Long Paper

Explanation Variability in Text Classification: Humans vs. LLMs

Jérémie Bogaert*[1], Louis Escouflaire**[2], Antonin Descampe[2], Cédrick Fairon[2], Marie-Catherine de Marneffe[2], François-Xavier Standaert[1]

[1] ICTEAM, UCLouvain
firstname.lastname@uclouvain.be

[2] ILC, UCLouvain
firstname.lastname@uclouvain.be

This paper examines the variability of human and large language model (LLM) token-level explanations applied to the use case of French journalistic text classification between news and opinion articles. To do so, we compare human annotations (in the form of highlighted tokens indicating perceived decision-relevant words) with Layer-wise Relevance Propagation (LRP) explanations from fine-tuned transformer models. Using ten texts annotated by multiple readers and classified by equivalent instantiations of a LLM, we analyze qualitative patterns and quantitative trends with a similarity score. Results show that humans highlight fewer, longer spans focused on linguistically salient cues while LRP produces more diffuse, token-level attributions. Humans also tend to agree more on the most important tokens, whereas models align better when considering all tokens, reflecting divergent sensitivities to granularity. Prediction class matters as well: humans are more consistent on opinion texts, while models show greater stability on news. To refine our variability analysis, we apply discretization schemes aligning LRP values with categorical human judgments. Both linear and human-aligned discretization increase similarity with human explanations and thus improve visual plausibility in aggregated attention maps without altering model predictions. These findings suggest that model explanations are not systematically more variable than human ones but follow different dynamics depending on representation and scope. They also highlight the Rashomon effect in LLM explainability, showing that agreement on outputs does not imply convergence in reasoning. Our work demonstrates how methodological choices shape explanation variability and offers practical insights for bridging faithfulness and plausibility in explainable NLP.

* Equal contribution
** Equal contribution and Corresponding author

Action editor: {Miguel Ballesteros}. Submission received: 4 November 2025; revised version received: 21 January 2026; accepted for publication: 12 March 2026.

1. Introduction

Transformer-based language models such as BERT (Devlin et al. 2019) or GPT-4 (Achiam et al. 2023) have significantly advanced the state of the art in many natural language processing (NLP) tasks. However, despite their efficiency, these models remain largely opaque, raising concerns around the transparency and interpretability of their decisions (Lipton 2018). A growing body of research has proposed *explainability* methods, such as model perturbation (Cavalcanti et al. 2024) and layer-wise relevance propagation (LRP) (Chefer, Gur, and Wolf 2021), to shed light on model behavior. These methods produce model explanations, which are representations of which parts of the input most influenced a given prediction. In NLP, these explanations often take the form of token-level importance maps, sometimes visualized as highlighted text indicating the most influential words or phrases (Lyu, Apidianaki, and Callison-Burch 2024). Such visualizations allow both researchers and lay users to qualitatively inspect what the model "attends to" when making a decision.

While these methods can provide useful insights, they also raise questions about the reliability of the explanations they produce. More specifically, it has been observed that in the presence of randomness in the training process, a large set of models with similar performances may arise from training (Breiman 2001). This phenomenon, usually called the Rashomon effect in social sciences, has recently gained interest in the explainable machine learning community (Müller et al. 2023). In particular, the work of Bogaert et al. (2023) showed that language models belonging to the same Rashomon set (also referred to as equivalent models, i.e. models whose differences in efficiency for a given task are not statistically significant) may explain their decisions using different features, raising questions about their agreement or variability. Watson, Hasan, and Moubayed (2022) showed that similar effects can be observed in the context of image classification. This paper focuses on this type of variability. We investigate whether explanations extracted from large language models (LLMs) are more or less *stable* than explanations provided by humans and if they share similar trends. Most prior work has evaluated the quality of these explanations by comparing single explanations from humans and models on the same input. For example, Sen et al. (2020) compared transformer-derived token attention maps with so-called "human attention maps" of the same predicted texts, measuring the degree of overlap between the sets of highlighted tokens. While informative on the textual information captured by humans and models, such approaches overlook an important factor: the *variability* of explanations across multiple annotators or across multiple functionally equivalent models producing explanations for the same prediction of the same text.

We also aim to address broader questions about explainability in complex models. Specifically, we explore whether the decisions of highly complex models such as BERT-like LLMs can be explained simply enough to be interpretable by humans, and whether the variability observed in the explanations of equivalent models' decisions is similar to the inter-annotator disagreement. In contrast to ensemble explanation approaches such as the one proposed by Krishna et al. (2022), which aggregate explanations from different methods and models, we do not attempt to construct consensus maps. Instead, we treat the raw distribution of explanations (across both multiple humans and models) as the central object of study. This allows us to directly measure and compare the intrinsic variability of explanations in each group.

Two key criteria are often discussed in the explainability literature: *plausibility* (how convincing an explanation is to a human observer) and *faithfulness* (how accurately the explanation reflects the actual decision-making process of the model) (Jacovi and Goldberg 2020). We restrict our analysis to token-level post-hoc model explanations obtained via Layerwise Relevance Propagation (LRP)(Chefer, Gur, and Wolf 2021), a deterministic explanation method for a given model instance, which we consider, despite not being perfect, to be sufficiently faithful (Lyu, Apidianaki, and Callison-Burch 2024). At the same time, we investigate to what extent such explanations can also be judged as plausible. Importantly, we deliberately use this deterministic explanation method and do not use generative explanations (such as chain-of-thought reasoning) because these are non-deterministic by nature: the same model may generate different verbal justifications for identical predictions (Turpin et al. 2023). Our focus, instead, is on measurable stability, a property that cannot be properly assessed in generative outputs.

After presenting the state-of-the-art on variability in transformer and human explanations of text classification tasks, we present the data and methods used for our case experiment, which consists in classifying journalistic excerpts as belonging to the *news* or *opinion* genre (henceforth referred to as 'classes' in the context of the text classification task). While this experiment was carried out with French texts, the method presented in the paper is language-agnostic.

Our contributions are threefold: (1) we introduce a systematic framework for quantifying the variability of explanations within and across groups of equivalent models and human annotators; (2) we provide empirical evidence that the stability of model explanations differs markedly from that of human explanations, and that these differences vary across text classes and explanation formats; (3) we show that simple post-processing strategies, inspired by human annotation patterns, can substantially reduce the gap in stability between model and human explanations.

The rest of the paper is structured as follows. Section 2 summarizes prior works in related fields. Section 3 presents methods and data. Results are detailed in section 4, then discussed in section 5. Attention map visualizations of the text examples are also used throughout the paper.

2. Prior Works

2.1 LLM Explainability and Sensitivity to Randomness

Large language models (LLMs), particularly transformer-based architectures, have achieved impressive performance across a range of natural language processing tasks, such as text classification (Acheampong, Nunoo-Mensah, and Chen 2021). However, crucial concerns are still raised about the explainability of the predictions made by transformer models. Two widely discussed criteria in the explainability literature are plausibility (how convincing an explanation is to a human observer) and faithfulness (how accurately the explanation reflects the actual decision-making process of the model) (Jacovi and Goldberg 2020; Agarwal, Tanneru, and Lakkaraju 2024). Many proposed explanation methods for LLMs, such as input perturbation or attention probing, fall short of at least one of these criteria (Lyu, Apidianaki, and Callison-Burch

2024).

Beyond these criteria, several studies highlight the notion of sensitivity as an additional dimension for evaluating explanation quality (Adebayo et al. 2018; Yeh et al. 2019; Manna and Sett 2024). In a broad sense, sensitivity captures how explanations respond to variations in the input or in the modeling process, and is often used as a criterion for their reliability. In this perspective, explanations should be sensitive to variations that affect the model's prediction, and invariant to those that do not.

In this work, we focus on a specific case of sensitivity: the sensitivity of explanations to randomness in the training process, such as changes in hyperparameters like learning rate, batch size, or even random seed. As suggested in Bogaert et al. (2023)[1], even when models remain functionally equivalent in terms of their predictions, their explanations of these predictions can diverge significantly under such variations. However, one should remain cautious: a given prediction may admit multiple plausible explanations, but explanations that vary in an essentially arbitrary or uniformly distributed way provide little insight into the model's actual reasoning.

2.2 Variability in Human Annotations

Many experiments in formal linguistics or NLP rely on human-annotated texts that serve as training data or as a gold standard for evaluating models. Tasks such as sentiment analysis or text classification often depend on labeled datasets, manually annotated by human participants or experts (Wankhade, Rao, and Kulkarni 2022). However, human annotation of linguistic data presents several challenges (Mieleszczenko-Kowszewicz et al. 2023).

Crowd-sourced data may suffer from inconsistent quality due to annotators' varying levels of expertise or dedication to the task (Basile et al. 2021). Moreover, it is crucial to remain aware that the representations, knowledge, and subjectivity of annotators always exert an influence on the results of their annotations (Sap et al. 2021), which may have an impact on the reliability of annotated datasets. Even expert annotators frequently disagree, especially on tasks involving subjective or ambiguous categories (Jiang and de Marneffe 2022). This phenomenon of disagreement is often discussed in terms of inter-annotator agreement (IAA), where low IAA can limit the reliability of the annotated data. More recently, this variability in labels has been formally conceptualized as Human Label Variation, a term popularized by Barbara Plank (Plank 2022). This reframing highlights that the variability of labels assigned by different annotators to the same item is not merely "noise" or a failure of annotation, but rather a valuable piece of linguistic information. Human Label Variation may reflect the inherent nuance or complexity of a task (Wiebe et al. 2004; Weber-Genzel et al. 2024), or genuine differences in human interpretation that models should potentially be designed to capture.

2.3 Opinion vs. News Classification

The journalistic world traditionally separates their production in two main classes: *opinion* articles, such as columns and editorials, which are overtly subjective,

1 This paper was translated into English in Bogaert et al. (2024)

and *news*, which includes all press content intended to follow the standards of journalistic objectivity (Schudson 2001; Esser and Umbricht 2014). In today's media ecosystem, where information is disseminated via the (mostly) uncontrolled and unprofessionalized channels of digital social networks and where users are placing less and less trust in traditional media (Latkin et al. 2023; Newman et al. 2024), it is important to be able to clearly distinguish between *opinion* and *news* content. To this end, automated text classification can improve citizens' ability to navigate online information, improving their understanding of societal issues, and helping them form their own opinions.

In the NLP field, classifying press articles according to their genre is a popular automation task, in English as well as in less-resourced languages such as French (Katari and Myneni 2020; Vernier, Monceaux, and Daille 2009; Todirascu 2019). It can be accomplished to some degree with traditional feature-based or dictionary-based bag-of-words approaches (Krüger et al. 2017), or through more complex dynamic systems also capturing discursive information such as argumentation or figurative language (Benamara, Taboada, and Mathieu 2017). Studies have demonstrated that transformer models such as BERT or RoBERTa (Devlin et al. 2019; Liu et al. 2019) are able to predict text genre, in particular distinguishing between journalistic genres, with higher accuracy than previous state-of-the-art NLP models (Alhindi, Muresan, and Preoţiuc-Pietro 2020; Singh, Chun, and Atluri 2020).

