NLP Explainability Techniques

Maria Ribalta i Albado
mribalta001@ikasle.ehu.eus

Abstract

In (Danilevsky et al., 2020), five explainability techniques were defined: Feature Importance, Example Driven, Surrogate Model, Provenance-Based and Declarative Induction under the context of XAI for NLP.

This project undertakes a comprehensive review of each technique, both in a theoretical and experimental framework. We systematically examine each method, implementing at least one corresponding library or algorithm and evaluating its efficacy in terms of explainability. Our assessment encompasses theoretical considerations such as advantages and disadvantages previously stated in the literature, alongside practical aspects such as speed, scalability and reliability of the tested libraries.

Conclusions show that even though some of the techniques seem more solid than others, there is still a lot of room for improvement in the XAI field and there are still lots of open questions that can be addressed.

1 Introduction

In (Danilevsky et al., 2020), the authors reviewed the state of XAI under the scope of the NLP field. They defined five **explainability techniques** to classify the explanations that could be extracted from a model: Feature Importance, Example Driven, Surrogate Model, Provenance-Based and Declarative Induction.

The same year, (Xie et al., 2020) also made a great report describing the state of XAI in a wider context, providing great and deep insights regarding the techniques used and their reliability.

In this project, we understake a comprehensive review of each technique by experimenting with, at least, one library or algorithm. Our assessment encompasses theoretical considerations such as advantages and dis- advantages previously stated in the literature, alongside practical aspects such as speed, scal- ability and reliability of the tested libraries.

The work is available and reproducible in https://github.com/maria-ribalta/nlp_explainability_techniques.

2 NLP Explainability Techniques

In this section, we describe the five explainability techniques in a theoretical framework and the main advantages and concerns found in the literature.

Later on, in section 3 we experiment with specific libraries and derive our own conclusions.

2.1 Feature Importance

The Feature Importance technique derives an explanation by investigating the importance scores of different features used to output the final prediction (e.g. latent features learned by neural networks, attention scores, lexical features, etc).

This technique is interesting since it takes features that are intrinsically from the model to show how it reacts or behaves given an input. Nonetheless, an interpretable feature does not imply explainability in the prediction, in other words, the fact that we can plot a feature does not mean that is relevant to the explanation or the prediction of a given classification or feature.

As mentioned in (Xie et al., 2020), especially for attention scores, there are several concerns in the literature about its reliability regarding using it for understanding models' predictions.

2.2 Example-Driven

Example-driven approaches explain the prediction of an input instance by identifying and presenting characteristics of the given sample or similar ones.

This technique differs from the previous one in the sense that the elements used for the explanations are not part of the model but rather metrics or features extracted from the input samples themselves.

This technique works well but might need specific implementation for each example or input we pass to the model.

2.3 Surrogate Model

In Surrogate Models, predictions are explained by learning a second, usually more explainable model, as a proxy, this is: explaining the decisions of a first model through a second. Surrogate model-based approaches are model-agnostic which allows great flexibility and usage.

However, it is mentioned in the literature that learned surrogate models and the original models may have completely different mechanisms for making predictions, which leads to concerns about the fidelity of surrogate model-based approaches.

2.4 Provenance-Based

Provenance-Based explanations are provided by illustrating some or all of the prediction derivation process. This process is intuitive and effective and the final prediction is the result of a series of reasoning steps.

This technique is understandable but it also implies having a step of the derivational process that can be explained in a human-readable way. The resources to achieve these explanations can be expensive in time, development and/or require human assessment.

2.5 Declarative Induction

The Declarative Induction technique works with human-readable representations, such as rules, trees, and programs. Those themselves are induced as explanations.

Some good examples of those are decision trees (which generate a series of rules that step by step derive the data onto branches) or any other type of rule-based systems.

This technique is the most simple and effective in terms of explainability. Since the model is already self-explanatory, we don't need to put extra effort into understanding predictions or behaviours.

However, these kinds of models are quite limited; for instance: rule-based systems are not scalable and they are high-maintenance and tree models can be too simple for certain complex tasks.

