Explaining Transformers Through Similarity Difference and Uniqueness Masks – A Pilot Study

1st Marco Parola[§]
University of Pisa, *Pisa*, Italy
marco.parola@ing.unipi.it
0000-0003-4871-4902

2nd Mohammad Naser Sabet Jahromi[§]
Aalborg University, Aalborg, Denmark
mosa@create.aau.dk
0000-0002-6332-7567

3rd Giovanni Bergami[§]
University of Pisa, *Pisa*, Italy
g.bergami@studenti.unipi.it

4th Mario G. C. A. Cimino
University of Pisa, *Pisa*, Italy
mario.cimino@unipi.it
0000-0002-1031-1959

5th Thomas B. Moeslund
Aalborg University, Aalborg, Denmark
tbm@create.aau.dk
0000-0001-7584-5209

***Abstract*—The widespread deployment of transformers in text classification creates the need for interpretable AI systems, particularly in regulatory-sensitive domains where transparency is mandatory. This pilot study presents the first adaptation of the Similarity Difference and Uniqueness (SIDU) method, originally developed for CNN explainability, to transformer architectures. We address the challenge of bridging spatial feature maps to sequential token representations by exploring two masking strategies: Persistent Homology masking, which utilizes angular distances between the [CLS] token and context tokens, and Cosine Similarity masking, based on semantic relationships. Our approach operates on final hidden layer representations, requiring multiple forward passes to evaluate different mask configurations and compute similarity-uniqueness scores for token-level explanations. Through quantitative and qualitative evaluation across diverse text classification scenarios, from movie reviews to legal document processing, we investigate how transformer hidden states can be leveraged for explainability. To support this evaluation, we introduce a novel metric called Average Token Activation, which captures the mean activation of individual tokens without relying on any threshold mechanisms typical of XAI plausibility evaluation metrics. Our findings reveal robust performance across different domains and classification setups, providing the first insights into the potential and limitations of this cross-domain XAI adaptation approach.**

***Index Terms*—XAI, Transformer Interpretability, NLP Classification, Token-Level Importance, SIDU**

I. Introduction

The digital transformation of business processes has created an unprecedented reliance on automated text classification systems across critical sectors [1]. Financial institutions process millions of regulatory documents daily, legal firms depend on AI systems to categorize case files, healthcare organizations analyze patient feedback at scale, and e-commerce platforms make real-time content moderation decisions. The widespread deployment of transformer-based models, such as BERT [2] and RoBERTa [3], has achieved remarkable accuracy improvements; however, it has simultaneously created a transparency crisis that threatens the sustainable adoption of AI in high-stakes environments [4]. The regulatory landscape has fundamentally shifted the requirements for AI deployment. The European Union's AI Act [5] mandates explainability for high-risk AI systems, while GDPR's ”right to explanation” extends to automated decision-making processes [6]. Similar regulatory frameworks emerging globally are transforming explainability from a desirable feature into a legal necessity. Organizations can no longer justify deploying ”black box” systems solely based on performance metrics; they must demonstrate that their AI systems make decisions for the right reasons.

To address such needs, Explainable AI (XAI) has come into play as a dedicated field, with methods initially developed in computer vision that have since been extended to text and NLP applications. At the core of many XAI methods are saliency maps (also known as heatmaps in literature), which provide visual representations of feature importance by assigning relevance scores to input elements. In text classification, these maps highlight which words or tokens most influence the model's predictions, offering an intuitive way to understand model behavior. Various techniques have been developed to generate these saliency maps, each with distinct advantages and limitations. Gradient-based approaches like Grad-CAM [7], for instance, provide effective spatial explanations for CNNs and have been adapted for text classification, though they suffer from gradient instability issues that make them problematic for production environments. Model-agnostic methods such as LIME [8] and SHAP [9] offer broader applicability but are computationally complex and often highlight contextually irrelevant tokens in a given text.

Of particular interest is the Similarity Difference and Uniqueness (SIDU) method [10], which operates on the principle that important regions should be both similar to the overall prediction pattern and unique compared to other regions. SIDU has demonstrated success in computer vision and has been extended to text classification using 1D-CNNs, called SIDU-TXT [11]. However, to go beyond the specific architecture limitations of CNNs and leverage the full potential of modern

[§] Equal contribution

transformer models, we investigate adapting SIDU to transformer architectures. This presents fundamental challenges: while CNNs process information through hierarchical spatial feature maps where each channel corresponds to distinct input regions, transformers use attention mechanisms to create dynamic relationships between tokens, with the [CLS] token aggregating information through complex attention patterns that vary across layers and heads.

While CNNs naturally provide spatial feature maps that can be directly associated with input pixels and masked to create explanations, transformers lack this explicit spatial structure. However, we observe that in the final hidden layer, the [CLS] token, which aggregates global context for classification, exhibits meaningful relationships with individual context tokens. These relationships, captured through geometric and semantic similarity measures in the final hidden space, can serve as a proxy for the spatial feature relationships that SIDU exploits in CNNs. By identifying which tokens have strong similarity patterns with the [CLS] token in this final representation and exhibit unique characteristics compared to other tokens, we can construct meaningful masks that preserve the contextual dependencies inherent in transformer representations. Drawing inspiration from SIDU as a simple and effective XAI method that uses feature maps to generate masks in CNNs, we develop novel masking strategies that exploit the [CLS] token's learned relationships with context tokens in the final hidden space. These relationships, captured through geometric and semantic similarity measures, serve as a proxy for the spatial feature relationships that SIDU exploits in CNNs, enabling us to construct meaningful masks that preserve the contextual dependencies inherent in transformer representations.

Our **main contribution** is presenting the first adaptation of SIDU to transformer architectures by developing two novel masking strategies that leverage [CLS] token relationships: Persistent-Homology masking using angular distance matrices to identify connected token components, and Cosine Similarity masking based on semantic similarity thresholds. Through comprehensive evaluation across movie review sentiment analysis and legal document classification (including challenging long-text scenarios with paragraph selection strategies), we demonstrate varying performance characteristics that provide practical insights for deploying transformer explainability methods in production environments. Additionally, we also introduce a novel metric called Average Token Activation (ATA) to support the evaluation of our explainability experiments. Specifically, ATA quantifies the mean activation of individual tokens without relying on threshold-based criteria differently from existing plausibility measures [12], [13].