3. Methods

3.1 Data

Two different datasets are described in this section: the large dataset using as training corpus for fine-tuning the text classification models, and the small dataset which was both annotated by human participants and used as a test set for the models.

For the transformer-based classification experiment, the corpus used for fine-tuning the models is a subset of the corpus of 80,000 texts introduced by Escouflaire et al. (2024). This corpus contains 8,000 press articles written in French, published by the four Belgian media mentioned above, balanced between the *opinion* and *news* classes. In total, the fine-tuning corpus contains 5.3 million tokens. The preprocessing steps follow the steps described in Bogaert et al. (2023), which include dynamic truncation of the text before the 512-token limit of the model.

The sample of texts analyzed in this paper contains 60 texts selected from the test set (1,000 texts) of the Belgian journalistic texts corpus introduced by Escouflaire et al. (2024). All texts are excerpts consisting of the first paragraph or first lines of press articles written in French, published between 2014 and 2024 by four different Belgian media: public service news website *RTBF Actus*, regional daily newspaper *L'Avenir*, and national daily newspapers *Le Soir* and *La Libre*. The texts contain between 69 and 129 tokens and were selected following several criteria: 30 are excerpts of texts considered by the media in which they were published as belonging to the *opinion* class, 30 to the *news* class. Of those 60 excerpts, 40 were predicted by the linguistic feature-based classification model (logistic regression) used in Bogaert et al. (2023) as very close to their corresponding class (regression score between 0 and 0.2 and between 0.8 and 1), while 20 were not as clearly classified by the model (regression score between 0.3 and

0.7).

All code and data is available on Github.[2]

3.2 Human Annotations

We collected manual annotations of the texts through an experiment involving 42 human annotators, all native French speakers, who were tasked with reading journalistic texts, classifying them as either *news* or *opinion* and explaining their decisions by highlighting relevant segments of the text. Of the 42 participants, 26 are master's students of journalism in Belgium, aged between 21 and 24 (except two students aged 29 and 43). All students willingly participated in the experiment, in the context of a course on information literacy. They were given the choice between participating in this experiment or writing a short essay on journalistic objectivity. Respect of deadlines and serious participation in the experiment were the only explicit criteria for the evaluation of the course. In addition to the 26 students, 16 additional annotators were recruited through social networks. They are 27 years old on average, occupy a variety of professions (e.g., lawyers, IT specialists, researchers), and have also agreed to take part in our experiment willingly. All 42 participants signed a consent form for their participation in the experiment. The design of the experiment was approved by the ethical board of the Institute for Language & Communication (UCLouvain).

Annotations were collected using a custom made platform inspired by PADDLe (Pirali, François, and Gala 2022). The platform was designed to enable participants to create an account, log in and out, and annotate excerpts at their own pace. For each excerpt, participants completed three annotation steps, similarly to what was done by Escouflaire, Descampe, and Fairon (2024): (1) reading the excerpt and selecting the journalistic genre it belongs to (*opinion* or *news*); (2) indicating their level of confidence in this choice on a scale from 1 (*not confident at all*) to 5 (*absolutely confident*); and (3) highlighting the textual elements that led them to choose one class over the other. In the third step, annotators could use three colors to indicate the importance of each highlighted element: yellow (*slightly important*), orange (*important*), and red (*very important*). Additionally, an optional free-text field allowed annotators to provide further comments or considerations. Figure 1 provides an example annotation made by a student for a given text.[3]

The first step of the annotation campaign, involving the 26 student participants, lasted four weeks (from October 21st to November 17th, 2024). Each Monday, students

2 https://github.com/louisescouflaire/HumanSvsLLMS

3 English translation of the annotation interface in Figure 1:
Read the excerpt below carefully.
[Excerpt]
According to you, is this excerpt from an opinion article or a news article?
How confident are you with your classification? [Not confident at all 1 2 3 4 5 Absolutely confident]
Highlight in the excerpt below the elements that led you to choose the opinion category. Choose the highlighting color according to the importance of the element in your choice.
Keys: Slightly important (yellow), Important (orange), Very important (red), Erase selection (white), Cancel all highlights.
Would you like to explain your decision (opinion or news) in another way? Or do you have additional considerations about this excerpt?

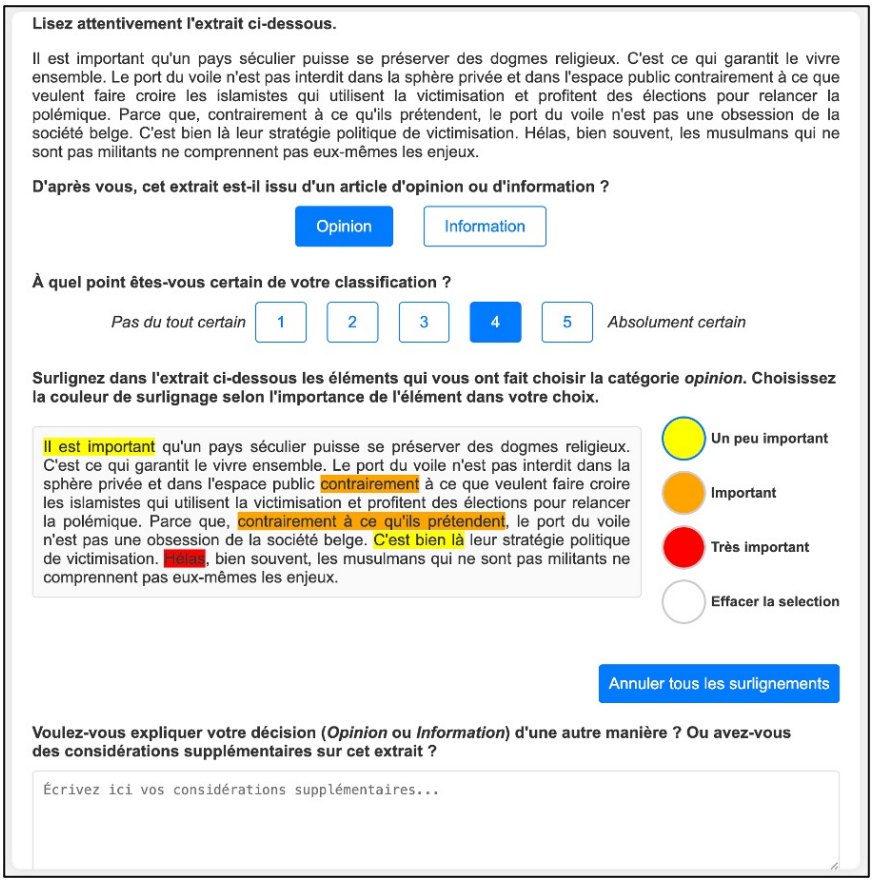


Figure 1: Screenshot of an annotation made using the custom interface used for the experiment.

were assigned 15 random texts from the corpus, which they had to annotate by the following Sunday. By the end of the campaign, all 26 students had annotated each of the 60 texts. Only 16 texts were unanimously categorized under the same class by all 26 annotators (6 as *opinion*, 10 as *news*). From these 16 texts, 10 (5 per class) were randomly selected for further annotation. The second step of the experiment, involving the 16 additional participants, lasted between December 10th, 2024, to January 31st, 2025. They were given the 10 selected texts in random order and as much time as they needed to read and annotate them. Most participants annotated all the texts in a single session, a few spread them out over several days. One of the 16 annotators completed only 6 out of 10 annotations, while all others annotated all 10 excerpts. In the end, 6 texts were annotated by 42 different annotators, and 4 texts by 41.

3.3 Attention-based Explanations

The model used to predict the class of the texts and from which explanations were generated is the case-insensitive base version of CamemBERT, containing 110 million parameters (Martin et al. 2020). CamemBERT relies on the architecture of RoBERTa

and was pre-trained on French data (Liu et al. 2019). We chose CamemBERT over other French or multilingual transformer models because of its compatibility with the method used for providing explanations: *Layer-wise Relevance Propagation* (LRP) (Chefer, Gur, and Wolf 2021).

LRP belongs to the class of post-hoc local explanation methods, which aim to attribute a model's prediction to specific input features (here, individual tokens in the text) by back-propagating relevance scores from the output layer through the network. Originally proposed by Bach et al. (2015) and adapted to NLP settings by Arras et al. (2017), LRP redistributes the scores layer by layer according to conservation constraints, assigning to each token a relevance value that reflects its contribution to the final decision. While several categories of local explanation methods exist, back-propagation based methods such as LRP offer a direct view into the model's internal reasoning and allow for the representation of explanations in the form of attention maps suited for human insight (see section 3.4). This makes LRP particularly suited for analyzing the sensitivity of explanations to hyper-parameters in transformer models. Finally, a crucial feature of LRP for our study is that it produces deterministic explanations for a given model instance: the same input and the same model parameters always yield the same explanation, on the contrary of other post-hoc local methods, such as perturbation-based explanations (**?**). This determinism is crucial, as it allows us to isolate the effect of randomness introduced during model training. Moreover, in line with previous works, we assume that back-propagation provides explanations that are sufficiently faithful and interpretable at the token level to give meaningful insights, while acknowledging that this faithfulness is not perfect.

For this experiment, we fine-tuned CamemBERT on the *opinion* vs. *news* classification task using the training set of 8,000 texts, perfectly balanced between the two classes (4,000 *opinion* and 4,000 *news* texts). The hyper-parameters used are the same as those used by Bogaert et al. (2023) learning rate of 2×10^{-5}, batch size of 4, and 2 epochs. To generate 100 equivalent versions of the fine-tuned model, we replicate the fine-tuning step on the same corpus separately, using 100 different random seeds (0 to 99).

Then, we have all 100 equivalent models predict the class of the 10 texts which constitute our target sample. Finally, using the LRP method, we generate an explanation of each prediction made by each model for each text, resulting in a set of 1,000 attention-based explanations.

3.4 Comparable Explanation Formats

Our goal is to produce explanation formats that are as comparable as possible between human annotations and model-based attention scores, while being explicit about the limitations of this approach. We therefore represent both human and model-generated explanations using attention maps, formatted as vectors associating each token with an importance value, and allowing for visualizations which are easily interpretable and enhance the plausibility of explanations to human readers (Sen et al. 2020). Although attention maps primarily offer localized, token-level insights and may not capture higher-level features such as long-range dependencies or syntactic relations, their clarity makes them well-suited for comparing explanations across different sources (human vs. LRP explanations).

For human explanations, attention maps are computed by aggregating the highlights of all annotators who selected the same class (*opinion* or *news*) for the same text. For example, the human attention map presented in Figure 12a (News 2) represents a vector combining the annotations of 34 participants. Each token is assigned a score depending on whether it was not highlighted (0), slightly important (0.33), important (0.66), or very important (1). These scores are then averaged and normalized between 0 and 1 across the tokens of the text. We acknowledge that these values are chosen arbitrarily and may not perfectly reflect the relative importance perceived by annotators (e.g., an "important" highlight may not actually be twice as relevant as a "slightly important" one), but they provide a consistent and reproducible way to aggregate human explanations.