3 Experiments

3.1 Procedure

For the experiments, we have chosen a library and/or algorithm for each explainability technique and coded the steps to showcase the explanations provided by it given a model.

To evaluate it, we have performed **human evaluation** and focused on the following aspects:

- Addressing the advantages and disadvantages of each technique (in a model and library agnostic focus). This has already been discussed in section 2.

- Addressing strengths and weaknesses of the specific libraries or method selected (implementation details, speed of the algorithm, limitations, etc). Here we focus on the model and library used.

In general terms, we will evaluate if the explanation allows us to see if the model predictions are aligned with a human interpretation. In no case, we have judged the quality of the model used or the accuracy of the predictions but rather how the explanations help to understand the results.

3.2 Models

In this project we have used:

- A `google-bert/bert-base-uncased` model which we have fine-tuned for sentiment classification. This has been used in the experiments of the following techniques: Feature Importance, Example Driven and Surrogate Model.

- For the Provenance-based approach, we required a generative model so we used `meta-llama/Llama-2-7b-chat-hf`.

- We have used a `Random Forest` classifier from sklearn in the experiments of Surrogate Model and Declarative Induction.

3.3 Data

The data used to fine-tune the Bert model is the poem_sentiment dataset (Sheng and Uthus, 2020) from Hugging Face. It consists of verses of different poems classified from 0 to 3 according to the sentiment: 0 for negative, 1 for positive, 2 for neutral and 3 for mixed feelings.

Since the dataset was very small, the original test and train datasets were merged. For our experiments and examples, we have used the song All Too Well (10 Minute Version) (Taylor's Version) (From the Vault), which we have manually labelled with the training's dataset criteria.

Specifically in this report, we will show how each technique behaves with the verse number 93: *"From when your Brooklyn broke my skin and bones"*, which is labelled with the tag 0 (negative sentiment). Other verses and their explanation results are available in the public code.

3.4 Feature Importance

Attention scores are easy to plot and are one of the most common features to show when explaining certain models.

The BertViz library (Vig, 2019) allows to automatically plot the attention scores and heads of any Transformer architecture. We have plotted the attention scores of our fine-tuned Bert model with the input sentence mentioned before. The results can be seen in Figure 1.

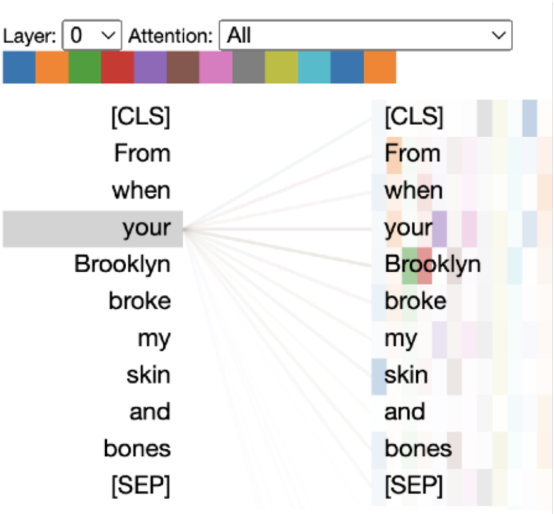


Figure 1: BertViz Attention plot

The BertViz library is smooth and easy to use. The results are fast to compute and the interactivity allows the user to play around. A worth mentioning limitation is that the usage of this library is restricted to transformer architectures.

As for the results obtained, the attention plots are not fully interpretable and very few conclusions can be derived from the charts obtained (especially from the head's plot, available in Appendix A).

As a human, it is hard to check how the plots provided by the library are interpretable or related to the output of the network. From this experiment, we have learnt that the selection of bad features can lead to poor or no explainability at all.

3.5 Surrogate Model

For this technique, we have used the LIME library. Lime (Ribeiro et al., 2016) is the abbreviation for Local Interpretable Model-agnostic Explanations and it supports surrogate model explanations for classifiers.

We provide the code for using LIME's algorithm with a neural network, which was not available online by the time this project was done.

We checked the explanations given by the library comparing two models: our fine-tuned Bert model, which provided a correct classification of the verse (label 0: negative, with 0.99 probability); and an undertrained Random Forest, which provided an incorrect classification of the verse (label 2: neutral, with 0.75 probability).

