# Explainable AI Final Project

Luca Maas, s1096050

April 16, 2025

## Contents

## GitHub

The accompanying use case code notebook and PDF can be found on GitHub

## Word Count

This report contains approximately 1540 words, excluding figures and references.

# 1 Introduction

Large language models like BERT and GPT have proven incredibly capable in the domain of natural language processing, powering a wide range of tasks such as sentiment classification, summarisation, and translation. These models are utilised by many big tech companies, like Google and Microsoft. However, due to their complexity, LLMs often behave like black boxes, making it difficult to understand how they arrive at certain outputs. For this reason, methods that attempt to explain their decision-making processes are crucial (Zhao et al., 2024).

In this report, two categories of explanations will be discussed: attention-based methods, like attention heatmaps, which use the internal attention weights of LLM models to showcase what tokens the model attends to; and post-hoc methods, like LIME (Ribeiro et al., 2016), which generate an explanation after the model has made its prediction.

I describe how both methods work and can be evaluated with respect to faithfulness—whether explanations reflect the model's reasoning—and plausibility—whether they align with human intuition. A use case in sentiment analysis is proposed—where both LIME and attention heatmaps are applied to the IMDB movie review dataset.

The research question thus is: *"To what extent can attention-based and post-hoc explanation methods provide plausible and faithful explanations for LLM predictions?"*

# 2 Methods

## 2.1 Attention-based methods

Attention is a crucial component of transformer-based LLM models, introduced by Vaswani et al. (2017). As shown in 1, it computes pairwise interactions between tokens using scaled dot-products of query and key vectors, normalised by the square root of the dimension of the key vector.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

This mechanism, known as self-attention, allows each token to attend to every other token, including itself. The resulting attention weights indicate how much each token attends to other tokens and are used to compute weighted output vectors passed to the next layer of the model, capturing contextual relationships. With multi-head attention, multiple sets of attention weights are computed simultaneously, allowing the model to attend to information at different locations in input (Vaswani et al., 2017). Explanations typically rely on the weights from the last layer, either from the first head or the average across heads.

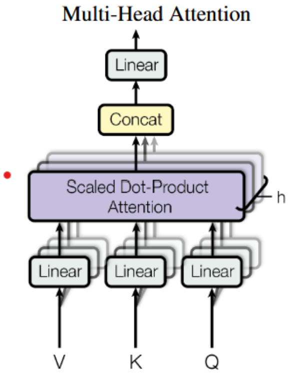


Figure 1: Illustration of the multi-head attention mechanism, where multiple attention layers (heads) run in parallel to capture different types of relationships between tokens. The outputs of all heads are concatenated and linearly transformed to form the final representation. Figure adapted from Vaswani et al. (2017)

There are multiple ways to represent attention weights for explanation: Visualisations, function-based methods, and probing-based methods (Zhao et al., 2024). This report focuses on visual representations. Examples are bipartite graphs, which indicate how much each token attends to all other tokens, and attention heatmaps, which show the weights as a matrix with tokens on both axes. Examples of heatmaps appear in the results section.

## 2.2 Post-Hoc Methods

Post-hoc methods are applied after a model is trained. They only require a model's input(s) and output(s), and can thus be applied on black box models, without needing their internal architecture. This report focuses on LIME.

Local interpretable model-agnostic explanations (LIME) proposed by Ribeiro et al. (2016), is a method that creates local explanations, i.e., it explains a single prediction. Given an input, LIME creates a large amount of perturbed variations of it, for example, by masking certain words. These are passed through the original classifier model via a wrapper function that outputs label probabilities. A model, usually linear, is then fitted to approximate the behaviour of the black-box model around the original input. The learned weights are then interpreted as the importance of input features, in this case words or tokens. These can be negative or positive, indicating their contribution to the predicted class. Example LIME outputs can be seen in the results section and the accompanying code notebook and PDF.

# 3 Evaluation

When discussing the evaluation of explanations for LLMs, two terms often arise: faithfulness and plausibility. Faithfulness refers to how closely an explanation reflects the model's internal reasoning, while plausibility concerns whether the explanation aligns with human intuition (Jacovi and Goldberg, 2020), i.e., whether the output makes sense. Although both are important, prioritising faithfulness helps avoid misleading conclusions that only *appear* intuitive.

