Evaluating self-attention interpretability through human-grounded experimental protocol

Milan Bhan, Nina Achache, Victor Legrand, Annabelle Blangero, Nicolas Chesneau

Evaluating self-attention interpretability through human-grounded experimental protocol

Milan Bhan[1,2], Nina Achache[1], Victor Legrand[1], Annabelle Blangero[2,3], and Nicolas Chesneau[1]

[1] Ekimetrics
[2] Sorbonne Université, Paris, France
[3] Aix-Marseille Université, Aix-Marseille, France

Abstract. Attention mechanisms have played a crucial role in the development of complex architectures such as Transformers in natural language processing. However, Transformers remain hard to interpret and are considered as black-boxes. In this paper we assess how attention coefficients from Transformers help in providing classifier interpretability when properly aggregated. A fast and easy-to-implement way of aggregating attention is proposed to build local feature importance. A human-grounded experiment is conducted to evaluate and compare this approach to other usual interpretability methods. The experimental protocol relies on the capacity of an interpretability method to provide explanation in line with human reasoning. Experiment design includes measuring reaction times and correct response rates by human subjects. Attention performs comparably to usual interpretability methods and significantly better than a random baseline regarding average participant reaction time and accuracy. Moreover, data analysis highlights that high probability prediction induces great explanation relevance. This work shows how self-attention can be aggregated and used to explain Transformer classifiers. The low computational cost of attention compared to other interpretability methods and its availability by design within Transformer classifiers make it particularly beneficial. Finally, the quality of its explanation depends strongly on the certainty of the classifier's prediction related to it.

Keywords: Interpretability · NLP · XAI · Attention · Human-grounded ML

1 Introduction

The field of machine learning (ML) has recently witnessed great advances. ML algorithms have achieved high levels of performance in a wide variety of tasks due to their rapid development. Natural language processing (NLP) has also taken advantage of recent breakthroughs in ML with the development and democratization of Transformer-type models [29]. Attention mechanism [3] is a crucial component in Transformer architecture, enabling models to focus on specific parts of the input text. The complexity of these new models led to an increasing difficulty in understanding and interpreting their predictions. The field

of eXplainable Artificial Intelligence (XAI) has emerged to overcome this lack of transparency by developing methods of "interpretability" or "explainability" [18]. Such a gap needs to be filled in many areas ,e.g., health care [13] and finance [32]. Most commonly used interpretability methods are computationally greedy and based on strong hypothesis such as features' independence and linear approximation [21]. More specifically, attention interpretability from Transformers has been debated [7] and remains questionable. The diversity of interpretability methods raises the need for their comparison.

Human-grounded protocols have been proposed to experimentally address the assessment and the comparison of interpretability methods [31, 24, 5, 25]. These empirical approaches consist in asking humans to interact with a machine and to perform a specific task under the influence of interpretability methods. An XAI method will be considered as better than another should it yield to an improved human-making performance.

This paper aims at assessing how self-attention from Transformer classifiers can be used to build reliable local feature importance. Self-attention is aggregated in a specific way that we call CLaSsification-Attention (CLS-A) for convenience in the following. The main contributions of this paper are summarized as follows:

1. CLS-A, an easy-to-implement way of aggregating self-attention in Transformer classifiers is presented.
2. CLS-A is experimentally compared to other interpretability methods with a human-grounded protocol.
3. The dependency between prediction certainty and explanation reliability is highlighted.

In this paper we first introduce key notions of Transformers architecture and local feature importance. We then present CLS-A and the experimental protocol used to assess its interest compared to a baseline and other XAI approaches. Finally, we analyze the data produced by the three experiments.

2 Background & related work

This section introduces some key notions about Transformers architectures and XAI.

2.1 Background

Transformers, attention and BERT In NLP, Transformer-like models have achieved high levels of performance in a variety of tasks, such as text summarizing, question answering or named entity recognition. These models are particularly complex, with a number of parameters that can exceed the billion [23]. The Transformer architecture is based on multi-head self-attention mechanisms, aiming at making learning more efficient [3] by encoding the relations between words. The model attends to different parts of the input in parallel, using multiple

self-attention heads. A self-attention head takes as input a triplet (Q, K, V) and outputs a representation as formalized in the following formula :

$$\texttt{Attention}(Q, K, V) = \texttt{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Where:

- Q (query) is a matrix that represents the input in which the attention mechanism focuses on.
- K (key) is a matrix that represents the different elements in the input that the attention mechanism can attend to.
- V (value) is a matrix that represents the output of the attention mechanism.
- d_k is the dimension of matrices Q and V and allows to stabilize the model during the training phase

Hence, each head has its own set of parameters, allowing the model to learn different types of attention patterns. The attentions resulting from each of the heads are then concatenated and projected on a dense layer.

Bidirectional Encoder Representations from Transformers (BERT) [11] is a stack of n encoders from the Transformer architecture. Each BERT layer contains h attention heads with its own set of weights, which have been learned during training. These weights determine how the model will attend to different parts of the input when making a prediction. In this way, words are related to each other even in the case of long-term dependency. BERT has been widely adopted and has achieved state-of-the-art performance on a variety of benchmarks.

One of the key features of BERT is its bidirectional nature. Unlike previous models that were only trained to look *before* or *after* a word, BERT is trained to look *before* and *after* a word at the same time. This allows BERT to understand the full context of a word and improve its performance on NLU tasks. Moreover, BERT has several advantages such as its scalability, its compatibility with parallelization and its ability to capture long-distance dependencies. The BERT *CLS* (for "classification") token is a special token added at the beginning of an input text. This token is used as a representation of the entire input sequence by the classifier to perform prediction. The final hidden state of the *CLS* token, which is a fixed-size vector, is typically used as the input to a classifier or other downstream task. This allows BERT to take into account information from the entire sequence when making a prediction by computing only one token.

