Benchmarking Post-Hoc Interpretability Approaches for Transformer-based Misogyny Detection

Giuseppe Attanasio[1,2], **Debora Nozza**[1], **Eliana Pastor**[2], **Dirk Hovy**[1]

[1]Bocconi University, Milan, Italy
[2]Politecnico di Torino, Turin, Italy
{giuseppe.attanasio3,debora.nozza,dirk.hovy}@unibocconi.it,
eliana.pastor@polito.it

Abstract

Warning: This paper contains examples of language that some people may find offensive.

Transformer-based Natural Language Processing models have become the standard for hate speech detection. However, the unconscious use of these techniques for such a critical task comes with negative consequences. Various works have demonstrated that hate speech classifiers are biased. These findings have prompted efforts to explain classifiers, mainly using attribution methods. In this paper, we provide the first benchmark study of interpretability approaches for hate speech detection. We cover four post-hoc token attribution approaches to explain the predictions of Transformer-based misogyny classifiers in English and Italian. Further, we compare generated attributions to attention analysis. We find that only two algorithms provide faithful explanations aligned with human expectations. Gradient-based methods and attention, however, show inconsistent outputs, making their value for explanations questionable for hate speech detection tasks.

1 Introduction

The advent of social media has proliferated hateful content online – with severe consequences for attacked users even in real life. *Women* are often attacked online. A study by Data & Society[1] of women between 15 to 29 years showed that 41% self-censored to avoid online harassment. Of those, 21% stopped using social media, 13% stopped going online, and 4% stopped using their mobile phone altogether. These numbers demonstrate the need for automatic misogyny detection systems for moderation purposes.

	You	**are**	**a**	**smart**	**woman**
ΔP (10^{-2})	-0.1	1.1	-0.0	0.8	-47.6
G	0.11	0.10	0.09	0.25	0.27
IG	-0.17	0.18	-0.09	-0.35	-0.20
SHAP	0.00	-0.14	-0.04	-0.03	0.78
SOC	0.07	-0.13	0.03	0.03	0.52

Table 1: Explanations generated by benchmarked methods. A fine-tuned BERT wrongly classifies the text as misogynous. Darker colors indicate higher importance.

Various Natural Language Processing (NLP) models have been proposed to detect and mitigate misogynous content (Basile et al., 2019; Indurthi et al., 2019; Lees et al., 2020; Fersini et al., 2020a; Safi Samghabadi et al., 2020; Attanasio and Pastor, 2020; Guest et al., 2021; Attanasio et al., 2022). However, several papers already demonstrated that hate speech detection models suffer from unintended bias, resulting in harmful predictions for protected categories (e.g., *women*). Table 1 (top row) reports a very simple sentence that a state-of-the-art NLP model misclassifies as misogynous content.

This issue shows the need to understand the rationale behind a given prediction. A mature literature on model interpretability with applications to NLP-specific approaches exists (Ross et al., 2021; Sanyal and Ren, 2021; Rajani et al., 2019, inter-alia).[2] As explanations become part of legal regulations (Goodman and Flaxman, 2017), a growing body of work has focused on the *evaluation* of explanation approaches (Nguyen, 2018; Hase and Bansal, 2020; Nguyen and Martínez, 2020; Jacovi and Goldberg, 2020, inter-alia). However, little guidance on which interpretability method suits

[1] https://www.datasociety.net/pubs/oh/Online_Harassment_2016.pdf

[2] We refer the reader to Danilevsky et al. (2020) and Madsen et al. (2021) for a recent, thorough perspective on explanation methods for NLP models.

Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP, pages 100 - 112
May 26, 2022 ©2022 Association for Computational Linguistics

best to the sensible context of misogyny identification has been given. For instance, some explanations in Table 1 hint to which token is wrongly driving the classification and even highlight a potential bias of the model. But not all of them.

