

Study on Explainability concerning Black-Box systems

Nidhir Bhavsar

Universität Potsdam

Matrikel-Nr.: 819171

nidhir.bhavsar@uni-potsdam.de

Abstract

Through this paper, we offer a comprehensive overview of experiments investigating Explainable AI (XAI). We delve into analyzing a wide variety of prevalent black-box systems. We study these model systems using a specific set of dataset and utilizes interpretation techniques like SHAP, LIME, and other generative methods to reveal decision-making processes. We highlight the application of these techniques in extracting intelligence related to input features. Furthermore, we explore methods to enhance the efficacy of the modeling approach, allowing us to delve more profoundly into the utilization of interpretability techniques. The paper highlights the effectiveness of these XAI techniques and suggests improvements by identifying gaps in the current systems.

1 Introduction

Previously, Natural Language Processing (NLP) systems primarily relied on inherently understandable methods. These methods are known as white box techniques and include approaches such as conditional rules, decision trees, hidden Markov models, and logistic regressions, among others. Nonetheless, the rise of recent advancements has led to the widespread adoption of black-box techniques, including deep learning and transformer-based language models, among others. Achieving interpretability through these methods is particularly captivating. Despite obtaining reasonably satisfactory outcomes across various tasks using these techniques, comprehending the decision-making process and identifying the precise functions that elicit the outputs remains challenging. Even when employing white box methodologies, surpassing a certain threshold of complexity makes it tough to derive a sense of rationale, as observed in cases like Random Forest. This could also cause a decline in the trust of human users who adequately express their daily subtleties while interacting with models

in systems such as virtual assistants, information retrieval algorithms, search engines, recommendation systems, etc. All these factors have led to the emergence of Explainable AI (XAI).

The concept of explainability may necessitate a comprehensive elucidation, and it will be a central theme throughout this paper. In the existing literature, the term XAI encompasses a range of descriptors like transparency, intelligibility, interpretability, and explainability. Nonetheless, the objective of this revolves around discerning the rationale behind the model's outcomes from the standpoint of the end user, referred to as the *outcome explanation problem* (Guidotti et al., 2018). Explanations hold significant importance across various applications, particularly in critical domains like healthcare, law, transportation, and finance (Adadi and Berrada, 2018). Numerous researchers have investigated the value of explanations for humans, including their role in enhancing human decision-making (Lai and Tan, 2018; Lertvittayakumjorn et al., 2021) and fostering greater trust in AI systems (Jacovi et al., 2021). Conversely, AI systems also stand to gain from explanations, as they can be employed to validate model reasoning (Caruana et al., 2015) and identify potential sources of errors (Han et al., 2020).

Although the task of Explainable AI encompasses more than just NLP and its related modalities, we choose to confine our efforts to this domain in the current work. We investigate diverse domain-specific tasks with significant nuances, employing various modeling techniques. All of these approaches encompass the essence of AI. We present our analysis of these particular tasks, highlighting their individualities, and use this analysis to extract insights into the favored modeling strategy. We break down our work in this paper into four distinct parts, as outlined below. It should be noted that we select a distinct dataset of our choice for all the below-mentioned tasks.

- We begin by examining the notion of interpretability in connection with a tree algorithm, which serves as the primary modeling approach for the dataset. Our emphasis is on comprehending the subtleties of features that contrast with a complex representation within a limited set of sparse inputs.
- Subsequently, since most NLP-oriented systems currently utilize language models in various forms to generate contextually informed outcomes, we delve into examining the impact of this approach using a specific dataset. Our focus on interpretability for feature analysis is confined to encoder-based systems.
- Expanding our investigation, we explore the fusion of multimodality with a zero-shot classification strategy to analyze its behavioral modeling on a dataset. As outlined in the following section, leveraging multimodality aids in extracting more understandable features, thereby enhancing the generation of rationales.
- Ultimately, we delve into the potential for explainability offered by the recent advancements in Natural Language Processing (NLP) concerning Large Language Models (LLMs). In this regard, we investigate the possibilities of enhancing prompt engineering-based systems' explainability. Our exploration focuses on the intricacies of crafting prompt instructions and generating explanatory text (TG explanations), allowing us to discern specific aspects from a multimodal data perspective.

We discuss the aforementioned concepts in greater depth in the upcoming sections. Additionally, We make our code accessible here¹.

2 Background

2.1 Transformer; Language Models

A language model that utilizes Transformers is composed of stacked Transformer blocks (Vaswani et al., 2017). Each of these blocks is accompanied by a multi-head self-attention layer, which is primarily responsible for applying attention to adjacent token segments while concurrently processing them. In recent times, a series of large-scale Pre-trained Language Models (PLMs) based on the

Transformer architecture have gained prominence. Noteworthy examples include GPT (Radford and Narasimhan, 2018; Brown et al., 2020), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019c), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), T5 (Raffel et al., 2019), and ERINE (Zhang et al., 2019). These PLMs have been fine-tuned using task-specific labels and created a new state of the art in many downstream natural language processing (NLP) tasks (Liu et al., 2019b; Minaee et al., 2020; Jiang et al., 2020; Shen et al., 2020).

Relative Position Embeddings, The standard self-attention lacks inherent word position information. To address this, current methods add positional bias to input word embeddings, creating vectors that consider both content and position. This bias can be implemented using absolute position embedding (Vaswani et al., 2017; Radford and Narasimhan, 2018; Devlin et al., 2018) or relative position embedding (Huang et al., 2018). However, Dai et al. (2019) showed relative position representations to be more effective for natural language tasks, enhancing transformer efficiency at various processing stages. He et al. (2020), through their proposed model aims to overcome this problem by utilizing a disentangled context representation, resulting in substantial improvements over previous methods.

Large Language Models (LLMs), similar to all language models, predicts token probabilities based on preceding and surrounding context (Bengio et al., 2003; Radford et al., 2021). Modern LLMs use self-attention architectures akin to transformers, with hundreds of billions or more parameters (Ganguli et al., 2022). Unlike earlier models on smaller datasets, current LLMs train on massive datasets with billions or even trillions of tokens (Borgeaud et al., 2021), necessitating significantly more computation. This expansion grants LLMs heightened sophistication compared to predecessors.

2.2 Topic Modeling

Researchers have increasingly demonstrated the success of neural topic modeling in utilizing neural networks to enhance existing topic modeling methods (Terragni et al., 2021; Larochelle and Lauly, 2012). By incorporating word embeddings into traditional models like LDA, the feasibility of employing these powerful representations was showcased (Liu et al., 2015). A recent upsurge

¹https://gitup.uni-potsdam.de/bhavsar/swathi_nidhir_delwar

in topic modeling techniques has been centered around embedding models, highlighting the potential of embedding-based approaches (Bianchi et al., 2020). Some methods have begun to streamline the topic creation process by clustering both word and document embeddings (Angelov, 2020). In their work, Grootendorst et al. (Grootendorst, 2022) aggregate this concept by utilizing sentence embeddings that are contextually aware and grouped into topics.

2.3 Interpretability

Throughout this entire work, we utilize a range of distinct techniques to attain interpretability and generate explanations for the results of diverse modeling strategies. We utilize these methods to create visual illustrations that help us extract reasons behind a produced output. Additionally, we depend on various neural models to offer explanations for the generated output using textual methods of reasoning and attribution (Discussed in section 7). Thus, to ensure clarity, we present a thorough introduction to these utilized methods in the following section.

2.3.1 Shapley Values

In the specifics of game theory, Shapley Values emerge as a fundamental concept. The origin of Shapley values was subject to equitably apportioning rewards among a group of participants who collectively influence a particular result (Shapley, 1953). Transposing this into the context of ML, the participants in question correspond to the input features, while the outcome corresponds to the model’s decision. As stated by Lundberg and Lee (2017), Shapley values play a role in assigning importance ratings to distinct parts of the input, attributing to the significance held by each segment.

Given the a set of input features $\mathbf{X} = \{x_1, x_2, \dots, x_n\}$, all the features in a certain coalition $S \subset \mathbf{X}$ contribute towards the outcome $f(S)$, with $f(\emptyset) = 0$. Note, coalition in terms of game theory is when two or more participants/features work together in a group to obtain a specific output. Shapley values redistribute the total outcome value $f(\mathbf{X})$ among all features based on their average marginal contribution across all possible coalitions S . Thus, the contribution of feature i given the coalition S :

$$\Delta_f(i, S) = f(S \cup \{i\}) - f(S)$$

is averaged across all $S \subset \mathbf{X} \setminus \{i\}$. The marginal contribution measures the value that participant/feature i added when included to the coalition S . Do note this contribution can be null, positive or negative. Hence, the corresponding Shapley values $\phi_f(i)$ measures its contribution based on the equation below;

$$\phi_f(i) = \sum_{S \subseteq \mathbf{X} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \Delta_f(i, S)$$

Here, the coefficient $\frac{|S|!(n - |S| - 1)!}{n!}$ is used as normalization term based on the number of choices for the subset S .

2.3.2 Shapley Values Approximation and SHAP

In their work, Lipovetsky and Conklin (2001) suggested using Shapley value imputation for comparative importance analysis of predictors in the model. Additionally, the study by Song et al. (2016) emphasized the application of Shapley value for global sensitivity analysis. They pointed out that when dealing with dependent inputs, computing the Shapley value might provide an approximate measure of the output’s sensitivity to a specific input segment. Both the concepts involve using the results of game prediction from an ML model denoted as f , and Shapley values $\phi_f(i)$ gauge the impact of each feature i based on its present value.

Additionally, there were efforts concerning on devising approximation strategies since the vast multitude of coalitions makes it impractical to precisely calculate Shapley values (Štrumbelj and Kononenko, 2014; Datta et al., 2016). The primary objective involves the computation of $\phi_f(i)$ solely for a reduced set of subsets $S \subset \mathbf{X}$ and estimating the impact of feature removal by integrating across training samples.

The work from Lundberg and Lee (2017) introduces a new perspective that unifies Shapley value estimation with other explainability methods. They suggest that SHAP values can be used to measure the importance of features and show that these values are a unique solution that meets the criteria of accuracy, handling missing data, and consistency. They also offer a set of methods in a library to calculate SHAP values efficiently in different situations. A further detailed analysis on Shapley values for explanations is provided by Mosca et al. (2022).

2.3.3 LIME

Local surrogate models involve providing input data to the black box model and receiving predictions without the necessity of training the original model for interpretative purposes. These surrogate models are specifically crafted to imitate the predictions of the underlying black box model. The objective is to gain an understanding of how the machine learning model makes decisions for specific input data. Overall, surrogate models are categorized as *post-hoc* methods. They necessitate the execution of extra operations following ML model predictions (Sydorova et al., 2019). The formulation of the local surrogate model, while adhering to constraints on interpretability, can be expressed as follows:

Given an instance \mathbf{X} , the explanation model g is determined by minimizing the function:

$$f_{exp}(\mathbf{X}) = \arg \min_{g \in G} L(f, g, \pi_{\mathbf{X}}) + \Omega(g)$$

Here, L quantifies the proximity between the explanation provided by model g and the prediction generated by the original model f . The model complexity $\Omega(g)$ is maintained at a low level to ensure minimal feature preference. The proximity metric $\pi_{\mathbf{X}}$ defines the extent of the neighborhood around the instance \mathbf{X} that is considered for the purpose of generating the explanation.

Ribeiro et al. (2016) demonstrate a practical implementation of a surrogate model. They introduce the concept of Local Interpretable Model-Agnostic Explanations (LIME), which involves training local surrogate models to elucidate individual predictions, as opposed to directly constructing a global surrogate model. LIME employs a comparable approach to that of Alvarez-Melis and Jaakkola (2017). It generates perturbed samples from the given data and acquires predictions from the black box model associated with these perturbed samples. These newly generated datasets with perturbations are then used to train an interpretable model. This model assigns weights based on the proximity of the sampled instances to the instance of interest. The objective of this interpretable model is to replicate the behavior of the machine learning model for the specific data instance, without concentrating on approximating global predictions.

3 Survey: Multimodal Explanations

We highlight the study conducted by Park et al. (2018) towards multimodal explainability. Their contributions are two-fold. Firstly, they enhance multimodal explainability through the introduction of the VQA-X and ACT-X datasets. These datasets are adaptations of the previous VQA (Antol et al., 2015) and MPII MHP (Zhao et al., 2018) datasets, with additional incorporation of explanations that pertain to both modalities, namely vision, and text, i.e. encompassing textual rationales paired with visual indicators that highlight the elements described in the rationale. These modifications are aimed at addressing various challenges, such as distinguishing between inconspicuous image attributes, employing common-sense reasoning, discerning subtleties within image-text pairs, distinguishing among complementary images that reflect the representation of intricate details, and most importantly, enhancing the reasoning ability of machine learning models to deduce rationales that are comparable to those derived by actual humans. Next, they also suggest a benchmark modeling strategy capable of justifying a decision with natural language and points to the evidence. They propose a Pointing and Justification model (PJ-X) trained on two different task stages. They consider the target output and the corresponding rationales as the only source of supervision and learn to point in a latent manner. They compare the results of their proposed modeling strategy with various different configurations. Additionally, they assessed the pointing technique that was learned latently, comparing it with the then-prevalent off-the-shelf method, by Das et al. (2016).

They examine the models' learning capabilities by selecting specific examples and demonstrating them using both textual explanations and the model's ability to point out masked elements in images, as shown in figure 15. They demonstrate the model's capability to accurately associate and elucidate the justification for a choice. This is illustrated by their focus on diverse supplementary images sourced from the VQA-X dataset. In these instances, the model effectively produces common-sense-driven rationales and accurately identifies the relevant object within the visual representation. This proficiency is also evident when contrasting images of activities that share similarities in the activity itself but differ in their surroundings. Remarkably, the model adeptly meets this challenge

by providing appropriate responses. Additionally, they present a scenario in which the model demonstrates consistency with inaccurately predicted target choices. Therefore, the authors provide a thorough analysis, elucidating the significance of producing multimodal explanations. They emphasize instances where visual explanations offer deeper insights than textual justifications, and conversely.

4 Tree Algorithms

4.1 Task Formulation

The focus of this task is to utilize currently accessible Tree algorithms and carry out interpretability analysis at both the global and local scale distinctively. This involves selecting a dataset for our experiments and applying a suitable tree-based modeling framework that generates reliable predictions. We then closely examine these predictions, analyzing the input features and their impact on the model's decision-making process.

We can divide the derivation of this task into the following objectives:

- Consider an input instance $X = \{x_{f_1}, x_{f_2}, \dots, x_{f_M}\}$, where f_M represents the size of the feature set. Using this input, we generate a prediction with the model we are examining. The process is illustrated in equation 1, which takes both X and a predefined set of hyperparameters H and generates a hypothesized output Y_h .
- Next, we are prompted to analyze the generated predictions. Here, the task is to recognize the significance of features within the provided input set. This, obligates us to calculate how each $X_{f_i} \in X$ affects the model's predictions. Afterward, the objective is to pinpoint the relevant context C_X surrounding the instance, as this context provides valuable insights into the model's behavior. Finally, the aim is to craft an explanation $E(X)$ that precisely showcases the influence of different features within the specified context. (See equations 2, 3)

$$Y_h = f_T(X, H) \quad (1)$$

$$\text{Contribution}_{X_{f_i}} = \frac{\Delta f_T(X_{f_i}, H)}{\Delta X_{f_i}} \quad (2)$$

$$E(X) = \{\text{Contribution}_{X_{f_i}}\}_{i=1}^{f_M} \quad (3)$$

4.2 Dataset & Preprocessing

We utilize the Litcovid (Chen et al., 2020) track 5 multi-label text classification dataset, originally made available through the Biocreative VII workshop (Chen et al., 2021). It is designed for proficiently categorizing scientific articles about the COVID-19 pandemic and offers multiple labels for text classification. These labels encompass Prevention, Transmission, Mechanism, Treatment, Case Report, Diagnosis, and Epidemic Forecasting. The dataset is part of the PubMed Corpus² of scientific articles and is maintained by the National Center of Biotechnology Information (NCBI) of NIH.