Local feature importance. There are several ways to interpret black-box systems such as BERT models [15]. One of the main approaches consists in computing local feature importance [18]. When the model to explain is a classifier, contributions to the probability score of the predicted class are computed and assigned to input features. It can be done by considering ML models as black-boxes and explaining their predictions *ex post*, without referring to their inherent parameters. This kind of approach is called *post-hoc model-agnostic*[8]. Another way to compute local feature importance is to go through inherent

model parameters[27, 28], which can be significantly less computationally greedy. However, this requires access to the model parameters, which is not always the case. This kind of approaches is referred to as *post-hoc model-specific*. The large number of local feature importance methods can make it difficult to choose the most appropriate one.

Due to their flexibility and the plurality of data types on which they can be applied, Linear Interpretable Model-agnostic Explainer (LIME) [22] and SHapley Additive exPlanations (SHAP) [16] are the most frequently used interpretability methods in the industry [6]. LIME offers to explain a prediction locally using models that are interpretable by design such as sparse regressions. The algorithm artificially generates data points in a neighborhood around the instance to explain and fits an interpretable model on these new examples. The SHAP method is inspired by the Shapley values[26] from economics and game theory. It aims to distribute fairly the rewards from a set of games to all the players. Feature importance of a specific prediction is computed by associating the features of a model to the players to whom the gains are distributed.

2.2 Related work

Substantial amount of linguistic and syntactic knowledge can be found in Transformer attention [10]. However, the interpretability of the attention coefficients is still an open question [7]. Several methods based on self-attention coefficients have been proposed to explain the predictions made by Transformer-type models, such as attention flow and attention roll-out [1]. These methods are based on complex aggregators to synthesize the information contained in the attention layers. Visualization tools allowing to dive in detail into self-attention have been developed as well [30]. Visualizing attention is the basis of saliency map approaches specific to Computer Vision for Vision Transformers [4, 9]. If these approaches enable to compute local feature importance, the quality of the explanation produced is not rigorously assessed. This raises the question of the evaluation of interpretability.

One way to assess the interpretability of a given method is to compare it quantitatively [2] to common interpretable approaches such as SHAP. A given method would thus be considered interpretable if it strongly correlates with one target local feature importance method. One way to measure such correlation is to use Pearson or Kendall coefficients in order to compare feature importance rankings resulting from these interpretability methods. This approach has its limitations because it would imply that LIME and SHAP would be unique ground truths to replicate.

When no ground truths are available, a given method can be evaluated with function-grounded metrics [2]. *Faithfulness* measures the impact on the probability score by perturbing the features considered as important or unimportant by a specific method. The higher the faithfulness, the more relevant the interpretability method. Another measure called *stability* (or *robustness*) assesses the explanation sensitivity to changes in features or model parameters. Finally, *fidelity* measures how closely an explanation reflects the model prediction. If these approaches

bring rich information about an interpretability method, they say nothing about the intelligibility of the resulting explanations to a human.

Human-grounded evaluation and experimental approaches can be alternatives to reach a rigorous science of interpretable machine learning [12]. Since local feature importance methods can have significant positive effect on human performance [31, 24], these methods can also be compared by evaluating how they make human more effective during a specific annotation task [21]. Such experimental protocols obtain contrasted outcomes, resulting in a more positive effect of XAI assistance on text than tabular data [24]. In the context of NLP, this type of protocol consists in asking humans to perform text annotation under the influence of local feature importance[25, 14]. The response time and the average accuracy are measured and compared between the different methods. This typology of experimental protocol has the advantage of quantifying the quality of an explanation, as long as the explanation is intended to support human decision.

Attention-based explanations in particular have been experimentally compared to LIME on a BERT model classifying genuine and deceptive hotel reviews [14]. In this work, attention is aggregated by averaging the attention coefficients over the whole 12 attention heads of the last BERT layer. This results in a higher human annotation performance with LIME compared to aggregated attention. The simplicity of this aggregation, however, does not enable one to conclude about the interest of using attention. Furthermore, attention-based explanations are not compared to a simple baseline, such as a random generator.

3 Methodology

We present an easy-to-implement way of aggregating attention from Transformer models that we call CLS-A. Then we define an experimental protocol inspired from the ones introduced in Section 2 that we apply on three different annotation tasks of binary classification. Finally we introduce the evaluation protocol with data description, linear and non linear modeling.

3.1 CLS-A

We introduce the way we aggregate Transformer self-attention to build a local feature importance metric. This approach can be used in every Transformer-like classifier such as BERT, as long as the *CLS* token is used to perform classification. We use the attention coefficients related to the *CLS* token. We call context the distribution of attention between an input word and the rest of the sequence. This way, CLS-A represents the average context of the *CLS* token.

We focus on the last layer of the BERT architecture. Figure 1 shows the global process of the CLS-A computation from an initial text. Since this last layer contains h self-attention heads, the coefficients are aggregated by averaging to build a one-dimensional local feature importance explanation. The proximity of the attention head within a specific layer [10] justifies our choice of aggregating

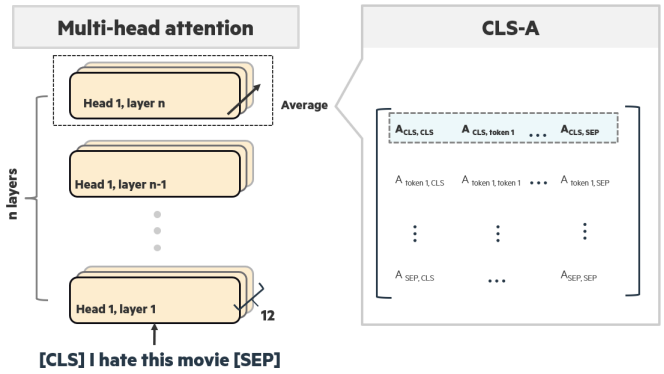

Fig. 1: Scheme of CLS-A. Average attention related to the *CLS* token is computed on the last attention layer.

through the whole layer. This results in an average context of the classification token in the last layer of BERT. A weight is assigned to each word of the input text, representing its importance in the context that induced the prediction of the classifier. The interest of this approach lies in the focus on the relationship between the CLS token and its context.