Model-based attention maps, in contrast, are inherently continuous: each token is associated with a value between 0 and 1. To allow direct comparison with human explanations, we discretized these continuous values into four categories matching those used for human annotations: *not highlighted* (0.0–0.25), *slightly important* (0.25–0.50), *important* (0.50–0.75), and *very important* (0.75–1.0). While these ranges are again arbitrary and may be replaced by alternative thresholds, they ensure a consistent discretization across models. To better match the color distributions of human annotations, we also experimented with text-specific thresholds (detailed in section 3.6).

As illustrated in Figures 12 and 14, both human and LRP explanations are visualized using color-coded attention maps. For *opinion* predictions, tokens are highlighted in shades of red (most important), orange, and yellow (least important); for *news* predictions, shades of blue, teal, and green are used. This formatting is inspired by Sen et al. (2020), and adapted by Escouflaire, Descampe, and Fairon (2024) to aggregate multiple explanations of a given text, making variability or disagreement between annotators visually salient.

3.5 Explanation stability and Measures of Similarity

The main observation in Bogaert et al. (2023) is that functionally equivalent models can provide the same label but different explanations, raising a need to characterize their sensitivity to randomness in training. For this purpose, one can construct an explanation matrix of m rows (corresponding to different models) and n columns (corresponding to words in the text), where each $a_{s,w}$ corresponds to the relevance value assigned by the s-th model to the word at the w-th position, as showed in Figure 2.

$$\overset{n \ words}{\longleftrightarrow} \qquad \overset{m \ models}{\updownarrow} \qquad \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{m,1} & \cdots & \cdots & a_{m,n} \end{bmatrix}$$

Figure 2: Explanation Matrix

Since we hypothesize that human annotators won't consider that all the words are relevant, we can post-process our LRP explanation in order to mimic this phenomenon.

To do so, we can evaluate so-called k-word explanations, obtained by keeping only the $0 \leq k \leq n$ highest relevance values of each explanation, and by assigning a relevance of 0 to all the other tokens.

$$a_{s,w} = \begin{cases} a_{s,w} \text{ if } a_{s,w} \in TOP_k(a_{s,:}), \\ 0 \quad \text{oth.} \end{cases} \tag{1}$$

Finally, to measure the similarity between the different explanations, we use the **M**ean **C**orrelation **W**ith **M**ean **E**xplanation (MCWME) metric introduced in Bogaert, Descampe, and Standaert (2025). This metric is obtained by repetitively computing the correlation between one explanation and the average of all the others, and by averaging these correlations in the end. Figure 3 shows a scheme of how the metric is computed. The MCWME metric allows us to evaluate the general similarity of all the explanations. The choice of this metric arises from our explanation format, and we note that working with more complex explanations (e.g. displaying multiple relevance values per word, using combinations of words or studying other statistical moments) may require to use other metrics to study their similarity.

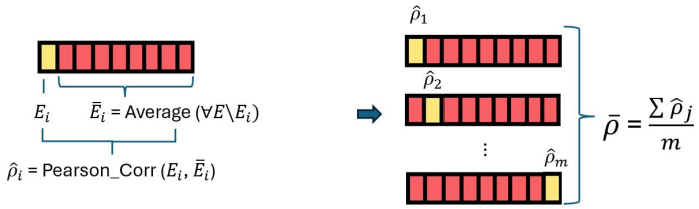


Figure 3: To obtain the MCWME, we first compute the Pearson correlation between an explanation and the average of all the others. We then repeat the process for all the explanations, and end by averaging all the obtained correlations.

As long as we can sort the words in an explanation by their importance (which is the case for explanations displaying continuous values for each word), we can compute a plot of the MCWME with respect to k and evaluate if the explanations differ or not on the most relevant words. Figure 4 shows such a plot. For human explanations, which are discrete by design (only 4 different categories or colors can be used to identify relevant words), we can not sort the words among each of the colors. It means that it is not possible to get a unique explanation containing the k most relevant values for all possible values of k. The only easy way to obtain explanations is the trivial *top-N* one, were we keep all the relevance values given by all the annotators. In order to have a similarity value for shorter explanation as well, we also compute the *top-1* MCWME by applying the procedure illustrated in Figure 5: We first create all variations of an explanation, that we denote as a block, by considering that each of the words having the highest relevance values is the *top-1* word. For example, for an explanation where x words would be highlighted in red by the human annotators, we generate x variations, each of them having one of these x words assigned with the highest relevance value

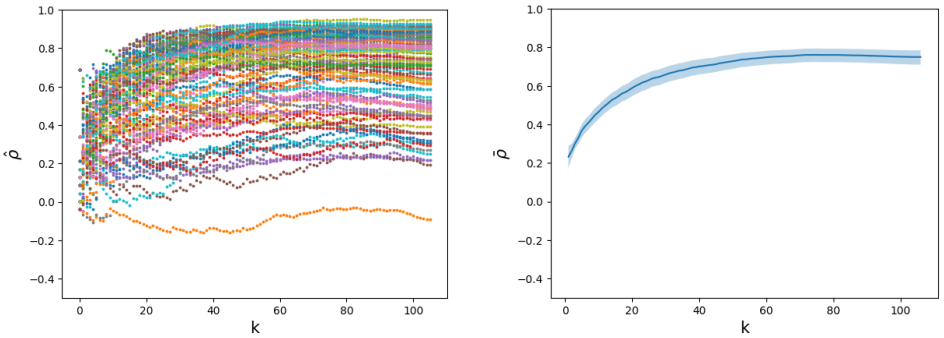


Figure 4: Correlations with Mean Explanation for 100 seeds (left) and Mean Correlation With Mean Explanation, with confidence intervals (right), for *News 5*.

and all the others being assigned a relevance of 0.[4] We then obtain the MCWME by first computing the Pearson correlation between each explanation and the average of the explanations in the other blocks. We finally compute the average correlation by block before averaging the values obtained for all the blocks.

It follows that for human explanations, we only compute the *top-1* and *top-N* MCWME values (the first and last points of the plots in Figure 4).

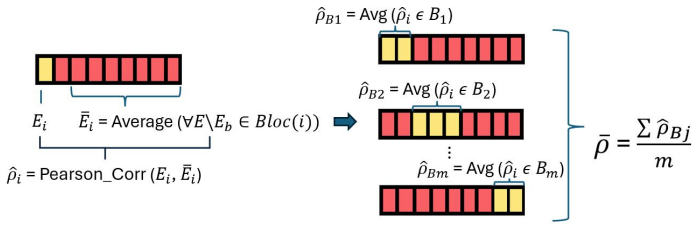


Figure 5: An example of MCWME computation for *top-1* explanations coming from discrete explanations with respectively 2, 3, ... and N words with the highest relevance score.

3.6 Explanation discretization

To compare the LRP and the human explanations, we post-process them in order to have a similar format. Since we can not turn the discrete human explanations (using 4 colors) into continuous ones (between 0 and 1), we need to discretize the LRP explanations. To do so, we use two different methods.

The first one, previously mentioned in section 3.4 and referred to as linear discretization, is a natural way to map continuous values between 0 and 1 to 4 values

4 This process scales very poorly when considering more than 1 top word, and is the reason why we did not consider more than *top-1* explanations as the process would quickly become too computationally intensive

using equals spans: values between [0 - 0.25] are assigned to white (not highlighted), [0.26 - 0.50] to yellow (slightly important), [0.51 - 0.75] to orange (important) and [0.75 - 1.00] to red (very important).

The second one, referred to as human-aligned discretization, aims at mimicking the same color distribution as the one used by the human annotators. Since this distribution varies from one text to another, the sizes of our 4 spans vary as well. By applying this process, we ensure that, for each color, the same amount of words are highlighted in the LRP and human explanations. To obtain the different color breakpoints, we first compute the amount of words highlighted in each of the colors by human annotators for a given text, denoted as c_r, c_o, c_y and c_w. We then sort all relevance values from the LRP explanations in descending order and take as breakpoints the value at the c_r, $c_r + c_o$ and $c_r + c_o + c_y$ positions.

4. Results

4.1 Classification Results

Although our focus is on the sensitivity of explanations to randomness in training rather than on the accuracy and variability of predictions, we first examine Table 1 to get an overview of the models' and humans' classifications for our target sample. The 10 texts in Table 1 are separated based on their "true label", which is the class attributed to the article (from which the piece of text was extracted) by the media that published it, *News* or *Opin(ion)*. We removed from the table all predictions that were not explained by the annotator by highlighting at least one token in the text (which explains why some texts contain less than 42 predictions, even though there were 42 annotators in the experiment).

	Human labels (*Opin*, *News*)	LLM labels (*Opin*, *News*)
News 1	1, 39	0, 100
News 2	6, 34	0, 100
News 3	1, 38	0, 100
News 4	2, 34	0, 100
News 5	0, 41	0, 100
Opin 1	42, 0	100, 0
Opin 2	42, 0	43, 57
Opin 3	40, 1	96, 4
Opin 4	40, 0	100, 0
Opin 5	40, 0	100, 0

Table 1: Predictions made by humans and CamemBERT models for the 10 texts.

Table 1 shows that human annotators appear more unanimous when predicting the class of texts labeled as *opinion*: except for *Opin 3*, all *opinion* texts were classified as such by all annotators. On the other hand, only one *news* text (*News 5*) was classified as *news* by all annotators. This may suggest that the discursive style of the *opinion* class in journalism appears more obvious to human readers than that of the *news* class. However, disagreements are not radical: *News 2* shows the highest imbalance in terms of human predictions, with 6 annotators identifying it as an *opinion* text and 34 as *news*.

Regarding the predictions made by the 100 equivalent fine-tuned CamemBERT models, it appears that they are almost always unanimous: for 8 out of 10 texts, all 100 models agree on their prediction. This was expected, as all models were trained on the same data, with the different seeds only impacting the order in which the 8,000 texts appeared in the training set. Interestingly, and as opposed to humans, the models never disagree on excerpts of *news* texts. Opin 2 shows a very high discrepancy between models, with 43 models predicting the *opinion* class and 57 going for *news*. Because this paper focuses on the variability of human and LRP explanations of similar predictions, we will leave out this text from qualitative analysis, as predictions made by the model are particularly divergent. For quantitative analysis, we will only consider explanations of predictions corresponding corresponding to the ground-truth class (43 explanations for Opin 2 and 97 for Opin 3).

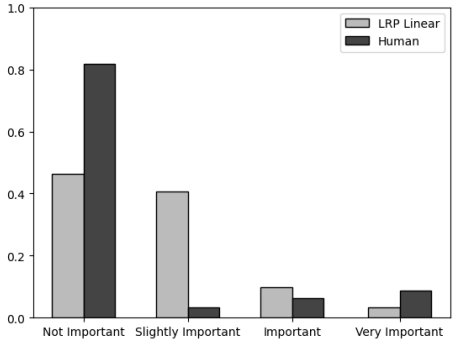
4.2 Qualitative Analysis

Before turning to quantitative measures of explanation stability, we first examine qualitative differences between human and LRP explanations. The goal here is to highlight structural contrasts in how importance is allocated across texts, both in terms of distribution of token-level importance values and of distribution of explanation segments.