For the purpose of this task, we perform a dataset factorization procedure. We eliminate instances that manifest in a multi-label structure, consequently narrowing our focus to treating the dataset as a multi-class classification task. As both the training and testing datasets encompass instances with multiple labels, we apply the preprocessing procedure to each partition of the data individually. This leads to the creation of a fresh dataset partitioning, comprising a collective total of 16,000 instances within the training subset and 1,500 instances within the testing subset.

Originally, the dataset provides with extensive meta-data information useful for efficient classification of the data. However, for the purpose of this task, we solely consider the combination of `title` and `abstract`. We featurize the data reflecting either a combination of the text fields, by deriving the TF-IDF (Term Frequency - Inverse Document Frequency) or BOW (Bag-of-Word) representation. Furthermore, we carry out data preprocessing by employing methods such as eliminating stop words, converting sentences to lowercase, limiting the vocabulary, and applying normalization to the input data.

4.3 Experiments

In this section, we are showcasing the methods we employ to explore the explainability of the current task. Given that the task revolves around the utilization of Tree algorithms, we consciously avoid incorporating any further implementations that make use of advanced machine learning methods or modeling techniques. Moving forward, we employ both global and local interpretability techniques to rein-

²<https://pubmed.ncbi.nlm.nih.gov/>

force the produced output with graphical explanations.

In the following subsection, we elucidate the overall architecture/pipeline that we employ to construct the classification outcome.

4.3.1 Methodology

For the purpose of modeling this task, we employ an ensemble learning technique called a Random Forest, also referred to as a Random Decision Forest (Ho, 1995). This method involves the utilization of the Bagging procedure (Breiman, 1996), wherein various subsamples of the given dataset are created. These subsamples encompass diverse combinations of randomly selected features, which are then simultaneously processed through distinct decision trees, ultimately culminating in a Random Forest.

In the context of Bagging, we select random samples from the original dataset, thereby generating multiple sets of these samples. This specific procedure is termed bootstrap. Consequently, the training process is executed independently on each sample to generate target outputs. The final output is determined through majority voting, achieved by aggregating the outcomes of all the models.

Similar to a Decision Tree, we structure the Random Forest Classifier (since we employ a classification task) based on the maximum permissible nodes, the total count of decision trees, and the number of features to choose, depending on this we model the classifier.

Following this, we perform multiple feature importance analysis procedures that relate to the predicted output. This analysis is conducted using open-source interpretability software tools.

4.3.2 Training

We conduct experiments with various hyperparameter settings to enhance our ability to achieve the best possible output from the chosen dataset. After careful analysis, we decided upon a specific set of attribute values that would optimize the output for our modeling strategy, as outlined in table 6. We derive this by performing a cross-validation procedure over a combination of different configurations (23 configurations to be precise).

We set up a cross-validation configuration by considering certain parameters that contribute to the process. Firstly, we have the option to adjust the total count of decision trees in a given set, represented as $\{50, 100, 150, 200, 300, 400\}$. Following

that, we need to decide on the max-features selection from the input data. This selection can either be $\sqrt{N_x}$ or $\log_2(N_x)$, where N_x indicates the size of max-feature in the input representation. Lastly, to address the dataset's notable imbalance, we include a scheme that utilizes preference weights based on class frequency. These weights are determined based on whether they are calculated from the complete dataset or from a dataset subsample. After passing these combinations through a Random Forest classifier, we obtain the optimal configuration mentioned in table 6.

4.4 Result & Analysis

In this section, we will assess the performance of the architecture described earlier. We will explore various combinations of input data and present a comprehensive analysis of the most effective configuration. We will achieve this by employing several global and local interpretability techniques and algorithms, as previously discussed. Our objective here is to utilize Tree algorithms for data modeling, employing the modeling strategy to create visual rationales. These visual rationales will assist in effectively discerning the significance of each feature. Consequently, we can emphasize the importance of features and understand how the model's output might change through the introduction of counterfactual instances or features.

In Table 1, we present the outcomes of assessing the test dataset with a Random Forest classifier. This analysis involves the utilization of a modeling approach across various input data segments, which are subsequently subjected to featurization using either TF-IDF or BOW techniques. It also demonstrates the utilization of diverse data segments and the application of distinct featurization methods for evaluation.

The table clearly demonstrates that utilizing the abstract as input while excluding stopwords yields the most optimal results among all the presented settings. This could be attributed to the fact that the dataset primarily revolves around scientific articles, featuring input fields such as title, abstract, and keywords. As a result, a significant portion of valuable information for efficient classification can be harnessed by directly leveraging the abstract of the article. This is owing to the abstract serving as a comprehensive summary of the entire article. Furthermore, TF-IDF representations offer an additional advantage compared to standard bag-of-word

featurization. This advantage stems from TF-IDF's ability to give greater weight to words that are infrequent across the training sample of the dataset. Moreover, TF-IDF also places emphasis on the repetition of words within a specific document.

4.4.1 Explainability Analysis

In this section, we thoroughly examine the explainability and interpretability aspects of the task under consideration by using the previously mentioned test split of the Litcovid track 5 dataset. We conduct this analysis by employing various independent interpretability techniques, namely SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). Additionally, we incorporate visual performance analysis to establish connections between categories and extract feature importance across the dataset.

In this task, we conduct an analysis at two distinct levels. Through this, we aim to attain global interpretability, while simultaneously examining individual instances. This approach allows us to identify anomalies and delve deeply into a specific set of features.

We analyze the general globally distinguishable characteristics of the provided input data by examining the visual representation derived from SHAP. However, before delving into that, we present the model's performance across various categories using a confusion plot/matrix. This is to compare the predictions generated by the utilized modeling approach with the actual labels obtained from the dataset. (See figure 1). The figure vividly illustrates that the model is struggling to generalize effectively for the minority classes due to a significant imbalance across various categories. Moreover, it highlights a discernible connection between the categories prevention and treatment, particularly beyond the substantial bias introduced by features associated with the prevention category. This suggests that the model might be implicitly computing relationships among specific sets of features that span across the provided classes.

In Figure 1, we conduct a feature importance analysis. This analysis involves determining the overall impact of the top-K ($K=10$) features accessible from the feature set used to represent the input data. To achieve this, the average Shapley values of each distinct feature across the dataset for a specific target are computed. The figure clearly illustrates that certain word features, such as {pandemic, sars, cov}, which can be grouped together as related

to the Covid-19 pandemic, exert a significant influence on the output. It's worth noting that the importance scale is not standardized for each category, resulting in a distribution that is somewhat affected by the majority class. Nonetheless, this approach aids in identifying both the correlation between a specific class and the features, as well as the significance they contribute.

Furthermore, to offer a comprehensive insight into how the model's output is affected by the key features, we showcase importance plots for each class (see figure 6). These plots illustrate the influence of individual instances in the dataset through their associated Shapley values. These values clarify the degree to which they impact the model's probability regarding a specific category, represented by positive impacts (in red) or negative impacts (in blue).

Figure 6 reveals that the feature "pandemic" exerts a somewhat balanced influence on the model's output. It has both positive and negative effects on individual instances across the four displayed categories. Next, if we were to assign topics to certain features unique to a class, we could establish a stronger connection between them and their respective categories. This connection is evident in their semantic relation to the parent class, and their consistent positive impact on the output. Additionally, we observed a shift in pattern for the category prevention. The main observation is that the prominent features that play a significant role in the other categories now exhibit a detrimental effect on this particular class (i.e., prevention). Moreover, due to an imbalance in the available data, the model consistently seeks lower feature values for these specific terms.

Next, in Figure 7, we present the local interpretation of a specific instance taken from the test split of the dataset. The figure illustrates a comparison between the plots generated while taking the classes prevention and diagnosis into account. The corresponding textual description of this comparison can be found in Table 9. However, one can easily observe that even though the originally assigned category is diagnosis, the model is currently assigning a greater probability to the counterpart category. This discrepancy could be attributed to the data imbalance issue, as well as the model's inclination towards features with lower Shapley values, leading to this outcome.

Finally, we examine a hypothetical situation. We

Model	W.P	W.R	W.F1	Accuracy
TF-IDF				
RFClassifier (title)	0.77	0.77	0.76	0.774
RFClassifier (abstract)	0.82	0.81	0.8	0.811
RFClassifier (abstract w/o stopwords)	0.84	0.83	0.82	0.833
BOW				
RFClassifier (title)	0.79	0.79	0.78	0.789
RFClassifier (abstract)	0.82	0.81	0.8	0.811
RFClassifier (abstract w/o stopwords)	0.83	0.83	0.81	0.825

Table 1: Illustrates the diverse scores acquired by applying TF-IDF or BOW representation using the recognized stages of the Random Forest Classifier. We employ weighted Precision (W.P), Recall (W.R), and F1 (W.F1) scores a text-modality-based to demonstrate these outcomes.

manipulate the textual data to divert the model’s decisions towards a particular outcome, as depicted in figure 12. We observe that the model is potentially giving greater importance to specific COVID-19 terminology. As the dataset aligns with this focus, it could lead to a divergence from the original outputs. By treating these terms as stopwords, we can potentially achieve improved results.

4.5 Conclusion

We use an ensemble-learning Tree algorithm to fit our selected dataset. The model effectively captured patterns, except for some minority classes, yielding moderately interpretable results. Simple data visualization techniques helped identify category correlations and exemplify prevalent anomalies. Both SHAP (utilizing Shapley values) and LIME interpretation methods provided insights into manipulated features, revealing the major overall and class-specific impactful feature in the input data interpretation. Furthermore, we observed that specific terminology began to alter the outcomes and could potentially be equated with stop-words.

5 Lanaguge Task and Feature Attribution Analysis

5.1 Task Formalisation

This task centers on exploiting the capabilities of contemporary transformer-based architectures (Vaswani et al., 2017). These architectures are applied to a specific impromptu dataset. The refined model obtained from this process assists in enhancing understandability and interpretability. Consequently, it facilitates the analysis of overall patterns and characteristics within the dataset. We

are allowed to select a text-modality-based dataset. However, the optimization problem for adjusting the transformer architecture given the input, must follow either the classification or regression approach.

We can divide the derivation of this task into the following objectives:

- Let’s examine an input instance denoted as $X = x_1, x_2, \dots, x_M$, with M representing the maximum sequence length. To process this input using a transformer architecture, we perform a tokenization procedure to map to relevant embeddings. The encoded input is then fed into the relevant model to generate predictions as targets. This is given in equation 4, which utilizes the input X , in conjunction with a tokenization function and hyper-parameter configuration H to generate target output Y_h accordingly.
- Next, we analyze the generated predictions to determine the significance of the features obtained from the input set that contribute to the output. This process includes evaluating how each individual word/token, denoted as $\{\forall x_i \in X\}$, affects the model’s predictions. Additionally, we identify the relevant context C_X (equation 5) that surrounds each feature and form a concise explanation $E(X)$ (equation 6) that demonstrates the influence of features within this context. (See section 4.1 and 2 for further details)

$$Y_h = F_{transformer}(X, f_{tokenize}, H) \quad (4)$$

$$\text{Contribution}_{X_{f_i}} = \frac{\Delta f_T(X_{f_i}, H)}{\Delta X_{f_i}} \quad (5)$$

$$E(X) = \{\text{Contribution}_{X_{f_i}}\}_{i=1}^{f_M} \quad (6)$$

5.2 Dataset & Preprocessing

We opt for the Twitter Topic classification dataset³ (Antypas et al., 2022), which is the first extensive dataset for classifying topics on social media, distinct from conventional sources like news articles (Greene and Cunningham, 2006) or scientific literature (Lazaridou et al., 2021). The dataset

³https://huggingface.co/datasets/cardiffnlp/tweet_topic_single

contains tweets that were collected from September 2019 to October 2021. To filter the tweets, a process called *pre-filtering* is used to eliminate non-English tweets, and another process known as *near-deduplication* helps in removing duplicates. Additionally, the collection comprises 11K cleaned tweets, which have been categorized into 23 different topics presented in a taxonomical sequence.

Since the dataset provides multiple different tweet segments aggregated based on the year of curation, we choose to utilize the 2021 split. These splits further consist of train, test, and validation sets. The selected split of the dataset includes the following class categories, namely pop-culture, daily-life, sports & gaming, arts & culture, business & entrepreneurs, and science & technology. The data distribution across various subsets is characterized by the subsequent ratio of 70:20:10 for training, testing, and validation sets. Additionally, we select only the single-label instances from the original dataset. Since the authors of the dataset already preprocess the data to eliminate irrelevant details that could deviate from the topic classification procedure, we adhere to their approach and utilize the data directly for additional experiments and modeling, as outlined in the upcoming sections.

5.3 Experiments

In this section, we demonstrate the modeling strategies we use to successfully complete the current task. These strategies help us achieve interpretability, which contributes to a clearer understanding of the dataset under consideration. Since the stated problem pertains to modeling the selected data using a transformer architecture thus we restrict ourselves from any additional processing or modeling modifications.

5.3.1 Methodology

We choose the DeBERTa (Decoding-Enhanced BERT with Disentangled Attention) model architecture (He et al., 2020) to encode the textual information offered through the selected dataset. We choose this model since it has a greater resemblance to a BERT architecture (Devlin et al., 2018), however, it improves over the prior or other state-of-the-art pretrained language models through the induction of two novel techniques: a disentangled attention mechanism, and an enhanced mask decoder.

Unlike BERT, DeBERTa uses two distinct vectors to portray tokens present in the input. It employs disentangled matrices to calculate attention weights between these representations. This is motivated by the fact that the dependency of attention weights calculated for specific tokens should be prioritized in context to their relative positions too. Similar to BERT, DeBERTa uses the strategy of masked language modeling (MLM) for training its weights/parameters. However, additionally, DeBERTa incorporates absolute position token embeddings just before the softmax layer. This allows the model to decode the masked words using combined contextual embeddings of word contents and positions in conjunction with the absolute word position representations.

We select this model since it exhibits off-the-chart performance across several benchmarks including CoLA (Warstadt et al., 2018), MNLI-m/mm (Williams et al., 2018), STS-B (Cer et al., 2017), MRPC (Dolan and Brockett, 2005) among others. We employ a finetune procedure on the DeBERTa model using the selected dataset. The finetuning is executed as a sequence classification technique which optimizes the target labels accordingly.

5.3.2 Training

To facilitate sequence classification on top of a pretrained DeBERTa language model, we employ a certain specific configuration of the hyperparameters intended to optimize the overall performance. The hyperparameter configuration we select is showcased in Table 8. For encoding input tweets from the Tweet Topic Classification dataset, we utilize the standard DeBERTa tokenizer. This tokenizer applies word-piece tokenization, similar to BERT, to embed word sequences.

