

Hybrid Approach to Explain BERT Model: Sentiment Analysis Case

Aroua Hedhili^{1,2} ^a and Islem Bouallagui^{1,2} ^b

¹National School of Computer Sciences, Manouba University, Manouba 2010, Tunisia

²Research Lab: Laboratory of Research in Artificial Intelligence LARIA, ENSI, University of Manouba, Tunisia

Keywords: Counterfactual Explanation, Domain Knowledge, User-Centric Approach, BERT.

Abstract: The increasing use of Artificial Intelligence (AI), particularly Deep Neural Networks (DNNs), has raised concerns about the opacity of these 'black box' models in decision-making processes. Explainable AI (XAI) has emerged to address this issue by making AI systems more understandable and trustworthy through various techniques. In this research paper, we deal with a new approach to explain model combining counterfactual explanations and domain knowledge visualization. Our contribution explores how domain knowledge, guided by expert decision-makers, can improve the effectiveness of counterfactual explanations. Additionally, the presented research underscores the significance of collecting user feedback to create a human-centered approach. Our experiments were conducted on a BERT model for sentiment analysis on IMDB movie reviews dataset.

1 INTRODUCTION


Currently, Artificial Intelligence (AI), and more specifically Machine and Deep Learning technologies, are widely used in various fields such as healthcare, finance, and business. However, the opacity of AI decision-making processes presents a significant challenge, especially in high-stakes scenarios where the consequences of AI decisions can be critical. This opacity often results in AI models being perceived as 'black boxes', causing users to question and distrust the technology. To bridge the gap in comprehension between users and AI technology, Explainable AI (XAI) is used to enhance transparency, accountability, fairness, and user trust in AI systems. It is particularly important in applications where the consequences of AI decisions are significant and where human understanding of those decisions is crucial. (Dikmen and Burns, 2022). It is important to recognize that XAI is not a novel concept. Its roots can be traced back to earlier efforts in the field of artificial intelligence and machine learning, where interpretability and transparency were considered desirable qualities. However, what sets the contemporary discourse on XAI apart is the renewed and profound exploration of these principles within the context of deep learning models like BERT (Bidirectional Encoder Representations from Transformers). While BERT has demonstrated remarkable natural language un-


derstanding capabilities, its inherent complexity has raised questions about how it arrives at its predictions. Our objective in this paper is to bridge the gap between the incredible predictive power of BERT and the need for transparency in AI decision-making. We delve into the techniques, methodologies, and tools that have been developed at the intersection of XAI and BERT, aiming to make BERT predictions more understandable and accountable. For this purpose, we propose a hybrid approach that combines counterfactual explanations and domain knowledge visualization.

We start with a thorough literature review of XAI, focusing on BERT, to identify existing methods and areas needing further development. We then outline our method, and discuss our contribution after presenting our experiments results. Finally, we suggest potential directions for future research.

2 RELATED WORK

In the early days of AI, rule-based systems and simpler machine learning algorithms often provided inherently interpretable outputs. These systems were favored in applications where human understanding of the decision-making process was crucial, such as expert systems in medicine or finance. However, as AI evolved, particularly with the advent of deep learning, models grew in complexity, leading to a trade-off between predictive performance and

^a  <https://orcid.org/0000-0002-6918-0797>

^b  <https://orcid.org/0009-0007-2055-873X>

interpretability. In the literature, interpretability and explainability are used interchangeably. There is no fixed definition for each one. We consider **interpretability**, the ability of a model to be interpretable by a human without any need for a specific technique to explain the model in advantage. It means the model is interpretable by its intrinsic architecture such as decision trees. **Explainability** in the other hand refers to the ability to explain in human terms the internal mechanics of the machine or deep learning system considered as 'black box' and finding the causal or correlation relationship between its input and output (Minh et al., 2022). Explainable models are interpretable by default and the opposite is not always true (Angelov et al., 2021).

In addition, XAI methods can be distinguished based on three main characteristics (Van der Velden et al., 2022):

Intrinsic and Post-Hoc: This means either explainability comes directly from the model through its structure and design (intrinsic) or the method is applied to the model after the prediction is made (post-hoc). Black-box models need post-hoc explanation methods.

Model-Specific and Model-Agnostic: This means the explainability method is applied to a specific family of machine learning models or can be applied to a wide range classes of models regardless of their type or complexity. Model-specific methods are often able to exploit the inherent structure and properties of the model and it requires an understanding of the model.