II. Preliminaries and Related work

In this section, we introduce basic concepts in natural language processing useful for a better understanding of the methods and experiments presented in the remainder of the paper. We also provide an overview of recent developments in XAI techniques tailored to the text domain.

A. *Text classification*

Text classification is a fundamental supervised learning task where models assign input text documents to one or more predefined categories. The task involves learning a mapping function from textual input space to a discrete label space, enabling automated categorization of documents across diverse domains such as sentiment analysis, topic classification, and legal document processing. Early approaches relied on handcrafted features like bag-of-words and TF-IDF representations, treating text as unordered collections of terms. Modern deep learning has revolutionized this field through neural architectures that capture word order and contextual meaning. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) provided initial improvements, but transformer-based models have achieved state-of-the-art performance [2]. BERT [2] and its variants, like RoBERTa [3], exemplify this advancement, learning rich contextual representations through self-attention mechanisms. These models are pre-trained on large corpora using masked language modeling and next sentence prediction objectives, then fine-tuned for downstream classification tasks. The bidirectional nature of BERT allows it to consider both left and right context when encoding each token.

However, the quadratic computational complexity of self-attention constrains these models to approximately 512 tokens, creating significant challenges for real-world applications involving lengthy documents. Legal cases, research articles, and regulatory texts—domains where explainability is often mandatory—routinely exceed several thousand tokens. To address this limitation, researchers have proposed three main approaches: **(i) Long-sequence transformers** like Longformer [14] and BigBird [15] employ sparse attention patterns to process thousands of tokens, though at increased computational cost; **(ii) Hierarchical methods** [16]–[18] encode document segments independently before aggregating representations; and (iii) **Text selection strategies** [19], [20], including both simple truncation and more sophisticated paragraph selection methods, extract relevant portions before classification.

Our work adopts paragraph selection following Tuteja et al. [21], who demonstrated that strategically selecting paragraphs—such as concatenating first and last paragraphs—achieves near full-document performance while maintaining efficiency. Unlike simple truncation that blindly cuts at token boundaries, paragraph selection preserves complete semantic units and leverages document structure. However, this preprocessing introduces additional complexity for interpretability, as the decision-making process becomes distributed across both selection and classification stages.

B. *XAI methods for NLP*

The remarkable performance of transformers comes with reduced interpretability, creating urgent demand for explainable AI methods in high-stakes applications [22]. Existing XAI approaches for NLP can be categorized into three main families, each exhibiting distinct limitations that become particularly problematic in long-text scenarios: **(i) Gradient-**

based methods like Grad-CAM variants and Integrated Gradients [23] compute feature importance through backpropagation. While computationally efficient, they suffer from gradient instability and produce noisy attributions with many false positives, especially in transformer architectures with multiple layers and attention heads. **(ii) Perturbation-based approaches** such as LIME [8] and SHAP [9] treat models as black boxes, systematically masking inputs to measure prediction changes. Though model-agnostic, these methods have high computational complexity and often violate the contextual dependencies crucial to transformer effectiveness, resulting in fragmented explanations where semantically related tokens are incorrectly separated. **(iii) Attention-based explanations** initially appeared promising, but extensive research has demonstrated that attention weights do not reliably indicate feature importance [24]. Moreover, the multi-head, multi-layer architecture of transformers makes it unclear which attention patterns should be interpreted as explanations. A fundamental challenge across all these methods is granularity: token-level attributions produce scattered explanations that fail to preserve semantic context, a problem that compounds when paragraph selection strategies may exclude crucial context from the explanation process entirely.

The computer vision community has addressed similar challenges through the Similarity Difference and Uniqueness (SIDU) method [10], which identifies spatially coherent regions that are both similar to the overall prediction pattern and unique compared to other regions. SIDU-TXT [11] successfully adapted these principles to 1D-CNNs for text classification by generating meaningful masks rather than isolated token scores. However, transformers lack the explicit spatial feature maps that SIDU exploits in CNNs, instead creating dynamic token relationships through attention mechanisms. This pilot study investigates whether novel masking strategies based on [CLS] token relationships can produce finer-grain token-level explanations that minimize irrelevant highlighting while preserving contextual dependencies, particularly in challenging long-text classification scenarios.

III. Methodology

This section presents our explainability method adapted to Transformer-based text classifiers, producing token-level saliency maps for model predictions. The explainability pipeline consists of the following stages:

- **Mask crafting** deals with generating a set of masks that will be used to alter the input at the token level.
- **Input masking** involves applying the generated masks to the input and evaluating the model's prediction, allowing us to measure the contribution of each token.
- **SIDU score computation and Heatmap generation** calculates the Similarity Difference and Uniqueness scores for each mask and combines the results from the previous stages to produce a token-level saliency map. This map assigns an importance score to each token, indicating its relevance to the model's prediction.

For the reason described in Section I, among these three stages, the first two (Mask crafting and Input masking) require a deeper analysis to be adapted, as they are tightly coupled with the underlying Transformer architecture. In contrast, the third stage (SIDU score computation and Heatmap generation) can be more directly derived from the original SIDU formulation. Section III-A and Section III-B delve into the details of the first two stages, while Section III-C will focus on the SIDU score computation and heatmap generation, illustrating how it builds upon the previous stages. Finally, the evaluation metrics adopted to assess the effectiveness of the proposed methods are described in Section III-D.

A. Mask crafting

As it will be better described in Section III-C, SIDU-transformer relies on masks to produce heatmap explanations as the original SIDU implementation for computer vision. This stage in the pipeline is referred to as Mask Crafting and is responsible for generating a set of masks M at the token level, which are later applied to the input to assess the impact of individual tokens on the model's output. As highlighted in Section I, classification using Transformers is performed on a specific token called `[CLS]`, which is appended to the input sequence. However, this presents a challenge for the explanation: the model's decision is encoded in the `[CLS]` token, while the goal of our explanation method is to attribute importance to the individual context tokens. Therefore, an important step in the SIDU adaptation is to establish a principled mechanism for mapping the influence of each input token onto the final `[CLS]` token. This mapping is the focus of the masking process strategies we propose in this section. To address this, we explore two distinct strategies: Persistent Homology Mask Crafting (PHMC) and Cosine Similarity Mask Crafting (CSMC), both designed to identify structurally or semantically similar tokens; then, we exploit these geometric properties to create masks.