We bridge this gap. We benchmark interpretability approaches to explain state-of-the-art Transformer classifiers on the task of automatic misogyny identification. We cover two benchmark Twitter datasets for misogyny detection in English and Italian (Fersini et al., 2018, 2020b). We focus on single-instance, post-hoc input attribution methods to measure the importance of each token for predicting the instance label. Our benchmark suite comprises gradient-based methods (Gradients (Simonyan et al., 2014) and Integrated Gradients (Sundararajan et al., 2017)), Shapley values-based methods (SHAP (Lundberg and Lee, 2017)), and input occlusion (Sampling-And-Occlusion (Jin et al., 2020)). We evaluate explanations in terms of plausibility and faithfulness (Jacovi and Goldberg, 2020). Table 1 reports an example of token-wise contribution computed with these methods. Furthermore, we study attention-based visualizations and compare them to token attribution methods searching for any correlation. To our knowledge, this is the first benchmarking study of feature attribution methods used to explain Transformer-based misogyny classifiers.

Our results show that SHAP and Sampling-And-Occlusion provide plausible and faithful explanations and are consequently recommended for explaining misogyny classifiers' outputs. We also find that, despite their popularity, gradient- and attention-based methods do *not* provide faithful explanations. Outputs of gradient-based explanation methods are inconsistent, while *attention does not provide any useful insights for the classification task*.

Contributions We benchmark four post-hoc explanation methods on two misogyny identification datasets across two languages, English and Italian. We evaluate explanations in terms of plausibility and faithfulness. We demonstrate that not every token attribution method provides reliable insights and that attention cannot serve as explanation. Code is available at `https://github.com/MilaNLProc/benchmarking-xai-misogyny`.

2 Benchmarking suite

In the following, we describe the scope (§2.1) of our benchmarking study, the included methods (§2.2), and the evaluation criteria (§2.2).

2.1 Scope

We consider *local* explanation methods (Lipton, 2018; Guidotti et al., 2019). Given a classification model, a data point, and a target class, these methods explain the probability assigned to the class by the model. *Global* explanations provide model- or class-wise explanations and are hence out of the scope of this work.

Among local explanation methods, we focus on *post-hoc* interpretability, i.e., we explain classification models that have already been trained. We leave out *inherently interpretable* models (Rudin, 2019) as they do not find widespread use in NLP-driven practical applications.

We restrict our study to input attribution methods. In Transformer-based language models, inputs typically correspond to the tokens' input embeddings (Madsen et al., 2021). We, therefore, refer to *token attribution* methods to generate a contribution score for each input token (or word, resulting by some aggregation of sub-word token contributions).

2.2 Methods

We benchmark three families of input token attribution methods. First, we derive token contribution using gradient attribution. These methods compute the gradient of the output with respect to each of the inputs. We compute simple gradient (G) (Simonyan et al., 2014) and integrated gradients (IG) (Sundararajan et al., 2017). Then, we attribute inputs using approximated Shapley values (SHAP) (Lundberg and Lee, 2017). Finally, following the literature on input perturbation via occlusion, we impute input contributions using Sampling-And-Occlusion (SOC) (Jin et al., 2020). See appendix A.2 for all implementation details.

Attention There is an open debate of whether attention is explanation or not (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020). Our benchmarking study provides a perfect test-bed to understand if attention aligns with attribution methods. We compare standard self-attention with effective attention (Brunner et al., 2020; Sun and Marasović, 2021). Further, we measure attribution between input tokens and

Dataset	# Train	# Test	Hate %	F1
AMI-EN	4,000	1,000	45%	68.78
AMI-IT	5,000	1,000	47%	79.79

Table 2: Summary of datasets in terms of the number of training, validation and test tweets, percentage of hateful records within the training split, and F1-score of BERT models on test sets.

hidden representations using Hidden Token Attribution (HTA) (Brunner et al., 2020).

2.3 Evaluation criteria

We use *plausibility* and *faithfulness* as evaluation criteria (Jacovi and Goldberg, 2020). A "plausible" explanation should align with human beliefs. In our context, the provided explanation artifacts should *convince* humans that highlighted words are responsible for either misogynous speech or not.[3] A "faithful" explanation is a proxy for the true "reasoning" of the model. Gradient attributions are commonly considered faithful explanations as gradients provide a direct, mathematical measure of how variations in the input influences output. For the remaining attribution approaches, we measure faithfulness under the *linearity assumption* (Jacovi and Goldberg, 2020), i.e., the impact of certain parts of the input is independent of the rest. In our case, independent units correspond to input tokens. Following related work (Jacovi et al., 2018; Feng et al., 2018; Serrano and Smith, 2019, inter-alia), we evaluate faithfulness by erasing input tokens and measuring the variation on the model prediction. Ideally, faithful interpretations highlight tokens that change the prediction the most.