Global and Local: The method explains the behavior of the entire model and the data that it represents or it focus on explaining the behavior of a machine learning model for a specific instance or prediction.

After reviewing existing research, we will next present our point of view on XAI methods for BERT, as this is the goal of our paper.

2.1 XAI for BERT

BERT is a language representation model used in a wide variety of NLP (Natural Language Processing) tasks (Devlin et al., 2018). In BERT, each representation of the token input is formed by summing token, segment, and position embeddings. Special tokens like (CLS) stands for "classification" and it plays a central role in encoding and representing the semantics of the input text. While (SEP) stands for "separator". It is used to separate sentences or segments of text when multiple sentences or segments are included in the input. These embeddings cap-

ture context. BERT employs self-attention, which indicate word importance in context. These representations pass then through feed-forward layers. The complexity of BERT, with over 100 million parameters, makes it an opaque black-box model, requiring the use of eXplainable AI (XAI) methods to provide transparency and establish trustworthiness. We present next the research conducted to deal with XAI applied to BERT.

2.1.1 Feature Relevance Methods

Feature relevance methods cover gradient-based methods (Niranjan et al., 2023) and perturbation-based methods (Ivanovs et al., 2021), (Borys et al., 2023) notably LIME (Local Interpretable Model-Agnostic Explanations) (Dieber and Kirrane, 2020) and SHAP (SHapley Additive exPlanations) (Salih et al., 2023) techniques.

Gradient-Based Methods

The article (Ali et al., 2022) points out that existing interpretability methods based on gradient information, such as Layer-Wise Propagation (LRP), are unreliable when it comes to identify the contribution of input features to predictions in Transformers and preserving the critical conservation property for Transformer models. It means the sum of relevance scores of inputs should be equal to the output. This unreliability is attributed to components like attention heads and LayerNorm in Transformers.

To address this issue, the article proposes a modification of the LRP method. The modification incorporates specific rules within the LRP framework to handle attention heads and LayerNorms correctly in Transformer models. As a result, the proposed method offers more precise and informative explanations, highlighting relevant input features while ensuring the conservation of information as relevance scores propagate through the network.

Perturbation Based Methods

One of the perturbation based methods is SHAP. To be used to explain BERT prediction, some changes are needed since it has problems with subword input because credits for an output cannot be assigned to units such as words. There is a research (Kokalj et al., 2021) aiming to address the problem of incompatibility of SHAP, and the pretrained transformer BERT for text classification. It proposes an approach called TransSHAP (Transformer SHAP).

The TransSHAP's classifier function first converts each input instance into a word level representation. SHAP perturbs the representation and generates new locally similar representations that we will use for explanation. BERT tokenizer uses these instances and converts the sentence fragments to sub-word tokens.

Finally, the predictions for the new locally generated instances are produced and returned to the Kernel SHAP explainer. The limit of this approach is that it supports random sampling from the word space, which may produce grammatically wrong sentences, and uninformative texts.

2.1.2 Explanation by Example

A BERT Case-Based Reasoning (CBR) System

A CBR model system is based on the idea that an ANN black box model can be abstracted into a more interpretable white-box. It retrieves cases with similar and equally important words in both predictions. It looks for the similar training instance of the query that contains words similar to the words existing in the query currently working on. The output helps understand the factors that influenced BERT positive classification and provides insights into the decision-making process of the twin-system (Kenny and Keane, 2021).

Counterfactual Explanation

A new method called Token Importance Guided Text Counterfactuals (TIGTEC) has been developed to generate counterfactual explanations for textual data, specifically for BERT predictions. TIGTEC identifies important tokens that significantly influence the classifier prediction and creates counterfactual examples by replacing these tokens. These examples are evaluated based on a cost function balancing the probability score and semantic distance. A candidate is accepted if the prediction of the classifier changes and makes margin high. The next iteration starts from the nodes with the lowest cost value. The solution space uses a beam search where candidate sequences are generated and sorted based on their probabilities, and only the top-k candidates with the highest probabilities are kept. Then, it iterates until it arrives at a stopping condition. While TIGTEC achieves high success, it may have limitations in human comprehension due to automatic evaluation (Bhan et al., 2023).

2.1.3 Based Rules Methods

XAI Linguistic Rules

A new XAI approach utilizes linguistic rules based on NLP building blocks to globally reconstruct predictions made by BERT for token-level classification (Binder et al., 2022). These rules are structured to capture the underlying logic used by the language model in assigning class labels and understanding the patterns and relationships between tokens and their class labels. This method maintains both fidelity and comprehensibility in global reconstructions.