Persistent Homology Mask Crafting. The PHMC method leverages topological data analysis to generate masks aiming to capture the semantic structure encoded within Transformer tokens. The intuition is to measure and exploit the semantic alignment between each context token and the `[CLS]` token, which aggregates information for the final prediction.

We begin by computing the cosine similarity between the embedding of each token h_i and the `[CLS]` embedding h_{cls}, as shown in Equation 1. This similarity is then transformed into an angular distance θ_i, as shown in Equation 2.

$$\text{sim}(h_i, h_{\text{cls}}) = \frac{h_i \cdot h_{\text{cls}}}{|h_i||h_{\text{cls}}|} \tag{1}$$

$$\theta_i = \arccos(\text{sim}(h_i, h_{\text{cls}})) \tag{2}$$

These angular distances serve as a representation for semantic dissimilarity: smaller angles indicate stronger semantic relevance to the prediction. To capture the pairwise relationships between all token embeddings in a way that emphasizes these angular differences, we define a similarity matrix D using an exponential decay over angular differences, as shown in Equation 3.

$$D[i,j] = \exp(-|\theta_i - \theta_j|) \tag{3}$$

This matrix defines a fully connected weighted graph over the tokens, where the weights reflect semantic proximity in the hidden space. To extract meaningful structural information from this graph, we apply a topological thresholding process. Specifically, we construct a sequence of simplicial complexes Complex_ϵ by thresholding the edge weights in D. At each threshold ϵ, an edge between tokens i and j is included in the complex if $D[i,j] \geq \epsilon$, grouping connected tokens into clusters. Finally, we compute the 0-dimensional homology group H_0 at each step, identifying connected components of semantically similar tokens. Tokens that belong to singleton components, those that remain isolated at a given threshold, are considered less semantically relevant and are excluded from the mask. Conversely, tokens that consistently form stable groups across multiple thresholds are deemed semantically significant and are included in the mask.

Cosine Similarity Mask Crafting The CSMC method is designed to exploit the alignment between each context token and the `[CLS]` token to construct a sequence of masks. The intuition behind this method is that the `[CLS]` token acts as an aggregate representation of the entire input sequence, and its embedding is influenced more strongly by the most semantically relevant tokens. By quantifying this semantic alignment through cosine similarity, we aim to infer which tokens contribute more to the final prediction.

Specifically, we start by computing the cosine similarity between each contextualized token embedding h_i and the contextualized `[CLS]` embedding h_{cls}. This produces a similarity score for each token, reflecting its contribution to the information encoded in the `[CLS]` representation.

Based on these similarity scores, we generate a sequence of binary masks by progressively lowering a cosine similarity threshold used to determine which context tokens are included. Initially, only the token with the highest similarity to the `[CLS]` embedding is retained. As the threshold decreases, additional tokens whose similarity to `[CLS]` exceeds the current threshold are included in the mask. This process continues until the threshold reaches its minimum value, at which point all context tokens are included.

B. Input masking

To adapt SIDU to Transformer-based architectures, we explore two strategies for input masking. The first approach, referred to as Token Zeroing, operates directly on the input embeddings. Specifically, for each mask, we replace the embeddings of the selected tokens to be masked with zero vectors of the same dimensionality, nullifying their semantic contribution to the model.

The second strategy, which we denote as Attention Suppression, intervenes at the level of the model's attention mechanism. Instead of altering the token embeddings, this method modifies the self-attention weights during the forward pass. For each mask, we set the attention weights corresponding to the masked tokens to zero across all heads and layers.

C. SIDU score computation and Heatmap generation

Following the mask generation strategies described in Section III-A and the input masking approach from Section III-B, we now compute the SIDU scores to generate the final token-level explanation heatmap. The SIDU method evaluates each mask based on two key metrics: **Similarity Difference (SID)** and **Uniqueness (U)**.

Let c be the target class for which we generate explanations. Since Transformer predictions are class-specific, all computations are relative to c. For each mask M_i^c and masked input A_i^c, we obtain prediction scores P_i^c and P_{org}^c from the masked and original inputs. The Similarity Difference SID_i^c measures how closely the masked prediction matches the original.

$$SID_i^c = \exp\left(-\frac{\|P_{\text{org}}^c - P_i^c\|^2}{2\sigma^2}\right), \tag{4}$$

where σ controls the sensitivity of the similarity measure. This metric assigns higher scores to masks that preserve the model's original prediction behavior. The Uniqueness score U_i^c measures how distinct each mask's prediction is compared to all other masks:

$$U_i^c = \sum_{j=1}^{N} \|P_i^c - P_j^c\|, \tag{5}$$

where N is the total number of masks. This encourages the selection of masks that capture unique aspects of the model's decision-making process. The final SIDU weight for each mask combines both metrics:

$$W_i^c = SID_i^c \cdot U_i^c. \tag{6}$$

To generate the explanation heatmap S^c, we select the top K masks with highest weights and compute a weighted average:

$$S^c = \frac{1}{K}\sum_{i=1}^{K} W_i^c \cdot M_i^c. \tag{7}$$

This formulation ensures that the final heatmap highlights tokens that are both similar to the overall prediction pattern (high SID) and unique in their contribution (high U), providing a balanced explanation that captures the most semantically important features for the model's decision.

D. Evaluation metrics

Dhaini et al. [13] proposed an XAI evaluation framework that organizes explanation quality into three dimensions: faithfulness, reflecting how well the explanation represents the model's predictions; plausibility, its alignment with human intuition; and complexity, its interpretability and conciseness. Given the growing emphasis on human-centered explainability and the need for human-understandable explanation in different domains [25]–[28], this pilot study of the SIDU algorithm adaptation places particular focus on plausibility metrics. For this purpose, we rely on human ground truth explanations. Specifically, the subsentences that annotators identify as most relevant to the model's prediction. We can use some segments to identify these relevant sub-sentences. These segments provide a reference for assessing how well the model architecture aligns with human understanding of what constitutes a valid explanation.