For attention-based explanations, it is often assumed that higher attention weights mean higher importance. However, Jain and Wallace (2019) found that attention weights are often uncorrelated with gradient-based feature importance measures, and that a variety of attention distributions could lead to the same predictions. Furthermore, Serrano and Smith (2019) found that removing high-attention tokens did not always affect the prediction, suggesting that attention is not necessarily equal to importance.

Yet, Wiegreffe and Pinter (2019) argue that attention can still be used as an explanation when evaluated accompanied by other methods. As for plausibility, attention (heat)maps are intuitive for humans, especially when represented as text with high attention score words highlighted. However, since attention scores come from the model's internal architecture, they may highlight tokens which do not appear intuitively important to humans.

LIME is also often evaluated on its plausibility and faithfulness. LIME explanations are generally seen as very intuitive, because importance scores are presented in an interpretable, visually appeasing, human-friendly way. This increases users' trust in the prediction (Ribeiro et al., 2016).

However, because LIME locally assumes linear decision boundaries, it may produce oversimplified explanations for complex models. This indicates a trade-off between faithfulness and plausibility. Moreover, LIME is highly sensitive to small changes in input. Leaving out tokens with high importance scores can result in significantly different predictions, suggesting limitations in robustness and faithfulness (Alvarez-Melis and Jaakkola, 2018).

# 4 Use Case

A domain in which explainability methods like LIME and attention maps could be useful is sentiment analysis. In sentiment analysis, NLP models extract sentiments or emotions from text (Birjali et al., 2021). These models are often used to classify text as either positive, neutral, or negative, and are applied in areas such as crowd analysis and customer feedback. Their outputs can directly influence decisions in organisations such as governments and businesses. Given this impact, understanding how these models arrive at their conclusions is crucial.

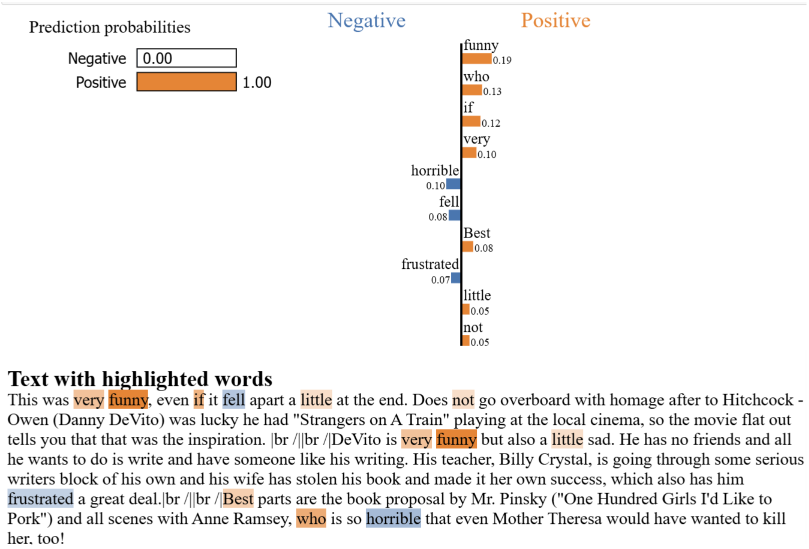
Explainability methods can reveal why a model is producing a certain output.

In customer feedback, explanations can highlight to which product aspects customers react positively or negatively, giving inspirations for possible improvements. Attention maps, for example, could show what negative terms such as “bad” are most associated with. Additionally, explainability can uncover biases, such as models associating negative sentiment with language patterns common in dialects or certain communities, reinforcing harmful stereotypes.