XAI Global Decision Tree

In another research paper (Binder, 2021), a global decision tree is constructed to explain BERT star rating predictions. This approach seeks to strike a balance between fidelity (matching BERT predictions) and interpretability (human understanding). The resulting decision tree can be analyzed to extract interpretable rules that offer actionable business insights. These rules provide information about how specific features relate to BERT star rating predictions, and this approach is model-agnostic, making it adaptable to other deep learning models.

2.1.4 Hybrid Approches

LIME and Anchors

A research (Szczepański et al., 2021) that suggests a local surrogate type approach that does not require extensive changes to the system and can be attached as an extension instead of redesigning the model to make it more transparent. The approach is based on two XAI techniques, LIME which represents the model behaviour in the neighbourhood of the predicted sample and Anchors which is based on 'if-then' rules. Their key advantage is that it can be rapidly deployed within the frameworks of already existing solutions based on BERT without the need of making changes in the architecture.

Integrated Gradient, SHAP and LIME

A research (Rietberg et al., 2023) has been conducted to employ XAI methods, SHAP, LIME, and Integrated Gradient (IG), to identify the words that are important for the classifier's decision. SHAP and LIME succeeded in identifying relevant features compared to IG that demonstrated shortcomings in the same task. Also, a domain knowledge-based test was performed to assess the alignment of model explanations with domain knowledge and it resulted in consistency of LIME and SHAP explanations with domain knowledge.

2.2 Limits

We show, in Table 1 Comparative study., the limitations of the most commonly used methods, as well as the aspects that should be taken into account in our solution. In addition, many researches encourage working on counterfactual methods because of their ability to guide users for attending specific outcome. The provided explainable methods applied to BERT model explain the model without suggesting alternative scenarios. We choose to develop a counterfactual explanation method. Counterfactuals offer a user-friendly and easily understandable method. They establish causal links between inputs and outputs and enable users to investigate hypothetical sce-

Table 1: Comparative study.

Method	Limits	Our objectives
Gradient based	Depend on a particular architecture and can be sensitive to specific parameters.	Can be applied to other transformers models.
Perturbation based	Only provide the reason behind specific result	Provide causal relationship between input and output and explores hypothetical scenarios on how to achieve specific result.
Rule-based	May oversimplify or omit certain aspects and also can struggle to capture context.	Maintain fidelity of the original context
Local explanation	Can not be used as a general interpretation for the whole model.	Provide a global understanding of the behaviour of the model by using domain knowledge visualisation

narios with varying input changes to arrive to different outcomes. However, a potential limitation of this method (as TIGTEC) is its automatic evaluation metrics for counterfactual examples. The algorithm proved its performance but did not guarantee human understanding. We believe that working on human-grounded experiments would be more appropriate to assess the relevance of the generated text and its explanatory quality. This will be our purpose when developing our method.

3 OVERVIEW OF THE PROPOSED APPROACH

Counterfactual explanation methods have their limitations. In fact, each counterfactual is tailored to a specific data point, making it a local method and it is essential to provide accompanying context whenever possible to communicate the boundaries of its generalizability. Also the method is susceptible to what is known as the Rashomon effect. This manifests when it has different counterfactuals, each counterfactual tells a different story of how a certain outcome was reached. One counterfactual might say to change feature A, the other counterfactual might say to leave A and change feature B. Although this seems contradictory, it is important to acknowledge that all of these suggestions can play a part in reaching the desired decision. Which raises the question of which counterfactual can we consider.

Domain knowledge can enhance interpretability and evaluation of the explanation method when it is integrated in XAI methods. In addition, visualizing knowledge can be valuable in addressing the

”Rashomon” effect in XAI. While it may not fully eliminate it, foundational knowledge assists in evaluating and choosing suitable counterfactuals. Since counterfactual explanations are tied to specific data points, domain knowledge aids in their generalization. Contextual information enhances user understanding of the limitations, preventing misinterpretation in inappropriate situations.

4 PROPOSED APPROACH PIPELINE

The pipeline of our work is illustrated in Figure 1 Explanation method pipeline.. To realize this pipeline, we first fine-tuned BERT, a state of the art language model for sentiment analysis task (Yalçın, 2020). We used IMDB movie review data set (Maas et al., 2011), a publicly available data set. Then, we used our approach to explain the model behavior for the prediction.