We train the sequence classifier utilizing the aforementioned model, while iteratively optimizing the weights in correspondence to the subjected input data. We train the model using the train split, followed by validation on the development set, as detailed in section 5.2. Finally, we evaluate using the test split of the dataset. We conduct the entire process using suitable GPU support to enhance computational speed.

5.4 Result & Analysis

In this section, we present the outcomes we achieve by modeling the input dataset using the task-mandated settings detailed in section 5.1. More-

Category	P	R	F1	Support
arts & culture	0.55	0.12	0.20	48
business & entrepreneurs	0.53	0.69	0.69	78
pop culture	0.93	0.93	0.92	671
daily life	0.57	0.84	0.68	178
sports & gaming	0.98	0.94	0.96	630
science & technology	0.67	0.43	0.52	88

Table 2: Shows the scores obtained by the fine-tuned DeBERTa classifier on the Tweet Topic classification dataset. We present class-wise scores measured using metrics including Precision (P), Recall (R), and F1. Additionally, we also provide the class-wise instance count.

over, as part of the task requirements, we conduct an interpretability analysis to determine feature importance. In addition, we are presenting a comparative analysis that examines further factors influencing the data.

In Table 2, we present the results obtained by evaluating the employed modeling strategy across the predefined class categories or topics. The evaluation yields a weighted-F1 score of 0.8318, indicating a commendable performance in the context of multi-class classification. However, a closer analysis of the table reveals that the model faces challenges due to the imbalance induced by majority classes like pop culture and sports & gaming. It becomes evident that the model struggles to adequately address all the categories. This situation arises potentially due to the model being overfitted by the majority classes. Notably, the minority classes should have been easier to distinguish, as they significantly differ in context. For instance, the context of the category arts & culture greatly contrasts with that of business & entrepreneurs. This discrepancy might contribute to the observed performance issues.

5.4.1 Explainability Analysis

In this section, we are performing a comparative analysis of input features representation of tokenized words. Our goal is to enhance our comprehension of predictions generated by the employed modeling strategy in relation to the provided dataset. To achieve this, we utilize various distinct techniques. We employ interpretability methods like SHAP (SHapley Additive exPlanations) to delve into the insights. Additionally, we integrate data-specific analysis to address pre-conceived notions from result analysis and to identify anomalies. We perform the analysis on two distinct levels:

global and local.

In Figure 4, we present the performance of the aforementioned modeling strategy using a confusion matrix/plot. This display showcases the model’s predictions compared to the actual target values extracted from the dataset. The figure reveals a precise alignment between the model’s predictions and the target values, particularly evident in the class category labeled as sports & gaming. Similarly, the pop-culture category demonstrates minimal instances of false-positive results. Nonetheless, this pattern shifts when we evaluate the performance of the remaining classes. Notably, the daily life class exhibits a high rate of misclassifications, as do the other two classes. It is crucial to emphasize that the bias primarily originates from the pop culture class; the other predominant class lacks a significant number of true negatives. Moreover, a potential connection seems to exist between the classes labeled as art & culture and daily life, despite their comparatively higher misclassification rate. It is interesting to note that the category art & culture showcases the lowest performance across other classes, however, it exhibits a pronounced association with the category daily life.

The examples in Table 14 exhibit instances extracted from the test subset of the provided dataset. These instances illustrate the utilization of Shapley values to perform a contextual interpretation process on a modeling function, predicting the output topic class. The figures in the table demonstrate the varying effects of individual features on maintaining the desired probability for the output class.

The table presents the initial two figures illustrating a situation of positive classification. The model’s confidence in classification varies by class, as indicated by the differing final probabilities assigned. This variation highlights a significant distinction. Examining the distinct characteristics of each example reveals that the model possesses a degree of awareness regarding the essential information needed to differentiate named entities within a tweet. For example, it adeptly highlights attributes like Chelsea FC, Man City, and Premier League, signifying an association with a sports organization, and appropriately links them to the category of sports & gaming. The model also possesses the capability to compute attention among particular groups of tokens, such as novel, On Being Yukiko, #japanesecanadian. It correlates these to-

kens and links them to the class category known as arts & culture.

Following this, we highlight instances from the table (rows three and four) where there are inaccuracies in classification. The transformer heavily depends on its attention-based mechanism to connect contextual features and word tokens, allowing it to effectively differentiate and produce classification results. Nevertheless, in the cases mentioned, the model incorrectly identifies the intended objective and produces erroneous outcomes. For example, in both cases, we observe the model incorrectly categorize them as daily life and business & entrepreneurs instead of the accurate categories of art & culture and science & technology. Jargon related to the accurate class holds greater significance or exerts a stronger influence on the misclassified category. Consider the word poetry, which bears greater importance within the category of daily life rather than art & culture. Similarly, a technical term like internet garners more attentional engagement within the business & entrepreneurs category. Next, the model’s choices in assigning a category to the output based on an input tweet seem quite reasonable. From a human standpoint, reading the provided tweet might indeed establish a connection with the misclassified class (especially row four).

Furthermore, we offer a comprehensive visualization of the top-k (where k=10) influential words for the categories of arts & culture and business & entrepreneurs in figure 14, based on the distribution of data among these categories. It is apparent that words with a positive impact on each class exhibit clustering and a strong thematic relevance. Conversely, the token attributes exerting negative influence represent an aggregation of terminologies impacting all other classes except the current class.

We closely analyze how the model’s assessment of class preference is influenced by specific words/tokens used in input tweets. To demonstrate this, we introduce a counterfactual scenario where we manipulate the input data to observe the resultant effects. The information regarding this is found in the lower rows of table 14. It’s clear that, at the outset, the model is classifying a given input sequence as business & entrepreneurs. However, upon closer examination, it becomes apparent that the model is directing some of its attention towards

the tweet-tags, some of which are related to marketing products. Analyzing the Shapley value representation reveals that this emphasis is not particularly pronounced. Nevertheless, upon removing these specific tags, or tags in general, from the input tweet, it becomes evident that the model’s inclination towards the arts & culture category significantly increases, as depicted in the second-to-last and last figures. In general, we recognize that the data contains some inconsistencies causing the model to not adequately address features found in minority classes.

5.5 Conclusion

In this task, we analyze interpretability using a language model-driven technique. We examine the selected dataset, establish correlations among target categories through visualizations, and extend the analysis with Shapley values. Our findings suggest that while SHAP effectively generates visual interpretations, its utility is limited with transformer-based models. Initial experiments reveal no support for interpreting Named Entity Recognition systems. Furthermore, we infer that constrained parameter settings in transformer models lead to their narrower general knowledge compared to recent models. We suggest to experiment with a knowledge base (KB) augmentation which could aid in recognition of entities representing current events.

6 Multimodal Task and Feature Attribution

6.1 Task Formalisation

This task aims to create feature attributes that can be easily interpreted by a modality-aware ML model. This model helps extract feature encodings, which in turn facilitate the classification of a desired output through a Boosting algorithm classifier. The central objective involves examining these derived features and conducting both local and global explainability analyses. The task permits the selection of a multi-modal dataset for experimentation. However, we are constrained to utilize an XGBoost classifier to train the formulated feature representation from the textual and image modalities of the dataset under consideration.

We choose to focus on the problem of Natural Language Inference (NLI) for this work. In this, we classify pairs of a premise and a hypothesis into three distinct categories: contradiction, entailment, and

neutral. The NLI problem involves determining the validity of a hypothesis—whether it is true (*entailment*), false (*contradiction*), or indeterminable (*neutral*)—based on a given premise.

We can divide the derivation of this task into the following objectives:

- Consider an input with two modalities: visual (V) and textual (T). The visual data, represented by $V \in R^{W \times H \times 3}$ (where W and H are image dimensions encoded in RGB), and the textual data $T = \{x_{T_1}, x_{T_2}, \dots, x_{T_M}\}, T \in R$ (with T_M as max sequence length). We define feature attributes $F \in \{f_1, f_2, \dots, f_n\}$.
- Given the input pair and the formulated feature attributes, we employ a distinct modality-aware function to derive feature values. as shown in equation 7, 8.
- Finally, We employ an XGBoost model (refer to equation 9) for producing the desired output. Subsequently, we utilize this generated output to conduct an interpretability analysis (refer to equations 10 and 11). This analysis allows us to compute the significance of each feature with respect to the provided input $X = [F_V, F_T]$.

$$F_V = f_V(V, F) \quad (7)$$

$$F_T = f_T(T, F) \quad (8)$$

Here f_V, f_T are modality-aware ML models, which utilize the input and feature set F to generate encodings.

$$Y_h = f_{XGBoost}(F_V, F_T, H) \quad (9)$$

$$\text{Contribution}_{X_{f_i}} = \frac{\Delta f_T(X_{f_i}, H)}{\Delta X_{f_i}} \quad (10)$$

$$E(X) = \{\text{Contribution}_{X_{f_i}}\}_{i=1}^{f_M} \quad (11)$$

Here, we have $X_{f_i} \in X$, and we are evaluating the impact of each feature using the objective outlined in equation 10. Our ultimate goal is to create the explanation $E(X)$, which is further elaborated upon in section 4.1.

6.2 Dataset & Preprocessing

For the purpose of this task, we choose the e-SNLI-VE dataset (Do et al., 2020), which represents a multimodal extension of the original Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015). This dataset predates the multimodal implementation by Xie et al. (2019), which introduced the SNLI-VE (Visual Entailment) dataset. The primary contribution of this dataset lies in its reduction of errors within the neutral class of the previous dataset. Furthermore, the dataset includes textual explanations crafted by human annotators, centered around the language entailment task.

This dataset presents a unique variation of the usual Natural Language Inference (NLI) task, incorporating multimodal input characteristics. In the typical NLI task, two textual sentences are taken as a premise-hypothesis pair. The objective is to determine if they exhibit entailment, contradiction, or a neutral relationship. However, in this dataset, the author innovates by substituting the hypothesis segment of the text with a corresponding visual image depicting a comparable situation. These image-text pairs are sourced from the Flickr-30K dataset (Young et al., 2014).

To accomplish this task, we use the train split extracted from the original iteration of the e-SNLI-VE dataset. We extract a subset comprising 10% of the total 500K instances. To maintain consistency with the original dataset, the sampling procedure ensures a comparable ratio of entailment, neutral, and contradiction instances. Subsequently, we partition the obtained sample into a training set and a test set using a seeded approach, with a division ratio of 80:20 respectively. Since the task requires us to create feature representations for both the vision and text modalities, we utilize the obtained samples and apply further processing as detailed in subsequent sections.

6.3 Experiments

In this section, we present various modeling strategies that we investigate to effectively accomplish the task of multimodal feature attribution. We then proceed to conduct a feature importance analysis using interpretability techniques that enhance our ability to examine the selected multimodal data. In this approach, we completely rely on the zero-shot capabilities of contemporary ML models to extract novel features, aiming for optimal outcomes.

According to the given task requirements, we

must employ a boosting algorithm like XGBoost to function upon the extracted features and establish their relation with the target output. Consequently, the primary innovation in this approach lies in the process of refining these features into an optimal representation. Given the task’s multimodal nature, we adhere to the strategy of categorizing features in a way that aligns well with existing off-the-shelf models. The subsequent section elaborates precisely on the sequential actions we have undertaken to attain this objective.

6.3.1 Methodology

We begin by employing the BERTopic⁴ architecture (Grootendorst, 2022) to create topic-models from the following text segment of the dataset. These topic models form a network of document clusters, where each cluster denotes a specific group of document instances from the original dataset. In order to avoid introducing any human-judgmental bias into the features, we opt not to define our own feature attributes, given the extensive diversity of the dataset. Thus the process of topic modeling can be identified as initial pre-processing step which aids in recognizing the resemblance between documents and thus can aid in formulating attributional features.

The BERTopic framework involves three main sequential components. First, it generates contextual embeddings. Then, it performs dimensionality reduction and clustering. Finally, document word vectorization and keyword extraction are carried out. The sentence-transformer⁵ (Reimers and Gurevych, 2019) is used to derive contextual embeddings, offering a comprehensive representation of the entire sequence or document. This is distinct from BERT (Devlin et al., 2018), which produces individual sequence-aware word embeddings. Dimensionality reduction is employed to map the original document representation to a reduced space, mitigating the challenges of high dimensionality. UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2020) is utilized for this purpose. The clustering technique HDBSCAN (Hierarchical DBSCAN) (McInnes et al., 2017) is then applied to form clusters from the condensed embedding representation, relying on data point reachability. Subsequently, as proposed by the authors of the original paper, c-TF-IDF methods is employed for document vectorization in the dataset.

⁴<https://maartengr.github.io/BERTopic/>

⁵<https://www.sbert.net/>

This represents documents based on the word frequency within the current cluster and inverse frequency across all clusters (equation 12). Finally, Maximal Marginal Relevance (MMR) is used to calculate the relevance between the keywords and the document to derive a top-k word representation of the subsequent cluster.

$$W_{t,c} = \text{tf}_{t,c} \cdot \log\left(1 + \frac{A}{\text{tf}_t}\right) \quad (12)$$

Here, we have A which defines the average word count for each class/cluster category. $\text{tf}_{t,c}$ indicates the term-frequency of word t for a particular cluster c , and tf_t represents the overall frequency of word t across all clusters.

After generating clusters, we proceed to use a semi-automated method for obtaining feature attributes. From each cluster, we extract the top-k ($k=20$) representative words. These words are then inputted into an interactive generative transformer (ChatGPT), with a specific prompt instruction. This process results in the generation of mutually aware keywords, which serve as feature attributes for the data.

After formulating the feature attributes, we employ modality-specific zero-shot ML models to produce feature scores for each attribute. Specifically, we use the BART model (Lewis et al., 2020), fine-tuned for zero-shot capability, to extract textual features. For the visual modality, we rely on the CLIP architecture (Radford et al., 2021), pretrained to perform zero-shot image classification with a provided set of feature attributes. We circulate each pair of premise (image), and hypothesis (text) through the aforementioned model subjected to the target attributes. Later, we combine these attributional values to depict each instance.

Finally, We employ the XGBoost architecture to model the derived feature attributes. XGBoost, a gradient-boosted tree algorithm, employs supervised learning to model input data. This strategy, part of ensemble learning approaches, aims to predict a target variable accurately by combining estimates from simpler, weak models using a Gradient Boosting technique (Friedman, 2001). Figure 8 shows the entire pipeline employed by us, to model the data and generate target predictions.

6.4 Training

We conduct experiments using different hyperparameter configurations and provided meta-data details to determine the optimal approach for mod-

eling the selected dataset according to the task requirements. Under this section, we enlist all these details accordingly.

To facilitate topic modeling, for grouping of the documents according to their shared attributes, we apply specific hyperparameter configurations. Our analysis is conducted on a 0.1% subset of the original train split of the e-SNLI-VE dataset. Through our experimentation with different versions of sentence transformers to generate contextual embeddings, we ultimately select the all-MiniLM-L5-v2⁶ model checkpoint. Our main objective is to create a focused collection of 20 topic clusters, refraining from further reductions in this aspect. The derived features are illustrated in Table 15. It is important to note that out of the 20 extracted topic clusters, we eliminate 2 overlapping clusters.

To achieve feature values tailored to each modality, we employ the zero-shot classification pipeline. This pipeline leverages either BART⁷ or CLIP⁸ architecture to create textual and visual encodings. Notably, the textual model facilitates multi-label classification, a feature absent in the visual model. Consequently, the textual model derives feature probabilities by assuming the presence of multiple true features for an instance. This approach isn't viable with the visual model, which exclusively accommodates multiclass classification. We perform the feature extraction process in batches (batch_size = 5) using GPU support.

