and *ChatGPT-adv*. In this example, we chose LIME as a data interpretation model. According to the rules defined in the prompt, the explanations will be approximately 100 words with words taken from the reviews highlighted in quotation marks. The review example is:

`Review (x)` = "Very comfortable hotel rooms nicely furnished 2mins walk to the metro. Many restaurants around for different tastes. Staff are so helpful I would definitely go back to stay at this hotel when i visit Paris again"
`Sentiment (s)` = Positive, `Interpretation (i)` = [('comfortable', 0.056), ('helpful', 0.046), ('nicely', 0.040), ('furnished', 0.033), ('staff', 0.031), ('stay', 0.029), ...], `Topic representation (t)` = ['hotel', 'location', 'staff', 'great', 'close', 'station', 'good', 'room', 'friendly', 'nice'], `Emotion (e)` = Joy

The commentaries generated are:

ChatGPT-base = Based on the review provided, the classification of "positive" is appropriate. The reviewer mentions that the hotel rooms are very comfortable and nicely furnished, which indicates a positive experience with the accommodation. The proximity to the metro station and the presence of many restaurants nearby are highlighted as convenient features, adding to the positive impression. The reviewer also praises the helpfulness of the staff, which is a key factor in guest satisfaction. The tone of the review is enthusiastic and the reviewer expresses a desire to return to the hotel in the future, indicating a high level of satisfaction with their stay.

ChatGPT-interm-LIME = The model explanation for the positive classification of the review highlights words like "nicely" "helpful," "comfortable," and "staff," indicating the positive aspects of the hotel experience. These words contribute significantly to the overall positive sentiment expressed in the review. The presence of terms like "definitely" and "go back" further reinforces the positive nature of the feedback. The review also mentions the convenient location near the metro and the availability of restaurants, adding to the positive experience. Overall, the explanation emphasizes the positive attributes of the hotel, such as comfort, helpful staff, and convenient amenities, aligning with the positive classification.

ChatGPT-adv-LIME = The review expresses satisfaction with the hotel comfort, helpful staff, and convenient location near the metro and restaurants. The posi-

tive classification is supported by words like "comfortable," "helpful," and "nicely furnished," which align with the topic of a positive hotel experience. The emotion associated with the review is also positive, further confirming the customer satisfaction. The review emphasis on comfort, helpfulness, and location resonates with the topic representation of a hotel with great staff, good location, and friendly service. Overall, the review reflects a positive experience, supported by the words highlighted in the explanation and the underlying topic representation.

Comparing the outputs, the standard ChatGPT approach provides an explanation drawing inspiration from the extensive knowledge within ChatGPT rather than directly from the review content. In contrast, both the intermediate and advanced enhanced approaches exhibit significantly higher accuracy, pinpointing the specific words that influenced the sentiment classification. Between the intermediate and advanced levels, the difference is subtle; however, the advanced approach provides a more nuanced understanding. Specifically, in *ChatGPT-base*, the explanation broadly acknowledges positive aspects like comfortable rooms and helpful staff, whereas *ChatGPT-interm*-LIME delves into specific words like "nicely," "helpful," and "comfortable" that contribute to the positive sentiment. Moving to *ChatGPT-adv*-LIME, the explanation further analyzes the positive aspects of the review, emphasizing words like "satisfaction," "comfort," and "helpful staff" while aligning them with the underlying topic representation of a positive hotel experience.

However, we measure the quality of the explanations generated by the different approaches in three ways:

1. *Textual and Semantic Metric Analysis*: This approach involves analyzing the commentaries using various metrics derived from textual analysis [13]. We also utilize semantic similarity scores such as Dice, TF-IDF, Rouge-L, and S-BERT, which are typically employed to evaluate the information content of a summary compared to the original text. Here, we adapt these metrics for evaluating explanation commentaries.
2. *ChatGPT Evaluation*: In this approach, ChatGPT evaluates the commentaries based on criteria such as informativeness, quality, coherence, attributability, and overall impression.
3. *Domain Expert Evaluation*: This approach involves obtaining evaluations from experts who assess the commentaries and collectively choose the best one based on their expertise and judgment.