2.4 Data

Automatic misogyny identification is the binary classification task to predict whether a text is misogynous or not.[4] We focus on two recently-released datasets for misogynous content identification in English and Italian, released as part of the Automatic Misogyny Identification (AMI) shared tasks (Fersini et al., 2018, 2020b). Both datasets have been collected via keyword-based search on Twitter. Table 2 reports the dataset statistics.

3 Experimental setup

Among the Transformer-based models, we focus on BERT (Devlin et al., 2019) due to its widespread usage. We fine-tuned pre-trained BERT-based models on the AMI-EN and AMI-IT datasets. We report full details on the training in appendix A.1. Table 2 reports the macro-F1 performance of BERT models on the test splits.

We explain BERT outputs on both tweets from test sets[5] and manually-generated data. On real data, we address two questions: 1) *Is it right for the right reason?*, i.e., we assess if the model relies on a plausible set of tokens; 2) *What is the source of error?*, i.e., we aim to identify tokens that wrongly drive the classification outcome. By explaining manually-defined texts, we can probe for model biases.

Tables 3-6 report token contributions computed with benchmarked approaches (§2.2). We report contributions for individual tokens.[6] We define table contents as follows. Separately by explanation method, we first generate raw contributions and then L1-normalize the vector. Finally, we use a linear color scale between solid blue (assigned for contribution -1), white (contribution 0), and solid red (contribution 1). For all reported examples, we explain the `misogynous` class. Hence, positive contributions indicate tokens *pushing* towards the misogynous class, while negative contributions push towards the non-misogynous one. Lastly, the second top row reports the variation on the probability assigned by the model when the corresponding token is erased (ΔP).

4 Discussion

Error analysis Table 3 shows the explanations for a tweet incorrectly predicted as misogynous. IG, SHAP, and SOC assign a negative contribution to the word *boy*. This matches our expectations since the target of the hateful comment is the male gender. These explanations are thus plausible. Still, the tweet is classified as misogynous. The tokens *pu* and *##ssy* mainly drive the prediction to the misogynous class, as revealed by all explainers (SHAP and SOC in a clearer way). Ex-

[3] In this study, the human expectation corresponds to the authors'.

[4] Characterizing misogyny is a much harder task, possibly modeling complex factors such as shaming, objectification, or more. Here, we simplify the task to focus on benchmarking interpretability.

[5] We rephrase and explain rephrased versions of tweets to protect privacy.

[6] While several work average sub-word contributions for out-of-vocabulary words, there is no general agreement on whether this brings meaningful results. Indeed, an average would assume a model that leverages tokens as a single unit, while there is no clear evidence of that.

	You	pu	##ssy	boy
ΔP (10^{-2})	-0.3	-0.2	-35.6	0.8
G	0.11	0.19	0.32	0.18
IG	0.26	0.00	0.14	-0.60
SHAP	-0.03	0.52	0.28	-0.17
SOC	-0.01	0.03	0.51	-0.14

Table 3: Example from AMI-EN test set, anonymyzed text on first row. Ground truth: `non misogynous`. Prediction: `misogynous` ($P = 0.78$).

planations suggest the model is failing to assign the proper importance to the targeted gender of the hateful comment. These plausible explanations are also faithful. Removing the term *boy* increases the probability of the misogynous class while omitting tokens *pu* and *##ssy* decrease it.

We further analyze the term *p*ssy* and its role as a source of errors. Almost all tweets of the test set containing the term *p*ssy* are labeled by the model as misogynous. The false-positive rate on this set of tweets is 0.93 compared to the 0.49 of the overall test set. Similar considerations apply to English words typically associated with misogynous content as *b*tch* and *wh*re*.