4.1 Data Preparation

The data preparation process begins with the data set importation. We used the IMDB movie reviews data set, a large number of text expressions reflect different positive and negative feelings. Then we preprocessed this data set, unlabeled data is excluded as it is unnecessary for the fine-tuning phase. Subsequently, the data is split into two segments: 20000 files are used for training, while 5000 files are designated for validation. Following this, the text-based data is transformed into a suitable format that the BERT model can work with. Tokenization is applied to convert textual data into numerical representations, resulting

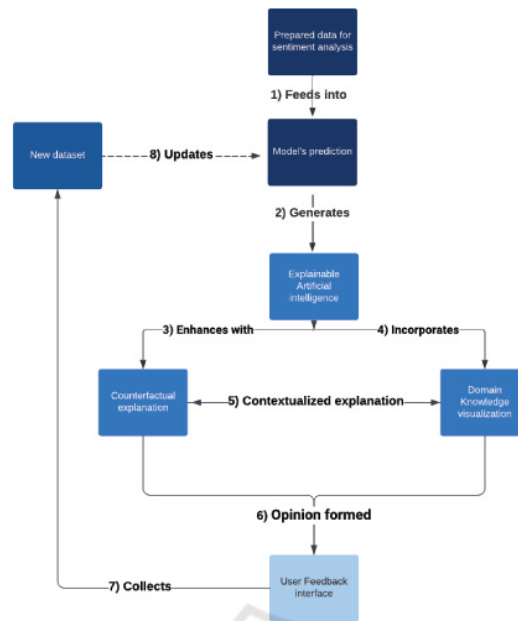


Figure 1: Explanation method pipeline.

in input datasets designed to feed the model effectively. Finally, we proceed to configure and fine-tune our model specifically for the sentiment analysis task. We compile the model with Adam optimizer, sparse categorical cross-entropy loss, and accuracy metrics. then train the model for 2 epochs using training and validation data. During the two training epochs, the model processed 1250 batches of data per epoch. In the first epoch, each batch took about 1217s (919ms/step) to process. The training loss was 0.2652 with an accuracy of 0.8895. Validation data had a loss of 0.3112 and an accuracy of 0.8846. In the second epoch, batch processing time reduced to around 1135s (908ms/step). The training loss significantly decreased to 0.0767 with an improved accuracy of 0.9733. However, the validation loss increased to 0.4833, while validation accuracy remained relatively high at 0.8822.

4.2 Prediction

In our work, we implement a user-friendly interface in which the user inputs text and the model predicts the sentiment. The process begins by tokenizing the input sentences using the BERT tokenizer, converting the text into a numerical format. The tokenized input is then passed through the pre-trained BERT model, which rigorously processes it, taking into account contextual information and relationships between tokens. Next, the model generates prediction scores for various sentiment classes. These scores are further

transformed into probabilities, allowing the model to assign probabilities to each sentiment class. Finally, the model selects the sentiment label with the highest probability as the predicted sentiment, and this label is presented as the output within the user interface as shown in Figure 2 Prediction interface..

4.3 Explainable AI Method

Now, our goal is to understand our model predictions using the counterfactual explanation. To enhance understanding and evaluation, we use domain knowledge visualization. Additionally, we present a user feedback interface to collect feedback for future model improvements.

4.3.1 Counterfactual Explanation

Our counterfactual explanation method follows the depicted steps in Figure 3 Steps of counterfactual explanation method.. First, we use BERT to identify the most important tokens, while excluding (CLS) and (SEP) special tokens. We choose to find the top 3. For example, in the input text important tokens are identified as ('fears', 'reducing', 'disconnected') and their indices are [16, 14, 21] as shown in Figure 4 Important tokens identification..

These tokens are replaced with the (MASK) token, and the BERT model is used to predict the most probable tokens that would fill their positions. BERT selects tokens that can integrate within the text, align with the surrounding context and produce meaning-

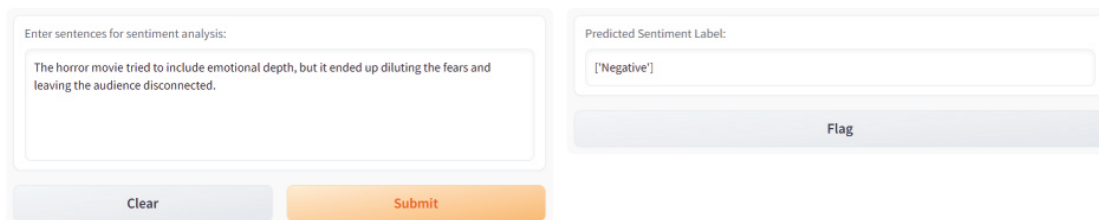


Figure 2: Prediction interface.