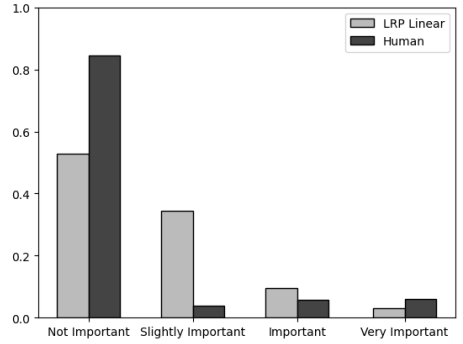
4.2.1 Distribution of Importance Values in Explanations. We begin by investigating how importance values are distributed across tokens in human and LRP explanations. This approach allows us to directly compare how annotators and LRP allocate importance over an entire text. Figure 6 allow us to compare the distribution of the different importance values assigned by annotators and LRP in terms of the proportion of tokens in the text. We can see in plots 6b, 6a and 6c, which compare importance distribution in the human explanations of the three texts with the models' discretized attention, that annotators tend to give importance to fewer tokens than models do, regardless of the text. Plot 6d shows how this computation is used to format the 'human-aligned' versions of the CamemBERT explanation vectors and attention maps. On top of it, the distributions of the three degrees of importance are inverted between humans and CamemBERT models, further confirming the disparities later observed in Section 4.2.2. These results suggest that human annotators and transformer models importantly differ in terms of how they explain, at token level, the same predictions.

4.2.2 Distribution of Explanation Segments. We now analyze how explanations are distributed across contiguous spans of text, so-called explanation segment (for example, in Figure 1, *Il est important* is the first explanation segment highlighted in yellow by the annotator).

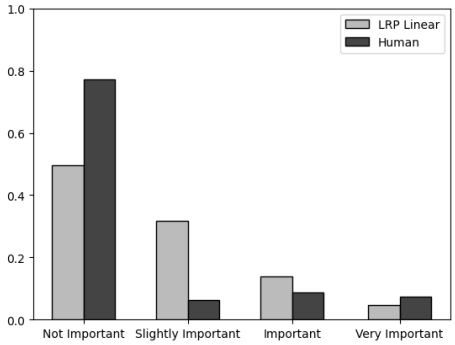
The goal of this analysis is to compare how humans and LRP explanations behave in terms of segment distribution. To make such a comparison possible, we apply the discretization scheme introduced earlier, aligning LRP attention values with the categorical scales used in human annotations. The barplots in Figure 7 allow us to compare the distribution of the different values attributed to the segments of texts (spans of at least one token) highlighted by humans and given attention by the CamemBERT models, following the two discretized explanation formats that render them comparable to human explanations. For human explanations, we include under the value *not important* all segments that were not highlighted at all.



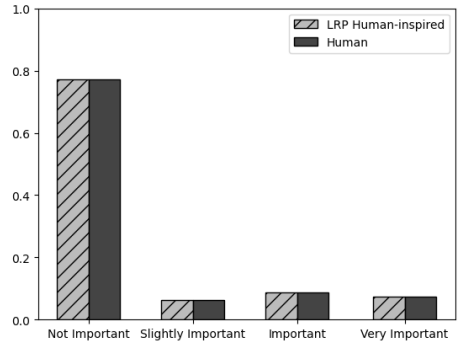
(a) *Opin 1* (LRP Linear)



(b) *News 2* (LRP Linear)



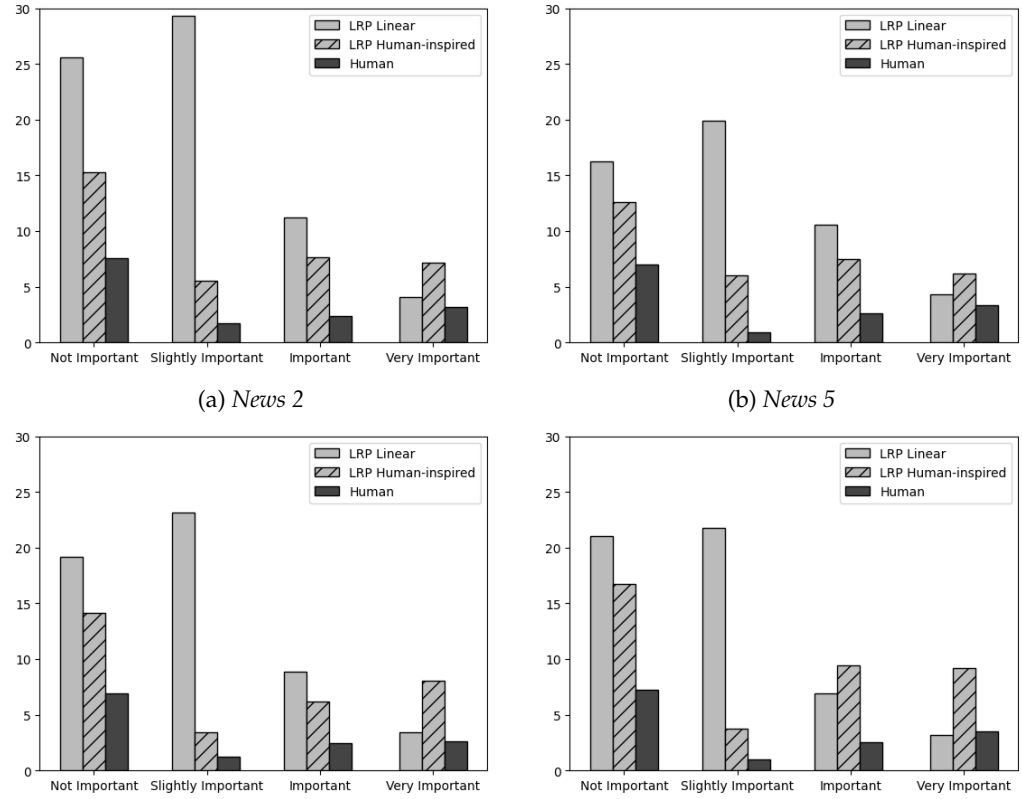
(c) *News 5* (LRP Linear)



(d) *News 5* (LRP human-aligned)

Figure 6: Proportion of tokens (words) associated to each relevance level.

One expected observation arising from this comparison is that humans tend to highlight significantly fewer segments of text as more or less important, compared to CamemBERT. This is due to the fact that the attention values of CamemBERT are a lot more sparse and volatile from one token to another than the highlights made by the annotators, who tend to consider longer segments of text as either somewhat important or not important at all. For example, this appears clearly in the attention maps of Figure 14, where humans seem to focus their explanations on the beginning and the end of the text, while LRP gives attention in a more dispersed way. Another major outcome of Figure 7 is that, proportionally, CamemBERT models have a tendency to attribute low or very low attention values to a lot of segments, and to give a lot of attention to only a small number of them. On the other hand, annotators leave a lot of segments blank (they do not highlight them as important at all), but regarding the highlighted segments, they are more likely to highlight segments as *very important* than *important*, and *slightly important* is the least used value by humans. One possible explanation for this phenomenon is the difficulty of discerning the relative importance of three levels of importance for human readers. Readers may favor one or two levels of importance over three when the explanation of a choice through our highlighting task appears more complex.



(a) *News 2* (b) *News 5*

(c) *Opinion 1* (d) *Opinion 4*

Figure 7: Average number of highlighted segments for two *Opinions* and two *News*

4.3 Quantitative Analysis

While the previous sections focused on qualitative contrasts between human and LRP explanations, we now turn to quantitative measures of their variability. The goal here is to assess, in a systematic way, whether machine-generated explanations are more or less stable than human ones, and to examine how methodological choices such as discretization affect this stability.

4.3.1 Variability of Human vs. Automated Explanations. In this section, we examine how consistent human explanations are compared to those produced by transformer models. The goal is to assess whether humans and LRP explanations highlight similar textual cues when justifying the same predictions, and to what extent the level of agreement varies depending on the class of the text. To do so, we use the MCWME similarity score (Bogaert, Descampe, and Standaert 2025), introduced in Section 3.5, which quantifies the similarity across a set of explanations. We first compare human and LRP explanations directly, before analyzing how similarity evolves when considering an increasing number of highlighted tokens (k). For humans, we compute MCWME of the most unanimous prediction (*news* or *opinion*), as presented in Table 1. This analysis allows us to characterize not only the general variability in explanations, but also their

relative stability across the two classes, and to identify cases where humans and models diverge most strongly.

In Figures 8, 9, 11, we compare the MCWME similarity scores between all formats of human and LRP explanations described in sections 3.4 and 3.6, using boxplots representing the confidence intervals around the MCWME values. The explanation modes and formats being compared are: human explanations, original (non-discretized), linear and human-aligned LRP explanations. All MCWME measures are computed both for *top-1* and *top-N* (all tokens) explanations. For both models and humans, we only included explanations of predictions of the ground-truth class for each text.

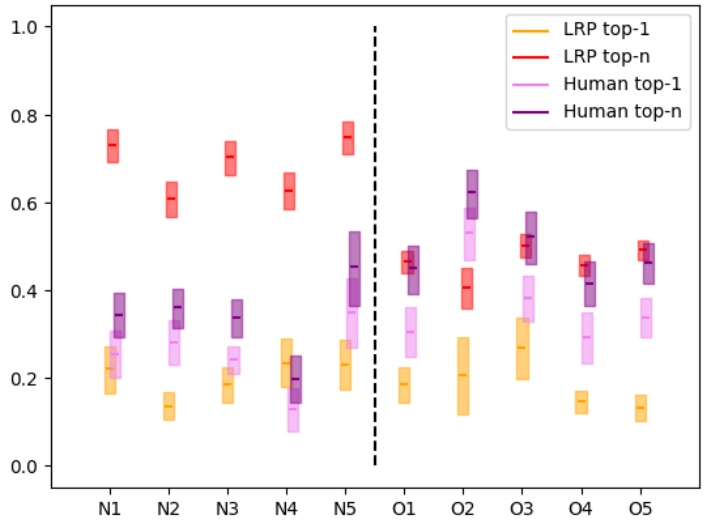


Figure 8: Boxplots of MCWME with confidence intervals for LRP (original) and Human explanations of the 10 texts, computed for the top-1 or for all tokens (top-N).

We first focus on a comparison of human and non-discretized LRP explanations (Figure 8). The first observation arising from Figure 8 is that taking only the *top-1* token with the most attention in each explanation systematically produces a lower MCWME score than when all tokens are considered in the calculation. However, this drop is much more severe for CamemBERT than for humans, which suggests that humans tend to agree more on the most important tokens in their explanations than CamemBERT models do.