Is it right for the right reason? Table 4 shows the explanation of a correctly predicted misogynous tweet. Gradient, SHAP, and SOC explanations assign a high positive contribution to slurs (*b*tch*, *s*ck*, and *d*ck*). These explanations align with human expectations. However, not all slurs impact the classification outcome. Explanations on *b*tch* are faithful but they are not for *s*ck* and *d*ck*. Differently, IG does not highlight any token with a positive contribution. This goes against expectations as the predicted class is misogynous and therefore we cannot draw conclusions.

Unintended bias We study explanations to search for errors caused by unintended bias, a known phenomenon affecting models for misogynous identification. A model suffering from unintended bias performs better (or worse) when texts mention specific identity terms (e.g., *woman*) (Dixon et al., 2018).

Table 1 reports the non-misogynous text "You are a smart woman" incorrectly labeled as misogynous. SHAP, SOC, and, to a lesser extent, Gradient explanations indicate the term *woman* as responsible for the prediction. This result matches with recent findings on the unintended bias of hateful detection models (Nozza et al., 2019; Dixon et al., 2018; Borkan et al., 2019) and therefore explanations are plausible. Removing the term *woman* causes a drop of 0.48 to the probability of the misogynous class. This validates the insight provided by the explanations. Similar to the previous examples, the explanation of IG is difficult to interpret.

Table 5 shows another example of unintended bias. The text “Ann is in the kitchen” is incorrectly labeled as misogynous. Gradients, SHAP, and SOC assign the highest positive contribution to the (commonly) female name *Ann*. Interestingly, the second most important word for Gradients and SHAP is *kitchen*, reflecting stereotypes learned by the classification model (Fersini et al., 2018). These explanations are faithful: the model prediction drops by a significant 0.40 and 0.24 when erasing the tokens *Ann* and *kitchen*, respectively. We substitute the name *Ann* with *David*, a common male name. We observe that the prediction and the explanations drastically change. The text is correctly assigned to the non-misogynous class and IG, SHAP, and SOC assign a high negative contribution to the word *David*. The all-positive contributions of Gradients do not provide useful insights.

Bias due to language-specific expressions Table 6 (left) shows an example of incorrectly predicted misogynous text in Italian: "p*rca p*ttana che gran pezzo di f*ga" ("holy sh*t what a nice piece of *ss"). The expression "*p*rca p*ttana*" (literally *pig sl*t*) is a taboo interjection commonly used in the Italian language and does not imply misogynous speech.

The interpretation of the gradient explanation is hard since all contributions are positive and associated with the misogynous class. All explanation methods assign a positive contribution to the word *f*ga* (**ss*). SHAP, SOC, and, to a lesser extent IG, indicate that the main reason behind the non-misogynous prediction is the term *p*rca*. The bias of the model towards this expression was firstly exposed in (Nozza, 2021) and it thus validates IG, SHAP, and SOC explanations as plausible. When one of the two terms of the expression is removed, the probability increases significantly. This suggests that explanations by IG, SHAP, and SOC are faithful. Further, we inspect the behavior of explanation methods when we erase one of the terms. We omit the word *p*rca* and we report its explanations on Table 6 (right). The text is correctly assigned to the misogynous class and the word

	s*ck	a	d*ck	and	choke	you	b*tch
ΔP (10^{-2})	-0.02	0.2	0.8	0.3	-0.1	0.03	-13.4
G	0.10	0.08	0.14	0.07	0.08	0.10	0.25
IG	-0.14	-0.16	-0.08	-0.05	-0.20	-0.22	-0.16
SHAP	0.24	-0.03	0.07	-0.05	0.05	-0.06	0.50
SOC	0.20	-0.02	0.26	-0.02	0.07	0.00	0.29

Table 4: Example from AMI-EN test set, anonymyzed text on first row. Ground truth: `misogynous`. Prediction: `misogynous` ($P = 0.90$).