To show how LIME and attention-based methods can be applied to sentiment analysis, I used the IMDB dataset. It contains movie reviews, which are labled either as positive or negative. I applied a DistilBERT model on the dataset and generated both LIME and attention weights for multiple example reviews. The implementation and full results are showcased in the attached code notebook and PDF, which can be accessed on GitHub

# 5 Results

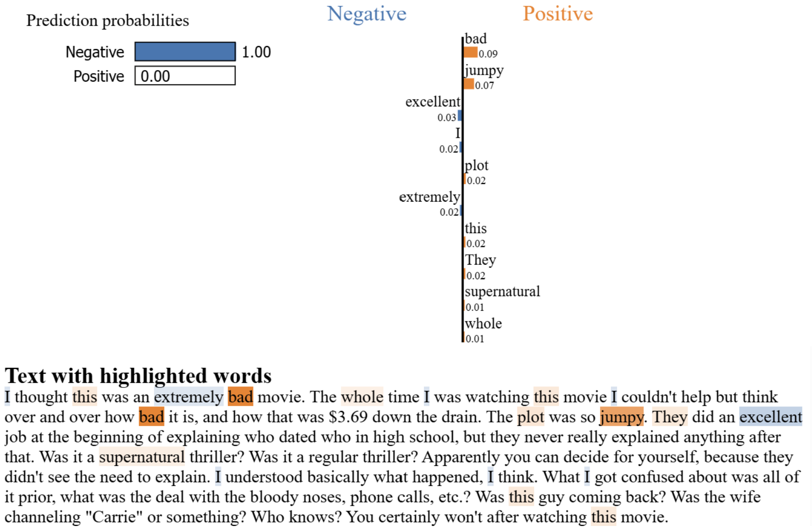
## 5.1 LIME



**Text with highlighted words**

This was very funny, even if it fell apart a little at the end. Does not go overboard with homage after to Hitchcock - Owen (Danny DeVito) was lucky he had "Strangers on A Train" playing at the local cinema, so the movie flat out tells you that that was the inspiration. |br /||br /|DeVito is very funny but also a little sad. He has no friends and all he wants to do is write and have someone like his writing. His teacher, Billy Crystal, is going through some serious writers block of his own and his wife has stolen his book and made it her own success, which also has him frustrated a great deal.|br /||br /|Best parts are the book proposal by Mr. Pinsky ("One Hundred Girls I'd Like to Pork") and all scenes with Anne Ramsey, who is so horrible that even Mother Theresa would have wanted to kill her. too!

Figure 2: An example of a LIME explainer output on a positive IMDB review.



**Text with highlighted words**

I thought this was an extremely bad movie. The whole time I was watching this movie I couldn't help but think over and over how bad it is, and how that was $3.69 down the drain. The plot was so jumpy. They did an excellent job at the beginning of explaining who dated who in high school, but they never really explained anything after that. Was it a supernatural thriller? Was it a regular thriller? Apparently you can decide for yourself, because they didn't see the need to explain. I understood basically what happened, I think. What I got confused about was all of it prior, what was the deal with the bloody noses, phone calls, etc.? Was this guy coming back? Was the wife channeling "Carrie" or something? Who knows? You certainly won't after watching this movie.

Figure 3: An example of a LIME explainer output on a negative IMDB review.

## 5.2 Attention-based Methods

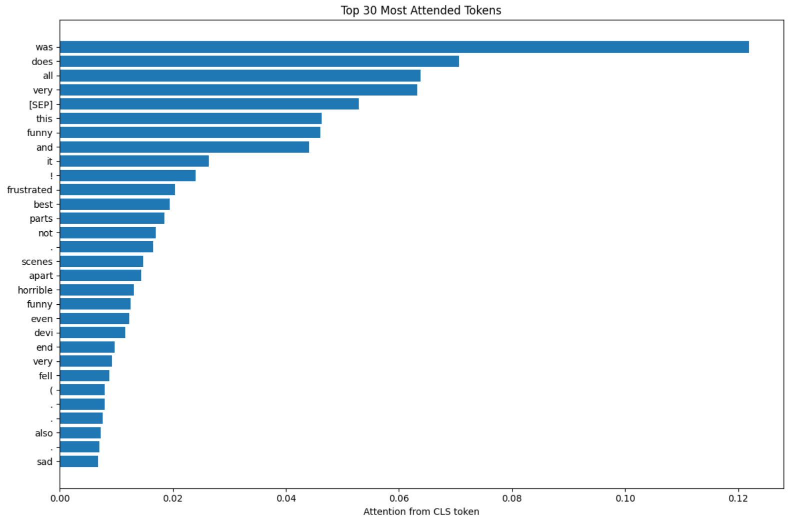


Figure 4: The tokens which were most attended to by the model for a positive IMDB review.