Explanations can be checked in Figure 2. In 2a we can check how the verb "broke" had a big relevance towards the classification of the verse. In 2b we can see the effects of the undertraining since the words highlighted do not seem to have any relevance regarding the classification of the verse. In fact, they are mostly stop-words, which proves that the model wrongly assigns relevance to words that have little to no meaning for classification and at a lexical level. This technique is useful because we can quickly depict which model works logically and which acts more confusing.

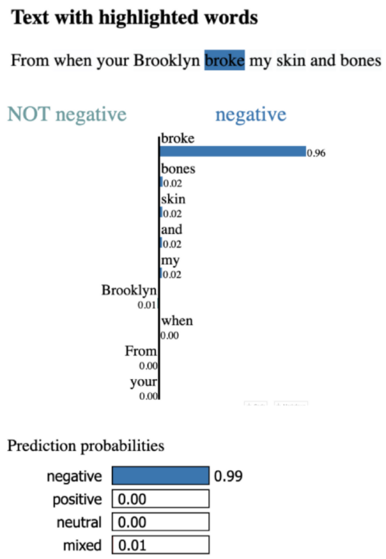
LIME is restricted to classification algorithms so this limits many other models: regression models, generative models, etc. The explainer takes long to process, the explanations for one single verse took 3 minutes when using our neural model.

3.6 Example Driven

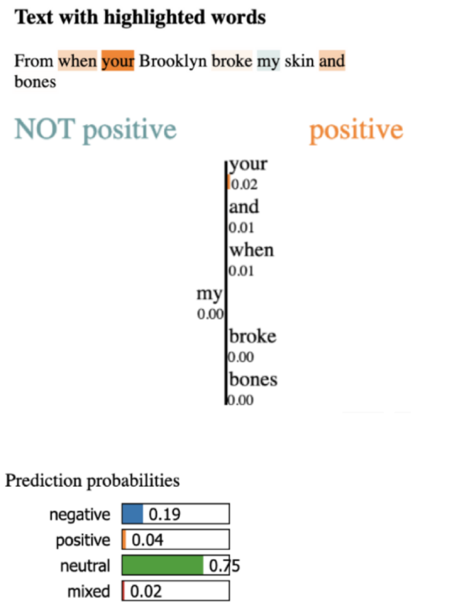
LRP (Bach et al., 2015) stands for Layerwise Relevance Propagation. This method assigns a score to the classification layer and propagates back to each layer until it reaches the input. This way, we get a score that explains the relevance of an input given the predicted class. [1]

To test it, we have used the library interpreting_nlp. We have checked the explanations of our fine-tuned Bert. Results can be checked in Figure 3.

[1]Address to Appendix B for mathematical details



(a) LIME's explanation of fine-tuned Bert



(b) LIME's explanation of undertrained random forest

Figure 2: Comparison of LIME's explainability with two different models

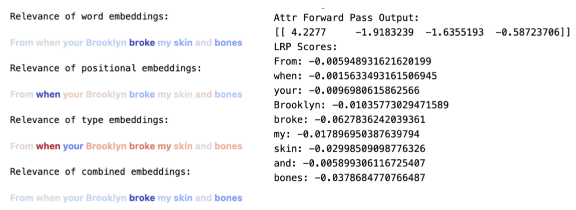


Figure 3: LRP score results

It can be seen that the verb *broke* has the highest LRP score, contributing to the classification of the first class, which is the label assigned to “negative".

The algorithm has proved to detect the verb form “broke" as crucial for the verse to be labelled as negative. As humans, given the sentence, it is logical to think that the verb contributes to a bad connotation in the verse. In terms of library efficiency, the algorithm has run fast.

Example-Driven explanations are one of the most clear and understandable of all the techinques we have tested. They can also have a wide variety of implementations and contexts.

We believe that this is the best technique of the ones we have tried since it finds a balance between good and understandable explanations, reliability and solidity backed by mathematical explanations; and good latency and ease of use of the library.