	Ann	is	in	the	kitchen	David	is	in	the	kitchen
ΔP (10^{-2})	-40.4	15.4	12.7	-12.6	-24.3	-1.0	8.0	-1.3	-5.8	-6.7
G	0.25	0.16	0.08	0.10	0.21	0.19	0.18	0.09	0.09	0.28
IG	-0.15	0.18	0.12	-0.33	-0.22	-0.36	0.14	0.09	-0.25	-0.17
SHAP	0.27	-0.31	-0.15	-0.01	0.27	-0.29	-0.38	-0.19	-0.05	0.09
SOC	0.28	-0.19	-0.06	0.10	0.07	-0.25	-0.11	-0.03	0.04	0.05

Table 5: Manually-generated example. Text starts with a female (left) and male (right) name. Ground truth (both): `non-misogynous`. Prediction: `misogynous` ($P = 0.53$) (left), `non-misogynous` ($P = 0.14$) (right).

	p*rca	p*ttana	che	gran	pezzo	di	f*ga	p*ttana	che	gran	pezzo	di	f*ga
ΔP (10^{-2})	94.7	79.7	-0.8	-0.6	0.3	-0.7	-0.6	1.0	-2.3	-1.3	0.4	0.3	-22.9
G	0.17	0.15	0.06	0.07	0.11	0.07	0.13	0.20	0.08	0.10	0.14	0.08	0.21
IG	-0.25	-0.10	-0.09	-0.16	-0.04	0.21	0.13	-0.12	-0.03	-0.25	0.11	0.17	0.32
SHAP	-0.69	-0.01	0.01	0.05	0.05	0.05	0.14	0.15	0.10	0.13	0.10	0.10	0.43
SOC	-0.56	-0.07	0.00	0.04	0.05	-0.05	0.22	0.00	0.05	0.07	0.04	-0.12	0.57

Table 6: Manually-generated example. Complete text (left) and text without initial "p*rca" (right). Non-literal translation: "*holy sh*t what a nice piece of *ss*". Ground truth (both): `misogynous`. Prediction: `non-misogynous` ($P = 0.03$) (left), `misogynous` ($P = 0.97$) (right).

*f*ga* (**ss*) has the highest positive contribution for all the approaches.

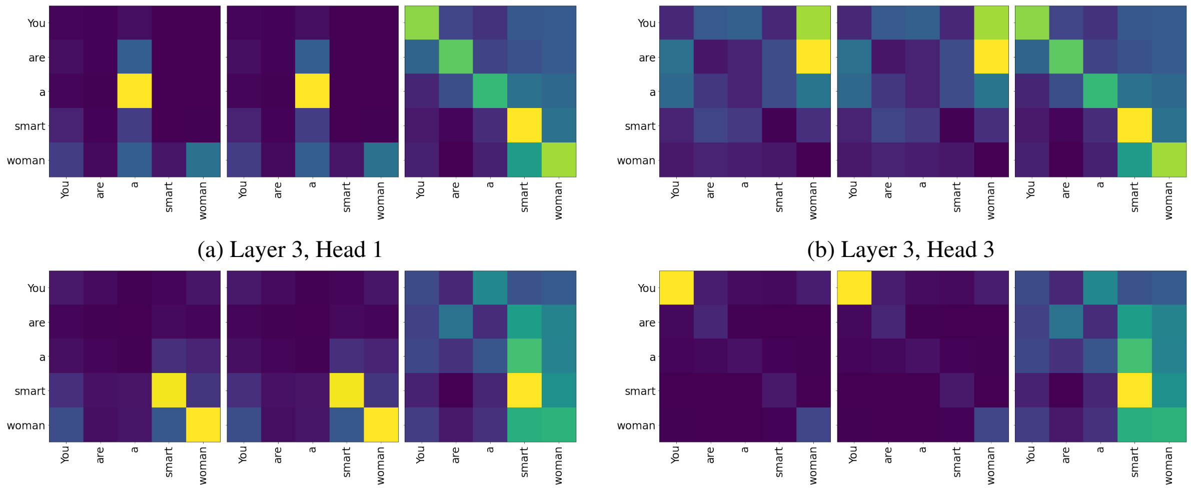
4.1 Is attention explanation?

We follow up on the open debate on attention used as an explanation, providing examples on the misogyny identification task. Figure 1 shows self-attention maps in our fine-tuned BERT at different layers and heads for the already discussed sentence "You are a smart woman". Based on our previous analysis (§4), we know that the model has an unintended bias towards the token "woman".