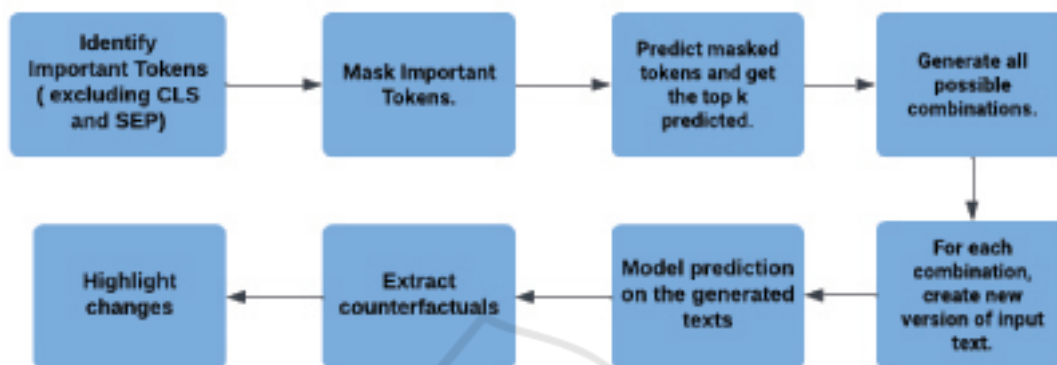


Figure 3: Steps of counterfactual explanation method.

This is not expected if you are emphasizing on emotional depth from the snippet.
 Important Tokens: ['f e a r s', 'r e d u c i n g', 'd i s c o n n e c t e d']
 Important Token Indices: [16, 14, 21]

Figure 4: Important tokens identification.

ful and coherent counterfactual variations. The top-8 predicted replacement tokens for each important token are stored in a list.

All possible combinations of replacement tokens are generated. For each combination, a counterfactual text is created by replacing the original important words with words from the combination. The counterfactual texts are stored in a list and then filtered based on the prediction made on each one. If the original text input was negative then we only select counterfactual texts that have a positive prediction and vice versa. It means only those counterfactual texts that lead to a desired prediction are represented. And finally we highlight changes comparing to the original input text. This process helps users understand why a certain prediction was made and how changes to the input text can influence the model’s output. This could be valuable for explaining and interpreting model decisions, especially in sensitive domains that demand justifications for model outputs like finance or business or healthcare. The result of the counterfactuals generation is illustrated in Figure 5 Counterfactual explanation..

4.3.2 The Effect of Domain Knowledge

This section performed a series of tasks centered around NLP applied to specific data within the movie-review industry. These tasks include entity extraction, relation extraction, and the visualization of a knowledge graph. An expert will define the data knowledge, ensuring the inclusion of comprehensive movie industry knowledge to facilitate a profound understanding of the field. The domain knowledge will be visualized with counterfactual explanation. Stakeholders will then utilize it to identify suitable counterfactual candidates. The figure 6 Domain knowledge graph. shows the visualized graph.

We should evaluate the generated counterfactuals based on our knowledge graph, which illustrates connections from "good movie" to "horror scenes," further linked to "emotional depth", linked to "more fear," and "more emotions". Given this graph, we should prioritize counterfactuals that align with the idea of the horror movie attempting to include emotional depth, which results in an increase in fear. Specifically, we should favor counterfactuals like: "The horror movie tried to include emotional depth, but it ended up increasing the fears and leaving the audience grounded."

- ☐ the horror movie tried to include emotional depth, but it ended up **less** the fears and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **greater** the fears and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **greater** the fears and leaving the audience **emotional**.
- ☐ the horror movie tried to include emotional depth, but it ended up **increasing** the fears and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **less** the **concerns** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **greater** the **concerns** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **greater** the **emotions** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **greater** the **emotions** and leaving the audience **connected**.
- ☐ the horror movie tried to include emotional depth, but it ended up **more** the **emotions** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **increasing** the **emotions** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **increasing** the **emotions** and leaving the audience **connected**.
- ☐ the horror movie tried to include emotional depth, but it ended up **increasing** the **emotions** and leaving the audience **different**.
- ☐ the horror movie tried to include emotional depth, but it ended up **increasing** the **terror** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **greater** the **feelings** and leaving the audience **grounded**.
- ☐ the horror movie tried to include emotional depth, but it ended up **increasing** the **feelings** and leaving the audience **grounded**.