Regarding the predicted class being explained, the boxplots in Figure 8 show another interesting difference between human and LRP explanations. On the one hand, CamemBERT models tend to display a higher MCWME score when explaining *news* predictions than for explanations of *opinion* (this is only true for *top-N* MCWME scores, not for *top-1*). On the contrary, human annotators seem to agree more on explanations of *opinion* than *news* texts. This may be because *opinion* articles typically contain clear markers of subjectivity and strong stylistic choices, which provide clearer and more consistent cues for interpretation. By contrast, *news* articles tend to follow a more

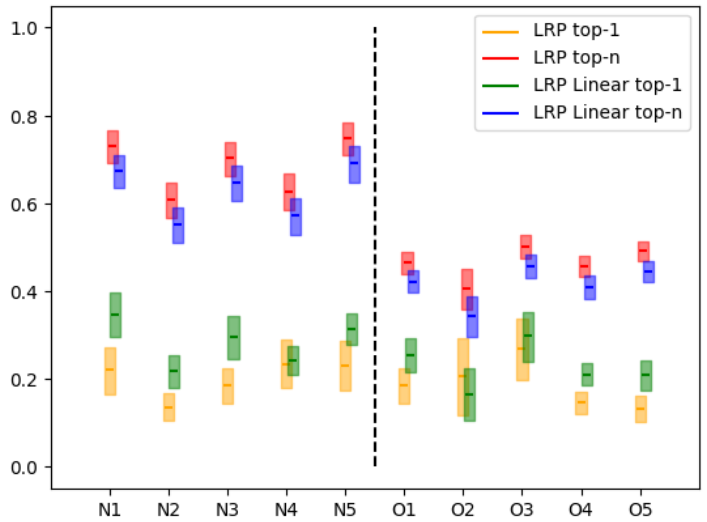


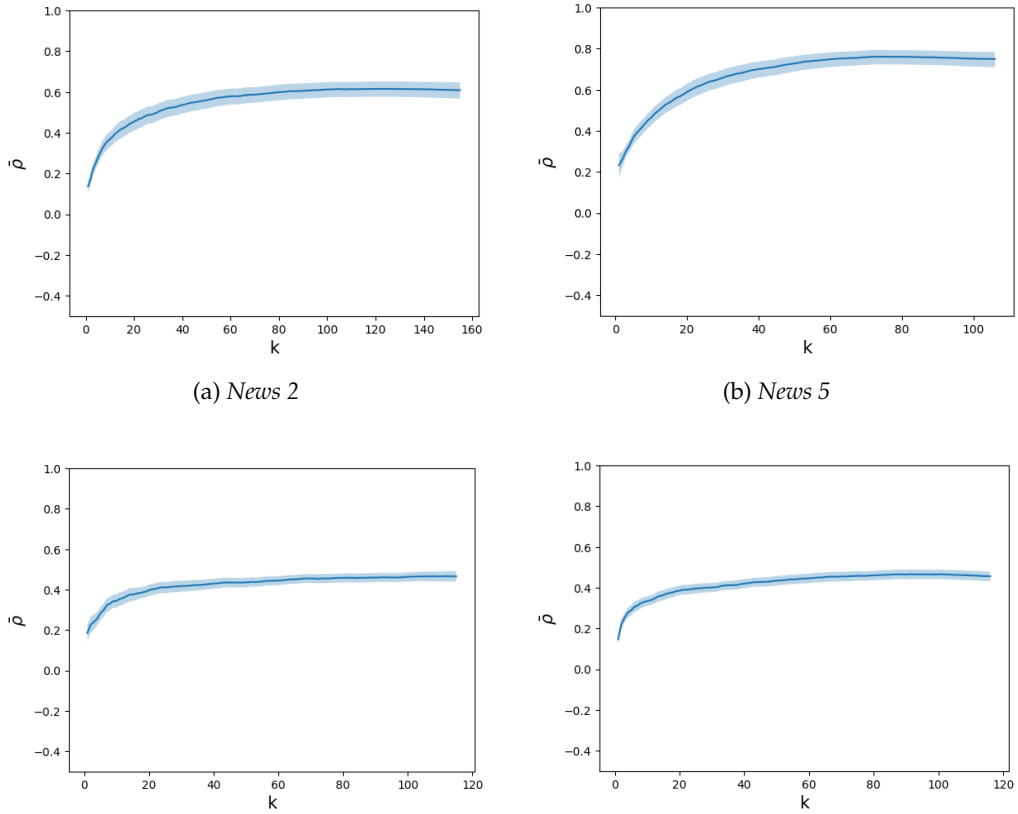
Figure 9: Boxplots of MCWME with confidence intervals for LRP (original) and LRP Linear of the 10 texts, computed for the top-1 or for all tokens (top-N).

neutral style, less marked by linguistic markers, which may make it harder for human annotators to identify consistent cues for their classification (Koren 2004).

The plots in Figure 10 display how the similarity between CamemBERT explanations improves in terms of MCWME when increasing the amount of most highlighted words considered among each explanation (k). We observe that tendency for all the texts used in our experiment. Moreover, this plot further highlights the difference in stability between the LRP explanations for *news* and *opinions*. First, while all the curves display a plateau, it is reached much earlier (around $k = 20$) for *opinions* than for *news* (around $k = 60$). On top of it, the MCWME value at the plateau is higher on average for *news* (between 0.6 and 0.8) than for *opinions* (between 0.4 and 0.5).

4.3.2 Impact of Discretization on Variability of Explanations. We now examine how the variability evaluated in the previous section is affected when continuous LRP explanations are discretized into categorical values. To address this, we compare LRP explanations in their continuous form (LRP) with their linear versions (LRP Linear), and further with discretization informed by human annotation distributions (LRP human-aligned). This step-by-step comparison makes it possible to assess whether discretization reduces sensitivity to the number of most highlighted tokens (k), and whether human-aligned thresholds bring LRP explanations closer to human patterns of variability.

The comparison between the original model (LRP) and linear (LRP Linear) explanations, which is displayed for the 10 texts in Figure 9, reveals a consistent pattern: for $k = 1$, the MCWME values for the linearly discretized explanations are higher than the one for their continuous counterparts, while for $k = N$, the MCWME values for the discretized explanations are lower. This dual shift effectively reduces the gap between the stability scores for $k = 1$ and $k = N$, bringing them closer to the



(a) *News 2*

(b) *News 5*

(c) *Opinion 1*

(d) *Opinion 4*

Figure 10: MCWME of four texts with respect to the amount of selected top words in each of the explanations.

values observed in human explanations (Figure 8). The introduction of human-based discretization (LRP human-aligned) accentuates this effect, as shown in Figure 11. Discretizing attention values based on human-aligned thresholds yields even smaller gaps with the MCWME values observed for human explanations.

The boxplots in Figures 9 and 11 indicate that discretizing the attention values of the LRP explanation vectors, particularly when informed by human annotation, moderates the sensitivity of explanations when only focusing on a small number of most highlighted tokens. In our case study based on LRP-based explanations of CamemBERT predictions, similarity of explanations provided by functionnally equivalent models tend to vary more sharply with k compared to human annotations. Discretization smooths this variation, and human-aligned discretization brings the resulting curves closer to those of human explanations.

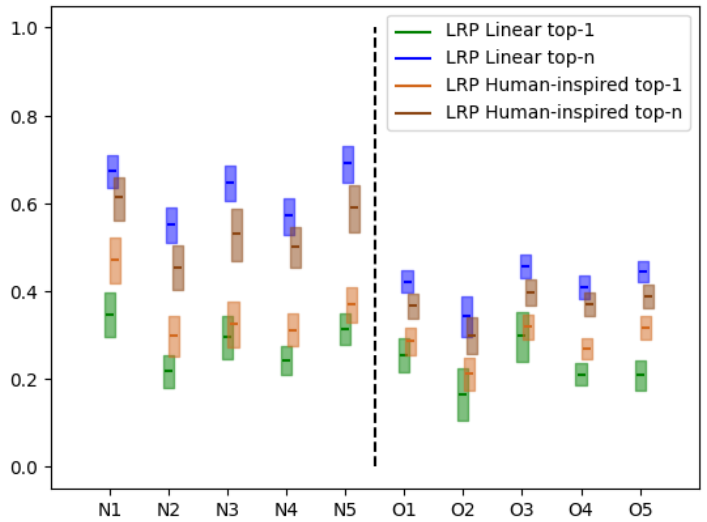


Figure 11: Boxplots of MCWME with confidence intervals for LRP Linear and LRP human-aligned explanations of the 10 texts, computed for the top-1 or for all tokens (top-N).

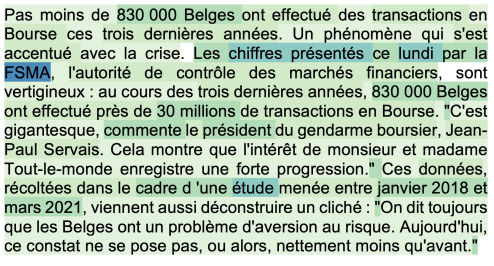
4.4 Linguistic Analysis of Texts

To complement the results presented in the previous sections, we now turn to a visual inspection of aggregated attention maps. The goal is to qualitatively assess how humans and LRP highlight different parts of the text when producing explanations, and to evaluate whether discretization or formatting choices bring LRP explanations closer to human-like patterns.

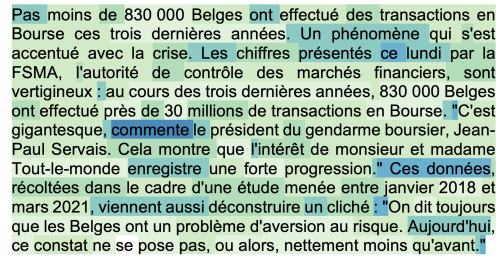
For this purpose, we compare attention maps that aggregate human explanations (top-N), continuous LRP explanations, their linear and human-aligned versions, as well as formats restricted to *top-1* tokens or to the most salient tokens in discretized vectors. Figures 12, 13 and 14 display examples for three representative texts: two classified as *news* (News 2 and 5) and one classified as *opinion* (*Opin 1*). These texts were selected because they allow us to illustrate clear contrasts between explanation formats. For the main part of this linguistic analysis, we deliberately avoided texts where the models disagreed or produced non-unanimous predictions, ensuring that the visualized comparison between model types is based on the same number of explanations for the same predicted label. At the end of the section, we will look at maps of an excerpt with high disagreement among model predictions (Figures 16 and 15).

First, the attention maps illustrate how human and LRP explanations diverge in terms of which tokens receive the most attention, even when aggregating across multiple explanations.For the text in Figure 12, humans (Figure 12a) appear to explain their predictions for *news* primarily with mentions of information sources (*FSMA*, *chiffres* [figures], *étude* [study]), as well as time markers (*lundi* [Monday], *janvier* [January]) and numerical digits. While some of these patterns align with those observed in Figure 12b (e.g., *ce lundi* [this Monday], *ces données* [these data]), the models assign

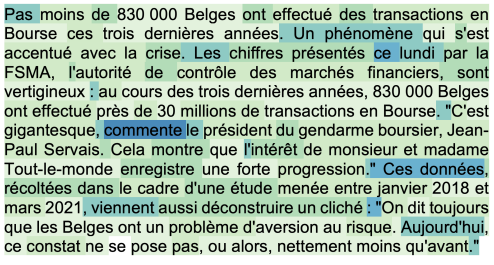
greater attention to quotation marks and to the quotation verb *commente* [comments]. Those are all recognized markers of the *news* class (Escouflaire et al. 2024), but not the same ones as the ones prioritized in human explanations.