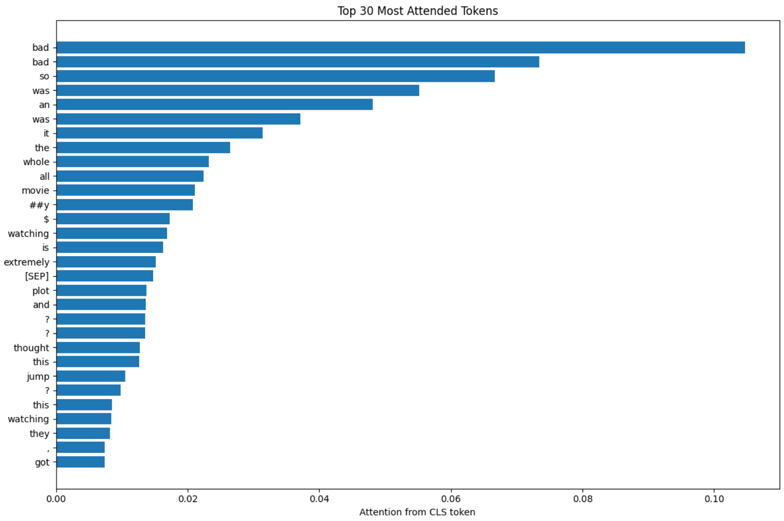


Figure 5: The tokens which were most attended to by the model for a negative IMDB review.

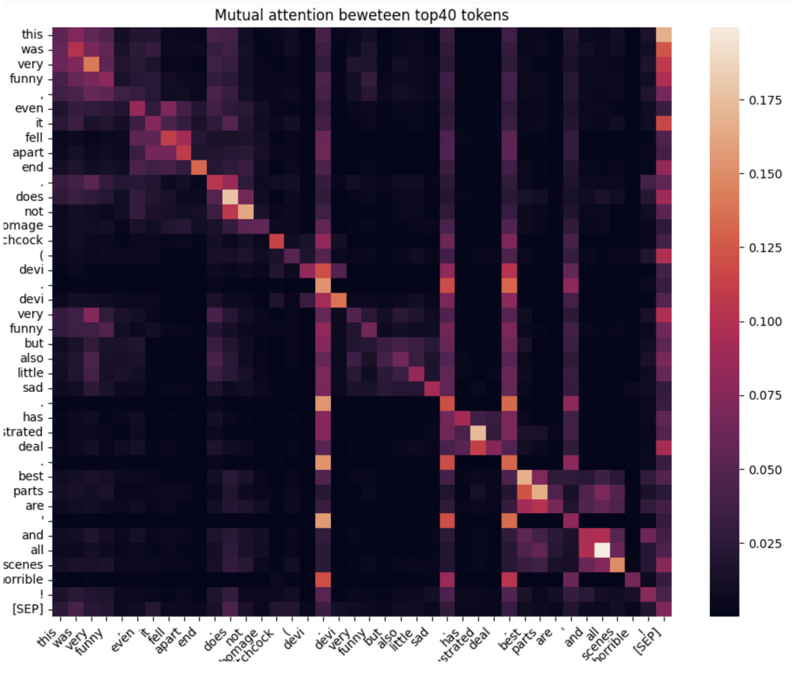


Figure 6: An attention heatmap showing the cross-attention between the top 40 most attended to tokens for a positive IMDB review. The rows are the attending tokens and the columns the attended-to tokens