Since the *CLS* token plays a central role in the computation of CLS-A, it is recommended that the BERT forward pass passes by the *CLS* token to perform its prediction. Therefore, the prediction has to be done by computing the embedding from the *CLS* token only. In the case where the BERT forward pass does not pass exclusively through the *CLS* token, a less satisfying alternative is to compute the average of all the coefficients of the attention heads (see Section 2).

3.2 Experimental protocol

Motivated by the proven utility of experimental protocols to compare XAI methods[25, 24, 31], we ask participants to annotate one hundred texts in a binary classification task. Each text has some words colored with a more or less intense shade of blue (see in the screen in Figure 2), based on an underlying interpretability method or a random generator. The higher the coefficient of the method, the stronger the highlighted blue shade. Accuracy and response time are measured to evaluate each method's ability to assist the participant in the annotation task. The higher the accuracy and the shorter the response time, the more relevant the method as it facilitates the human semantic processing of the text.

Setup and instructions. All participants take part in the experiment in the same room and can be up to three at the same time. They are isolated in order to limit any other exogenous influence (visual, sound) and are placed in front of

Fig. 2: Scheme of the experimental protocol. Each participant labels 100 different texts after reading the instructions. The participant has two possible answers, depending on the experiment he/she is participating in. The text is colored according to the interpretability method used to explain the classifier's prediction. The selected texts are all classified properly by the classifier.

a computer as depicted in Figure 2. An explanation of the protocol is displayed on the screen to put the participants in the right conditions. In order to perform the annotation task, participants are asked to press either one of two buttons corresponding to the two possible answers as shown in Figure 2. The buttons correspond to keys on the keyboard of the computer used. Two colored stickers are stuck on the keys to help locate them. When a text is annotated (response given/key pressed), the next one is displayed on the screen.

Three classification tasks are evaluated. The first one (Experiment 1) is to evaluate the global sentiment of a movie review. The participant must choose between the "positive" and "negative" sentiment. The second and third (Experiment 2 and 3) are film genre evaluations. In Experiment 2, the task is to distinguish between action and drama films, in Experiment 3 between horror and comedy. The information given in the explanation of the protocol differs depending on the classification task. Participants annotating film genres are asked to respond quickly. Participants are also told that displayed colors can potentially be useful in the annotation task. Participants annotating movie review sentiments have no information about the colors displayed and no incentive to respond quickly.

Experiment characteristics. All the 100 participants have a background in data science or statistics. None of the participants labels data as a profession. The first experiment involved 50 participants while the other two had 25 each. The participants were predominantly male: about 76% compared to 24% for women. They were between 22 and 40 years old and none of the subjects were cognitively impaired to our knowledge.

Every participant is asked to annotate 100 different texts in a binary classification task. The asked classification task remains the same during the whole experiment. A participant cannot see the same text twice during the experiment. Each text has its words colored in different shades of blue. This coloring is proportional to the coefficients chosen at random among a random baseline, SHAP, LIME and CLS-A built on the DistilBERT classifier attention. The random baseline assigns randomly a coefficient to each word. This way, participants are subjected to exogenous attentional orientation effects in order to compare the methods one-to-one. We show in Appendix, Table 6 the balanced distribution of the methods used to color the plotted texts. An example of the text displayed during the experiment is plotted in Figure 2.

The classes of the various classification tasks are all equally represented among the displayed texts. We assess the contribution of interpretability methods under the assumption that the prediction of a model is correct. Therefore, the instances selected for the study were all correctly predicted. We wanted to look at the effect of the review length and the prediction probabilities in Experiments 2 and 3. The reviews corresponding to the sentiment analysis task contain between 32 and 50 words. The text sequence lengths of the other classification tasks vary between 19 and 145 words. The probability scores of belonging to the target class are highly polarized for the sentiment analysis and the horror/comedy classification while probability score is more uniformly distributed for the action/drama classification task. We assume that an interpretability method provides good explanations to the extent that it helps an annotator to go faster and be more efficient. An explanation will then be the object of a semantic congruence between the label to be predicted and the words highlighted. Therefore, the response time is precisely measured for each text and the correctness of the answers is assessed.

3.3 Implementation details

The three classifiers analyzed during the three annotation tasks were based on a DistilBERT[23]. Each pre-trained DistilBERT was retrieved from Hugging Face[4]. A dense layer was added to perform the classification and fine-tune each model. The forward pass was defined as getting the embedding of the CLS token to perform the classification task. The library used to compile and fine-tune the models were Keras on the TensorFlow framework. Each model was trained with an initial learning rate of 10^{-5} and a reducing learning strategy when reaching a plateau. The number of epochs was for each model of 5 and the batch size was 32. The models were trained with a binary crossentropy loss and the Adam optimizer. The first model was fine-tuned for sentiment analysis on the IMDB database[17]. The second and third one were fine-tuned to perform movie genre classification on a Kaggle dataset[5].

For each text, SHAP was computed with the `shap` library [16]. The Shapley values were computed in a permutation way. Finally, LIME was computed with

[4] www.huggingface.co/

[5] www.kaggle.com/competitions/movie-genre-classification/overview

the `lime` [22] library. The whole experiment was performed on the `psychopy` [20] framework on Python.

3.4 Data analyses

In this section we define the followed methodology to analyze the data produced by the experimental protocol presented above. Each experiment produces $n \times 100$ answers, with n the number of participants, and 100 the number of plotted text samples during each experiment. The indicators of interest are the labeling time, which we call "reaction time", and whether or not the participant is wrong, which we call "accuracy". These variables of interest are then analyzed through their relationship with other characteristics such as features about the text (sequence length, probability score, trial number, relative position of impacting word) and the interpretability method used to color it.