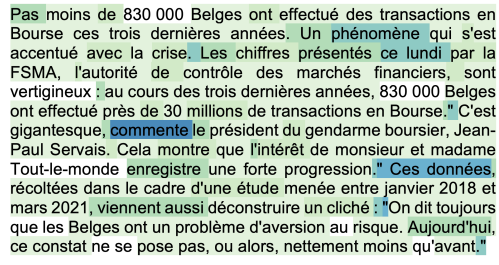
(a) Humans



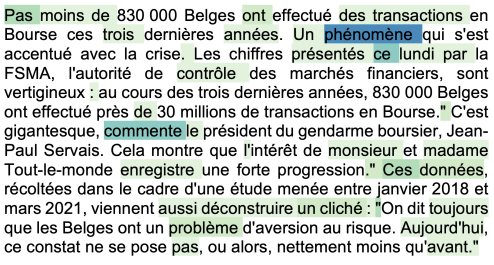
(b) LRP



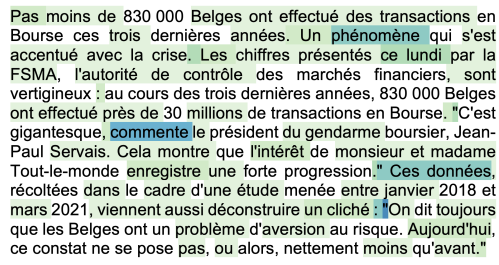
(c) LRP linear



(d) LRP human-aligned



(e) LRP Top 1



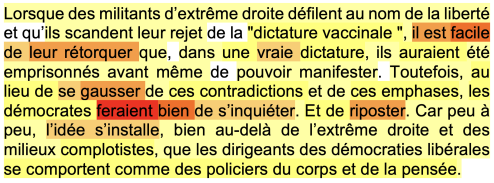
(f) LRP top color

No fewer than 830,000 Belgians have carried out stock market transactions over the past three years. A phenomenon that has become even more pronounced with the crisis. The figures presented this Monday by FSMA, the financial markets supervisory authority, are staggering: over the last three years, 830,000 Belgians have carried out almost 30 million stock market transactions. That's gigantic," comments the chairman of the stock market authority, Jean-Paul Servais. It shows that the interest of ordinary people is growing strongly. These data, gathered as part of a study conducted between January 2018 and March 2021, also deconstruct a cliché: "It's always said that Belgians have a problem with their aversion to risk. Today, this is not the case, or much less so than before."

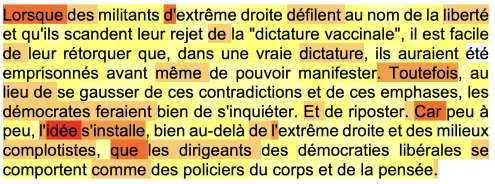
(g) Translation

Figure 12: Attention maps of different explanation methods for *News 2*, an excerpt of a *news* article.

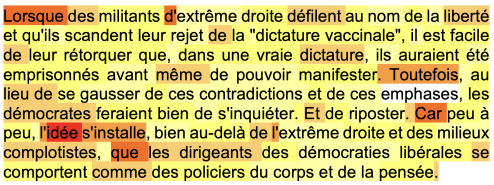
For *Opin 1*, where both humans and models classified the text as *opinion*, the differences in explanations remain notable. Humans (Figure 13a) give substantial importance to axiological verbs such as *gausser* [mock], *feraient* bien [would do well], *s'inquiéter* [worry], and *riposter* [retaliate]. In contrast, the tokens receiving the most attention in 13b are primarily argumentative connectives (*lorsque* [when], *toutefois* [however], *car* [because]) as well as the phrase *l'idée s'installe [...] que...* [the idea is taking hold [...] that...]. This text also illustrates the attention that transformer models allocate to sentence boundaries, i.e., the first and final tokens of sentences. This phenomenon, which was already documented by Clark et al. (2019), is not reflected in human explanations.



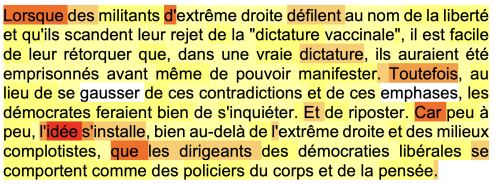
(a) Humans



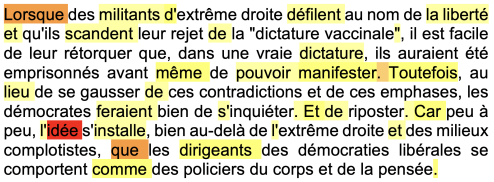
(b) LRP



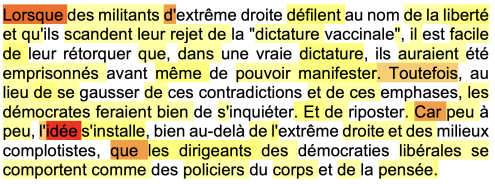
(c) LRP Linear



(d) LRP human-aligned



(e) LRP top 1



(f) LRP top color

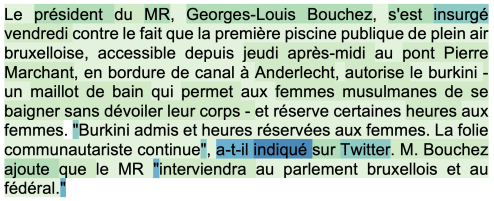
When far-right activists march in the name of freedom and chant their rejection of the "vaccine dictatorship", it's easy to retort that, in a real dictatorship, they would have been imprisoned before they could even protest. However, instead of laughing at these contradictions and emphases, democrats would do well to worry. And to fight back. Because little by little, the idea is taking hold, well beyond far-right and conspiracy circles, that the leaders of liberal democracies are behaving like police of the body and mind.

(g) Translation

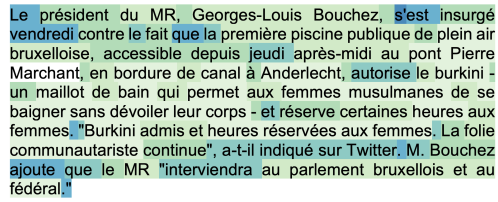
Figure 13: Attention maps of different explanation methods for *Opin 1*, an excerpt of an *opinion* article.

News 5, like *News 2* (Figures 14), was classified as news by both humans and models. This case further illustrates a recurrent difference between human and LRP

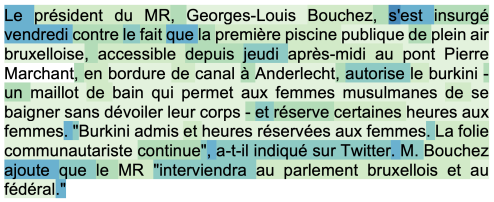
aggregated explanations: human attention often concentrates on a small number of tokens or on a specific section of the text, whereas LRP attention tends to be more diffuse. Here, both maps 14a and 14b highlight quotation marks and reporting verbs or clauses (e.g., ”a-t-il indiqué sur Twitter” [he said on Twitter], *s'est insurgé* [protested]), but the models also distribute attention across additional tokens such as *vendredi* [Friday], *que* [that], *jeudi* [Thursday], and *autorise* [authorizes].



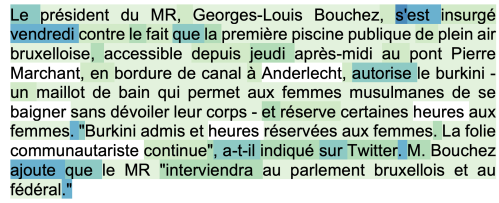
(a) Humans



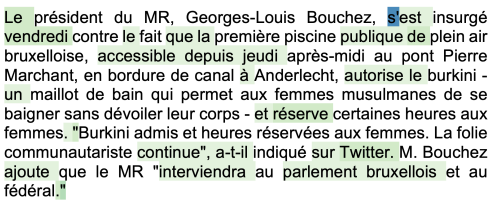
(b) LRP



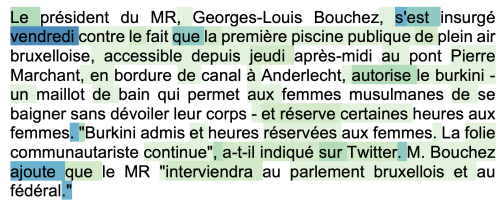
(c) LRP linear



(d) LRP human-aligned



(e) LRP top 1



(f) LRP top color

The president of the MR party, Georges-Louis Bouchez, protested on Friday against the fact that Brussels' first public open-air swimming pool, open since Thursday afternoon on the canal-side Pierre Marchant bridge in Anderlecht, allows burkini - a swimsuit that allows Muslim women to bathe without revealing their bodies - and reserves certain hours for women. "Burkini allowed and hours reserved for women. The communitarian madness continues," he posted on Twitter. Mr. Bouchez added that the MR “will intervene in the Brussels parliament and at federal level.”

(g) Translation

Figure 14: Attention maps of different explanation methods for *News 5*, an excerpt of a *news* article.

Comparing attention maps of the original LRP explanations with their discretized versions also yields interesting observations. For all three texts, we can observe that there is almost no difference in terms of visualization between the original map (for

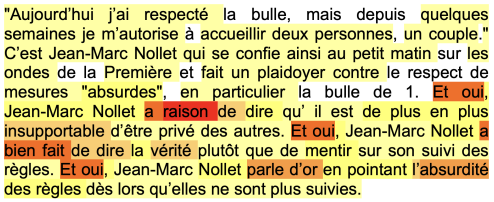
example Figure 12b) and the maps of the linear 12c and human-aligned 12d versions of the same aggregation of explanations. The difference is slightly more visible in maps 13d and 14d.

We then considered looking only at the *top-1* tokens of all original LRP maps, as well as at all tokens with the highest value in the discretized human-aligned maps (referred to as 'top-color' maps). These formatting choices produce much more noticeable changes in the visual appearance of the maps. *Top-1* maps, such as Figures 13e and 14e, show a very sharp focus on a few tokens that already stood out in the original maps (13b and 14b). In some cases, however, they also bring to the foreground tokens that were less salient in the original maps, as illustrated in Figure 12e with the token *phénomène*, which appears consistently across models despite being less emphasized in the non-discretized versions. By contrast, top-color maps (e.g., Figures 12f and 13f) are sparser and visually closer to human attention maps. They highlight the tokens that models considered most important, but in a way that preserves more continuity and readability than *top-1* maps. In this sense, such human-aligned top-color maps may represent a useful compromise: they remain faithful to the strongest signals in the LRP explanations, while at the same time producing patterns that are more plausible and easier to interpret for human readers.