Figure 5: Counterfactual explanation.

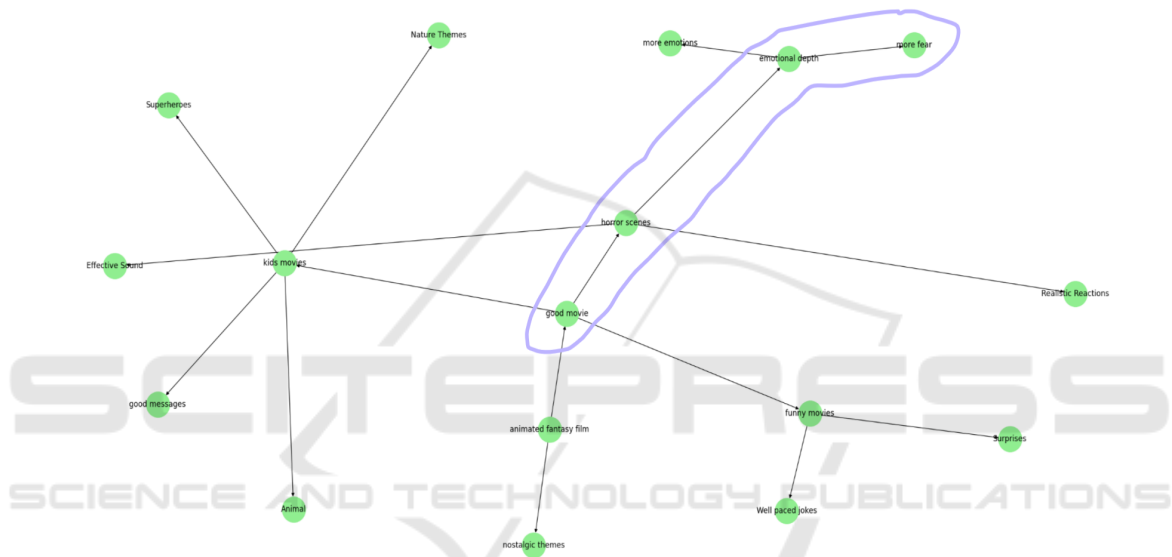


Figure 6: Domain knowledge graph.

”The horror movie tried to include emotional depth, but it ended up greater the fears and leaving the audience emotional.”

”The horror movie tried to include emotional depth, but it ended up increasing the terror and leaving the audience grounded.”

These counterfactuals resonate with the connections in our knowledge graph, suggesting that changes leading to increased fear align with the established relationships between narrative elements.

4.3.3 Feedback Interface

This section focuses on enhancing user interaction with the model after presenting the explanation. It involves gathering user feedback regarding their level of conviction with the prediction. A separate interface is introduced for this purpose, offering two checkbox options: ”agree with the prediction” and

”disagree with the prediction” as illustrated in Figure 7 Feedback interface.. These check-boxes are used to generate user feedback, which can be categorized as ”agree,” ”disagree,” or ”neutral” based on the selected checkbox values. Then we collect the data with their corresponding input text and predicted label collected at the beginning and store them for future enhancement of the model.

5 SYNTHESIS

The code for the components of the previously described pipeline is available. ¹

¹<https://www.kaggle.com/code/arouahedhili/hybridapproachtexplainbert>



Figure 7: Feedback interface.

5.1 Implication

Our implemented XAI method holds the potential to significantly assist various stakeholders within the film industry. It provides a level of confidence regarding the precision of sentiment analysis predictions generated by BERT and offers insights into achieving desired predictions. While our approach was initially applied to understanding the decisions made by the BERT model, it can be adapted to other transformer models as well. In practice, our approach can aid businesses in their decision-making processes, offer explanations for loan approval or rejection in the banking domain, and provide specific explanations for medical diagnoses, among other applications in various domains.