5.1 Textual and Semantic Metric Analysis

Regarding text metric analysis, Table 2 presents scores derived from the generated explanations on fifty positive and negative reviews using: *i*) TextDescriptives library for linguistic scores, and *ii*) semantic similarity scores, used to assess the informational alignment between two texts.

Based on the provided scores obtained from the TextDescriptives library, here is a description of each criterion and the results obtained:

- *Readability* of the explanations was assessed using the Coleman-Liau index, which estimates the U.S. grade level required to understand a text. Explanations generated by *ChatGPT-base* require a lower U.S. grade level compared to those generated by *ChatGPT-adv*. Other readability indices such as Gunning-Fog and SMOG also indicate similar trends, suggesting that the reports from *ChatGPT-adv* required a deeper understanding of linguistics.
- *Quality* was measured using repetitive text patterns, specifically the duplicate n-gram character fraction, which indicates the fraction of characters in a document that are contained within duplicate n-grams. Since the comments generated are very short (around 100 words), the level of repetition is almost always zero.
- *Coherence* was evaluated based on the cosine similarity between sentences, with their embeddings obtained as the average vector representation of words computed by Latent Semantic Analysis. Even in this case, the coherence of the texts remains consistent across both models.
- *Complexity* was assessed using the entropy of the text, which measures the level of randomness or unpredictability, with higher values indicating greater diversity and complexity of language use. Explanations from *ChatGPT-adv* demonstrate higher complexity, characterized by greater diversity and complexity of language use, compared to those from *ChatGPT-base*, which exhibit more repetitive or predictable language patterns.

	Textual analysis scores				*Semantic similarity scores*			
Approach	Readability	Quality	Coherence	Complexity	Dice_similarity	TF-IDF	Rouge-L	S-BERT
ChatGPT-base	14.62	0.00	0.85	3.80	0.19	0.15	0.57	0.26
ChatGPT-interm-IG	15.84	0.02	0.80	4.08	0.20	0.15	0.59	0.26
ChatGPT-interm-LIME	15.15	0.00	0.79	3.87	0.22	0.18	0.62	0.26
ChatGPT-adv-IG	15.79	0.02	0.85	4.08	0.24	0.24	0.86	0.35
ChatGPT-adv-LIME	16.22	0.01	0.84	4.14	0.24	0.26	0.86	0.38

Table 2. Scores derived from comparing reviews with the generated explanations using textual and semantic similarity metrics.

Regarding semantic similarity metrics, we adapted those traditionally used to evaluate the closeness between two texts (e.g., an original text and a summary) to compare user reviews with the generated explanations. Below is the description of each criterion used:

- *Dice* coefficients over the sets of words (excluding stop words) of the input review and the generated explanation. This metric measures the extent of overlap between the two sets, indicating the similarity in content.
- Cosine-based lexical similarity between *TF-IDF* vectors of the review and explanation, obtained after stop word removal and stemming. This metric quantifies the similarity in word usage and distribution, providing insight into the semantic correspondence.

- *Rouge* metrics, with a focus on *Rouge-L*, which evaluates the longest common subsequence between the review and explanation. Rouge-L assesses the overall coherence and adequacy of the generated explanation in capturing the essence of the review.
- Cosine-based semantic similarity between *S-BERT* embeddings of the review and explanation. This metric measures the semantic relatedness between the two texts, offering a deeper understanding of their contextual similarity.

In examining the semantic scores reported in Table 2, a consistent upward trend is observed across all metrics as we progress from *ChatGPT-base* to *ChatGPT-interm*, and finally to *ChatGPT-adv* (both with IG and LIME versions). Transitioning from basic to intermediate with the addition of sentiment interpretation brings improvements, and further incorporating multidimensional classification brings even greater enhancements. While the increases in Dice and S-BERT scores are relatively modest, TF-IDF and Rouge-L exhibit more pronounced improvements. This indicates that the explanations provided by ChatGPT-adv are more comprehensive and nuanced compared to the basic ones, effectively capturing the essence of the reviews. Regarding the two interpretability techniques used, LIME outperforms IG in terms of scores, albeit marginally, indicating slightly superior performance in generating explanations.