Finally, to model the XGBoost architecture we split the extract feature into train and test subsets. We decide upon the set of hyperparameters depicted in table 7. Finally following this we perform interpretability analysis using the techniques mentioned in section 2.

6.5 Result & Analysis

This section outlines the results obtained by us after modeling the input dataset termed with the specific task details provided at the beginning. Additionally, since the task pertains to performing feature importance analysis and interpretability analysis, we also conduct a comparative study in this regard.

In Table 3, we present the micro-average f1-

⁶<https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2>

⁷<https://huggingface.co/facebook/bart-large-mnli>

⁸<https://huggingface.co/openai/clip-vit-large-patch14>

Configuration	Contradiction	Entailment	Neutral
multimodal			
topic-oriented	0.65	0.61	0.58
features			
text based			
topic-oriented	0.59	0.59	0.56
features			
image based			
topic-oriented	0.35	0.35	0.29
features			

Table 3: presents the micro-average f1-score achieved through different modeling strategies utilized for the multimodal feature attribution task. These strategies encompass employing either an unimodal approach—text only or vision only—or a multimodal approach.

scores achieved across the three classes by training an XGBoost model with varying configurations. These configurations involve employing distinct feature sets: solely utilizing unimodal attributes, such as text or vision, and combining both modalities for multimodal analysis. It is evident from the table that the multimodal configuration significantly outperforms the other two by a definitive margin. However, It is important to highlight that a model trained solely on visual features struggles to adapt to the dataset, resulting in underperformance across all three categories, particularly in the neutral class. This discrepancy arises because the text-only features encode multi-label information, which is lacking in the case of the vision-only attributes due to the module's limited functionality.

6.5.1 Explanability Analysis

In this section, we conduct a comparative analysis of features using various techniques to enhance our understanding of predictions generated by the mentioned modeling approach for a given input set. We employ interpretability techniques, specifically SHAP, along with other data interpretation methods for improved deduction. This analysis occurs in two stages: global and local. This approach enables us to detect feature patterns and study individual instances to identify anomalies effectively.

Firstly, since we execute a topic-modeling strategy, we intend to analyze the aggregated clusters. Topic modeling operates with the assumption that the documents within the study encompass all relevant topics and that the terms contained in these documents suffice to define each topic. Consequently, topic models are typically constructed through ad-hoc analyses. The arrangement of documents into distinct clusters is illustrated in Figure 9,

which highlights the compact overall structure. Despite some overlapping clusters, they remain easily distinguishable, each capturing a unique conceptual entity. Additionally, these clusters exhibit correlations based on their respective representations. For instance, clusters associated with various activities appear in the lower half of the axes, whereas those depicting human forms are centered around the axes' intersection.

In Figure 11, we depict correlations between distinct feature attributes, derived via the zero-shot classification process. The inter-correlation among these attributes is intuitive, as they inherently pertain to themselves. However, the intensity of their interactions with other attributes varies when comparing text and visual features. Notably, the Transportation Modes feature exhibits a strong connection with Wheeled Sports and Transportation, owing to instances specific to these features conveying potentially opposing information, yet fundamentally encapsulating vehicular machinery-related details. Furthermore, within textual attributes, Nudity and Public Spaces displays a moderate association with Photography and Outdoor Activities, along with Water Activities and Recreation. Furthermore, in Figure 2, we depict the correlation among primary features obtained from the condensed embedding representation, which is employed for cluster creation. This method selects the top-k ($k=3$) words for topic representation. It is important to note that the inter-correlation remains consistent across all these figures.

Upon examination, the actively contributing features are collectively oriented toward textual attributes. This could be because the multi-class label features for the visual data segment are ineffective in describing the image, unlike the textual encodings. We can observe a recurring pattern when conducting a feature importance analysis in a generalized manner, as illustrated in figure 13. The figure showcases the top-K features that exert significant influence on the model's output. In order to overlay this information, we derive the corresponding Shapley value for every distinct feature. An analysis of the three global representations corresponding to each category reveals a consistent trend: the impact of image features is notably diminished. However, it should be noted that this is not an absolute rule. For

instance, certain image features, like Group Gathering or Waiting Outdoor, exhibit a neutral effect on the model's output within the contradiction class. In contrast, the image feature Music and Performance contributes positively, particularly when associated with higher values. Whilst not easily deducible, most of the significant attributes associated with image features have an almost neutral impact on the model's output. However, their presupposition does shift when aggregated with supplementary feature data. In addition to this, the textual characteristics like Snow Sports and Mountain Activities, Nudity and Public Spaces, and Animal Play and Activities hold significant influence over the model's output. These features also exhibit a similar distributional impact as the image features. In other words, when the Shapley value is higher, there is a greater tendency for a shift toward the target category, and vice versa (with a few exceptions).

Subsequently, we proceed to compare specific local interpretations of sampled instances within the selected dataset. To assess the impact of individual features on a given predefined category (i.e., contradiction, entailment, or neutral), we employ Shapley values as indicators of feature influence in the utilized model. In Table 13, we compile premise-hypothesis pairs and model predictions. We then employ the Shapley local construct (force-plot), to elucidate the significance of the primary influential features that contribute to a given output. This methodology is applied across all three distinct classes. The initial plot consistently portrays the predicted label, followed by subsequent plots in the same sequence. Additionally, we provide human-annotated ground-truth explanations that describe the reasoning behind the assigned target label. These annotations might aid in enhancing the visual interpretations of the data.

Table 13 presents a clear illustration of the prevailing trend where features predominantly exhibit contrasting effects on each other's contributions, whether positive or negative. For example, the feature *Smiling and Happiness* coexists with the feature *Crying and Emotional Distress*. While the former strongly forces the model towards predicting a specific category, the latter diminishes the overall proba-

bility, thereby decreasing the scores and vice versa. Similarly, certain features are actively backed by other homogeneous features. To illustrate, consider the feature Snow Sports and Mountain Activities, which exhibits relative compatibility with Water Activities and Recreation, among others.

6.6 Conclusion

Using a transformer-based topic-modeling approach, we generate feature attributes and exploit zero-shot modeling for efficient multi-modal representation. Our analysis demonstrates that this strategy for feature extraction produces decisive outcomes. Rather than manually crafting features and introducing human bias, we embrace unsupervised learning methods for better results. The model effectively establishes feature connections. However, the vision encoder’s inability to generate multi-label probabilities hindered proper model fitting. Shapley values effectively revealed this limitation, illustrating that the model didn’t adequately address visual features. Simple data visualization also yielded valuable insights into semantic and contextual feature correlations.

7 Multimodal Task with Prompting Large Language Models

7.1 Task Formalisation

The multimodal task of prompting large language models (LLMs) poses the problem of experimenting with the zero-shot capability of current ad hoc model architectures toward explainability. We are posed to choose a multimodal dataset that combines textual information with visual features. This task involves investigating various dataset configurations such as question-answering (QA), natural language inference (NLI), and others. We will utilize the target, along with the given input, to extract rationales.

We choose to focus on the question-answering task for this project, and through this, we explore the aspects of explainability in current LLM systems/architectures. A typical QA input comprises a Question (Q), an Image (V), and an Answer (A). These three components form a triplet, and they are subsequently matched with an Explanation (E) that has been annotated by a human.

We can divide the derivation of this task into three primary parts.

- Initially, we encode visual attributes from the dataset, transforming them into textual representations. Given an image, denoted as $V_x \in \mathbb{R}^{W \times H \times 3}$, where W and H represent its dimensions in RGB. We then structure the image contextually as $V_x = \{x_{v_1}, x_{v_2}, \dots, x_{v_M}\}$, with v_M indicating the maximum sequence length.
- Next, we utilize the Question $Q_x = \{x_{q_1}, x_{q_2}, \dots, x_{q_M}\}$ and craft a specific prompt instruction P_A , typically composed of textual tokens outlining the task objective. As shown in equation 13, we create a triplet containing the question Q_x , context V_x , and prompt P_A . This triplet is then fed into the modeling function to produce the hypothesized answer A_h .
- To conclude, we utilize the TG answer A_h , the question Q_x , context V_x , and a newly designed prompt P_E as inputs to the modeling function. This process generates the desired explanation E_h (Equation 14).

The sequence in which we mention these parts also indicates their interdependency. Each part relies considerably on the part preceding it. Therefore, making slight modifications to any of these parts can have an impact on the output of the overall model.

$$A_h = f_A(Q_x, V_x, P_A) \quad (13)$$

$$E_h = f_E(Q_x, V_x, A_h, P_E) \quad (14)$$

7.2 Dataset & Preprocessing

We choose the VQA Explanation dataset (VQA-X) (Park et al., 2018). VQA-X is an expansion of the initial Visual Question Answering (VQA) dataset (Antol et al., 2015), paired with both visual and textual explanations from human annotators. The original dataset consists of approximately 200K MSCOCO images (Lin et al., 2014), each accompanied by three questions and ten answers per question. Furthermore, VQA-X includes individual textual explanations for question-answer pairs within the training set and presents three textual explanations for question-answer pairs in the test/validation sets.

To accomplish this task, we extract 100 instances from the validation split of the VQA-X dataset.

We ensure an even distribution across 41 distinct question-type categories as formulated in the original version of the dataset. In Figure 3, we can see a detailed analysis of the questions. The questions are organized according to how often the initial word sequences are repeated. These questions aim to elicit logical and informative answers, covering a wide range of categories such as color, object, activity, and time. Moreover, they also display an interest in obtaining details about particular brands, distinguishing between animals, and posing logical queries related to object quantities or the overall appearance of the image. Thus the sample derived from the validation set encompasses all these attributes.

7.3 Experiments

In this section, we illustrate multiple approaches and carry out various experimental decisions to achieve definitive results for producing high-quality explanations for the input data. We examine the results of different strategies we employ and explain our choice of the suggested approach. As mentioned before, we do not engage in any fine-tuning of data-specific models for this task. Instead, we fully depend on the pretrained version of the model at hand to generate outcomes.

In this task, we employ a multimodal dataset that includes both textual and visual data as our input. The purpose of this task is to derive reasonable textual explanations. Therefore, our initial step involves transforming the visual data into its corresponding textual format. This conversion is crucial to extract optimal contextual information, enabling us to generate explanations effectively.

7.3.1 Methodology

Initially, we employ the BLIP-2⁹ (Li et al., 2023) architecture, which makes use of the FLAN-T5-XL large language model. Instead of relying upon visual data, we create captions from images since the task requires textual context rather than visual representation paired with a question to generate answers and explanations. This process can be likened to projecting multimodal features (such as visual data) into a textual domain for subsequent processing.

The BLIP-2 architecture comprises two main components: a Visual Encoder and a Large Language Model. These two components are intercon-

nected through a lightweight Q-former, a querying transformer that extracts visual features from the frozen image encoder. The Q-former acts as a bottleneck between the frozen image encoder and the frozen LLM. Its role is to select and transmit the most relevant visual features to the LLM, which then produces the desired text as output. The Q-Former is pretrained for vision-language representation learning, and vision-to-language generative learning (See figure 10).

To obtain the best possible textual depiction of the image, we engage the BLIP-2 architecture using a diverse set of prompt instructions that cover a wide range of categories such as scene description, emotional tone, elements, and specifics, as well as perspectives. These instructions work to extract a wealth of information that assists in crafting believable explanations. (See table 10 for a detailed overview).

The TG output derived by utilizing the prompts we create may lack cohesion and potentially contain repetitive details. To tackle this issue, we employ a paraphrasing module that operates upon a distinctive LLM (i.e. FLAN-T5-XXL). Within this module, we rephrase the provided set of individual visual captions, arranging them into a coherent paragraph. The outcome is a text that is understandable and retains a logical flow, emphasizing the importance of the captions.

We utilize the FLAN-T5-XXL¹⁰ large language model (LLM) to extract responses for a given pair of questions and visual context. Subsequently, we leverage these responses along with questions and context to construct credible explanations through an instruction-based approach (namely Rationale prompting). We formulate the instruction according to the guidelines outlined in Table 11. Our approach adheres to a fundamental structure for prompting, encompassing elements such as Instruction, Context, Input Data, and Output Indicator. The underlying FLAN-T5-XXL model is built upon an architecture reminiscent of the T5 (Raffel et al., 2019) model. It undergoes pretraining employing the instruction finetuning method introduced by Chung et al. (2022). This strategy involves training the model on diverse finetuning tasks presented as instructions (1.8K instruction phrased tasks), thereby empowering the model with robust few-shot capabilities.

⁹<https://github.com/salesforce/LAVIS/tree/main/projects/blip2>

¹⁰<https://huggingface.co/google/flan-t5-xxl>

7.3.2 Additional Experimentation

In addition to employing the modeling strategy for extracting explanations from a provided set of Context, Question, and Answer triplets, we further explore various alternative configurations. These configurations involve generating captions and presenting prompts to the model as variations to the arrangements mentioned earlier.

Initially, we conducted experiments by directly employing the BLIP-2 architecture, which employs a FLAN-T5-XL LLM to produce responses to questions and subsequently re-prompting the same for generating explanations. In this method, we formulate a comparable prompt as outlined in Table 11. However, instead of depending on context, we establish a direct pipeline that interacts with the BLIP-2 architecture.

Next, for effectively producing image captions, instead of depending on the complex procedure of multi-prompt captioning, we also explored an alternative approach inspired by PromptCap (Hu et al., 2023). Here, we directly prompt the visual encoder by utilizing the BLIP-2 architecture to craft contextually relevant captions tailored to specific questions. This strategy allows us to obtain contextual information that is closely tied to the questions and, consequently, facilitates efficient alignment with answers. However, this method necessitates an extensive tuning process. In our experimentation, we focused solely on the BLIP-2 architecture's zero-shot capability, refraining from making any modifications.

7.3.3 Annotation Startegy

We annotate the generated explanations manually. These explanations are derived by prompting the triplet of context, question, and answers, utilizing the method outlined in section 7.3.1. We prepare a questionnaire that considers three annotation schemes/metrics: Goodness, Satisfaction and Trust, similar to Hoffman et al. (2023). The following section discusses the importance of these annotation criteria in evaluating generated rationales.

Explanation Goodness, taking into account the claim regarding the elements of a good explanation, the common agreement emphasizes factors like clarity and precision. As a result, when an explanation is presented, one can make a proactive assessment to determine its quality. This process serves as an optimal framework for incorporating goodness into the explanations produced by XAI

systems.

Explanation Satisfaction, revolves around asserting the adequacy and goodness of explanations. While an explanation may receive a positive evaluation based on a prior judgment, it might not prove sufficient when viewed from the user's standpoint. This mirrors the assessment of explanations in terms of "human-likeness," meaning how closely the explanations resemble those provided by an actual human rather than a computational model. This assessment delves into factors such as understandability, reliability, and sufficiency. Consequently, it can be labeled as a posterior judgment of an explanation.

Trusting Explanation involves directly asking the user for their confidence level in the XAI system. This includes evaluating the predictability, efficiency, and believability of the XAI system. These measurement scales are taken directly from the ones specified by Cahour and Forzy (2009). This concept primarily pertains to whether the user places their interpersonal trust in the results produced by the XAI system.