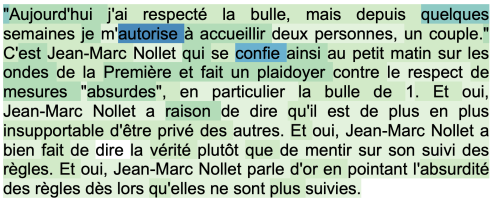
Finally, we also examined the human and model explanations of the most divisive excerpt in the experiment: *Opin 2* (Figures 15 and 16). As explained in section 4.1, this excerpt is the only one with a high disagreement in class predictions across the 100 equivalent models, as 43 predicted *opinion* while 57 predicted *news*. Interestingly, this excerpt was not ambiguous for human readers, as all 42 annotators classified the excerpt as *opinion*. Comparing the maps of the 43 *opinion* explanations in Figure 16 and those aggregating the 57 *news* explanations in Figure 15 provides insight into what kind of tokens drive our models to pick one class over the other.

We first observe that there is almost no overlap between the words with the highest attention in Figure 15b and 16b. This is also true for the other explanation formats, although two tokens appear strongly highlighted in both groups of explanations, especially visible when comparing the human-aligned maps (Figure 15d and 16d): the first two tokens of the *Opin 2* excerpt, the initial quotation mark and the adverb *Aujourd'hui* [Today]. The importance attributed to these tokens may again be due to their starting position in the text (Clark et al. 2019) rather than to the lexical and syntactic information they carry. Overall, observing that the most salient tokens may differ depending on the predicted label further suggests that the LRP explanation method provides relevant yet partial information about a model's reasoning towards a prediction.

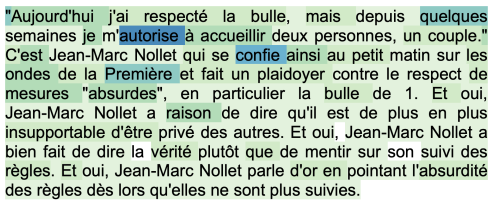
Looking at the most attended tokens for the *news* explanations first, we find that Figure 15e highlights the verb *autorise*, which is the head of the only quoted sentence in the text. It emphasizes that the presence of quotations in an article is an important *news* feature according to the models. Other words inside the quotation marks, as well as the reporting verb introducing the quote, *confie* [confide], strengthen this hypothesis. Additionally, the origin of the quotation (radio station *La Première*) is also given relative attention in Figure 15c (though this does not appear in Figure 15e).



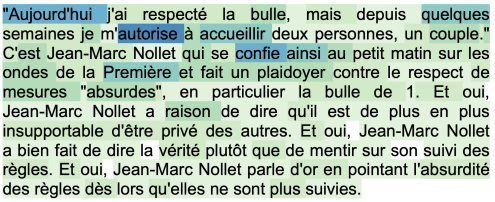
(a) Humans



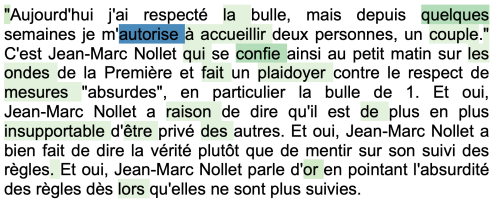
(b) LRP



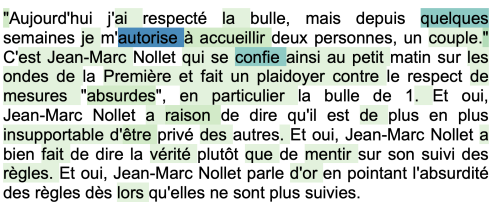
(c) LRP linear



(d) LRP human-aligned



(e) LRP top 1



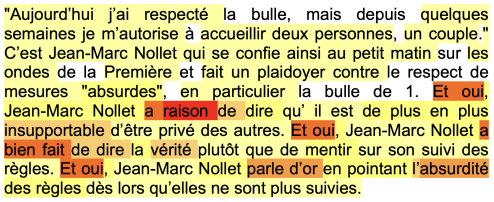
(f) LRP top color

"Today I respected the bubble, but for the past few weeks I have been allowing myself to welcome two people, a couple." Jean-Marc Nollet confided this early in the morning on La Première radio station and made a plea against respecting "absurd" measures, in particular the bubble of 1. And yes, Jean-Marc Nollet is right to say that it is more and more unbearable to be deprived of others. And yes, Jean-Marc Nollet did the right thing in telling the truth rather than lying about his compliance with the rules. And yes, Jean-Marc Nollet is spot on when he points out the absurdity of rules as soon as they are no longer followed.

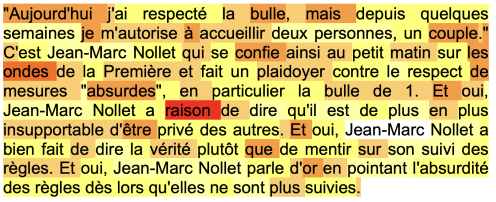
(g) Translation

Figure 15: Attention maps of different explanation methods for *Opin 2*, an excerpt of an *opinion* article. For this excerpt, all 42 human annotators agreed on *opinion*, while 57 out of the 100 equivalent models predicted *news*. Figures (b) to (g) show the aggregations of the explanations provided only by the models who predicted *news*.

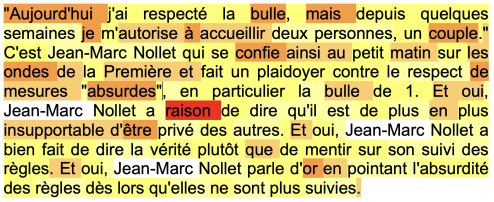
Finally, Figure 16e shows that the token being the most unanimously highlighted among the 43 model explanations for the *opinion* class is *raison*, main semantic part of the verbal idiom *avoir raison de* [to be right to]. This word is also the one that was considered the most important by the human annotators for classifying this excerpt as *opinion* (Figure 16a), as it is a strong piece of evidence of the author's implication in the text. The verb *être* [to be] is also given a lot of attention among models in Figure 16e, likely because it completes the axiological adjective *insupportable* [unbearable]. The



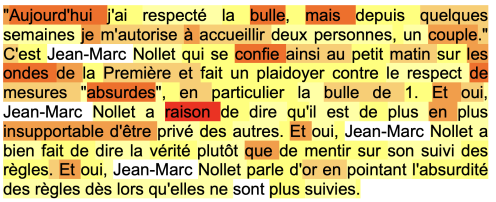
(a) Humans



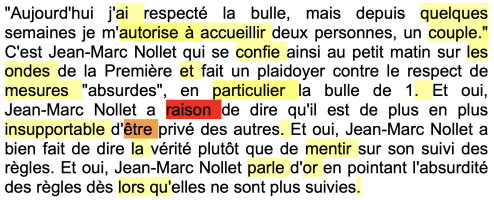
(b) LRP



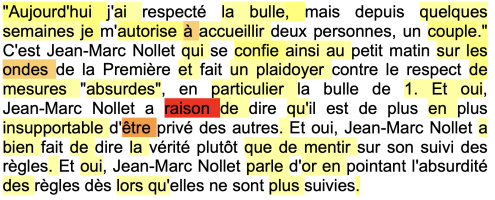
(c) LRP linear



(d) LRP human-aligned



(e) LRP top 1



(f) LRP top color

"Today I respected the bubble, but for the past few weeks I have been allowing myself to welcome two people, a couple." Jean-Marc Nollet confided this early in the morning on La Première radio station and made a plea against respecting "absurd" measures, in particular the bubble of 1. And yes, Jean-Marc Nollet is right to say that it is more and more unbearable to be deprived of others. And yes, Jean-Marc Nollet did the right thing in telling the truth rather than lying about his compliance with the rules. And yes, Jean-Marc Nollet is spot on when he points out the absurdity of rules as soon as they are no longer followed.

(g) Translation

Figure 16: Attention maps of different explanation methods for *Opin 2*, an excerpt of an *opinion* article. For this excerpt, all 42 human annotators agreed on *opinion*, while 43 out of the 100 equivalent models predicted *opinion*. Figures (b) to (g) show the aggregations of the explanations provided only by the models who predicted *opinion*.

human-aligned map (Figure 16d) supports this claim, as both words appear important. The same map highlights another strong adjective, *absurdes* [absurd]. Interestingly, the quotation marks around this token seem ignored by the models, while the human annotators attributed almost no importance to the word (probably because of the quotation marks marking the author's distance from the adjective). In Figure 16a, annotators almost systematically highlighted the threefold anaphoric repetition of *Et oui* [And yes] at the start of the last three sentences of the excerpt. This pattern also seems to be present

in the human-aligned map in Figure 16d, suggesting that the fine-tuned models may also capture stylistic constructions, which are characteristic of opinionated discourse and signal a stronger authorial stance.

5. Discussion

This study set out to compare the variability of human and LRP explanations for a French binary journalistic text classification task (news vs. opinion), focusing on a sample of 10 texts. Our expectation was that the variability of human explanations would be lower than that of LRP explanations, as a previous study by Bogaert et al. (2023) suggested that explanations provided by functionally equivalent models for a text were highly sensitive to random hyperparameters, even when trained on the same data. However, across all our analyses, both humans and fine-tuned CamemBERT models produced explanations that varied from one instance to another. The patterns and sources of this variability differed interestingly between the two.

We first observed in section 4.3.1 that humans tended to provide more concentrated explanations, highlighting relatively few tokens and often focusing on a small number of linguistically salient cues of *news* and *opinion* classes. In contrast, LRP explanations were more diffuse and dispersed, with attention values spread across a wide range of tokens. This difference in style was reflected in the MCWME similarity results: humans showed stronger agreement on the most important tokens (especially in *top-1* scores) than BERT models did, whereas models tended to align more when all tokens (especially the low-attention ones) were taken into account. Interestingly, the predicted class of the text also played a role in explanation similarity. For humans, explanations of *opinion* predictions were slightly more consistent than those of *news* texts, though the difference was small. For LRP, however, *news* explanations consistently showed higher similarity scores than *opinion* explanations, especially when considering all tokens. This divergence suggests that the humans and models likely rely on different information or cues in the texts for predicting their class. This was also reflected in our qualitative analysis of attention maps (see section 4.4), as we observed that humans and models did not focus on the same types of tokens for prediction *news* or *opinion*.

The key finding of our analysis of human vs. LRP explanation variability (see section 4.3.1) is that discretization of LRP explanations substantially reduced the human–model gap. Applying uniform breakpoints (linear discretization) to LRP explanation vectors already improved similarity scores with human explanations, and human-aligned discretization (matching the highlight distribution of human annotators for each text) narrowed the gap further. When visualizing them under the form of attention maps and through qualitative analysis (section 4.4), we showed that human-aligned maps (generated under human-aligned thresholds) more closely resembled human attention maps. This supports the idea that explanation post-processing can meaningfully improve the alignment between LRP outputs and human reasoning without altering model predictions.