To quantify this alignment, we introduce a new metric called Average Token Activation. This metric measures the mean activation score of each input token, derived from explanation heatmaps, and serves as a proxy for how much each token contributes to the model's final decision. Intuitively, ATA is similar to other existing plausibility metrics such as IoU-F1 or token-F1 score [12], [13], as it aims to identify the input text parts marked as relevant by humans. However, unlike threshold-based metrics, ATA captures token importance in a continuous manner without requiring predefined threshold values [12], [13]. By computing ATA over annotated relevant segments (ATArel) and comparing it with ATA over irrelevant parts (ATAirr), we assess the concentration of model attention on the key informative regions.

Formally, given a saliency map composed of different activation values per each toke $S = [s_1, s_2, \ldots, s_n]$ for an input sequence of length n, and a binary mask $M = [m_1, m_2, \ldots, m_n]$, where $m_i = 1$ if token i is within a human-marked relevant region and 0 otherwise, we define ATA$_{rel}$ and ATA$_{irr}$ in Equation 8. Ideally, a well-aligned explanation yields ATA$_{\text{rel}} \approx 1$ and ATA$_{\text{irr}} \approx 0$, indicating that the model focuses on the same parts as the human annotators.

$$\text{ATA}_{\text{rel}} = \frac{\sum_{i=1}^{n} s_i \cdot m_i}{\sum_{i=1}^{n} m_i}; \quad \text{ATA}_{\text{irr}} = \frac{\sum_{i=1}^{n} s_i \cdot (1 - m_i)}{\sum_{i=1}^{n} (1 - m_i)} \quad (8)$$

IV. Experiments

Datasets. Two publicly available datasets from two distinct application domains were employed to evaluate the performance of the proposed analysis: Movie Reviews (MR) [29] in the context of sentiment analysis, and the AsyLex [30] dataset from the legal domain. MR [29] consists of 2000 user-generated movie reviews characterized by an average text length of 132 characters. The output labels associated with each text review are a binary label corresponding to positive or negative sentiment. The AsyLex dataset [30] provides a large-scale long-text collection of 59,112 documents related to refugee status determination decisions in Canada, spanning the years 1996 to 2022. To evaluate the versatility of the proposed method, we considered two classification configurations: (i) a binary classification setup, where each case is labeled as either accepted or rejected considering the OUTCOME of the final decision; and (ii) a multi-class classification setup focused on NORP classification, in which each document is annotated based on mentions of nationalities, religious affiliations, political ideologies, or ethnic groups defined as legally relevant and curated in collaboration with legal experts.

A. Text classification results

To address the classification tasks, we employed two widely used Transformer-based architectures: BERT [2] and RoBERTa [3], both initialized with pretrained weights from the Hugging Face Transformers library. Specifically, we used `bert-base-uncased` and `roberta-base`, which are pretrained on large-scale English corpora including BooksCorpus and English Wikipedia (for BERT), and a broader collection of news, web text, and Common Crawl data (for RoBERTa). Each model was fine-tuned for 20 epochs on all datasets using the Adam optimizer [31]. We explored a range of learning rates $[1e^{-4}, 5e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}]$ and fixed the batch size to 32.

Given the length of the documents in the AsyLex dataset, which exceeds the input size limit of the employed Transformer models, we adopted paragraph selection strategies described in Section II. Specifically, we report results for three selection approaches: **First**, where the model processes only the initial paragraph of each document; **Last**, which uses the final paragraph; and **Random**, where a single paragraph is randomly sampled from the document. The classification results of MR, AsyLex-OUTCOME, and AsyLex-NORP are shown in Tables I, II, and III, respectively.

TABLE I
Classification evaluation on MR.

Model	Acc	Prec	Recall
BERT	92.46	92.47	92.46
RoBERTa	91.46	91.45	91.46

TABLE II
Classification evaluation on AsyLex-OUTCOME.

Parag. Sel.	Model	Acc	Prec	Recall
First	BERT	87.43	84.30	83.51
	RoBERTa	87.25	83.61	84.55
Last	BERT	98.23	97.90	97.60
	RoBERTa	97.74	97.16	97.11
Random	BERT	90.62	89.71	85.90
	RoBERTa	90.61	88.36	87.63

TABLE III
Classification evaluation on AsyLex-NORP.

Parag. Sel.	Model	Acc	Prec	Recall
First	BERT	78.94	74.06	75.95
	RoBERTa	78.36	73.62	76.03
Last	BERT	74.85	71.08	71.32
	RoBERTa	73.88	69.76	69.97
Random	BERT	78.75	74.71	73.86
	RoBERTa	78.75	76.10	75.84

B. XAI results

In this section, we report the results obtained from the proposed SIDU adaptation. We compare the SIDU version relying on persistent homology to generate masks (SIDU-PHMC) and the one based on cosine similarity (SIDU-CSMC) against established explainability methods, including LIME and SHAP. The explanation evaluation results of MR, AsyLex-OUTCOME, and AsyLex-NORP are shown in Tables IV, V, and VI, respectively.

TABLE IV
EXPLANATION EVALUATION ON MR

	BERT		RoBERTa	
Model	ATA_{rel} ↑	ATA_{irr} ↓	ATA_{rel} ↑	ATA_{irr} ↓
PHMC	0.821	0.814	0.892	0.899
CSMC	0.735	0.341	0.802	0.456
SHAP	0.583	0.030	0.678	0.042
LIME	0.312	0.011	0.356	0.009

V. DISCUSSION

First, the SIDU-PHMC variant appears to be ineffective in distinguishing between relevant and irrelevant tokens. This is evident from Tables IV, V, and VI where the difference between ATArel and ATAirr is almost absent across all datasets and architectures, indicating that the attributions generated by SIDU-PHMC are essentially random with respect to human-annotated relevance. Qualitative evidence from Figures 1 and 2 (first sentence) further supports this, showing spotty heatmaps with no clear separation between relevant and irrelevant regions. In contrast, SIDU-CSMC shows substantially better performance. Quantitatively, it achieves higher ATArel than ATAirr, confirming its ability to correctly identify tokens considered important by human annotators. This trend is consistent across Tables IV, V, and VI, and holds for SHAP and LIME as well. However, SIDU-CSMC yields comparatively higher ATA scores for both relevant and irrelevant tokens. This indicates that, although it tends to capture relevant information, it may also partially attribute importance to tokens that are not explicitly marked as relevant by experts. In contrast, SHAP and LIME typically highlight shorter phrases or isolated words with lower overall attribution intensity. From a qualitative standpoint, SIDU-CSMC demonstrates a closer alignment with human reasoning than SHAP and LIME. Human explanations often span entire phrases or sentences, rather than individual tokens. SIDU-CSMC reflects this behavior, producing more coherent and contiguous attribution maps across sequences of related tokens. As illustrated in Figures 1 and 2 (second sentence), the highlighted regions generated by SIDU-CSMC tend to follow a consistent semantic flow, unlike the more fragmented outputs produced by SHAP and LIME.