Data description. The descriptive analysis is first performed by calculating the average reaction time and the average accuracy. The one-tailed t-test is then used to compare the distributions of reaction times between interpretability methods and the random baseline in order to have statistically significant comparisons. This test is applied here to the average difference between each interpretability method and the random baseline, per participant, per experiment.

Linear modeling. The impact of interpretability methods on reaction time is estimated with a linear regression by incorporating the effect of other explanatory variables. The random baseline is used for reference. Thus, the coefficients of the linear regression associated with the method used to color the text are expressed with respect to this baseline. For each experiment, one linear model is built per participant to explain its reaction time to the labeling task. The explanatory variables of the models for an experiment are the same for all participants. The mean value of the regression parameters and their distribution are then analyzed with the one-tailed t-test presented above.

Non-linear modeling. Decision tree boosting algorithms enable to model complex and non-linear phenomena. We apply this type of algorithm to model the participant accuracy. This binary classification problem is addressed via Explainable Boosting Machine (EBM) [19]. EBM reaches performance levels equivalent to other boosting approaches based on decision trees, while decomposing its prediction into interpretable contributions of the explanatory variables. EBM is a generalized additive model (GAM) of the form:

$$g(\mathbb{E}[y]) = \beta_0 + \sum_i f_i(x_i) + \sum_{i,j,i \neq j} f_{i,j}(x_i, x_j) \tag{2}$$

Where:

Metrics	Experiment	CLS-A (mean ± std)	LIME (mean ± std)	SHAP (mean ± std)	Random (mean ± std)
Reaction Time (s)	Exp 1	**10.3 ± 4.2**	10.6 ± 4.6	10.5 ± 4.5	11.0 ± 3.9
	Exp 2	9.0 ± 3.7	**8.3 ± 3.5**	8.4 ± 4.1	8.6 ± 3.3
	Exp 3	**11.02 ± 6.8**	11.3 ± 6.8	11.8 ± 5.9	12.3 ± 7.5
Accuracy (%)	Exp 1	**97.0 ± 5.8**	96.3 ± 3.9	95.5 ± 5.3	95.0 ± 5.5
	Exp 2	**80.6 ± 9.9**	79.4 ± 12.4	79.9 ± 9.1	79.8 ± 10.1
	Exp 3	**86.1 ± 9.5**	85.4 ± 9.3	81.3 ± 12.8	84.1 ± 8.9

Table 1: Average reaction time and accuracy per experiment per method. The numbers in bold correspond to the best performance.

- y is the variable indicating whether a participant has successfully completed its labeling
- g is the link function
- β_0 is the intercept
- f_i is the feature function of the variable x_i,
- $f_{i,j}$ is the pairwise interaction function of the two variables x_i and x_j

A response curve represents the effect of a given explanatory variable by plotting the evolution of its contribution to the target variable. One model is fitted per method and per experiment to compare the response curves of the methods within a given experiment. Each model has to be trained with the same explanatory variables. Since the participants generally perform their annotation tasks accurately, the data are largely imbalanced. Sub-sampling is therefore performed to run the EBM on a balanced dataset with a balanced distribution between right and wrong answers. Since this sub-sampling induces sampling bias, this operation is run 50 times. Average response curves and standard deviations are finally calculated.

4 Results

In this section we compare CLS-A to LIME, SHAP and a random baseline following the methodology introduced in Section 3. We show that CLS-A improves both speed and accuracy of annotation in a statistically significant way compared to the random baseline. CLS-A, SHAP and LIME result in statistically similar response times and accuracy. Furthermore, we highlight the relationship between the quality of an explanation and the certainty of the classifier's prediction.

4.1 Data description

The first experiment consisted of responses from 50 participants while the other two had 25 each. Table 1 relates the average reaction time and the average accuracy per experiment and per interpretability method. This shows that the average reaction time related to CLS-A is lower for experiments 1 and 3. Accuracy

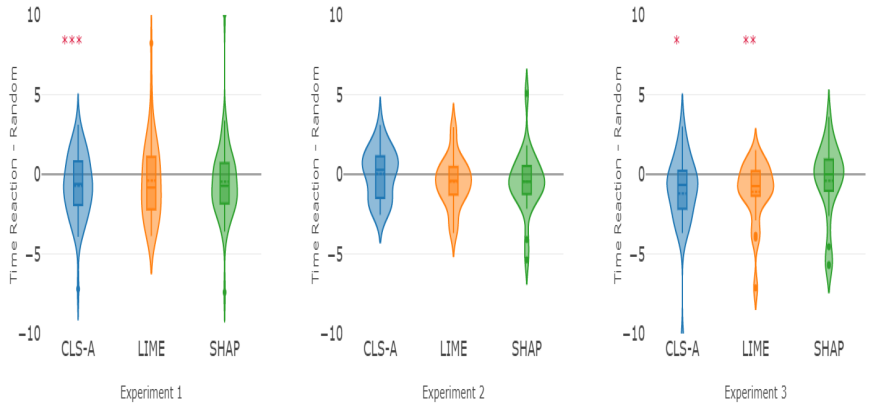

Fig. 3: Distribution of mean reaction time difference from random baseline by participant, by experiment. The results of the one-tailed t-test are represented with stars above the violin plots. With p as the p-value of the t-test, $^{*}p < 5\%$, $^{**}p < 1\%$, and $^{***}p < 0.5\%$

is also on average higher for participants who were exposed to CLS-A. The random baseline induces less accurate and slower responses overall.

We compare the distributions of the average response time of the CLS-A, LIME and SHAP methods in comparison to the random baseline. We perform this distribution comparison using the one-tailed t-test on the average difference between the XAI method and the baseline, per participant as presented in Section 3. Figure 3 plots the distribution of the average reaction time deviation from the random baseline with the results of the t-tests, by method and by experiment. The mean difference in reaction time between CLS-A and the random baseline is statistically significant in the first and third experiments. This difference is also statistically significant between LIME and the baseline in the third experiment.