We create a questionnaire that comprises of questions attending to three specific criteria of goodness, satisfaction, and trust. To evaluate the generated explanation, we employ three human annotators. The annotators are posed to compare the generated explanation with the ground-truth explanation accessed from the VQA-X dataset. Our assessment involves using a Likert scale ranging from 1 to 5, representing a spectrum from strongly disagree to highly agree.

7.4 Result & Analysis

In this section, we are evaluating how well the utilized LLMs can adapt to the given task within a zero-shot setting. As mentioned before, this task focuses on gauging the effectiveness of existing LLMs and related architectures in providing explanatory responses via textual prompts. We analyze the generated explanations using both automated computational metrics and human annotation methods. Thus, we can offer a comprehensive analysis that covers both quantitative and qualitative facets of the results and task under study.

7.4.1 Quantitative Analysis

We utilize the following scores to employ automated computational metrics in comparing the explanations generated with the ground-truth explanations: i) METEOR (Banerjee and Lavie, 2005), ii)

ROUGE (Lin, 2004), and iii) BLEU (Papineni et al., 2002). In Table 4, the scores achieved across the mentioned metrics are displayed. These scores are derived from a comparison of the generated explanations using either the predicted answers or the ground-truth answers in the given question-context pairs. It is observed that the scores for the explanations generated through the triplet of ground-truth answers, questions, and context are slightly higher than those obtained by utilizing the triplet of predicted answers, questions, and contexts. However, we can deduce that these scores do not provide the most dependable means of contrasting the explanations produced by the models. This is because the diversity and innovation introduced through output generated by the model are not captured by these scores. Furthermore, these scores solely emphasize string similarity with a specific reference, thus failing to assess quality in a manner aligned with human comprehension (Sulem et al., 2018).

To tackle this problem, we investigate different evaluation approaches through human annotation. This can aid us in better understanding the generated output and boost the model’s performance on the selected dataset. We formulate annotation strategies that cover three criteria, this includes, Goodness, Satisfaction, and Trust (See section 7.3.3). Table 5 presents the scores achieved by averaging annotations provided by individual annotators. One can readily conclude that the average annotation scores for TG explanations accompanied by predicted answers are slightly better than the scores derived for ground-truth answers. This pattern stands in contrast to the information presented in table 4. We deduce that the divergence could arise from the production of more intelligible TG predictions facilitated by prompting the aforementioned model pipeline, thereby providing supplementary context during rationale generation. In general, most annotators exhibit a neutral bias toward the majority of questions related to the specified criteria. However, there’s an exception in cases where annotators strongly favor the comprehensiveness of explanations, specifically focusing on sentence structure and grammaticality.

7.4.2 Qualitative Analysis

Table 12 presents specific instances taken from the original VQA-X dataset and processed through the mentioned pipeline. It becomes evident that in certain situations, the model fails to effectively adjust and provide suitable explanations based on the in-

put it receives. Consequently, in this section, we concentrate on conducting an analysis of instance-based attributes that impact the model’s output.

Firstly, by examining the data in table 12, we can observe that the model does not possess the capability to articulate common characteristics for the provided input data i.e., questions and answers coupled with visual descriptions for TG explanations. Instances arise where the model is subjected to relay upon depicting distinct attributes of objects present in the image. However, rather than incorporating external knowledge, the model depends solely on the utilization of the visual context. For instance, when we compare the cases in the first and second rows of the table 12, we can conclude that the model cannot accurately define the attributes of the object (in this case, the "Elephant"). Instead, it repeats the reasons in a situational manner. In contrast, for the latter case, the visual depiction effectively recognizes the objects within the image. As a result, the model manages to provide a more descriptive explanation compared to the ground-truth explanations.

The model assumes that the user or reader already possesses a certain level of mental knowledge and a firm grasp of reality. This assumption forms the foundation for its perspective. It is taken for granted that the user has a general understanding of various subjects and common sense. An illustration of this can be seen in the third and fourth rows of the table 12, where a scenario is presented based on these assumed foundations. In the former, a question is posed related to technological reasoning. The actual explanation relies on the information presented, while the model’s explanation draws parallels between the recognized object and its likely purpose. Similarly, in the second example, the model depends on common sense to deduce the meaning of the described situation.

Similarly, the model consistently offers straightforward explanations, specifically favoring direct TG rationales. Instead of constructing a hierarchical reasoning process from a given input’s sense and deriving a conclusion, the model leaps directly to findings and places the responsibility of context interpretation on the user. This characteristic is intriguing to consider because humans tend to describe things in an assertive manner while expecting machines to operate in an inverse or more formalized way. For example, as illustrated in the fifth row of the table 12, instead of contextualizing

Modeling Type	Meteor	RougeL	Rouge2	RougeL	Bleu
Explanation based on predicted answers	0.34	0.38	0.14	0.32	0.093
Explanation based on ground-truth answers	0.35	0.38	0.15	0.33	0.101

Table 4: Shows the diverse scores achieved by the generated explanation across different metrics. These explanations are derived based on the model’s utilization of predicted or ground truth answers.

the given scenario, the model highlights the activity and establishes a transitive dependency between overlaid factors, quickly arriving at a conclusion.

Furthermore, the current design of the visual context generation architecture lacks the capability to establish a correlation between the image and the TG output. Additionally, a diverse range of input data concentrates on extracting textual details from the given image, as it encompasses the qualities of an optimal visual encoder. However, due to our utilization of the off-the-shelf model version in a zero-shot scenario, the model is incapable of addressing such cases. When tasked with recognizing prevailing text, the model struggles to perform this task accurately, resulting in its dependence on the semantic context of the input for a solution.

7.5 Conclusion

In this task, we analyze the capabilities of the text-generation (TG) based explanations derived by using recent large language models (LLMs) on multi-modal data. We realized that generating meaningful visual captions is challenging when accounting for the used visual captioning systems. Our utilized BLIP-T5 model struggled to comprehensively depict images, prompting us to implement an aggregation process using manually crafted prompts and instructions for captions. Moreover, given the inclusion of Q&A in the task, the model proficiently predicted answers when supplied with informative captions.

Next, we noticed that the FLAN-T5 architecture exhibits limited reasoning abilities compared to other major LLMs. This stems from the models’ utilization of an Encoder-Decoder architecture, differing from the autoregressive approach in other models. Encoder-decoder models excel in tasks like Machine Translation and Summarization, where external information integration is less crucial. As a result, our chosen model showed underwhelming performance in logically straightforward scenarios.

Nevertheless, the model showcased adeptness in common-sense reasoning and general human knowledge. It is crucial to emphasize that the generated explanation heavily hinges on the end-users’ interpretative skills. To enhance forthcoming outcomes, we suggest incorporating instruction fine-tuning or training the model to reason based on a chain of thoughts. Additionally, we acknowledged the potential bias introduced by human annotation. Thus, a sturdier automated metric is required to evaluate outputs, proving both effective and economical.

References

- Amina Adadi and Mohammed Berrada. 2018. [Peeking inside the black-box: A survey on explainable artificial intelligence \(xai\)](#). *IEEE Access*, 6:52138–52160.
- David Alvarez-Melis and Tommi Jaakkola. 2017. [A causal framework for explaining the predictions of black-box sequence-to-sequence models](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.
- Dimo Angelov. 2020. [Top2vec: Distributed representations of topics](#). *CoRR*, abs/2008.09470.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: visual question answering](#). *CoRR*, abs/1505.00468.
- Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Vitor Silva, Leonardo Neves, and Francesco Barbieri. 2022. [Twitter topic classification](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3386–3400, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.
- Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020. [Cross-lingual contextualized topic models with zero-shot learning](#). *CoRR*, abs/2004.07737.

Annotation Scheme	Explanation based on predicted answers	Explanation based on ground-truth answers
The explanation is accurate and reliable.	2.84	2.67
The explanation is sufficiently complete.	3.26	3.28
How confident do you feel in your understanding of the model’s decision-making process after reading the explanation?	2.13	2.27
The explanation has sufficient details to explain the output.	2.47	2.56
You are satisfied with the clarity and comprehensibility of the explanation?	2.19	2.24
I feel safe relying on the explanations to make informed decisions.	2.64	2.60
Do the aspects of the explanation increase your confidence in the reliability of the model’s decision?	2.65	2.56

Table 5: The Table presents the averaged scores from three annotators for each devised scheme, showcasing the results of manually annotating the generated explanations.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *CoRR*, abs/2112.04426.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Leo Breiman. 1996. Bagging predictors. *Mach. Learn.*, 24(2):123–140.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Beatrice Cahour and Jean-François Forzy. 2009. Does projection into use improve trust and exploration? an example with a cruise control system. *Safety Science*, 47:1260–1270.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’15, page 1721–1730, New York, NY, USA. Association for Computing Machinery.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj Dogan, and Zhiyong Lu. 2021. Overview of the biocreative vii litcovid track: multi-label topic classification for covid-19 literature annotation.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2020. LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1):D1534–D1540.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-snli-ve-2.0: Corrected

- visual-textual entailment with natural language explanations. *CoRR*, abs/2004.03744.
- William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.
- Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.
- Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning, ICML ’06*, page 377–384, New York, NY, USA. Association for Computing Machinery.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5).
- Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.
- Tin Kam Ho. 1995. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR ’95, page 278, USA. IEEE Computer Society.
- Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5.
- Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided task-aware image captioning.
- Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, and Douglas Eck. 2018. An improved relative self-attention mechanism for transformer with application to music generation. *CoRR*, abs/1809.04281.
- Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, page 624–635, New York, NY, USA. Association for Computing Machinery.
- Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.
- Vivian Lai and Chenhao Tan. 2018. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *CoRR*, abs/1811.07901.
- Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- Angeliki Lazaridou, Adhiguna Kuncoro, Elena Grubovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomás Kociský, Susannah Young, and Phil Blunsom. 2021. Pitfalls of static language modelling. *CoRR*, abs/2102.01951.
- Piyawat Lertvittayakumjorn, Ivan Petanjek, Yang Gao, Yamuna Krishnamurthy, Anna Van Der Gaag, Robert Jago, and Kostas Stathis. 2021. Supporting complaints investigation for nursing and midwifery regulatory agencies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 81–91, Online. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.
- Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17:319 – 330.
- Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, page 2418–2424. AAAI Press.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.
- Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.
- Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.
- Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *CoRR*, abs/2004.03705.
- Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1802.08129.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.
- Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.
- L. S. Shapley. 1953. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton.
- Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. *CoRR*, abs/2004.14224.
- Eunhye Song, Barry L. Nelson, and Jeremy Staum. 2016. Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1060–1083.
- Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. *Interpretable question answering on knowledge bases and text*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4943–4951, Florence, Italy. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. *OCTIS: Comparing and optimizing topic models is simple!* In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. *CoRR*, abs/1706.03762.

Erik Štrumbelj and Igor Kononenko. 2014. *Explaining prediction models and individual predictions with feature contributions*. 41(3):647–665.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. *A broad-coverage challenge corpus for sentence understanding through inference*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. *Visual entailment: A novel task for fine-grained image understanding*. *CoRR*, abs/1901.06706.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *Xlnet: Generalized autoregressive pretraining for language understanding*. *CoRR*, abs/1906.08237.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. *ERNIE: Enhanced language representation with informative entities*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Jian Zhao, Jianshu Li, Yu Cheng, Li Zhou, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2018. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. *arXiv preprint arXiv:1804.03287*.

A Example Appendix

Hyperparameter setting for Random Forest

1. **n_estimators:** 100
2. **bootstrap:** True
3. **criterion:** gini
4. **min_samples_leaf:** 1
5. **max_samples_split:** 2

Table 6: Presents the hyperparameter settings for the Random Forest Classifier.

Hyperparameter setting for XGBoost

1. **eta:** 0.01
2. **num_class:** 3
3. **objective:** multi:softmax
4. **subsample:** 0.5
5. **base_score:** Mean(train-labels)
6. **eval_metric:** mgloss

Table 7: Present the hyperparameter setting for the XGBoost classifier.

Hyperparameter setting for fine-tuning DeBERTa sequence classifier

1. **epochs:** 10
2. **batch-size:** 2
3. **learning-rate:** 1e-6
4. **weight-decay:** 1e-4
5. **loss:** Cross-Entropy loss
6. **optimizer:** AdamW

Table 8: Present the hyperparameter setting for finetuning the DeBERTa classifier.

Example concerning to local interpretability specific to Tree algorithms.

1. Artificial intelligence (AI), a method of simulating the human brain in order to complete tasks in a more effective manner, has had numerous implementations in fields from manufacturing sectors to digital electronics. Despite the potential of AI, it may be obstinate to assume that the person-administered society would rely solely on AI; with an example being the healthcare field. With the ever-expanding discoveries made on a regular basis regarding the growth of various diseases and its preservations, utilizing brain power may be deemed essential, but that doesn't leave AI as a redundant asset. With the years of accumulated data regarding patterns and the analysis of various medical circumstances, algorithms can be formed, which could further assist in situations such as diagnosis support and population health management. This matter becomes even more relevant in today's society with the currently ongoing COVID-19 pandemic by SARS-CoV-2. With the uncertainty of this pandemic from strain variants to the rolling speeds of vaccines, AI could be utilized to our advantage in order to assist us with the fight against COVID-19. This review briefly discusses the application of AI in the COVID-19 situation for various health benefits.

Table 9: The illustrations showcase instances used for local-interpretation analysis, specifically pertaining to Tree algorithms. Illuminates the characteristics that the model examines in order to classify it with respect to a specific category.

Prompt Instructions for Image Captioning with BLIP-2 Model Architecture

1. **Main-caption:** Describe the scene in the image below with an elaborate caption. Pay attention to small details and make the description engaging and immersive.
2. **Object-enlisting:** List all the objects in the image.
3. **Standout-features:** What are the standout elements/features in the image?
4. **Color-enlisting:** Enlist all the colors in the image.
5. **Emotional-ambiance:** Describe the prevailing emotional ambiance depicted in the image.
6. **Location:** Does the setting take place inside or outside.
 - 6.1. **Outdoor-setting:** What weather conditions are depicted in this image?
 - 6.2. **Indoor-setting:** What is the lighting condition of the indoor setting in the image?
7. **Activity:** What kind of activity is occurring in the image?
8. **Human-presence:** Do you see humans in the image?
 - 8.1. **Human-attire:** Please provide a description of the clothing worn by the individuals in the image.
9. **Camera-perspective:** What is the source of camera perspective in the image?

Table 10: Presents the prompts we have designed for creating optimal image captions further used to for Q&A and explanation generation.

Prompts Instructions for Q&A and Explainability with the FLAN-T5-XXL Model Architecture

- 1. Paraphrasing Image Captions:** Paraphrase text and make it grammatically correct.
- 2. Question-Answering:** Answer the question using the provided context.
- 3. Explanation Generation:** Explain the reason behind the answer to the question using the context of the image.

Table 11: Presents the 3 distinct prompts used for the purpose of i) paraphrasing, ii) question-answering, and iii) explanation generation.