Recent work by Deutsch, Foster, and Freitag (2023) highlights that discretization and the resulting increase in tied scores can mechanically inflate correlation-based evaluation metrics, independently of any substantive improvement in alignment between compared objects. Applied to our setting, this observation suggests that part of the similarity gains observed after discretizing LRP explanations, both linear

and human-aligned, may stem from reduced sensitivity of the MCWME metric to fine-grained ranking differences once attribution values are grouped into fixed categories. We therefore interpret these quantitative improvements with caution: while discretization facilitates comparison between continuous model attributions and categorical human highlights, higher similarity scores alone do not necessarily imply deeper resemblance in reasoning, but may rather reflect how representation and metric design interact in explanation evaluation.

In light of both the qualitative trends (Section 4.4) and quantitative results (Section 4.3), the answer to the question "are model explanations more variable than human ones?" is not straightforward. On one side, we showed that LRP explanations are stylistically very different from human ones: they are more diffuse, spread over many tokens, and less concentrated on a few salient cues. Then, quantitative trends highlighted that LRP sometimes display higher sensitivity than humans (especially when considering only the most highlighted tokens) but this difference diminishes, or even reverses, when taking into account larger portions of the explanation. In other words, while machine explanations are not systematically more variable than human explanations, their variability follows different trends depending on how explanations are represented and compared.

These findings have implications for explainability research in NLP. Raw LRP attention values may be faithful indicators of internal model reasoning, but they are not necessarily well-aligned with human explanations, which are by nature plausible. Incorporating human-aligned constraints into explanation visualization of transformer models' predictions could improve plausibility without compromising faithfulness. Also, the divergences observed in our analyses suggest that humans and transformer models may prioritize different types of information when performing the same task, reflecting expected differences between cognitive and algorithmic processing.

6. Limitations

Our study, limited to the CamemBERT model and LRP explanations, may not be representative of the entire field of explainability applied to transformer architectures, but we believe that it still yields interesting insights to better understand the potential weaknesses of these approaches in general. In particular, our conclusions are drawn from a single RoBERTa-based architecture with fixed model size, pretraining objectives, and pretraining data. Explanation variability may differ for models trained under different conditions, even within the same family of models, as changes in pretraining data or scale are known to affect both internal representations and downstream behavior (Zhao et al. 2024). It is also possible that using another explainability method (Müller et al. 2023) or focusing on different use cases would lead to different conclusions. Furthermore, LRP explanations are restricted to token-level attributions and do not capture potential relations or dependencies between words, which may limit the alignment with human reasoning processes. Likewise, larger or differently pretrained models may rely on alternative textual cues or distribute relevance more evenly across tokens, potentially leading to distinct variability patterns compared to those observed for CamemBERT. In addition, human annotations in this experiment were themselves constrained by the in-house highlighting interface and by the predefined importance categories, which may have influenced the way annotators expressed their explanations. Finally, our analysis focused on a single classification task (distinguishing between *news* and *opinion*), and

it is possible that other tasks would yield different patterns of agreement or variability. This raises the inevitable broader question of whether human and model explanations, given their different formats and constraints, are ever fully comparable.

7. Conclusions

In this work, we compared the variability observed across explanations provided by equivalent models with that observed across annotations made by different human readers. We first started by showing that despite predicting the same label most of the time, the models and the annotators provide various explanations. We then highlighted that the annotations provided by human readers displayed less variability than model explanations when considering fewer words, but that the opposite trend was observed when considering all the words in the texts. More importantly, our observations also shed light on the balance between *faithfulness* and *plausibility* in explainable NLP. While propagation-based explanations are designed to be as faithful as possible to the model's internal decision process, plausibility remains a challenge. Bridging this gap through the use of discretization methods is a relevant direction for future work. We showed that, when using linear and human-aligned breakpoints for discretizing model explanations, their distribution was very different between humans and models. In particular, human annotators tend to assign a uniform level of importance to long continuous chunks of text, whereas models decompose their explanations at a much finer granularity, attributing varying contributions to individual tokens.

In future research, it would be relevant to explore whether, despite observed variability among explanations, some clusters of "typical" explanations may emerge (Mitrut et al. 2024). Using standard dimensionality reduction techniques (such as *t-SNE*) for vizualisation, we do not observe clear clusters in the explanations provided by equivalent models, but it would be interesting to investigate this possibility further. Another complementary direction would be to investigate explanations generated directly by large language models asked to annotate salient tokens. It could help reveal whether generative models converge on similar patterns of saliency across texts and classes, and how these patterns compare with human judgments. While such an experiment may be more aligned with the current trends in LLM research, it would introduce additional variability due to model stochasticity, which makes direct comparison with deterministic explanations challenging. Future work could therefore aim to design controlled setups for LLM-based annotations to enable more meaningful comparisons with human and propagation-based model explanations.

In the bigger picture, we hope that our study will help in understanding the impact of the Rashomon effect on model explainability in general, and is a step toward providing better explanations regarding model predictions.

Acknowledgments

Louis Escouflaire was a PhD student of the the Belgian National Fund for Scientific Research (FNRS-F.R.S.) when completing this work. Marie-Catherine de Marneffe and François-Xavier Standaert are respectively research associate and research director of the Belgian National Fund for Scientific Research (FNRS-F.R.S.). This work has been supported by the Service Public de Wallonie Recherche, grant n°2010235-ARIAC.

References

Acheampong, Francisca Adoma, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of BERT-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adebayo, Julius, Justin Gilmer, Ian Goodfellow, and Been Kim. 2018. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307*.

Agarwal, Chirag, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*.

Alhindi, Tariq, Smaranda Muresan, and Daniel Preoţiuc-Pietro. 2020. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th international conference on computational linguistics*, pages 6139–6149.

Arras, Leila, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *WASSA@EMNLP*, pages 159–168, ACL.

Bach, Sebastian, 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):e0130140.

Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21, Association for Computational Linguistics.

Benamara, Farah, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Bogaert, Jeremie, Antonin Descampe, and François-Xavier Standaert. 2025. Consolidating explanation stability metrics. In *World Conference on Explainable Artificial Intelligence*, pages 310–323, Springer.

Bogaert, Jérémie, Marie-Catherine de Marneffe, Antonin Descampe, Louis Escouflaire, Cédrick Fairon, and François-Xavier Standaert. 2023. Sensibilité des explications à l'aléa des grands modèles de langage: le cas de la classification de textes journalistiques. *Trait. Autom. Des. Langues*, 64(3).

Bogaert, Jeremie, Marie-Catherine de Marneffe, Antonin Descampe, Louis Escouflaire, Cedrick Fairon, and Francois-Xavier Standaert. 2024. Explanation sensitivity to the randomness of large language models: the case of journalistic text classification. *arXiv preprint arXiv:2410.05085*.

Breiman, Leo. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Cavalcanti, Anderson Pinheiro, Rafael Ferreira Mello, Dragan Gašević, and Fred Freitas. 2024. Towards explainable prediction feedback messages using bert. *International Journal of Artificial Intelligence in Education*, 34(3):1046–1071.

Chefer, Hila, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *CVPR*, pages 782–791, Computer Vision Foundation / IEEE.

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Deutsch, Daniel, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. *arXiv preprint arXiv:2305.14324*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805 version: 2.

Escouflaire, Louis, Antonin Descampe, and Cédrick Fairon. 2024. Unveiling subjectivity in press discourse: A statistical and qualitative study of manually annotated articles. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, 34.

Escouflaire, Louis, Antonin Descampe, Antoine Venant, and Cédrick Fairon. 2024. La subjectivité dans le journalisme québécois et belge: transfert de connaissance inter-médias et

inter-cultures. In *35èmes Journées d'Études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*, volume 2, pages 12–13, ATALA & AFPC.

Esser, Frank and Andrea Umbricht. 2014. The evolution of objective and interpretative journalism in the western press: Comparing six news systems since the 1960s. *Journalism & Mass Communication Quarterly*, 91(2):229–249.

Jacovi, Alon and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Jiang, Nan-Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Katari, Rohan and Madhu Bala Myneni. 2020. A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–5, IEEE.

Koren, Roselyne. 2004. Argumentation, enjeux et pratique de l'«engagement neutre»: le cas de l'écriture de presse. *Semen. Revue de sémio-linguistique des textes et discours*, (17).

Krishna, Satyapriya, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*.

Krüger, Katarina R, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede. 2017. Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering*, 23(5):687–707.

Latkin, Carl A, Lauren Dayton, Justin C Strickland, Brian Colon, Rajiv Rimal, and Basmattee Boodram. 2023. An assessment of the rapid decline of trust in us sources of public information about covid-19. In *Vaccine Communication in a Pandemic*. Routledge, pages 22–31.

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

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lyu, Qing, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723.

Manna, Supriya and Niladri Sett. 2024. Faithfulness and the notion of adversarial sensitivity in nlp explanations. *arXiv preprint arXiv:2409.17774*.

Martin, Louis, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *ACL*, pages 7203–7219.

Mieleszczenko-Kowszewicz, Wiktoria, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanislaw Wozniak, Ewa Dziecіol, Przemyslaw Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *NLPerspectives@ ECAI*.

Mitrut, Oana, Gabriela Moise, Alin Moldoveanu, Florica Moldoveanu, Marius Leordeanu, and Livia Petrescu. 2024. Clarity in complexity: how aggregating explanations resolves the disagreement problem. *Artif. Intell. Rev.*, 57(11).

Müller, Sebastian, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. 2023. An empirical evaluation of the rashomon effect in explainable machine learning. In *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part III*, volume 14171 of *Lecture Notes in Computer Science*, pages 462–478, Springer.

Newman, Nic, Richard Fletcher, Craig T Robertson, A Ross Arguedas, and Rasmus Kleis Nielsen. 2024. *Reuters Institute digital news report 2024*. Reuters Institute for the study of Journalism.

Pirali, Camille, Thomas François, and Núria Gala. 2022. Paddle: a platform to identify complex words for learners of french as a foreign language (ffl). In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 46–53.

Plank, Barbara. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Schudson, Michael. 2001. The objectivity norm in american journalism. *Journalism*, 2(2):149–170.

Sen, Cansu, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Association for Computational Linguistics, Online.

Singh, Roshan, Soon Ae Chun, and Vijay Atluri. 2020. Developing machine learning models to automate news classification. In *Proceedings of the 21st Annual International Conference on Digital Government Research*, pages 354–355.

Todirascu, Amalia. 2019. Genre et classification automatique en tal: le cas de genres journalistiques. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (78).

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Vernier, Matthieu, Laura Monceaux, and Béatrice Daille. 2009. Deft'09: détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique. In *Atelier Défi Fouille de Textes (DEFT'09)*, pages 101–112.

Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Watson, Matthew, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. 2022. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1524–1533, IEEE.

Weber-Genzel, Leon, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation. *arXiv preprint arXiv:2403.01931*.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Yeh, Chih-Kuan, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32.

Zhao, Yang, Li Du, Xiao Ding, Kai Xiong, Zhouhao Sun, Shi Jun, Ting Liu, and Bing Qin. 2024. Deciphering the impact of pretraining data on large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9386–9406.