Therefore, participants went faster on average in the text annotation task in Experiments 1 and 3 when they were exposed to CLS-A compared to the baseline. This difference from baseline is exclusive to CLS-A for Experiment 1, and shared with LIME for Experiment 3.

4.2 Linear modeling.

The target variable is reaction time and the explanatory variables (see Table 3 in Appendix A) are information about the assessed text on one hand and the interpretability method used to color it on the other. The performance metrics of the linear regressions on reaction times are presented in Table 2 Appendix A.

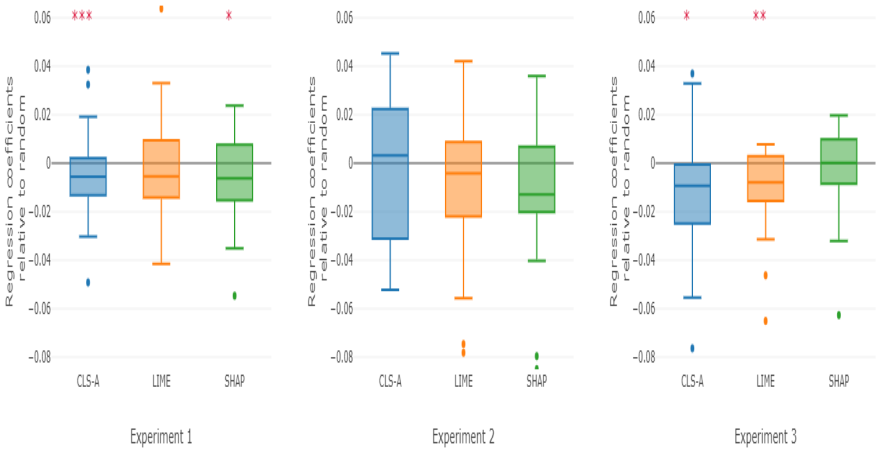

Fig. 4: Distribution of linear modeling on reaction times coefficients of each interpretability method variable with respect to the baseline. The results of the one-tailed t-test are represented with stars above the box plots. With p as the p-value of the one-tailed t-test, $^*p < 5\%$, $^{**}p < 1\%$, and $^{***}p < 0.5\%$

Method effect. Figure 4 shows the distributions of coefficients of linear regressions on reaction times associated with the interpretability method used to color the text. All coefficients are computed with respect to the random baseline.

The average linear regression coefficients on reaction times of CLS-A and LIME are significantly negative for Experiment 1 and 3. Average reaction time coefficient is negative for SHAP in Experiment 1. The results for the CLS-A method are broadly consistent with the previous exploratory analysis. Participants took less time on average to complete their annotation task on Experiment 1 and 3 when important words in the text were colored via the CLS-A method compared to the random baseline. However, the results differ for SHAP and LIME, and Experiment 2 does not show a statistically significant difference between these 3 methods compared to the random baseline.

Probability score and review length effects. We similarly assess the distributions of two more features in the linear regression model on reaction times, namely the probability of belonging to the target class, and the length of the review. The Figure 5 represents the distribution of these coefficients, by experiment. The significance of the means of the distributions is assessed with a one-tailed paired t-test.

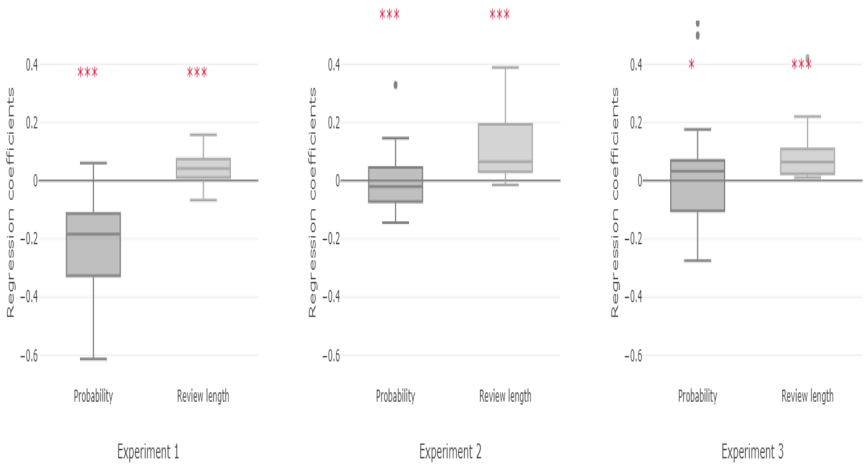

Fig. 5: Distribution of probability score and review length coefficients in linear modeling. The results of the one-tailed t-test are represented with stars above the box plots. Noting p as the p-value of the statistical test, the notations are as follows. $*p < 5\%$, $**p < 1\%$, and $***p < 0.5\%$

The sign of each of the two coefficients is consistent across all three experiments. The effect of the probability score variable on response time is negative on average, whereas the effect of the sequence length is positive. These impacts are statistically significant across all three experiments. Thus, all things being equal, the higher the probability score of belonging to the target class, the lower the reaction time. This highlights the relationship between the quality of an explanation and the certainty of a prediction from a time reaction perspective. In the same way, all things being equal, the annotation time increases with the length of the textual sequence processed.

To summarize, linear modeling highlights that CLS-A fosters quicker responses on average compared to the random baseline. Besides, the higher the prediction certainty, the lower the human reaction time.

4.3 Non-linear modeling

This section compares CLS-A and the random baseline through the prism of the participant's accuracy, which is modelled using non-linear Explainable Boosting Machines (EBM) introduced in Section 3. The explanatory variables used to explain participant's response to the experiment are presented in Appendix A, Table 3.

In this section we focus on the impact of the probability score and the reaction time on the annotation accuracy. Figure 6, 7 and 8 represent the EBM response curves of the probability score and the reaction time, by method, by experiment. These curves represent the contributions to the probability scores that the participant performs the correctly the annotation. The interval around the mean curve represented the standard deviation measured on 50 sampling iterations for a given model. For each of the analyzed variables, we focus on the comparison between CLS-A and the random baseline.