VI. CONCLUSION

In this pilot study, we introduced the first adaptation of the SIDU framework to transformer-based models, aiming to enhance the interpretability of text classification tasks through token-level explanations. Our investigation centered on two novel masking strategies, Persistent Homology-based Masking and Cosine Similarity-based Masking, applied to final hidden layer representations of transformers. While SIDU-PHMC failed to produce meaningful heatmaps to explain transformer predictions, SIDU-CSMC demonstrated promising results, effectively highlighting relevant tokens in text sequences.

Quantitative and qualitative analyses revealed that SIDU-CSMC not only achieves a clearer separation between relevant and irrelevant tokens than SIDU-PHMC, but is also more in line with human reasoning, favouring semantically more coherent token activations than isolated words. Despite its encouraging results, this study remains exploratory. Future work should focus on optimizing the mask crafting computation of SIDU, extending the evaluation to additional NLP tasks.

ACKNOWLEDGMENT

This research is supported by: (i) the Grundfos Foundation (”REPAI”, grant no. 83648813), (ii) Villum Foundation (”XAI-CRED”, grant no. 69198); (iii) the European Commission under the NextGenerationEU program, Extended Partnership PNRR PE1 - ”FAIR - Future Artificial Intelligence Research” - Spoke 1 ”Human-centered AI”, and PNRR - M4 C2, Investment 1.5 ”Creating and strengthening of ”innovation ecosystems”, building ”territorial R&D leaders”; (iv) the Italian Ministry of Enterprises and Made in Italy, in the framework of the ”Agreements for Innovation” Projet ”4DDS - 4D Drone Swarms” Ref. no. F/310097/01-04/X56. (v) the Italian Ministry of Education and Research (MIUR) in the framework of the FoReLab project (Departments of Excellence), in the framework of the ”Reasoning” project, PRIN 2020 LS Programme, Project number 2493 04-11-2021.

REFERENCES

[1] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, “Text classification: How machine learning is revolutionizing text categorization,” *Information*, vol. 16, no. 2, 2025.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies.*

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[4] P. Fantozzi and M. Naldi, “The explainability of transformers: Current status and directions,” *Computers*, vol. 13, no. 4, 2024.

[5] European Parliament and Council, “Artificial intelligence act.” Regulation (EU) 2024/1689, Official Journal of the European Union, June 2024. Laying down harmonised rules on artificial intelligence.

[6] S. Wachter, B. Mittelstadt, and L. Floridi, “Why a right to explanation of automated decision-making does not exist in the general data protection regulation,” *International data privacy law*, vol. 7, no. 2, pp. 76–99.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, “” 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*, pp. 1135–1144, 2016.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, 2017.

TABLE V

EXPLANATION EVALUATION ON ASYLEX-OUTCOME WITH DIFFERENT PARAGRAPH SELECTION STRATEGIES

		SIDU-PHMC		SIDU-CSMC		SHAP		LIME	
Parag. Sel.	Model	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$
First	BERT	0.901	0.882	0.670	0.411	0.386	0.002	0.154	0.000
	RoBERTa	0.897	0.871	0.705	0.348	0.415	0.004	0.159	0.000
Last	BERT	0.873	0.896	0.652	0.364	0.388	0.002	0.135	0.001
	RoBERTa	0.871	0.847	0.715	0.338	0.392	0.001	0.148	0.000
Random	BERT	0.846	0.839	0.644	0.392	0.402	0.002	0.153	0.000
	RoBERTa	0.887	0.823	0.700	0.345	0.347	0.001	0.129	0.001

TABLE VI

EXPLANATION EVALUATION ON ASYLEX-NORP WITH DIFFERENT PARAGRAPH SELECTION STRATEGIES

		SIDU-PHMC		SIDU-CSMC		SHAP		LIME	
Parag. Sel.	Model	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$	$\text{ATA}_{rel} \uparrow$	$\text{ATA}_{irr} \downarrow$
First	BERT	0.833	0.818	0.609	0.377	0.358	0.002	0.141	0.000
	RoBERTa	0.836	0.798	0.650	0.317	0.378	0.004	0.146	0.000
Last	BERT	0.803	0.832	0.597	0.337	0.358	0.002	0.124	0.000
	RoBERTa	0.810	0.782	0.660	0.308	0.364	0.001	0.137	0.000
Random	BERT	0.779	0.773	0.591	0.361	0.374	0.002	0.141	0.000
	RoBERTa	0.812	0.765	0.637	0.315	0.323	0.001	0.117	0.000

[10] S. M. Muddamsetty, M. N. Jahromi, A. E. Ciontos, L. M. Fenoy, and T. B. Moeslund, "Visual explanation of black-box model: Similarity difference and uniqueness (sidu) method," *Pattern recognition*, vol. 127, p. 108604, 2022.

[11] M. N. Jahromi, S. M. Muddamsetty, A. S. S. Jarlner, A. M. Høgenhaug, T. Gammeltoft-Hansen, and T. B. Moeslund, "Sidu-txt: An xai algorithm for nlp with a holistic assessment approach," *Natural Language Processing Journal*, vol. 7, p. 100078, 2024.

[12] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "Eraser: A benchmark to evaluate rationalized nlp models," *arXiv preprint arXiv:1911.03429*, 2019.

[13] M. Dhaini, K. Z. Hussain, E. Zaradoukas, and G. Kasneci, "Evalxnlp: A framework for benchmarking post-hoc explainability methods on nlp models," *arXiv preprint arXiv:2505.01238*, 2025.