5.2 Comparison with Existing Methods

We previously discussed the TIGTEC algorithm (Bhan et al., 2023), which, while effective in automatic evaluation, lacked human understanding. To address this limitation, we add a simpler, human-centered interaction for evaluation. We believe that domain knowledge can overcome this limitation in evaluating generated counterfactuals. Our algorithm shares the same objective as TIGTEC but differs significantly in approach and complexity. It offers advantages in terms of simplicity and directness. It utilizes BERT predictive capabilities to replace important tokens with appropriate alternatives, creating counterfactual explanations that are likely to be both syntactically and semantically correct. This straightforward approach does not use complex iterations, custom cost functions, or intricate search policies. It is easier to implement.

Another significant paper (Dikmen and Burns, 2022) proposes a comprehensive approach combining SHAP visualizations and domain knowledge to explain AI models. While SHAP is valuable, it has limitations in aligning with human understanding. This research advocates using causality-based methods like counterfactual explanations alongside domain knowledge to address these limitations. Our research focuses on integrating the counterfactual explanation method with domain knowledge.

Finally, our proposed approach in Explainable Artificial Intelligence (XAI) offers numerous advantages over existing methods:

- **Causality Over Feature Relevance:** Unlike feature relevance methods that focus on relevance scores, our approach delves into the causal relationship between input and output, providing insights into how input changes affect model predictions.
- **Fidelity Preservation:** In contrast to methods that may lose details, our approach maintains the fidelity of the original model, ensuring a comprehensive understanding.
- **Generalizability with Local XAI:** Thanks to domain knowledge visualization, our approach can be applied even with local XAI methods, offering context and enriching explanations.
- **Comprehensive Perspective:** Our approach surpasses example-based explanations by generating counterfactual scenarios with integrated domain knowledge. This enhances context, user validation, and offers a more thorough, causal, and user-centered view of AI decisions compared to example-based methods.

6 CONCLUSIONS

In our work, we explored the counterfactual explanation method, aligning with recommendations from multiple articles. Our research focused on the significance of domain knowledge in evaluating this method and achieving a human-centered evaluation. Additionally, we showed how our approach addressed challenges encountered in other XAI methods. We successfully implemented this method, demonstrating its effectiveness. Our future research directions include: Exploring alternative feature relevance methods to enhance token identification for better counterfactual explanations in the context of BERT. Developing a tool to assess information trustworthiness and enhance model reliability.

REFERENCES

- Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. (2022). Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pages 435–451. PMLR.
- Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., and Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.
- Bhan, M., Vittaut, J.-N., Chesneau, N., and Lesot, M.-J. (2023). Tigtec: Token importance guided text counterfactuals. *arXiv preprint arXiv:2304.12425*.
- Binder, M. (2021). But how does it work? explaining bert's star rating predictions of online customer reviews. In *PACIS*, page 28.
- Binder, M., Heinrich, B., Hopf, M., and Schiller, A. (2022). Global reconstruction of language models with linguistic rules—explainable ai for online consumer reviews. *Electronic Markets*, 32(4):2123–2138.
- Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., and Nensa, F. (2023). Explainable ai in medical imaging: An overview for clinical practitioners—saliency-based xai approaches. *European journal of radiology*, page 110787.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Dieber, J. and Kirrane, S. (2020). Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*.
- Dikmen, M. and Burns, C. (2022). The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792.
- Ivanovs, M., Kadikis, R., and Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234.
- Kenny, E. M. and Keane, M. T. (2021). Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in xai. *Knowledge-Based Systems*, 233:107530.
- Kokalj, E., Škrlić, B., Lavrač, N., Pollak, S., and Robnik-Šikonja, M. (2021). Bert meets shapley: Extending shap explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21.
- Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66.
- Niranjana, K., Kumar, S. S., Vedanth, S., and Chitrakala, S. (2023). An explainable ai driven decision support system for covid-19 diagnosis using fused classification and segmentation. *Procedia computer science*, 218:1915–1925.
- Rietberg, M. T., Nguyen, V. B., Geerdink, J., Vijlbrief, O., and Seifert, C. (2023). Accurate and reliable classification of unstructured reports on their diagnostic goal using bert models. *Diagnostics*, 13(7):1251.
- Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., and Lekadir, K. (2023). Commentary on explainable artificial intelligence methods: Shap and lime. *arXiv preprint arXiv:2305.02012*.
- Szczepański, M., Pawlicki, M., Kozik, R., and Choraś, M. (2021). New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.
- Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. (2022). Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470.
- Yalçın, O. G. (2020). Sentiment analysis in 10 minutes with bert and tensorflow. *Towards Data Science*.