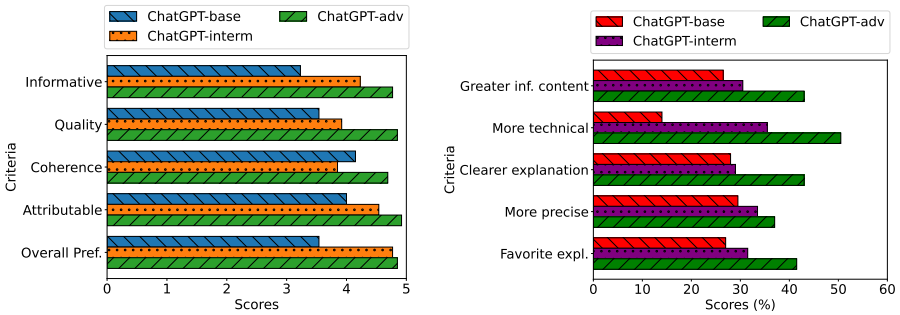
5.2 ChatGPT and Domain Expert Evaluation

The section describes and analyzes the explanations generated by assigning votes through ChatGPT and experts who are asked to choose the best explanation.

For the first rating, we provided ChatGPT with the review, sentiment, and explanation as input, and asked it to rate each on a scale from 1 (worst) to 5 (best) based on five criteria:

- *Informative*: The explanation encapsulates crucial details from the source, offering a precise and concise presentation.
- *Quality*: The explanation is understandable and comprehensible, demonstrating high quality.
- *Coherence*: The explanation demonstrates a sound structure and organization, ensuring coherence.
- *Attributable*: All information in the explanation is attributable to the source.
- *Overall preference*: The explanation succinctly, logically, and coherently conveys the primary ideas from the source.

Figure 4(a) shows the average scores achieved using the explanations generated by *ChatGPT-base*, *ChatGPT-interm*, and *ChatGPT-adv* on fifty positive and negative reviews. Due to space constraints, we only report the values obtained with the LIME interpretation (results are similar with IG). As shown, *ChatGPT-adv* offers the most preferable choice, with higher values than all other approaches across all criteria considered. It surpasses *ChatGPT-interm*, which improves on detail and organization, and *ChatGPT-base*, which provides a fundamental explanation with satisfactory clarity and coherence.



(a) ChatGPT Evaluation. (b) Domain expert.

Fig. 4. Scores obtained in the evaluation of explanations by ChatGPT and domain experts.

Regarding expert evaluations, we asked 20 experts to validate the explanations generated by the three approaches using 10 different reviews. In each test, we presented the review along with the three full explanations (in random order) and asked the experts to identify which explanation excelled in specific aspects. Specifically, they were asked to answer the following questions: (*i*) Which explanation do you believe offers *greater overall information content*? (*ii*) Which explanation contains *more technical* or specialized aspects? (*iii*) Which explanation provides a *clearer explanation* of the topics covered? (*iv*) Which explanation demonstrates *greater precision* and clarity in its contents? (*v*) Which explanation *do you prefer* for overall quality?

Figure 4(b) shows the average percentage of experts who preferred *ChatGPT-base*, *ChatGPT-interm*, and *ChatGPT-adv* for the five criteria considered. Domain experts consistently favored *ChatGPT-adv* over standard ChatGPT across all aspects. *ChatGPT-adv* received significantly higher ratings attributed to its greater information content, more technical nature, clearer explanations, and higher precision, as reflected in the criteria. Explanations generated by *ChatGPT-adv* were notably more detailed enhancing clarity and credibility. Furthermore, they exhibited superior coherence, organization, and grammatical consistency, resulting in higher evaluation scores compared to explanations generated with *ChatGPT-base* and *ChatGPT-interm*.