[14] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[15] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, vol. 33, pp. 17283–17297, 2020.

[16] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 838–844, 2019.

[17] D. Rau, M. Dehghani, and J. Kamps, "Revisiting bag of words document representations for efficient ranking with transformers," *ACM Transactions on Information Systems*, vol. 42, no. 5, pp. 1–27, 2024.

[18] J. Lu, M. Henchion, I. Bacher, and B. M. Namee, "A sentence-level hierarchical bert model for document classification with limited labelled data," in *Discovery Science*, Springer International Publishing, 2021.

[19] M. Tuteja and D. González Juclà, "Long text classification using transformers with paragraph selection strategies," in *Proceedings of the Natural Legal Language Processing Workshop 2023*, (Singapore), pp. 17–24, Association for Computational Linguistics, Dec. 2023.

[20] A. Jaiswal and E. Milios, "Breaking the token barrier: Chunking and convolution for efficient long text classification with bert," *arXiv preprint arXiv:2310.20558*, 2023.

[21] M. Tuteja and D. G. Juclà, "Long text classification using transformers with paragraph selection strategies," in *Proceedings of the Natural Legal Language Processing Workshop 2023*, pp. 17–24, 2023.

[22] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[24] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[25] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci, "Towards human-centered explainable ai: A survey of user studies for model explanations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2104–2122, 2024.

[26] M. Parola, A. L. Alfeo, and M. Cimino, "Human-centered xai via a concept-informed prompt-based validation framework for saliency maps [ciprova]," *Available at SSRN 5209213*.

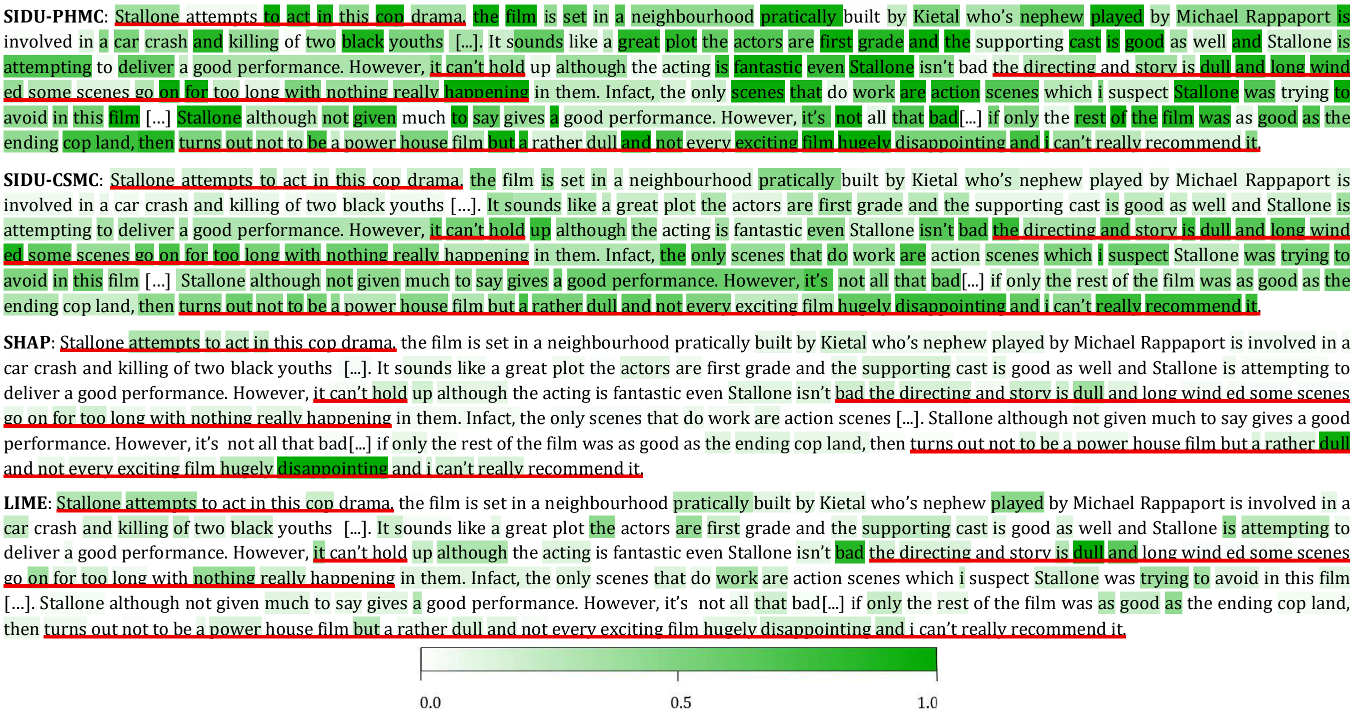
[27] M. Parola, F. A. Galatolo, G. La Mantia, M. G. Cimino, G. Campisi, and O. Di Fede, "Towards explainable oral cancer recognition: Screening on imperfect images via informed deep learning and case-based reasoning," *Computerized Medical Imaging and Graphics*, vol. 117, p. 102433, 2024.

[28] M. G. Cimino, G. Campisi, F. A. Galatolo, P. Neri, P. Tozzo, M. Parola, G. La Mantia, and O. Di Fede, "Explainable screening of oral cancer via deep learning and case-based reasoning," *Smart Health*, vol. 35, p. 100538, 2025.

[29] O. Zaidan, J. Eisner, and C. Piatko, "Using "annotator rationales" to improve machine learning for text categorization," in *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pp. 260–267, 2007.

[30] C. Barale, M. Rovatsos, and N. Bhuta, "Automated refugee case analysis: A NLP pipeline for supporting legal practitioners," in *Findings of the Association for Computational Linguistics: ACL 2023*, (Toronto, Canada), Association for Computational Linguistics, 2023.

[31] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.