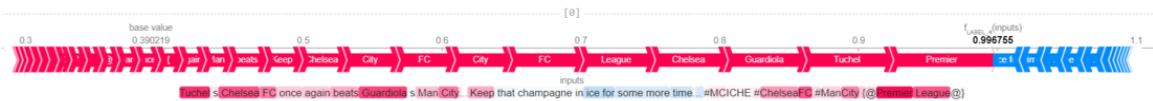
Image	Question	Answer	Explanations (GT)	Explanation (TG)
	What kind of animals are these?	Elephants	They have long trunks, big floppy ears, and weigh a ton.	The elephants are standing in a grassy field
	Is this a hotel room?	No	The large kitchen includes a full-size oven and stove.	The kitchen has a sink, door, stove, refrigerator, window, and a human
	What kind of controller is that?	Wii	The remote has wii written on the bottom of it.	The person is playing a video game.
	Is this a zoo?	No	The zebras are in a large field instead of a cage.	The zebras are in the field and there is no human presence.
	Is it cold?	Yes	They are dressed in winter clothes, walking through snow on skis.	The people are skiing in the snow which is cold.
	Is he getting ready to throw or catch a ball?	Throw	His arm is wound up and he is stepping forward.	The player is on the pitcher's mound.
	Has the sign been defaced? Yes		There is a sticker and graffiti on the back of it.	The sign has been defaced with a picture of flaming lips.

Table 12: Displays particular examples taken from the validation set of the VQA-X dataset, accompanied by TG explanations. It includes the question, answers (ground-truth), explanation (ground-truth), and explanations (generated via prompting LLM) in the same order.

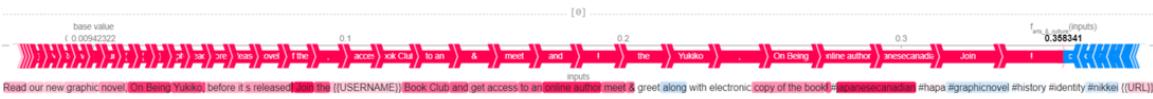
Premise (image)	Hypothesis (text)	Target Label	Predicted Label	Target Explanation
	The child is feeling good.	Neutral	Neutral	Just because the child is giving a donation does not mean that the child is feeling good.
				<p>higher ⇔ lower base value 1.80</p> <p>-1.017 -1.117 -0.6172 -0.1172 0.3828 0.8828 1.383 2.383 2.883 3.383</p> <p>0 = 0.3431 Children at Play_Id = 0.7361 Reading and Intellectual Pursuits_Id = 0.3326 Group Gathering or Waiting Outdoors_Id = 0.1742 Wheeled Sports and Transportation_Id = 0.387 Smiling and Happiness_Id = 0.9581 Cleaning and Household Chores_Id = 0.2498 Crying and Emotional Distress_Id = 0.00005919 Watching TV and Reading_Id = 0.0001612 Water Activities and Recreation_Id = 0.3326 Snow Sports and Mountain Activities_Id = 0.1565 Reading and Intellectual Pursuits_Id = 0.3326 Wheeled Sports and Transportation_Id = 0.387 Group Gathering or Waiting Outdoors_Id = 0.1742</p> <p>higher ⇔ lower base value 0.08</p> <p>-1.472 -0.9717 -0.4717 0.5283 1.428 1.528 2.028 2.528 3.028 3.528</p> <p>0 = 0.3431 Children at Play_Id = 0.7361 Reading and Intellectual Pursuits_Id = 0.3326 Group Gathering or Waiting Outdoors_Id = 0.1742 Wheeled Sports and Transportation_Id = 0.387 Smiling and Happiness_Id = 0.9581 Snow Sports and Mountain Activities_Id = 0.1565 Reading and Intellectual Pursuits_Id = 0.3326 Wheeled Sports and Transportation_Id = 0.387 Group Gathering or Waiting Outdoors_Id = 0.1742</p> <p>higher ⇔ lower base value 0.76</p> <p>-0.6952 -0.1052 0.3948 0.8948 1.395 1.895 2.395</p> <p>0 = 0.04599 Water Activities and Recreation_Id = 0.3326 Cleaning and Household Chores_Id = 0.2498 Snow Sports and Mountain Activities_Id = 0.1565 Animal Play and Activities_Id = 0.2883 Sleep and Rest_Id = 0.2874 Transportation Modes_Id = 0.1495 Group Gathering or Waiting Outdoors_Id = 0.1742 Women's Fashion and Shopping_Id = 0.0001612 Snow Sports and Mountain Activities_Id = 0.1565 Animal Play and Activities_Id = 0.2883 Sleep and Rest_Id = 0.2874 Transportation Modes_Id = 0.1495 Group Gathering or Waiting Outdoors_Id = 0.1742 Women's Fashion and Shopping_Id = 0.0001612</p>
	A woman plays golf.	Contradiction	Contradiction	There can either be a man or woman.
				<p>higher ⇔ lower base value 0.91</p> <p>-0.1717 0.02832 0.2283 0.4283 0.6283 0.8283 1.028 1.228 1.428 1.628 1.828 2.028 2.228</p> <p>0 = 0.0001612 Water Activities and Recreation_Id = 0.3069 Dining and Eating Out_Id = 0.00003636 Snow Sports and Mountain Activities_Id = 0.09001294 Sleep and Rest_Id = 0.00005957 Nudity and Public Spaces_Id = 0.003275 Animal Play and Activities_Id = 0.001335 Transportation Modes_Id = 0.000693 Sports and Athletics_Id = 0.056</p> <p>higher ⇔ lower base value 1.09</p> <p>-1.105 -0.6052 -0.1952 0.3948 0.8948 1.395 1.895 2.395 2.895</p> <p>0 = 0.0586 Transportation Modes_Id = 0.000693 Reading and Intellectual Pursuits_Id = 0.0001067 Watching TV and Movies_Id = 0.00004327 Crying and Emotional Distress_Id = 0.0001056 Snow Sports and Mountain Activities_Id = 0.00001294 Dining and Eating Out_Id = 0.00003636 Cleaning and Household Chores_Id = 0.000009106 Women's Fashion and Shopping_Id = 0.00006926 Sleep and Rest_Id = 0.0005997 Nudity and Public Spaces_Id = 0.003275 Cleaning and Household Chores_Id = 0.000009106 Crying and Emotional Distress_Id = 0.0001056 Smiling and Happiness_Id = 0.001729 Reading and Intellectual Pursuits_Id = 0.0001067 Watching TV and Movies_Id = 0.00004327</p> <p>higher ⇔ lower base value 0.78</p> <p>-0.6172 -0.1172 0.3828 0.8828 1.383 1.883 2.383</p> <p>0 = 0.005968 Women's Fashion and Shopping_Id = 0.1472 Snow Sports and Mountain Activities_Id = 0.001808 Dining and Eating Out_Id = 0.0003729 Cleaning and Household Chores_Id = 0.01569 Animal Play and Activities_Id = 0.00232 Photography and Outdoor Activities_Id = 0.003693 Transportation Modes_Id = 0.02823 Group Gathering or Waiting Outdoors_Id = 0.0005968 Crying and Emotional Distress_Id = 0.02652 Smiling and Happiness_Id = 0.4025 Nudity and Public Spaces_Id = 0.1692 Cleaning and Household Chores_Id = 0.01569 Watching TV and Movies_Id = 0.001402 Wheeled Sports and Transportation_Id = 0.005968 Group Gathering or Waiting Outdoors_Id = 0.4357 Nudity and Public Spaces_Id = 0.1692 Smiling and Happiness_Id = 0.4025 Animal Play and Activities_Id = 0.00232 Dining and Eating Out_Id = 0.0003729 Snow Sports and Mountain Activities_Id = 0.001808 Group Gathering or Waiting Outdoors_Id = 0.02194 Sleep and Rest_Id = 0.001788 Reading and Intellectual Pursuits_Id = 0.0001612 Water Activities and Recreation_Id = 0.3326 Snow Sports and Mountain Activities_Id = 0.1565 Animal Play and Activities_Id = 0.2883 Sleep and Rest_Id = 0.2874 Transportation Modes_Id = 0.1495 Group Gathering or Waiting Outdoors_Id = 0.1742 Women's Fashion and Shopping_Id = 0.0001612</p>
	A woman talking on a phone holds her daughter in a pink outfit.	Entailment	Entailment	A woman talking on a phone holds her daughter or it could be said a little girl is being held by a woman who is on her cellphone.
				<p>higher ⇔ lower base value 1.42</p> <p>-0.3052 -0.1052 0.0948 0.2948 0.4948 0.6948 0.8948 1.095 1.295 1.495 1.695 1.895 2.095</p> <p>0 = 0.005968 Women's Fashion and Shopping_Id = 0.1472 Snow Sports and Mountain Activities_Id = 0.001808 Dining and Eating Out_Id = 0.0003729 Cleaning and Household Chores_Id = 0.01569 Animal Play and Activities_Id = 0.00232 Photography and Outdoor Activities_Id = 0.003693 Transportation Modes_Id = 0.02823 Group Gathering or Waiting Outdoors_Id = 0.0005968 Crying and Emotional Distress_Id = 0.02652 Smiling and Happiness_Id = 0.4025 Nudity and Public Spaces_Id = 0.1692 Cleaning and Household Chores_Id = 0.01569 Watching TV and Movies_Id = 0.001402 Wheeled Sports and Transportation_Id = 0.005968 Group Gathering or Waiting Outdoors_Id = 0.4357 Nudity and Public Spaces_Id = 0.1692 Smiling and Happiness_Id = 0.4025 Animal Play and Activities_Id = 0.00232 Dining and Eating Out_Id = 0.0003729 Snow Sports and Mountain Activities_Id = 0.001808 Group Gathering or Waiting Outdoors_Id = 0.02194 Sleep and Rest_Id = 0.001788 Reading and Intellectual Pursuits_Id = 0.0001612 Water Activities and Recreation_Id = 0.3326 Snow Sports and Mountain Activities_Id = 0.1565 Animal Play and Activities_Id = 0.2883 Sleep and Rest_Id = 0.2874 Transportation Modes_Id = 0.1495 Group Gathering or Waiting Outdoors_Id = 0.1742 Women's Fashion and Shopping_Id = 0.0001612</p> <p>higher ⇔ lower base value 0.77</p> <p>-0.3172 -0.1172 0.08283 0.2828 0.4828 0.6828 0.8828 1.083 1.283 1.483 1.683 1.883 2.083</p> <p>0 = 0.1026 Photography and Outdoor Activities_Id = 0.003693 Crying and Emotional Distress_Id = 0.02652 Smiling and Happiness_Id = 0.4025 Nudity and Public Spaces_Id = 0.1692 Cleaning and Household Chores_Id = 0.01569 Watching TV and Movies_Id = 0.001402 Wheeled Sports and Transportation_Id = 0.005968 Group Gathering or Waiting Outdoors_Id = 0.4357 Nudity and Public Spaces_Id = 0.1692 Smiling and Happiness_Id = 0.4025 Animal Play and Activities_Id = 0.00232 Dining and Eating Out_Id = 0.0003729 Snow Sports and Mountain Activities_Id = 0.001808 Group Gathering or Waiting Outdoors_Id = 0.02194 Sleep and Rest_Id = 0.001788 Reading and Intellectual Pursuits_Id = 0.0001612 Water Activities and Recreation_Id = 0.3326 Snow Sports and Mountain Activities_Id = 0.1565 Animal Play and Activities_Id = 0.2883 Sleep and Rest_Id = 0.2874 Transportation Modes_Id = 0.1495 Group Gathering or Waiting Outdoors_Id = 0.1742 Women's Fashion and Shopping_Id = 0.0001612</p> <p>higher ⇔ lower base value 0.33</p> <p>-0.1717 0.02832 0.2283 0.4283 0.6283 0.8283 1.028 1.228 1.428 1.628 1.828 2.028 2.228</p> <p>0 = 0.1026 Photography and Outdoor Activities_Id = 0.003693 Crying and Emotional Distress_Id = 0.02652 Smiling and Happiness_Id = 0.4025 Nudity and Public Spaces_Id = 0.1692 Cleaning and Household Chores_Id = 0.01569 Watching TV and Movies_Id = 0.001402 Wheeled Sports and Transportation_Id = 0.005968 Group Gathering or Waiting Outdoors_Id = 0.4357 Nudity and Public Spaces_Id = 0.1692 Smiling and Happiness_Id = 0.4025 Animal Play and Activities_Id = 0.00232 Dining and Eating Out_Id = 0.0003729 Snow Sports and Mountain Activities_Id = 0.001808 Group Gathering or Waiting Outdoors_Id = 0.02194 Sleep and Rest_Id = 0.001788 Reading and Intellectual Pursuits_Id = 0.0001612 Water Activities and Recreation_Id = 0.3326 Snow Sports and Mountain Activities_Id = 0.1565 Animal Play and Activities_Id = 0.2883 Sleep and Rest_Id = 0.2874 Transportation Modes_Id = 0.1495 Group Gathering or Waiting Outdoors_Id = 0.1742 Women's Fashion and Shopping_Id = 0.0001612</p>

Table 13: Shows various instances sampled from the test split of the e-SNLI-VE dataset. It provides an interpretation view of how the extracted feature contributes to the classification.

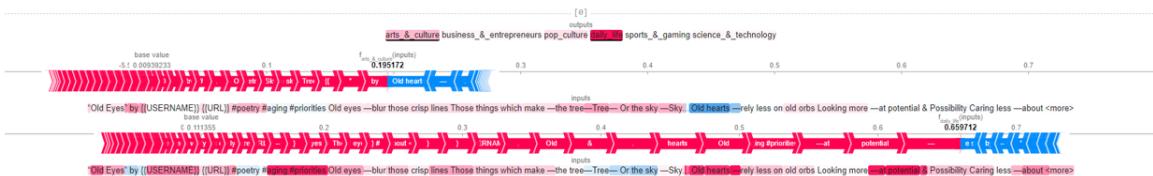
PREDICTION: sports & gaming (0.99553)



PREDICTION: arts & culture (0.35834)



PREDICTION-1: arts & culture (0.19517), PREDICTION-2: daily life (0.65971)



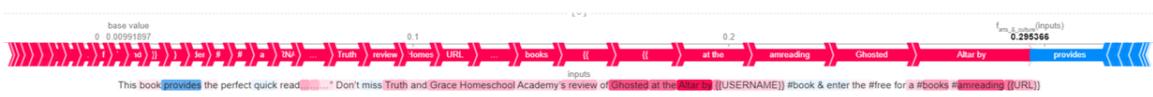
PREDICTION-1: business & entrepreneurs (0.61878), PREDICTION-2: science & technology (0.30818)



This book provides the perfect quick read..." Don't miss Truth and Grace Homeschool Academy's review of
 Ghosted at the Altar by {{USERNAME}} #book & enter the #free #giveaway for a \$25 Amazon gift card
 card#books #amreading {{URL}}
 PREDICTION: business & entrepreneurs (0.28900)



This book provides the perfect quick read..." Don't miss Truth and Grace Homeschool Academy's review of
 Ghosted at the Altar by {{USERNAME}} #book & enter the #free for a #books #amreading {{URL}}
 PREDICTION: arts & culture (0.28900)



This book provides the perfect quick read..." Don't miss Truth and Grace Homeschool Academy's review of
 Ghosted at the Altar by {{USERNAME}}
 PREDICTION: arts & culture (0.483304)



Table 14: The table displays several selected examples from the test subset of the Tweet Topic classification dataset. It offers an interpretive perspective on the words/tokens that contribute towards the target classification. We present the model's predictions for each figure along with their corresponding probabilities.