Sentiment analysis. The first experiment emphasizes a higher target class probability score contribution for CLS-A compared to the random baseline in Figure 6. The response curves tend to merge for the polarized probability scores and CLS-A falls below the baseline for the probability score distribution tail. Fast reactions induce higher accuracy contribution for CLS-A. Accuracy contribution tend to be the same for very long response times. Therefore, the interest of CLS-A compared to the random generator lies in the relatively low probabilities in the first experiment. Note however that the probability score distribution is very high in the first experiment, and covers very few non-polarized predictions. Moreover, the contribution of CLS-A is significant for fast predictions, and tends to vanish gradually.

Movie genre classification, action vs. drama. The second experiment has a more dispersed distribution of target class probability scores than the first experiment. The CLS-A response curve associated with the target class probability score variable is higher for polarized predictions as shown in Figure 7. Then, the contribution of CLS-A compared to the baseline seems to be related to the certainty of the classifier prediction. The area in which the CLS-A response curve is higher corresponds to the majority of the probability score distribution of the target variable. Finally, the accuracy contributions of the reaction time variables are higher for CLS-A for short and very long responses. Finally and similarly to the first experiment, CLS-A has a strong impact to form rapid responses when labeling a high target class probability score text.

Movie genre classification, horror vs. comedy. The distribution of the target class probability score variable is less dispersed in the last experiment than in the second one. Figure 8 depicts a higher CLS-A response curve for high probability scores and falls below it at the distribution tail. Finally, the effect of CLS-A on the response time variable with respect to the baseline in experiment 3 is relatively similar to Experiment 2. Short and very long answers are more accurate with CLS-A compared to the random baseline.

To summarize, the analysis of the response curves highlights the non-linear relationships between the explanatory variables and the target variable. The interest of CLS-A is strong for high certainty prediction and less important or even non-existent for texts whose probability scores are low. This highlights a

strong relationship between the quality of an explanation and the certainty of a prediction. Additional analysis in Appendix A Figure 9 , 10 & 11 and Table 5 generalize this link to SHAP and LIME.

5 Discussion and Conclusion

We applied an experimental protocol to compare a local feature importance method called CLS-A based on Transformer self-attention to SHAP, LIME and a random baseline. We found that CLS-A helps in the same proportions as SHAP and LIME to annotate text on three different tasks and is significantly better than a random baseline. This work adds to the literature aiming to evaluate the interpretability of attention coefficients in recent deep learning models. Attention is appropriate to explain attention-based classifiers in NLP when aggregated in the proper way. Like other XAI methods, however, the relevance of the explanations provided by CLS-A depends heavily on the certainty of their related prediction. The higher the probability score, the more relevant the explanation. As far as we know, this is the first time that the relationship between the quality of an explanation and the certainty of its associated prediction has come to light.

We believe that additional experimental studies analyzing texts with more distributed probability score would be enlightening. The link between prediction certainty and explanation relevance would be more precisely outlined. The results of our study must be evaluated considering that all the participants had a data science or statistics background. This may induce a bias in our results, insofar as the participants have an occupation requiring advanced analytical skills. Finally, other usual model-specific XAI methods comparison could be added in such an experiment.

Ethic statement

Each participant signed an informed consent form containing the project purpose and details and the intended use of the data they would generate. The data was anonymized and processed only by our team. The data produced is stored in a file in respect with the General Data Protection Regulation (GDPR) regulations in force. Participation in the study was fully voluntary. It was possible to stop performing the labeling tasks at any time. Consent form used is presented anonymized in Appendix A , Figure 12 & 13. The authors of this paper do not represent any organization or institution whose activity is data labeling. This study was conducted for research purposes only.

Acknowledgment

We thank Gariel Olympie for insights and methodology and Jean-Baptiste Gette for English proof reading.

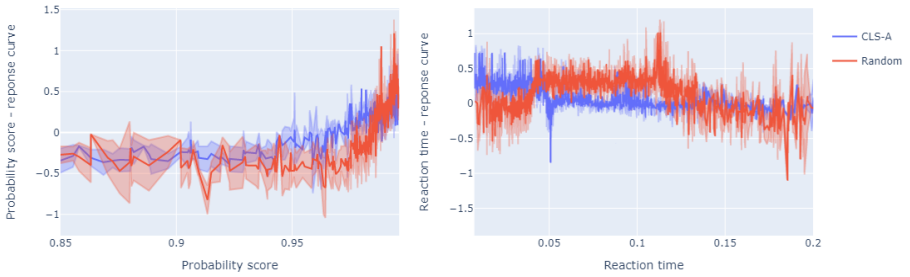

Fig. 6: Explainable boosting machine response curves of probability score and time reaction in the first experiment. The contribution of the probability score variable becomes positive at a lower probability threshold for CLS-A compared to the random generator. The contributions of the reaction time variables are positive for the fast reactions for the CLS-A method unlike the random generator.

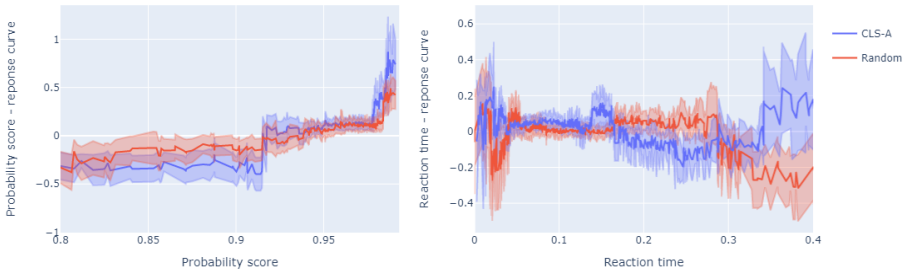

Fig. 7: Explainable boosting machine response curves of probability score and time reaction in the second experiment. The contribution of the probability score increases at a faster rate than the random generator. CLS-A favors more fast reactions.