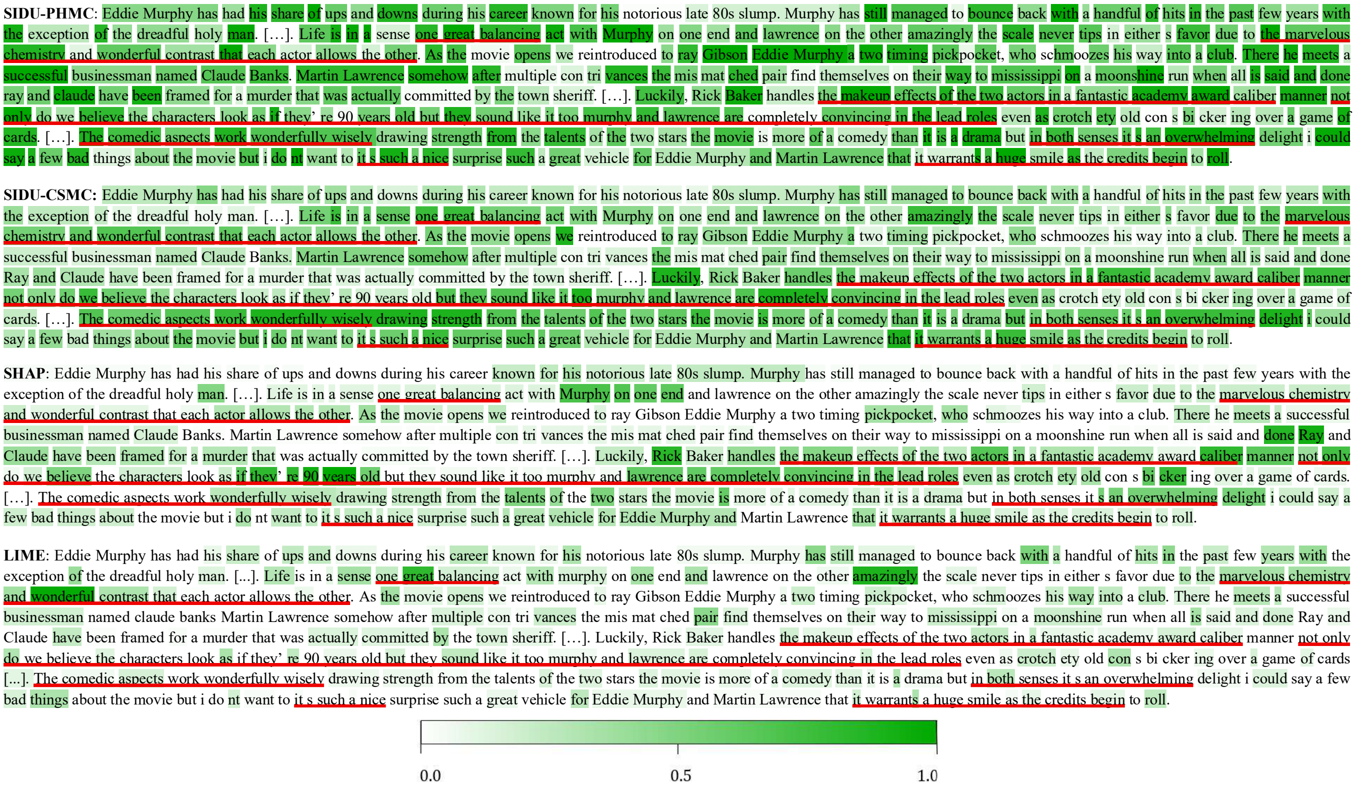
SIDU-PHMC: Stallone attempts to act in this cop drama, the film is set in a neighbourhood pratically built by Kietal who's nephew played by Michael Rappaport is involved in a car crash and killing of two black youths [...]. It sounds like a great plot the actors are first grade and the supporting cast is good as well and Stallone is attempting to deliver a good performance. However, it can't hold up although the acting is fantastic even Stallone isn't bad the directing and story is dull and long wind ed some scenes go on for too long with nothing really happening in them. Infact, the only scenes that do work are action scenes which i suspect Stallone was trying to avoid in this film [...] Stallone although not given much to say gives a good performance. However, it's not all that bad[...] if only the rest of the film was as good as the ending cop land, then turns out not to be a power house film but a rather dull and not every exciting film hugely disappointing and i can't really recommend it.

SIDU-CSMC: Stallone attempts to act in this cop drama, the film is set in a neighbourhood pratically built by Kietal who's nephew played by Michael Rappaport is involved in a car crash and killing of two black youths [...]. It sounds like a great plot the actors are first grade and the supporting cast is good as well and Stallone is attempting to deliver a good performance. However, it can't hold up although the acting is fantastic even Stallone isn't bad the directing and story is dull and long wind ed some scenes go on for too long with nothing really happening in them. Infact, the only scenes that do work are action scenes which i suspect Stallone was trying to avoid in this film [...] Stallone although not given much to say gives a good performance. However, it's not all that bad[...] if only the rest of the film was as good as the ending cop land, then turns out not to be a power house film but a rather dull and not every exciting film hugely disappointing and i can't really recommend it.

SHAP: Stallone attempts to act in this cop drama, the film is set in a neighbourhood pratically built by Kietal who's nephew played by Michael Rappaport is involved in a car crash and killing of two black youths [...]. It sounds like a great plot the actors are first grade and the supporting cast is good as well and Stallone is attempting to deliver a good performance. However, it can't hold up although the acting is fantastic even Stallone isn't bad the directing and story is dull and long wind ed some scenes go on for too long with nothing really happening in them. Infact, the only scenes that do work are action scenes [...]. Stallone although not given much to say gives a good performance. However, it's not all that bad[...] if only the rest of the film was as good as the ending cop land, then turns out not to be a power house film but a rather dull and not every exciting film hugely disappointing and i can't really recommend it.

LIME: Stallone attempts to act in this cop drama, the film is set in a neighbourhood pratically built by Kietal who's nephew played by Michael Rappaport is involved in a car crash and killing of two black youths [...]. It sounds like a great plot the actors are first grade and the supporting cast is good as well and Stallone is attempting to deliver a good performance. However, it can't hold up although the acting is fantastic even Stallone isn't bad the directing and story is dull and long wind ed some scenes go on for too long with nothing really happening in them. Infact, the only scenes that do work are action scenes which i suspect Stallone was trying to avoid in this film [...]. Stallone although not given much to say gives a good performance. However, it's not all that bad[...] if only the rest of the film was as good as the ending cop land, then turns out not to be a power house film but a rather dull and not every exciting film hugely disappointing and i can't really recommend it.