We cannot infer the same information from attention maps. Raw attention weights differ significantly for different layers and heads. In this example, there is a vertical pattern (Kovaleva et al., 2019) on the token "a" in layer 3 (Figure 1a). However, the pattern disappears from heads in the same layer (Figure 1b) and from the same head on deeper layers, where, instead, a block pattern characterizes "smart" and "woman" (Figure 1c). This variability hinders interpretability as no unique behavior emerges. Effective Attention (Brunner et al., 2020) is based on attention and shares the same issue.[7] These results further motivate the idea that attention gives only a *local* perspective on token contribution and contextualization (Bastings and Filippova, 2020). However, this does not provide any useful insight for the classification task. To further validate this limited scope, we use Hidden Token Attribution (Brunner et al., 2020) and measure the contribution of each input token (i.e., its first-layer token embedding) to hidden representations. On lower layers, there is a marked diagonal contribution, meaning that tokens mainly contribute to their own representation. Interestingly, on the upper layers, a strong contribution to "smart" and "woman" appears for all the tokens in the sentence. Different patterns between HTA and attention suggest that, even in the locality of a layer and a single head, attention weights do not measure token contribution.

We observed similar issues on other examples and for Italian models (see appendix B). We there-

[7] In most of our experiments, Effective Attention brings no perceptually different maps than simple Attention. The two methods are hence equivalent for local attention inspection.



(a) Layer 3, Head 1

(b) Layer 3, Head 3

(c) Layer 10, Head 1

(d) Layer 10, Head 3

Figure 1: Attention (left), Effective Attention (center), and Hidden Token Attribution (right) maps at different layers in fine-tuned BERT. Lighter colors indicate higher weights. Sentence: “You are a smart woman”.

fore cannot consider attention as a plausible nor a faithful explanation method and *discourage the use of attention to explain BERT-based misogyny classifiers*.

5 Related Work

Few works applied interpretability approaches to hate speech detection. Wang (2018) proposes an adaptation of explainability techniques for computer vision to visualize and understand the CNN-GRU classifier for hate speech (Zhang et al., 2018). Mosca et al. (2021) study both local and global explanations. They use Shapley values (Lundberg and Lee, 2017) to quantify feature importance on a *local* level and feature space exploration for a *global* explanation. Risch et al. (2020) analyze multiple attribution-based explanation methods for offensive language detection. The analysis includes an interpretable model (Naïve Bayes), model-agnostic methods based on surrogate models (LIME (Ribeiro et al., 2016), layer-wise relevance propagation (LRP) (Bach et al., 2015), and a self-explanatory model (LSTM with an attention mechanism). SHAP explainer is applied (Wich et al., 2020) to investigate the impact of political bias on hate speech classification. Sample-And-Occlusion (SOC) explanation algorithm has been used in its hierarchical version in different papers for showing the results of hate speech detection (Nozza, 2021; Kennedy et al., 2020).

In this paper, we specifically focus on hate speech against women. In this context, Godoy and Tommasel (2021) apply SHAP to derive global explanations with the aim of exploring unintended bias of Random Forest-based misogyny classifier.

While growing efforts are made for evaluating interpretability approaches for NLP models (Atanasova et al., 2020; DeYoung et al., 2020; Prasad et al., 2021; Nguyen, 2018; Hase and Bansal, 2020; Nguyen and Martínez, 2020; Jacovi and Goldberg, 2020), the evaluation is not domain-specific. Therefore, the benchmarking miss to consider specific sensitive problems and biases that are proper of the hate speech domain on which the explanation validation must focus. This paper fills this gap by focusing on post-hoc feature attribution explanation methods on individual predictions for the task of hate speech against women.

6 Conclusion

In this paper, we benchmarked different explainability approaches on Transformer-based models for the task of hate speech detection against women in English and Italian. We focus on post-hoc feature attribution methods applied to fine-tuned BERT models. Our evaluation demonstrated that SHAP and SOC provide plausible and faithful explanations and are consequently recommended for explaining misogyny classifiers’ outputs. In contrast, gradient- and attention-based approaches failed in providing reliable explanations.