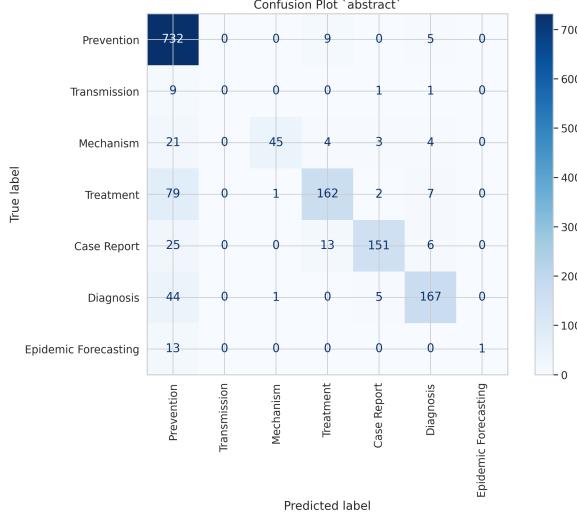


Figure 1: Show the confusion matrix/plot for the tuned Random Forest applied to the test-split of the pre-processed LitCovid track 5 dataset.

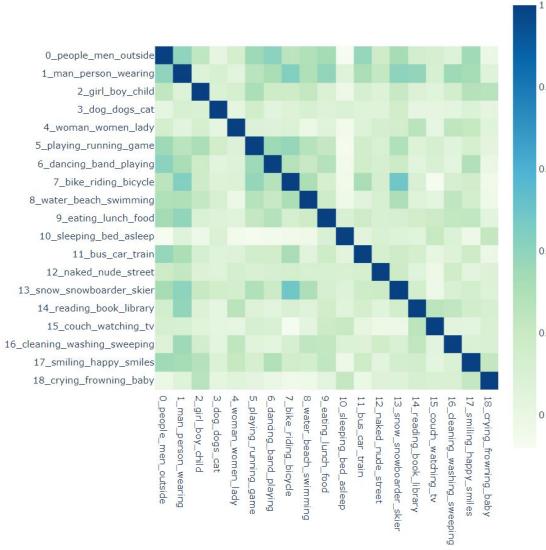


Figure 2: Shows the correlation between derived features using the topic-modeling clusters as shown in figure 9.

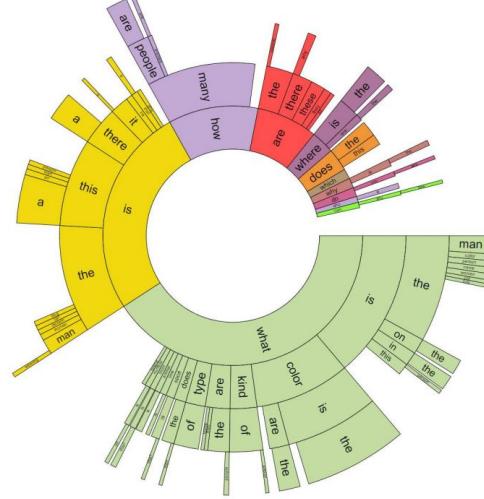


Figure 3: Represent the distribution of the question types across a random sample of 60K instances from the original VQA dataset. (Antol et al., 2015)

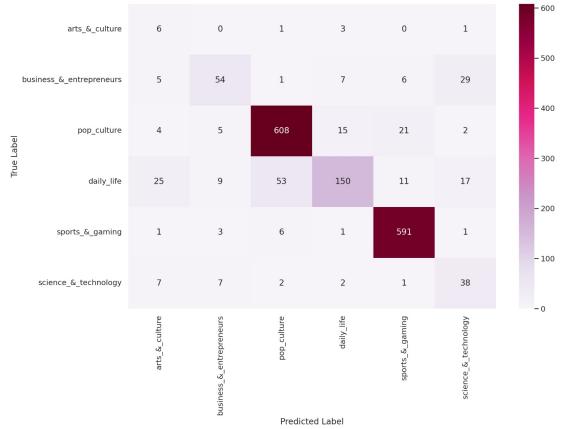


Figure 4: Shows the confusion matrix/plot for the fine-tuned DeBERTa sequence classifier applied to the test-split of the Tweet Topic classification dataset.

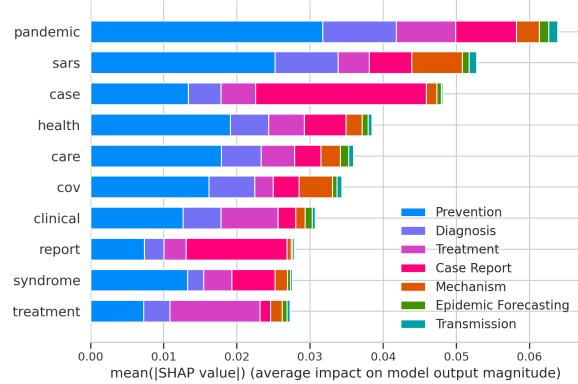


Figure 5: Showcasing the feature importance graph, generated by applying the optimized Random Forest model on the test subset of the pre-processed LitCovid track 5 dataset, reveals the ranked significance of input features.

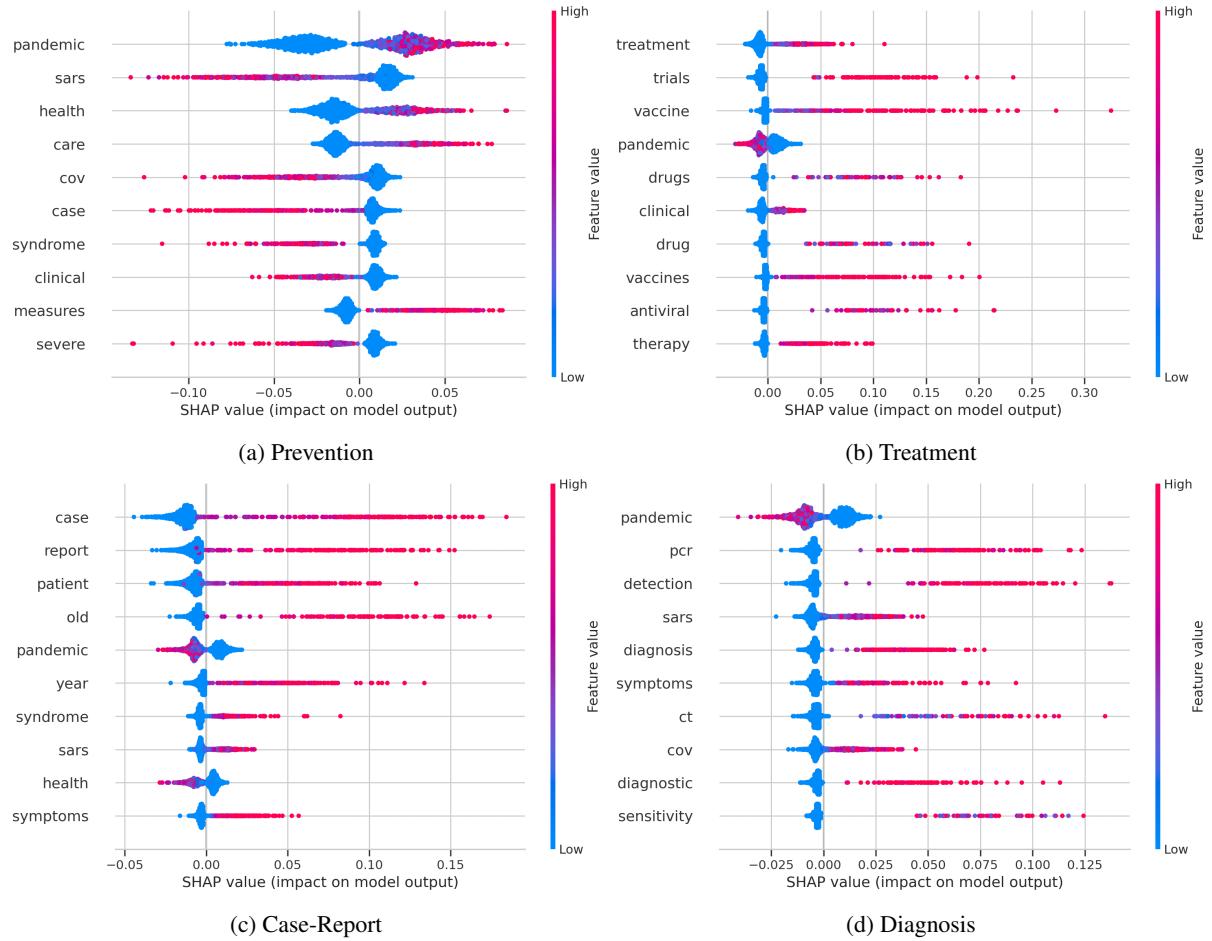


Figure 6: Exhibits the global interpretation of the displayed class categories from the LitCovid track 5 dataset (summary plot). These plots are a derivative of shapely values and emphasizes the importance of the top 10 features/words impacting the model output. Each instance explanation is represented by a single dot.

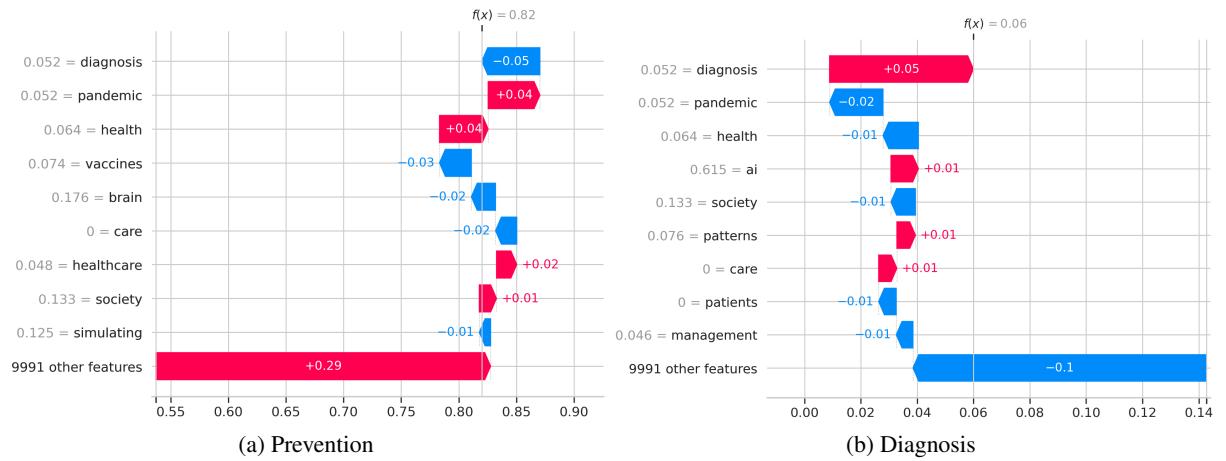


Figure 7: Contrast the local interpretation of an input sequence using a waterfall plot and illustrates the Shapley values of features influencing the computation of the final probability for each category.

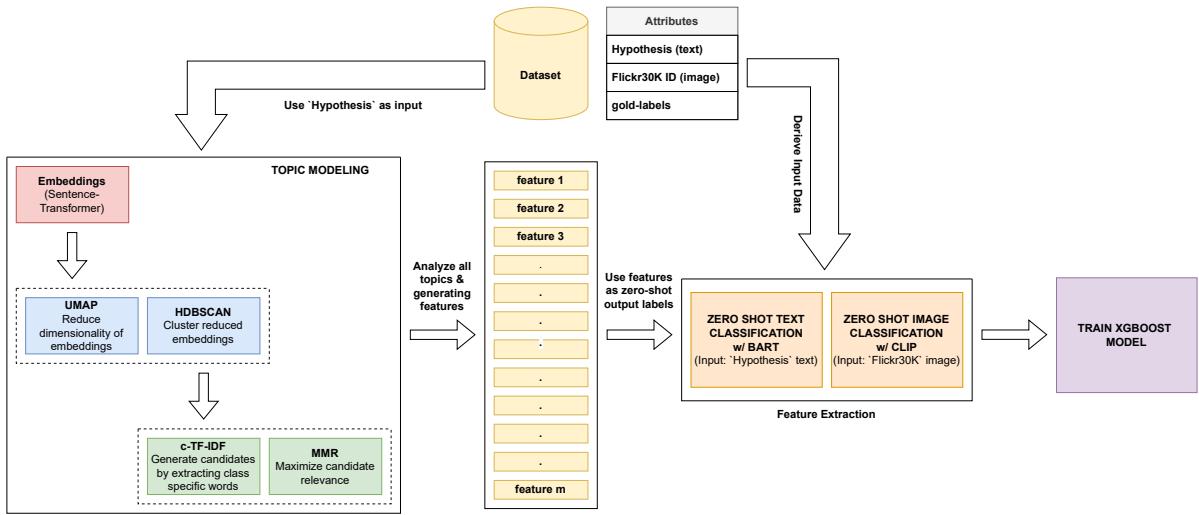


Figure 8: Presents the model architecture/pipeline we use for the multimodal feature attribution task. This consists of various modules, including Topic Modeling, Zeroshot Multimodal Classification, and XGBoost.

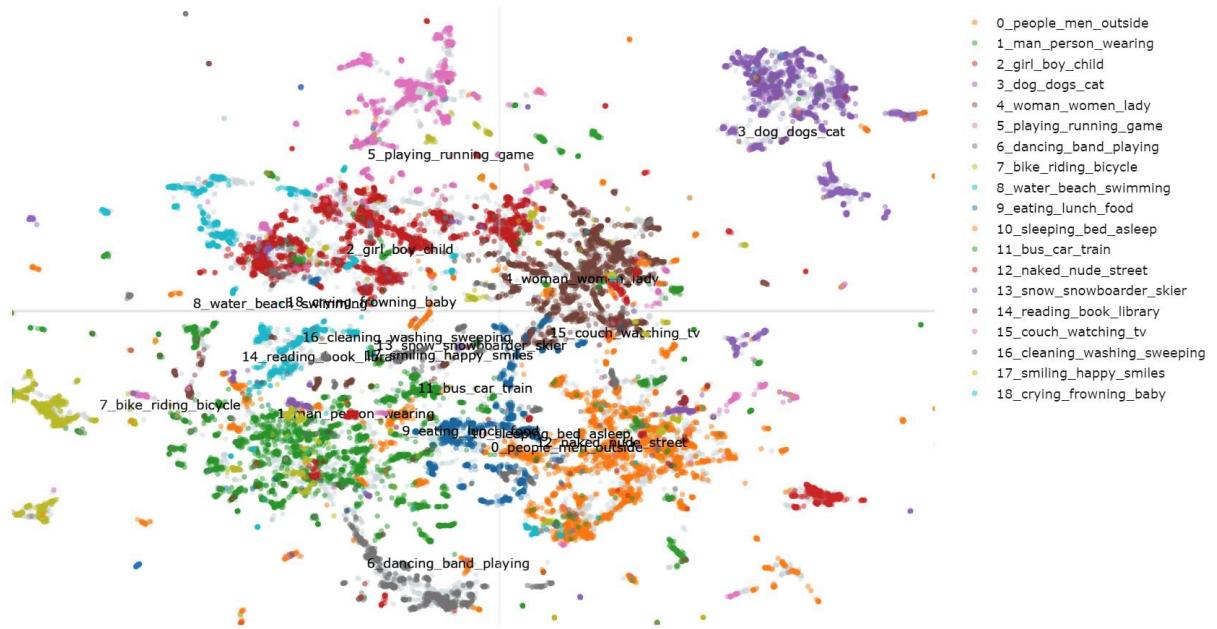


Figure 9: Illustrates how the utilized topic-modeling strategy forms various clusters for the multimodal feature attribution task. Each cluster is depicted using the top three frequently occurring words within the instances categorized under that cluster.