Fig. 1. Visual explanation of BERT [2] model prediction on a sample from the MR dataset belonging to the **negative** sentiment class. The explanations are generated using, from top to bottom: SIDU-PHMC, SIDU-CSMC, SHAP, and LIME. Underlined in red the sentences marked as relevant by humans.



SIDU-PHMC: Eddie Murphy has had his share of ups and downs during his career known for his notorious late 80s slump. Murphy has still managed to bounce back with a handful of hits in the past few years with the exception of the dreadful holy man. […]. Life is in a sense one great balancing act with Murphy on one end and lawrence on the other amazingly the scale never tips in either s favor due to the marvelous chemistry and wonderful contrast that each actor allows the other. As the movie opens we reintroduced to ray Gibson Eddie Murphy a two timing pickpocket, who schmoozes his way into a club. There he meets a successful businessman named Claude Banks. Martin Lawrence somehow after multiple con tri vances the mis mat ched pair find themselves on their way to mississippi on a moonshine run when all is said and done ray and claude have been framed for a murder that was actually committed by the town sheriff. […]. Luckily, Rick Baker handles the makeup effects of the two actors in a fantastic academy award caliber manner not only do we believe the characters look as if they' re 90 years old but they sound like it too murphy and lawrence are completely convincing in the lead roles even as crotch ety old con s bi cker ing over a game of cards. […]. The comedic aspects work wonderfully wisely drawing strength from the talents of the two stars the movie is more of a comedy than it is a drama but in both senses it s an overwhelming delight i could say a few bad things about the movie but i do nt want to it s such a nice surprise such a great vehicle for Eddie Murphy and Martin Lawrence that it warrants a huge smile as the credits begin to roll.

SIDU-CSMC: Eddie Murphy has had his share of ups and downs during his career known for his notorious late 80s slump. Murphy has still managed to bounce back with a handful of hits in the past few years with the exception of the dreadful holy man. […]. Life is in a sense one great balancing act with Murphy on one end and lawrence on the other amazingly the scale never tips in either s favor due to the marvelous chemistry and wonderful contrast that each actor allows the other. As the movie opens we reintroduced to ray Gibson Eddie Murphy a two timing pickpocket, who schmoozes his way into a club. There he meets a successful businessman named Claude Banks. Martin Lawrence somehow after multiple con tri vances the mis mat ched pair find themselves on their way to mississippi on a moonshine run when all is said and done Ray and Claude have been framed for a murder that was actually committed by the town sheriff. […]. Luckily, Rick Baker handles the makeup effects of the two actors in a fantastic academy award caliber manner not only do we believe the characters look as if they' re 90 years old but they sound like it too murphy and lawrence are completely convincing in the lead roles even as crotch ety old con s bi cker ing over a game of cards. […]. The comedic aspects work wonderfully wisely drawing strength from the talents of the two stars the movie is more of a comedy than it is a drama but in both senses it s an overwhelming delight i could say a few bad things about the movie but i do nt want to it s such a nice surprise such a great vehicle for Eddie Murphy and Martin Lawrence that it warrants a huge smile as the credits begin to roll.

SHAP: Eddie Murphy has had his share of ups and downs during his career known for his notorious late 80s slump. Murphy has still managed to bounce back with a handful of hits in the past few years with the exception of the dreadful holy man. […]. Life is in a sense one great balancing act with Murphy on one end and lawrence on the other amazingly the scale never tips in either s favor due to the marvelous chemistry and wonderful contrast that each actor allows the other. As the movie opens we reintroduced to ray Gibson Eddie Murphy a two timing pickpocket, who schmoozes his way into a club. There he meets a successful businessman named Claude Banks. Martin Lawrence somehow after multiple con tri vances the mis mat ched pair find themselves on their way to mississippi on a moonshine run when all is said and done Ray and Claude have been framed for a murder that was actually committed by the town sheriff. […]. Luckily, Rick Baker handles the makeup effects of the two actors in a fantastic academy award caliber manner not only do we believe the characters look as if they' re 90 years old but they sound like it too murphy and lawrence are completely convincing in the lead roles even as crotch ety old con s bi cker ing over a game of cards. […]. The comedic aspects work wonderfully wisely drawing strength from the talents of the two stars the movie is more of a comedy than it is a drama but in both senses it s an overwhelming delight i could say a few bad things about the movie but i do nt want to it s such a nice surprise such a great vehicle for Eddie Murphy and Martin Lawrence that it warrants a huge smile as the credits begin to roll.

LIME: Eddie Murphy has had his share of ups and downs during his career known for his notorious late 80s slump. Murphy has still managed to bounce back with a handful of hits in the past few years with the exception of the dreadful holy man. [...]. Life is in a sense one great balancing act with murphy on one end and lawrence on the other amazingly the scale never tips in either s favor due to the marvelous chemistry and wonderful contrast that each actor allows the other. As the movie opens we reintroduced to ray Gibson Eddie Murphy a two timing pickpocket, who schmoozes his way into a club. There he meets a successful businessman named claude banks Martin Lawrence somehow after multiple con tri vances the mis mat ched pair find themselves on their way to mississippi on a moonshine run when all is said and done Ray and Claude have been framed for a murder that was actually committed by the town sheriff. […]. Luckily, Rick Baker handles the makeup effects of the two actors in a fantastic academy award caliber manner not only do we believe the characters look as if they' re 90 years old but they sound like it too murphy and lawrence are completely convincing in the lead roles even as crotch ety old con s bi cker ing over a game of cards [...]. The comedic aspects work wonderfully wisely drawing strength from the talents of the two stars the movie is more of a comedy than it is a drama but in both senses it s an overwhelming delight i could say a few bad things about the movie but i do nt want to it s such a nice surprise such a great vehicle for Eddie Murphy and Martin Lawrence that it warrants a huge smile as the credits begin to roll.

Fig. 2. Visual explanation of a RoBERTa [3] model prediction on a sample from the MR dataset belonging to the **positive** sentiment class. The explanations are generated using, from top to bottom: SIDU-PHMC, SIDU-CSMC, SHAP, and LIME. Underlined in red the sentences marked as relevant by humans.