As future work, we plan to add to the benchmarking suite a systematic evaluation involving human annotators. We also plan to include recently introduced token attribution methods (Sikdar et al., 2021) as well as new families of approaches, like natural language explanations (Rajani et al., 2019; Narang et al., 2020) and input editing (Ross et al., 2021). Finally, we will assess explanations of the most problematic data subgroups (Goel et al., 2021; Pastor et al., 2021; Wang et al., 2021).

Acknowledgments

We would like to thank the anonymous reviewers and area chairs for their suggestion to strengthen the paper. This research is partially supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR), and by Fondazione Cariplo (grant No. 2020-4288, MONICA). DN, and DH are members of the MilaNLP group, and of the Data and Marketing Insights Unit at the Bocconi Institute for Data Science and Analysis. EP is member of the DataBase and Data Mining Group (DBDMG) at Politecnico di Torino. GA did part of the work as a member of the DBDMG and is currently a member of MilaNLP. Computing resources were provided by the SmartData@PoliTO center on Big Data and Data Science.

Ethical Considerations

We explain BERT-based classifiers using a controlled subset of a large, fast-growing collection of explanation methods available in the literature. While replicating our experiments with different approaches, or on different data samples, from different datasets or explaining different models, we cannot exclude that some people may find the explanations offensive or stereotypical. Further, recent work has demonstrated gradient-based explanations are manipulable (Wang et al., 2020), questioning the reliability of this widespread category of methods.

We, therefore, advocate for responsible use of this benchmarking suite (or any product derived from it) and suggest pairing it with human-aided evaluation. Moreover, we encourage users to consider this work as a starting point for model debugging (Nozza et al., 2022) and the included explanation methods as baselines for future developments.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL2022*. Association for Computational Linguistics.

Giuseppe Attanasio and Eliana Pastor. 2020. PoliTeam @ AMI: Improving sentence embedding similarity with misogyny lexicons for automatic misogyny identificationin italian tweets. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 48–54. Accademia University Press.

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

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association* *for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. 2020a. Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France. European Language Resources Association (ELRA).

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). volume 12, page 59, Turin, Italy. CEUR.org.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020b. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Daniela Godoy and Antonela Tommasel. 2021. Is my model biased? exploring unintended bias in misogyny detection tasks. In *AIofAI'21: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies*, pages 97–111.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42.

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. 2020. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Vijayasaradhi Indurthi, Bakhtiyar Syed, Manish Shrivastava, Nikhil Chakravartula, Manish Gupta, and Vasudeva Varma. 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 70–74, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65, Brussels, Belgium. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Bologna, Italy. CEUR.org.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2021. Post-hoc Interpretability for Neural NLP: A Survey. *arXiv preprint arXiv:2108.04840*.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. *arXiv preprint arXiv:2004.14546*.

An-phi Nguyen and María Rodríguez Martínez. 2020. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, , and Dirk Hovy. 2022. Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of the First Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Eliana Pastor, Luca de Alfaro, and Elena Baralis. 2021. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*, page 1400–1412, New York, NY, USA. Association for Computing Machinery.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2021. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 1–14, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 137–143, Marseille, France. European Language Resources Association (ELRA).

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 2021. Integrated directional gradients: Feature interaction attribution for neural NLP models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, Online. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014*.

Kaiser Sun and Ana Marasović. 2021. Effective attention sheds light on interpretability. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4126–4135, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Cindy Wang. 2018. Interpreting neural network hate speech classifiers. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 86–92, Brussels, Belgium. Association for Computational Linguistics.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018. Detecting hate speech on twitter using a convolution-GRU based deep neural network. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 745–760. Springer.

A Experimental setup

A.1 Training hyper-parameters

All our experiments use the Hugging Face transformers library (Wolf et al., 2020). We base our models and tokenizers on the `bert-base-cased` checkpoint for English tasks and on the `dbmdz/bert-base-italian-cased` checkpoint for Italian. We pre-process and tokenize our data using the standard pre-trained BERT tokenizer, with a maximum sequence length of 128 and right padding. We train all models for 3 epochs with a batch size of 64, a linearly decaying learning rate of $5 \cdot 10^{-5}$ and 10% of the total training step as a warmup, and full precision. We use 10% of training data for validation. We evaluate the model every 50 steps on the respective validation set. At the end of the training, we use the checkpoint with the best validation loss. We re-weight the standard cross-entropy loss using the inverse of class frequency to account for class imbalance.