3.7 Provenance Based

Techniques like COT (Wei et al., 2023) (Chain of Thought) or RAG (Lewis et al., 2021) (Retrieval Augmented Generation) could fall into provenance-based explanations.

In COT, the model (an LLM) provides step-by-step reasoning to reach an objective or decision.

In RAG systems, we send a query that, before reaching the LLM, is complemented with additional information extracted from an external resource (a database, the results of scraping the web, etc). RAG systems allow better contextualization and more accurate responses.

In this experiment, we have coded a toy-RAG system based on (gui). We asked a generative LLM (Llama 2) to classify the verses and provide an explanation for it. We have compared how the decision of the model changed if we provided additional information (using the RAG system querying additional information of the verse extracted from a self-made database) or not (sending only the verse). [2].

[2]Address to Appendix C for complete details on the implementation

Our results show that only in two cases the prediction of the model was the same with and without additional context, which reveals that adding information, (explainability) alters the model's behaviour.

However, our model had better classification results with the non-contextualized samples than those using the RAG system ($f1 = 0.66$ and $f1 = 0.54$ respectively). We attribute it to the simplicity of our system.

One example in which the model benefited from having context is the verse *"From when your Brooklyn broke my skin and bones"*. The LLM labelled it correctly with a 0 (Negative) with the additional context provided by the RAG and incorrectly with a 3 (Mixed) when no contextualized information was passed. However, their explanation showed that the LLM had a hallucination behaviour since it referred to the word "my" as the main cause of "Mixed" classification.[3]

In terms of explainability, even if our experiments didn't come as expected, prompts with additional information allow a better understanding and contextualization of a model's decision. However, this additional information that a human understands might confuse the model and affect the correctness of the answer (which has been our case).

3.8 Declarative Induction

Since this project was mainly focused on Deep Learning methods, we have simply provided an example of a Random Forest model to show how declarative induction works. However, no further contributions have been proposed for this technique.

Scikit Learn implementation allows the visualization of each of the estimators and their derivation rules, as shown in Figure 4. We have trained a Random Forest with our training set and tested it with our test data.

[3]You can check the prompt and the answer of the LLM in Appendix C

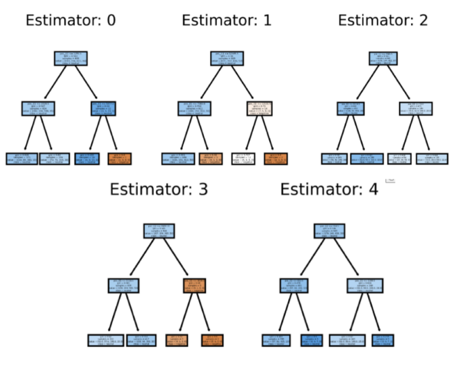


Figure 4: Random Forest

Our implementation of the Random Forest is explanatory but as mentioned, it might not be accurate enough for other tasks. Also, explanations might become incomprehensible if a large number of estimators is used, since it is not possible to check all the rules for each tree in a random forest.

In this case, the model was quite small and worked with bag of words, so the classification result and the rules learned by the model were not quite enlightening. However, the explanations are easy to follow.

4 Conclusions and Future Work

In this project, we have reviewed the five explainability techniques considered for XAI in NLP.

We have seen that **Feature Importance** approaches might involve features that are interpretable but not directly related to the prediction result. Plotting attention scores using BertViz, in our case, has not shown good results in terms of explainability.

In **Surrogate Models**, the explanations shown with LIME's algorithm are interpretable and aligned with the model's predictions. As a human, one can extract good insights using this library. However, there are concerns regarding the reliability of this approach, since the proxy model used for the explanations might have completely different procedures to do predictions than the original one. Therefore, this approach lacks solidity.

In **Example-Driven approaches**, LRP scores have proved to be useful and interpretable. The scores are given by a mathematical formula that involves knowledge regarding the neuron connections and activations, which makes this approach solid. We believe that this is the best approach of the ones we have tried since it has a good balance in speed, explainability and solidity.

Provenance-based approaches are easy to understand. They provide contextualization and clarity at certain steps of the derivation process. Our RAG toy system has shown that a contextualized input leads to a different decision-making by the machine. This approach, however, is not possible or suitable for any task, so its scope is limited.