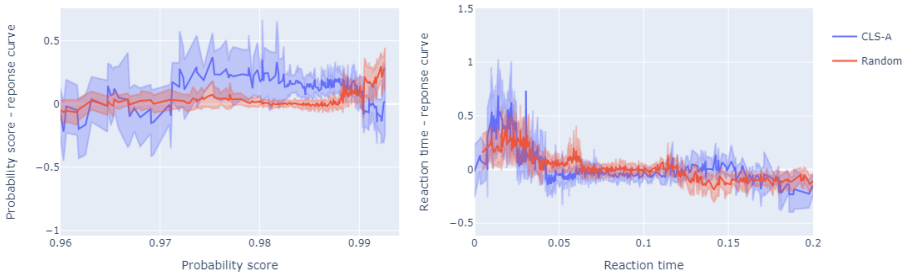

Fig. 8: Explainable boosting machine response curves of probability score and time reaction in the third experiment. The contribution of the probability score variable increases at a faster rate than the random generator and decreases for very high prediction probability score unlike the random generator. The contributions are slightly higher for fast response times for CLS-A.

Bibliography

[1] Abnar, S., Zuidema, W.H.: Quantifying attention flow in transformers. CoRR **abs/2005.00928** (2020), https://arxiv.org/abs/2005.00928

[2] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: Openxai: Towards a transparent evaluation of model explanations (2023)

[3] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014)

[4] Bastings, J., Filippova, K.: 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. pp. 149–155. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.blackboxnlp-1.14, https://aclanthology.org/2020.blackboxnlp-1.14

[5] Bell, A., Solano-Kamaiko, I., Nov, O., Stoyanovich, J.: It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. p. 248–266. FAccT '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3531146.3533090, https://doi.org/10.1145/3531146.3533090

[6] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P.: Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 648–657. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3375624, https://doi.org/10.1145/3351095.3375624

[7] Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., Watrin, P.: Is attention explanation? an introduction to the debate. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3889–3900. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.269, https://aclanthology.org/2022.acl-long.269

[8] Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8) (2019), https://www.mdpi.com/2079-9292/8/8/832

[9] Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. CoRR **abs/2012.09838** (2020), https://arxiv.org/abs/2012.09838

[10] Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does bert look at? an analysis of bert's attention. In: "Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP". pp. 276–286. Association for Computational Linguis-

tics, Florence, Italy (aug 2019). https://doi.org/10.18653/v1/W19-4828, https://aclanthology.org/W19-4828

[11] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Tech. Rep. arXiv:1810.04805, arXiv (May 2019), http://arxiv.org/abs/1810.04805, arXiv:1810.04805 [cs] type: article

[12] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017). https://doi.org/10.48550/ARXIV.1702.08608, https://arxiv.org/abs/1702.08608

[13] Farah, L., Murris, J.M., Borget, I., Guilloux, A., Martelli, N.M., Katsahian, S.I.: Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: What healthcare stakeholders need to know. Mayo Clinic Proceedings: Digital Health **1**(2), 120–138 (2023)

[14] Lai, V., Liu, H., Tan, C.: "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 1–13. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376873, https://doi.org/10.1145/3313831.3376873

[15] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)

[16] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[17] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://aclanthology.org/P11-1015

[18] Molnar, C.: Interpretable Machine Learning. Lulu.com (2020), google-Books-ID: jBm3DwAAQBAJ

[19] Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)

[20] Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K.: Psychopy2: Experiments in behavior made easy. Behavior research methods **51**(1), 195–203 (2019)

[21] Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI conference on human factors in computing systems. pp. 1–52 (2021)

[22] Ribeiro, M.T., Singh, S., Guestrin, C.: "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. p. 1135–1144. KDD '16, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778

[23] Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019)

[24] Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., Vössing, M.: A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. p. 617–626. AIES '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3514094.3534128, https://doi.org/10.1145/3514094.3534128

[25] Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. (2019)

[26] Shapley Ll, S.: A value for n-person games. Contributions to the Theory of Games II, Annals of Mathematical Studies **28** (1953)

[27] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 3145–3153. ICML'17, JMLR.org (2017)

[28] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 3319–3328. ICML'17, JMLR.org (2017)

[29] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[30] Vig, J.: A multiscale visualization of attention in the transformer model. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 37–42. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-3007, https://www.aclweb.org/anthology/P19-3007

[31] Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In: 26th International Conference on Intelligent User Interfaces. p. 318–328. IUI '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3397481.3450650, https://doi.org/10.1145/3397481.3450650

[32] Weber, P., Carl, K.V., Hinz, O.: Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. Management Review Quarterly pp. 1–41 (2023)

A Appendix

Metrics	Experiment	R^2
Reaction Time (s)	Exp 1	4.23e-1
	Exp 2	6.01e-1
	Exp 3	6.59e-1

Table 2: R-square of linear regression explaining participant reaction time

Task	**Target variable**	**Model**	**Explanatory variables**
Regression	Reaction time	Linear model	expected answer, probability score, review length, trial number, interpretability method, relative positions of 1^{st}, 2^{nd} and 3^{rd} most impacting words
Classification	Accurate	Explainable Boosting Machine	reaction time, probability score, review length, trial number, interpretability method, relative position of 1^{st} most impacting word

Table 3: Linear regression and explainable boosting machine explanatory variables. The variables of the relative positions of the second and third most important words were used only for reaction time modeling in the first .

Experiment	Method	Accuracy	Precision	F1-score	Recall
Exp 1	CLS-A	0.952	0.945	0.953	0.961
	LIME	0.992	0.991	0.992	0.993
	SHAP	0.961	0.955	0.961	0.976
	Random	0.982	0.988	0.982	0.976
Exp 2	CLS-A	0.889	0.894	0.888	0.884
	LIME	0.897	0.888	0.898	0.909
	SHAP	0.913	0.917	0.913	0.909
	Random	0.878	0.867	0.880	0.894
Exp 3	CLS-A	0.957	0.950	0.957	0.965
	LIME	0.946	0.943	0.946	0.949
	SHAP	0.920	0.916	0.920	0.925
	Random	0.920	0.921	0.920	0.919

Table 4: Average EBM performance per experiment, per method.