6 Conclusions

In today's digital age, user-generated reviews play a vital role in shaping business strategies by providing valuable insights into consumer experiences and preferences. Advanced machine learning techniques, such as BERT and GPT, have revolutionized the analysis of this vast pool of unstructured data, making it easier to extract actionable insights. However, in addition to achieving high classification accuracy, the demand for explainability has increased, underscoring the need to

understand the reasoning behind classification decisions. Our methodology introduces an innovative approach that leverages interpretable and multidimensional classification to generate comprehensive explanations from user reviews, outperforming basic approaches. Experimental results demonstrate the accuracy of our methodology in explaining review ratings. Future efforts will focus on analyzing review sets to identify product strengths and weaknesses, further improving our understanding of consumer sentiment and enabling companies to make informed decisions and improve their offerings.

Acknowledgements

This work was supported by the research project "INSIDER: INtelligent ServIce Deployment for advanced cloud-Edge integRation" granted by the Italian Ministry of University and Research (MUR) within the PRIN 2022 program and European Union - Next Generation EU (grant n. 2022WWSCRR, CUP H53D23003670006). It was also supported by the "National Centre for HPC, Big Data and Quantum Computing", CN00000013 - CUP H23C22000360005, and by the "FAIR – Future Artificial Intelligence Research" project - CUP H23C22000860006.

References

1. 515k hotel reviews data in europe (2024), `https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe`
2. Alam, M.H., Ryu, W.J., Lee, S.: Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences **339**, 206–223 (2016)
3. Ayanouz, S.e.a.: A smart chatbot architecture based nlp and machine learning for health care assistance. In: 3rd International Conference on Networking, Information Systems & Security (2020)
4. 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)
5. Belcastro, L., Carbone, D., Cosentino, C., Marozzo, F., Trunfio, P.: Enhancing cryptocurrency price forecasting by integrating machine learning with social media and market data. Algorithms **16**(12), 542 (2023)
6. Cantini, R., Cosentino, C., Marozzo, F.: Multi-dimensional classification on social media data for detailed reporting with large language models. In: 20th International Conference on Artificial Intelligence Applications and Innovations. pp. 100–114 (2024), `https://doi.org/10.1007/978-3-031-63215-0_8`
7. 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)
8. Davis, J.e.a.: The temperature feature of chatgpt: modifying creativity for clinical research. JMIR Human Factors **11**(1) (2024)
9. Devlin, J.e.a.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

10. Gilardi, F., Alizadeh, M., Kubli, M.: Chatgpt outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences **120**(30) (2023)
11. Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)
12. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM computing surveys **51**(5), 1–42 (2018)
13. Hansen, L.e.a.: Textdescriptives: A python package for calculating a large variety of metrics from text. Journal of Open Source Software **8**(84), 5153 (Apr 2023)
14. Hossain, M., Habib, M., Hassan, M., Soroni, F., Khan, M.M.: Research and development of an e-commerce with sales chatbot. In: 2022 IEEE World AI IoT Congress (AIIoT). pp. 557–564 (2022)
15. Kumar, P., Raman, B.: A bert based dual-channel explainable text emotion recognition system. Neural Networks **150**, 392–407 (2022)
16. Laato, S.e.a.: How to explain ai systems to end users: a systematic literature review and research agenda. Internet Research **32**, 1–31 (2022)
17. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
18. Makino, M.e.a.: The impact of integration step on integrated gradients. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 279–289 (2024)
19. Meng, W., Zaiter, F., Zhang, Y., Liu, Y., Zhang, S., Tao, S., Zhu, Y., Han, T., Zhao, Y., Wang, E., et al.: Logsummary: Unstructured log summarization for software systems. IEEE Transactions on Network and Service Management (2023)
20. Mudambi, S.M., Schuff, D.: Research note: What makes a helpful online review? a study of customer reviews on amazon. com. MIS quarterly pp. 185–200 (2010)
21. Myers, D.e.a.: Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. Cluster Computing **27**(1), 1–26 (2024)
22. Plumb, G., Molitor, D., Talwalkar, A.S.: Model agnostic supervised local explanations. Advances in neural information processing systems **31** (2018)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" Explaining the predictions of any classifier. In: 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
24. Saravia, E.e.a.: CARER: Contextualized affect representations for emotion recognition. In: 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3687–3697 (2018)
25. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
26. Wang, F., Xu, Z., Szekely, P., Chen, M.: Robust (controlled) table-to-text generation with structure-aware equivariance learning. arXiv:2205.03972 (2022)