Finally, **Declarative Induction** techniques rely on the fact that a model's architecture or decision path is explainable by itself which is highly desirable but not possible in most cases. Problems of scalability and/or simplicity in the algorithms are also limitations of this technique.

As of today, there is no standardized way of evaluating XAI approaches. Future work for the scientific community should be to answer questions like: *what is a good explanation? who is the end user of this explanation? what task or procedure are we trying to explain?* and agree to a standardization process for XAI projects.

In conclusion, even if promising, XAI is yet an underdeveloped field that has the potential to grow in the following years.

References

A beginner's guide to building a retrieval augmented generation (rag) application from scratch. https://medium.com/@wachambers/a-beginners-guide-to-building-a-retrieval-augmented-generation-rag-application-from-scratch-e52921953a5d. Accessed: 2024-03-11.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. *CoRR*, abs/2010.00711.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938.

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. 2020. Explainable deep learning: A field guide for the uninitiated. *CoRR*, abs/2004.14545.

A Feature Importance

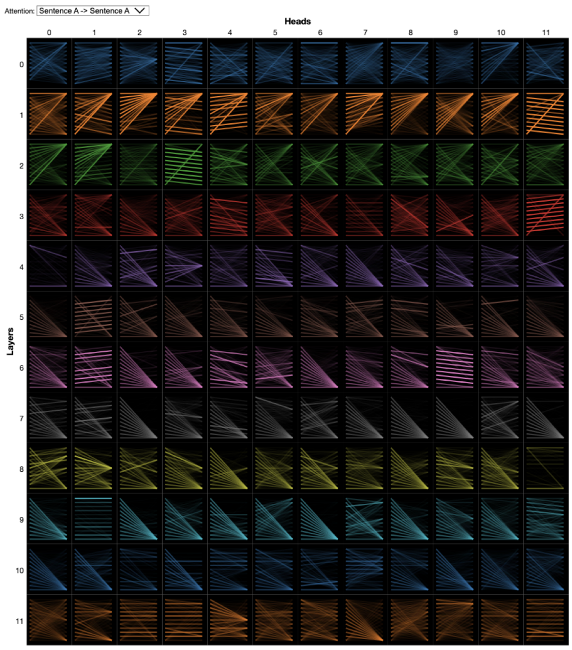


Figure 5: BertViz Attention Heads Plot

B LRP scores

LRP relevance score is defined by:

$$R_i = \sum_j \left(\alpha \cdot \frac{(w_i w_{ij})^+}{\sum_i (w_i w_{ij})^+} - \beta \cdot \frac{(w_i w_{ij})^-}{\sum_i (x_i w_i j)^-} \right) \cdot R_j \quad (1)$$

- * $(x_i)_i$ is the neuron activation at layer l
- $(R_j)_j$ is the relevance score associated to the neurons at layer $l+1$
- w_{ij} is the weight connectin neuron i to neuron j.
- Where $+$ and $-$ denote the positive and negative parts.

- α and β fulfill that: $\alpha - \beta = 1$

C RAG

The high-level components of a RAG system are a corpus, an input from the user and a similarity measure between the corpus and the user input.

In our case, we defined:

- The input: verses of the test data, lyrics from the song. The same as in the previous experiments.
- The corpus: the corpus works as the "additional information" that helps contextualise and enrich the query before reaching the LLM. In this case, we took the comments of the song from the Genius webpage. Genius provides context and explanations of the lyrics from songs.
- The similarity measure: the Jaccard similarity score.

Then, the process followed for the RAG system:

1. Receive a user input (the lyrics)
2. Perform a similarity measure (we chose the Jaccard similarity as in the article) to match the most suitable document (the context from Genius)
3. Send to the LLM the query with the added information. Our LLM is Llama2 and we call it via the ‘ollama‘ proxy.

The Jaccard similarity used for querying the corpus given the input lyric is defined as:

$$J(i, d) = \frac{|i \cap d|}{|i \cup u|}, i \in I, d \in D \qquad (2)$$

Where:

- D is the set of documents, in our case the additional information extracted from Genius.
- I is the user input, in our case, the verse from the test.
- The intersection $\cap$ represents the total number of words that appear in both: the verse and the document.
- The union $\cup$ represents the total number of words that appear in the verse or the document.