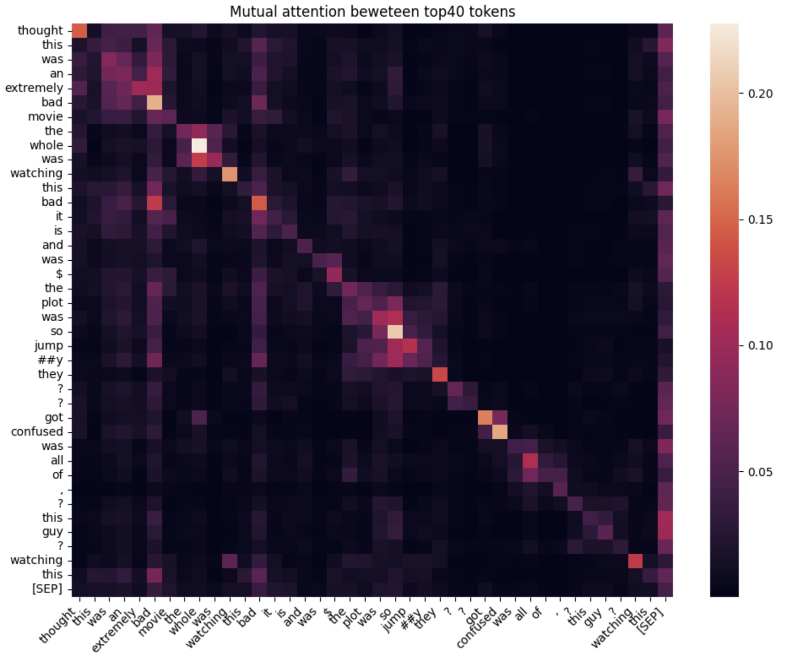


Figure 7: An attention heatmap showing the cross-attention between the top 40 most attended to tokens for a negative IMDB review. The row sare the attending token and the columns the attended-to tokens

## 6 Discussion

Figures 2 and 3 show LIME explanations for positive and negative IMDB reviews. The bar charts show which words contributed most to the prediction. The tokens most attended to by the model can be seen in figures 4 and 5. In figures 6 and 7 heatmaps are shown for the same two inputs. The rows are the attending tokens and the columns the attended-to tokens. Lighter colours indicate stronger attention.

From a human-intuition viewpoint, both methods perform well, though LIME stands out. Its importance weights are normalised, and visually linked to the highlighted and coloured words in the input text, making it highly interpretable. For this reason, LIME performs well when it comes to plausibility.

The attention-based methods were somewhat interpretable, but less so than LIME. In some aspects they performed well, like in the second example input 7. High attention weights were given to the word *bad*, which intuitively makes sense for a negative prediction. In other cases, like in 6, high attention scores were given to the words *was* and *does* which seems less informative. While the attention heatmaps show interesting patterns such as grouped or related words paying attention to each other, it is difficult to interpret, especially since the input was too long to include all words in the heatmap. Additionally, the presence of special characters and punctuation marks makes the explanation less interpretable. A more plausible explanation could be to highlight tokens in the input text such as in LIME.

Both methods also carry pitfalls. As mentioned previously, multiple researchers argue that attention weights do not properly reflect a model's reasoning (Jain and Wallace, 2019; Serrano and Smith, 2019). LIME, though highly interpretable, is at risk of oversimplifying complex models' behaviour due to the linear nature of its local approximations. Moreover, studies show that LIME is sensitive to small changes in its input, casting doubt over its robustness (Alvarez-Melis and Jaakkola, 2018).

Explainability also raises ethical concerns. LIME explanations being intuitive and straightforward might lead to overtrust or overdependence. A balance must be found between algorithmic aversion (distrust) and automatic complacency (over-reliance) (Zerilli et al., 2022). Additionally, care must be taken as to who has access to the outputs of explanation models. Otherwise, users with malicious intent could boost sentiment artificially, by using specific words they know the model is sensitive to. This could have consequences in domains such as movie or restaurant reviews.

# 7 Conclusion

This report aimed to evaluate to what extent attention-based and post-hoc explanation methods could provide faithful and plausible explanations. Insights were drawn from literature and a practical use case.

Both attention-based and post-hoc methods were shown to have advantages and limitations. Attention-based methods could provide valuable insights, but may not always reflect the model's internal reasoning. LIME, while aligning with human intuition, risks oversimplifying its complex model behaviour.

It has become clear that, although these methods can provide valuable insights, they should not be blindly trusted. Rather, their output should be interpreted with caution.

The trade-off between faithfulness and plausibility remains central to XAI, and work should focus on refining existing methods and developing new approaches to improve the balance between faithfulness and plausibility.

# References

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Jacovi, A. and Goldberg, Y. (2020). Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (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*, pages 1135–1144.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, 3(4).

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., and Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.