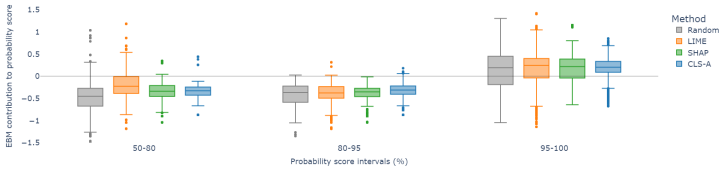

Fig. 9: Discretized EBM contributions of probability score in Experiment 1. High certainty prediction with probability higher than 95% have higher contributions.

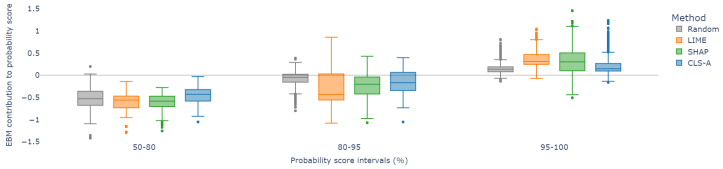

Fig. 10: Discretized EBM contributions of probability score in Experiment 2. High certainty prediction with probability higher than 95% have higher contributions.

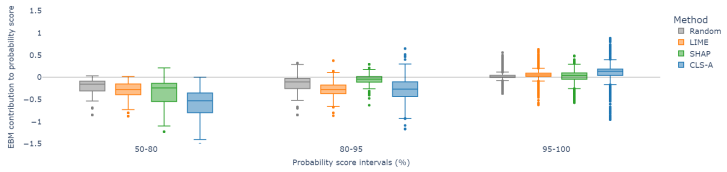

Fig. 11: Discretized EBM contributions of probability score in Experiment 3. High certainty prediction with probability higher than 95% have higher contributions.

Probability score interval	**Method**	**Average EBM contribution to probability score**		
		Experiment 1	**Experiment 2**	**Experiment 3**
50-80%	CLS-A	-0.34	**-0.46**	-0.59
	LIME	**-0.17**	-0.61	-0.30
	SHAP	-0.34	-0.60	-0.35
	RANDOM	-0.48	-0.53	**-0.21**
80-95%	CLS-A	**-0.32**	-0.16	-0.30
	LIME	-0.41	-0.30	-0.29
	SHAP	-0.38	-0.29	**-0.05**
	RANDOM	-0.40	**-0.08**	-0.15
95-100%	CLS-A	**0.19**	0.22	**0.10**
	LIME	0.18	**0.38**	0.05
	SHAP	0.17	0.31	0.02
	RANDOM	0.14	0.16	0.03

Table 5: Average EBM contribution to probability score. High certainty predictions lead to higher contribution. Highest average contributions per probability interval are highlighted in bold.

Experiments	CLS-A (mean ± std)	LIME (mean ± std)	SHAP (mean ± std)	RANDOM (mean ± std)
Exp 1	25.5 ± 3.8	25.5 ± 3.6	24.5 ± 3.9	24.5 ± 4.2
Exp 2	25.6 ± 4.2	25.0 ± 4.1	24.0 ± 4.8	25.4 ± 4.1
Exp 3	23.0 ± 4.9	24.3 ± 5.6	27.4 ± 5.4	25.3 ± 5.9

Table 6: Method distribution per participant per experiment.

INFORMED CONSENT FORM

Project Title : Evaluating self-attention interpretability through human-grounded experimental protocol

Research team : Omitted for anonymity

Research location : Omitted for anonymity

Project Presentation

Machine learning models for classification tasks on text are often black boxes. Today there are different methods to evaluate the factors (words) that were important for the decision of the algorithm. However the validity of these methods and their link with human semantics are not studied. The objective of this project is to establish the congruence between the results of different interpretability methods and human semantic analysis.

If you agree to participate in this study, we will ask you to read movie reviews and rate the category of the movie by pressing a key on the keyboard. The approximate duration of the experience is about fifteen minutes.

Your privacy rights

All the information collected during this experiment for the pursuit of the purposes set out in the previous paragraph will be processed by Omitted for anonymity, anonymously and will remain confidential. The legal basis for processing is your consent.

These will be kept in a computer file that complies with the applicable regulations in force (General Data Protection Regulations and Data Protection Act).

The data collected will be communicated only to the following recipients from the research team:

- Omitted for anonymity

The results obtained from the processing of this questionnaire may be the subject of scientific publications, but the identity of the participants will not be revealed, and no information that could reveal your identity will be disclosed.

The data is kept until the publication of an article or a maximum of 3 years.

Your rights to withdraw from this research at any time

Participation in this study is completely voluntary. Please note that even if you decide to complete this questionnaire, it is possible to stop completing it at any time, and as long as the final registration has not been made, none of your data will be processed.

You can access the data concerning you, rectify it, request its deletion or exercise your right to limit the processing of your data. You can withdraw your consent to the processing of your data at any time; you can also object to the processing of your data. Visit the cnil.fr website for more information on your rights.

To exercise these rights, you can contact Omitted for anonymity

If you believe, after contacting us, that your "Data Protection" rights are not respected, you can file a complaint with the CNIL.

Fig. 12: Consent form (1/2)

Diffusion

The results of this research may be published in scientific journals or be the subject of communications at scientific conferences.

You can ask questions about the research at any time by contacting Omitted for anonymity

Consent to participate

By checking the box below and **signing this consent form**, you certify that you have read and understood the above information and that you have been informed of your right to withdraw your consent or withdraw from this research at any time, without prejudice.

☐ ***I have read and understood the above information and I voluntarily agree to participate in this research.***

Done at : _____

On the : _____

Name, First Name : _____

Signature :

Fig. 13: Consent form (2/2)