We take the document d that has the higher Jaccard Similarity given the input i.

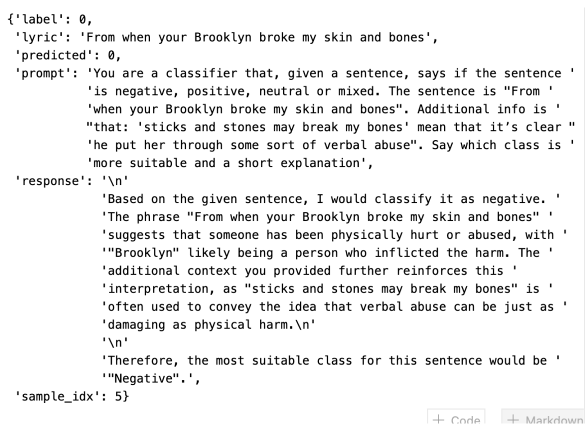
The prompt used as input for queries with contextualized information:

```
You are a classifier that, given a sentence, says if the sentence is negative, positive, neutral or mixed. The sentence is "{song_lyric}". Additional info is that: {additional_information}". Say which class is more suitable and a short explanation
```

The prompt used as input for queries without contextualized information:

```
You are a classifier that, given a sentence, says if the sentence is negative, positive, neutral or mixed. The sentence is "{user_input}". Say which class is more suitable and a short explanation
```

Our example of verse 93:



```
{'label': 0,
 'lyric': 'From when your Brooklyn broke my skin and bones',
 'predicted': 0,
 'prompt': 'You are a classifier that, given a sentence, says if the sentence '
           'is negative, positive, neutral or mixed. The sentence is "From '
           'when your Brooklyn broke my skin and bones". Additional info is '
           "that: 'sticks and stones may break my bones' mean that it’s clear "
           'he put her through some sort of verbal abuse". Say which class is '
           'more suitable and a short explanation',
 'response': '\n'
             'Based on the given sentence, I would classify it as negative. '
             'The phrase "From when your Brooklyn broke my skin and bones" '
             'suggests that someone has been physically hurt or abused, with '
             '"Brooklyn" likely being a person who inflicted the harm. The '
             'additional context you provided further reinforces this '
             'interpretation, as "sticks and stones may break my bones" is '
             'often used to convey the idea that verbal abuse can be just as '
             'damaging as physical harm.\n'
             '\n'
             'Therefore, the most suitable class for this sentence would be '
             '"Negative".',
 'sample_idx': 5}
```


Figure 6: Input and output of a prompt with additional information provided by a toy-RAG system. The LLM predicts the label correctly.

```
{'label': 0,
 'lyric': 'From when your Brooklyn broke my skin and bones',
 'predicted': 3,
 'prompt': 'You are a classifier that, given a sentence, says if the sentence '
           'is negative, positive, neutral or mixed. The sentence is "From '
           'when your Brooklyn broke my skin and bones". Say which class is '
           'more suitable and a short explanation',
 'response': '\n'
             'Based on the sentence provided, I would classify it as "Mixed" '
             'because it contains both negative and positive elements.\n'
             '\n'
             'The phrase "your Brooklyn broke my skin and bones" is negative '
             'in tone, as it describes physical harm caused by something '
             '(Brooklyn) that is presumably a person or entity. The use of the '
             'word "broke" implies damage or injury, which has a negative '
             'connotation.\n'
             '\n'
             'However, the sentence also contains positive elements, such as '
             'the use of the word "my," which suggests ownership or attachment '
             'to the thing being described. Additionally, the phrase "skin and '
             'bones" could be interpreted as a metaphor for something fragile '
             'or delicate, which has a neutral connotation.\n'
             '\n'
             'Therefore, based on the mixed nature of the sentence, I would '
             'classify it as "Mixed."',
 'sample_idx': 5}
```

Figure 7: Input and output of a prompt without additional information. The LLM does not predict the label correctly.