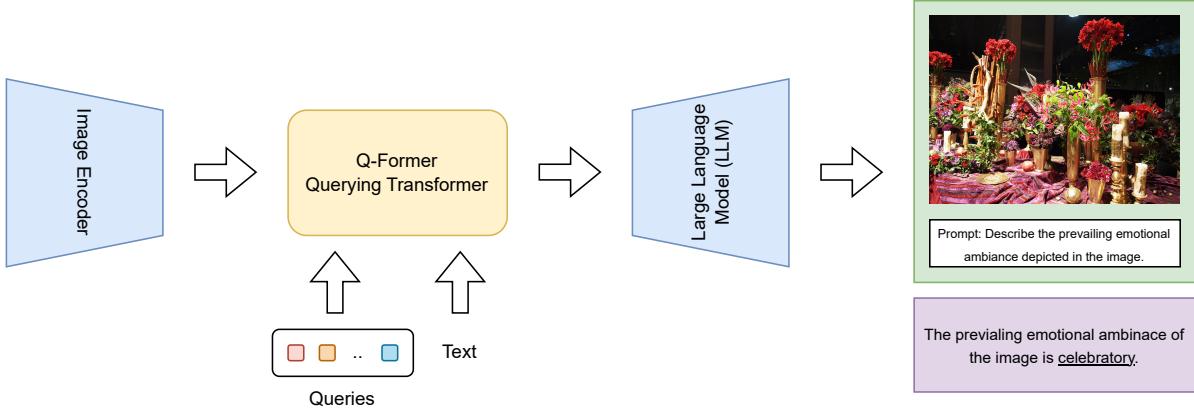


Figure 10: Illustrates the operational flow of the BLIP-2 architecture pipeline. It generates image captions by taking specific prompts as input.

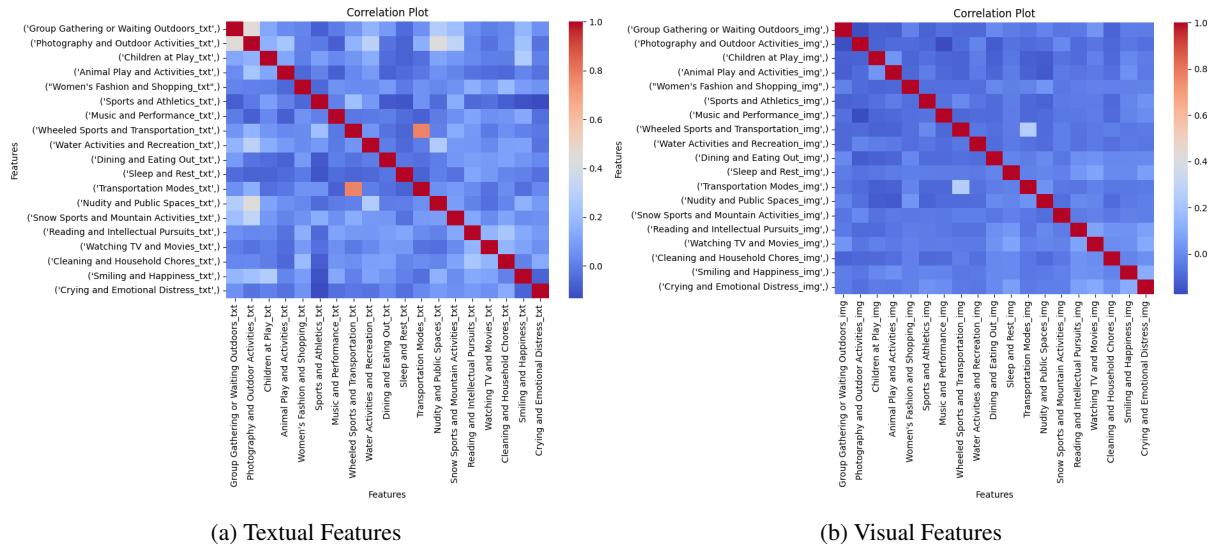


Figure 11: Displays the correlation plot extracted by comparing the attributional value of each of the features present in the image (right) and text (left) features set.

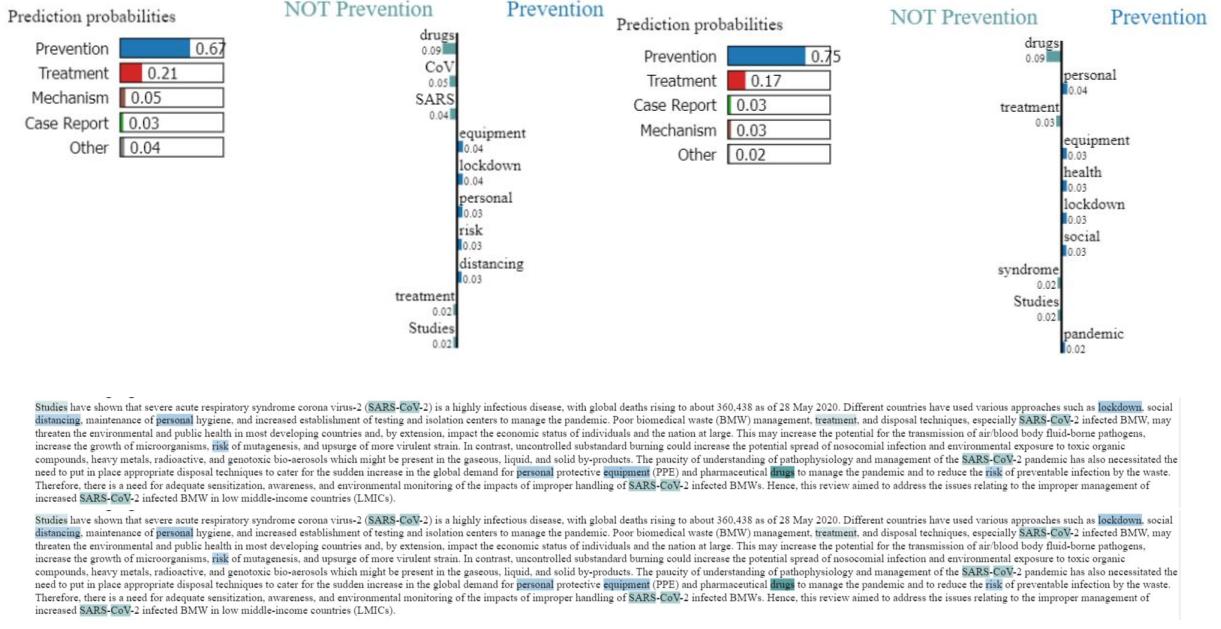


Figure 12: Showcase a counterfactual scenario from the Litcovid track-5 dataset studied using using a tree based algorithm. The text portions reflect the influence observed in the LIME plots. (Note, there might inconsistency while reading due to variations in image quality.)

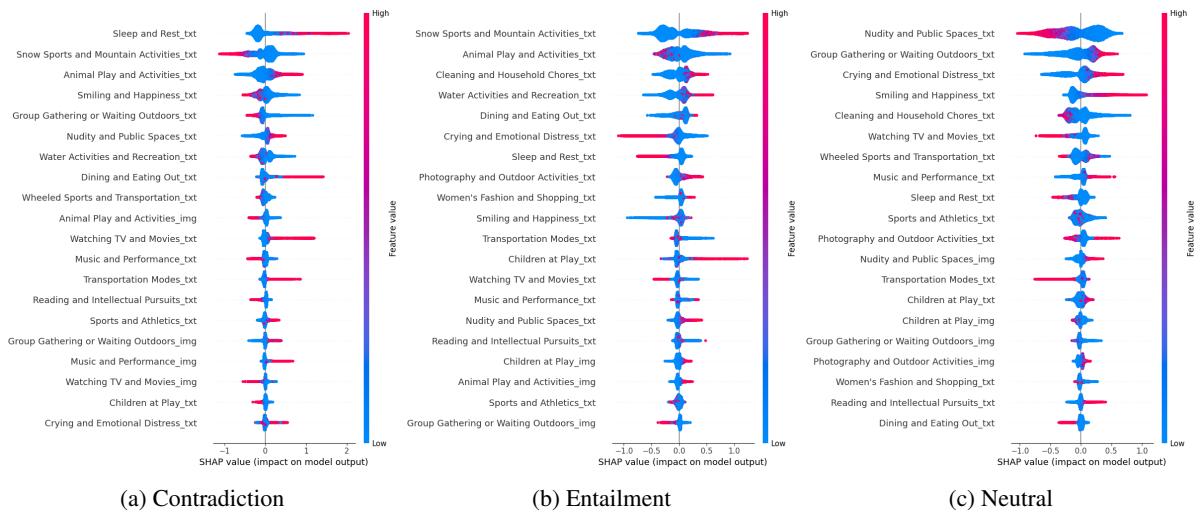


Figure 13: table presents an analysis of the overall importance of features and their impact corresponding to the model output.

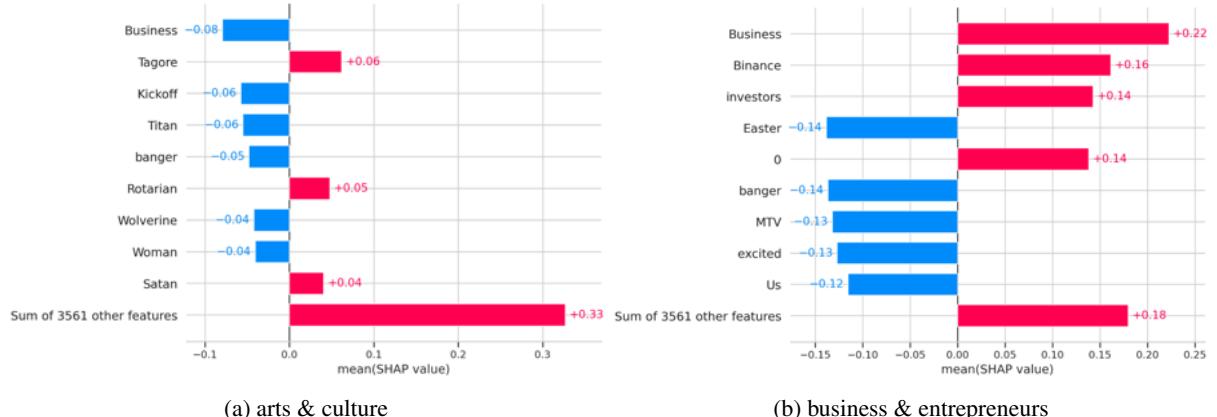


Figure 14: Exhibits the global interpretation of the displayed class categories from the Tweet Topic classification dataset. These plots are a derivative of Shapley values and emphasize the importance of the top 10 words/tokens impacting the overall class category.

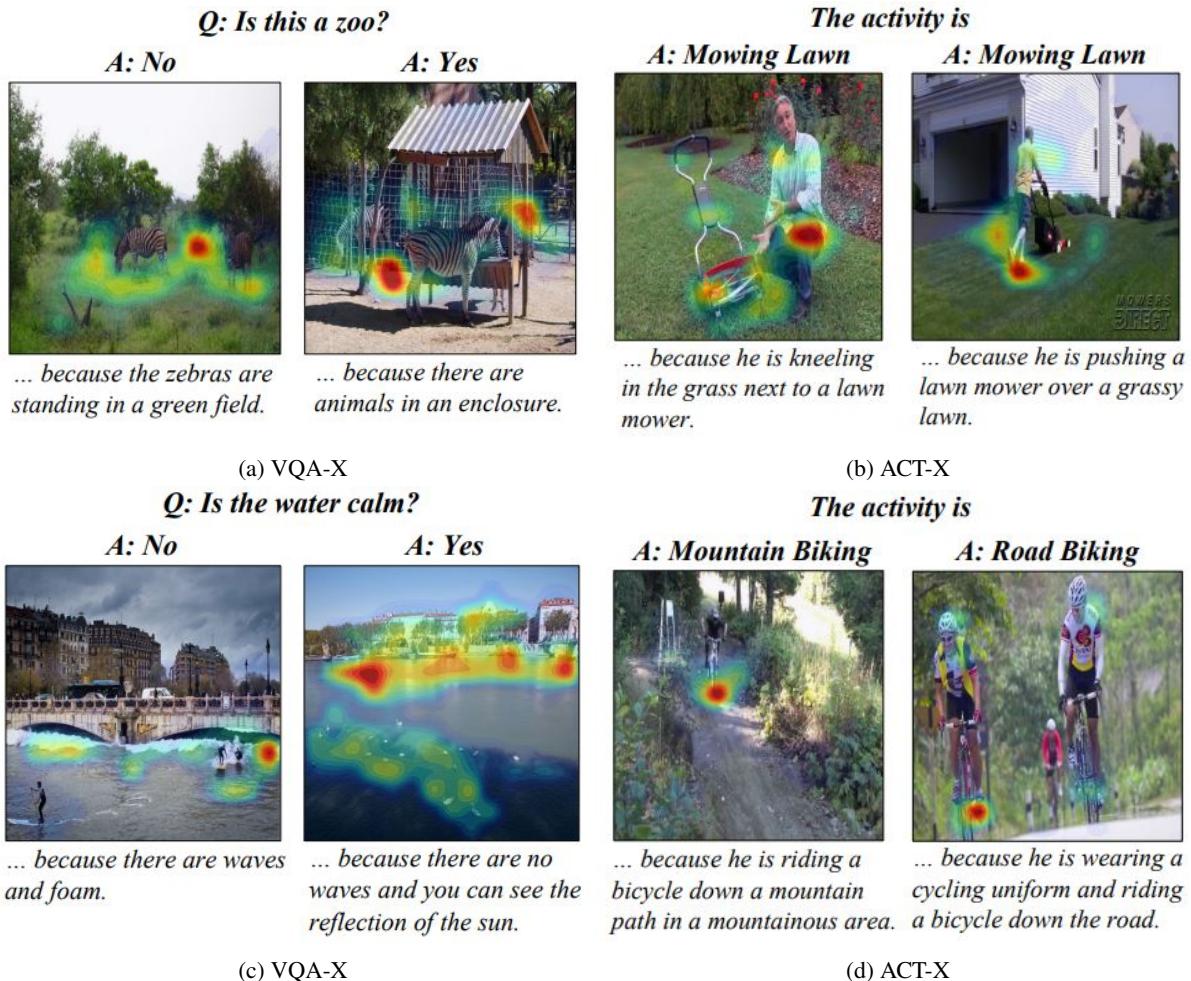


Figure 15: Presents illustrative cases utilized to analyze the PJ-X model, as indicated by Park et al. (2018). The authors highlight the model's positive explanatory characteristics in generating and visually indicating rationales.

Topic-Oriented Feature Attributes
<ol style="list-style-type: none"> 1. Group Gathering or Waiting Outside 2. Photography and Outdoor Activities 3. Children at Play 4. Animal Play and Activities 5. Women's Fashion and Shopping 6. Sports and Athletics 7. Music and Performance 8. Wheeled Sports and Transportation 9. Water Activities and Recreation 10. Dining and Eating Out 11. Sleep and Rest 12. Transportation Modes 13. Nudity and Public Spaces 14. Snow Sports and Mountain Activities 15. Reading and Intellectual Pursuits 16. Watching TV and Movies 17. Cleaning and Household Chores 18. Smiling and Happiness

Table 15: Presents the formulated set of features for the e-SNLI-VE dataset. These are treated as multimodal attributes and are derived using the method outlined in figure 8.