A.2 Explanation methods

We used the Captum library (Kokhlikyan et al., 2020) with default parameters to compute gradients (G) and integrated gradients (IG). Following (Han et al., 2020), for IG we multiply gradients by input word embeddings. For Shapley values estimation (SHAP), we use the shap library[8] with Partition-SHAP as approximation method. For Sampling-And-Occlusion (SOC), we used the implementation associated with Kennedy et al. (2020).[9] Please refer to our repository (`https://github.com/MilaNLProc/benchmarking-xai-misogyny`) for further technical details.

A.3 Attention maps

We used attention weights provided by the transformers library for visualization. We implemented Effective Attention and Hidden Token Attribution following Brunner et al. (2020). We release the implementation on our repository.

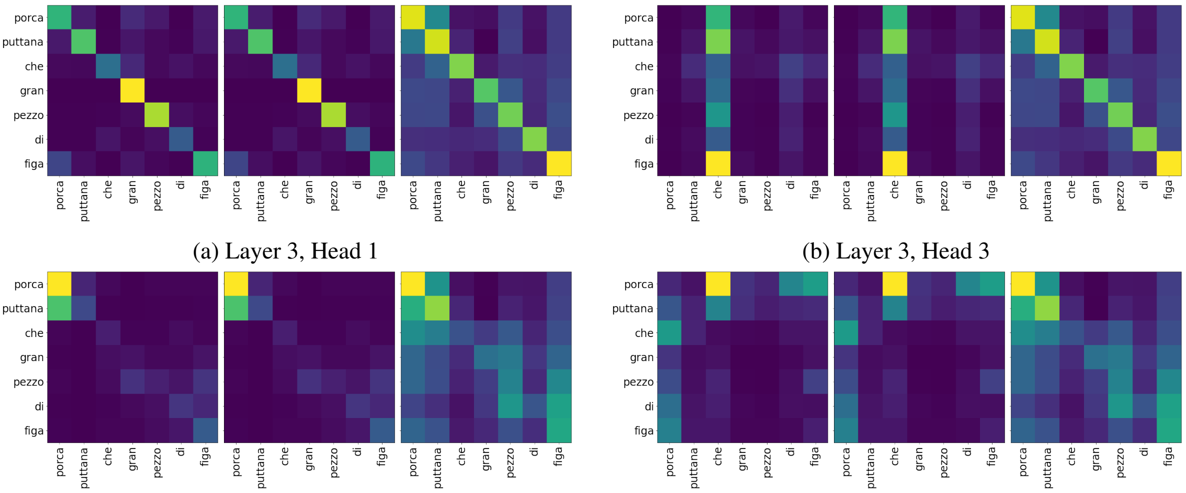
B Attention plots

Figure 2 shows attention visualizations for the sentence "p*rca p*ttana che gran pezzo di f*ga" (Non-literal translation: "*holy sh*t what a nice piece of *ss*"). As discussed in §4 (**Bias due to language-specific expressions**), the text is misclassified as `non-misogynous` and most of explanation methods correctly highlight the Italian interjection "p*rca p*ttana".

Similar to results reported in §2.2, we cannot find useful insights in attention plots. Attention in layer 3 has a diagonal pattern in head 1, and a diagonal pattern in head 3 on the word *che* ("*what*"). However, these patterns disappear in layer 10 where attention is focused on *p*rca*. At layer 10, HTA is more spread than attention, suggesting that the latter measures only a *local* token contribution.

[8] `https://github.com/slundberg/shap`

[9] `https://github.com/BrendanKennedy/contextualizing-hate-speech-models-with-explanations`



(a) Layer 3, Head 1

(b) Layer 3, Head 3

(c) Layer 10, Head 1

(d) Layer 10, Head 3

Figure 2: Attention (left), Effective Attention (center), and Hidden Token Attribution (right) maps at different layers in fine-tuned BERT. Lighter colors indicate higher weights. Sentence: “p*rca p*ttana che gran pezzo di f*ga”, non-literal translation: “*holy sh*t what a nice piece of *ss*”.