Explaining Explanations: Interpretability Methods for Discourse Analysis of Transformer Attention Maps

L. Escouflaire[1,2], **J. Bogaert**[2], **A. Descampe**[2], **C. Fairon**[2], **F.-X. Standaert**[2]
[1]Massachusetts Institute of Technology, [2]Catholic University of Louvain
escouflaire.louis@hotmail.fr

Abstract

While LLMs have achieved state-of-the-art performance in NLP, their opacity hinders a human understanding of their predictions. Standard explainability techniques often prioritize technical faithfulness over linguistic plausibility. This paper argues for an interdisciplinary approach that integrates discourse Analysis to critically interpret model explanations. We conduct a case study using CamemBERT, fine-tuned to classify French journalistic texts as *news* or *opinion*. We employ Layer-wise Relevance Propagation to generate attention maps for 1,000 test articles and analyze the token-level relevance scores through both in-depth qualitative analysis and a quantitative ranking of high-attention tokens. Our findings reveal that CamemBERT successfully captures genre-specific linguistic markers: it attends to cues of reported speech and temporal anchors in *news*; and to expressive punctuation, evaluative adjectives, and first-person pronouns in *opinion*. The discourse-analytic lens moves us beyond superficial observations, demonstrating how the model interprets features like punctuation as structural or stylistic conventions. We argue that integrating linguistic expertise into the explainability pipeline yields more nuanced, human-readable explanations.

Keywords: explainability, attention maps, journalistic discourse, discourse analysis

1. Introduction

Transformer-based large language models have pushed the performance limits on a variety of natural language processing (NLP) tasks. Despite their success, these models often function as 'black boxes', making it difficult to understand how they reach their predictions (Lipton, 2018). To address this challenge, researchers have developed various methods aimed at improving model interpretability, such as attention visualization (Clark et al., 2019) and layer-wise relevance propagation (LRP) (Chefer et al., 2021), which provide insights into the internal decision-making processes of these models. Yet, the plausibility and faithfulness of the explanations provided by these techniques are still debated, and most work has focused on technical metrics rather than human-centered understanding (Lyu et al., 2024). Very few studies explore what these explanations actually mean from a linguistic or discourse-analytic perspective.

In this paper, we argue that discourse analysis offers a valuable, underutilized lens for interpreting model explanations in the form of attention maps. Analyzing token-level explanations of predictions made by transformer models with human and linguistic insights enables us to link attention patterns to textual organization, pragmatic functions, and genre conventions. We propose that the intervention of experts can enhance explainability by uncovering the linguistic structures and stylistic signals that transformer models may implicitly rely on.

To demonstrate this, we conduct a case study on the model CamemBERT (Martin et al., 2020), fine-tuned on 8,000 to classify French-language journalistic texts into two genres: *opinion* and *news*. Using the LRP method to generate attention maps for 1,000 articles, as the one in Figure 1, we analyze which tokens the model attends to most strongly and how these patterns correlate with discourse-level features. We present two complementary approaches: on the one hand, an in-depth qualitative analysis of a sample of attention maps to identify trends in the discourse structures to which the model gives the most attention; on the other hand, an examination of the tokens that receive the most attention for each class in all 1,000 articles, allowing us to pinpoint broader patterns. This second method was successfully used and illustrated by Bogaert et al. (2024). The results of our case study reveal that the model captures genre-specific linguistic markers for our task: cues of reported speech, deictics, and temporal anchors in *news*; expressive punctuation evaluative adjectives, modal adverbs, and first-person pronouns in *opinion*.

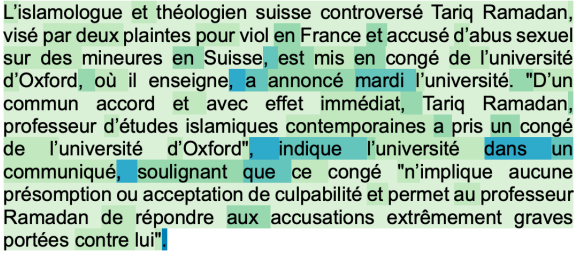

Figure 1: LRP attention map of an extract of a *news* article published in 2017 by *RTBF*, classified as *news* by CamemBERT.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 107–116
11–16 May 2026.

At the same time, our work illustrates the limits of purely visual or statistical explanation tools. High attention to structurally central or frequent tokens, such as punctuation marks, may appear superficial unless interpreted through a discourse-analytic framework. By integrating linguistic insight into the interpretability pipeline, we move from merely *explaining predictions* to critically *explaining explanations*. We argue that such an interdisciplinary approach can produce more nuanced, human-readable explanations, particularly in NLP applications that involve text classification tasks where explainability is an ethical requirement.

2. Prior Work

2.1. Explainability of Transformer-based Classifiers

LLMs built on the transformer architecture introduced by Vaswani et al. (2017) have achieved state-of-the-art performance across a wide range of NLP tasks, including text classification (Minaee et al., 2021; Acheampong et al., 2021). Unlike autoregressive models such as GPT-4, which are trained to predict the next token in a sequence using a unidirectional (left-to-right) context, BERT-based models rely on masked language modeling and are bidirectional, allowing them to capture contextual information from both sides of a token. This bidirectional architecture makes them particularly well-suited for document-level classification tasks, where understanding the full context is crucial for assigning an accurate label. However, despite their empirical success, these models remain largely opaque, raising ongoing concerns about the explainability of their predictions (Clark et al., 2019; Rogers et al., 2021). In response, a growing body of research has explored methods to generate post hoc explanations that can help interpret model behavior (Lyu et al., 2024).

Two widely discussed criteria in the explainability literature are plausibility and faithfulness. Plausibility is the degree to which an explanation is intuitive or convincing to human observers (Agarwal et al., 2024); faithfulness represents the extent to which an explanation accurately reflects the internal mechanisms used by the model to reach its prediction (Jacovi and Goldberg, 2020). Many explanation methods, such as input erasure, saliency maps, or attention weight visualizations, perform well on plausibility but fall short on faithfulness (Lyu et al., 2024). Among these, attention has often been proposed as a built-in explanation cue in transformer models, since self-attention mechanisms determine how tokens attend to each other during processing (Clark et al., 2019). However, several studies have shown that raw attention weights do not always correlate well with actual model decisions, casting doubt on their reliability as faithful explanations (Bibal et al., 2022). This trade-off between plausibility and faithfulness has motivated new approaches, such as gradient-based methods or Layer-wise Relevance Propagation (LRP), aimed at improving both dimensions simultaneously (Chefer et al., 2021). Attention maps, such as the one in Figure 2, have been used to visualize and analyze the behavior of models like BERT, offering insights into word importance (Sen et al., 2020).

In this paper, we aim to show that discourse-informed qualitative analysis of attention maps can uncover systematic behaviors in BERT, contributing to a deeper understanding of how transformer models make predictions, and helping to bridge the gap between plausibility and faithfulness in model explanations.

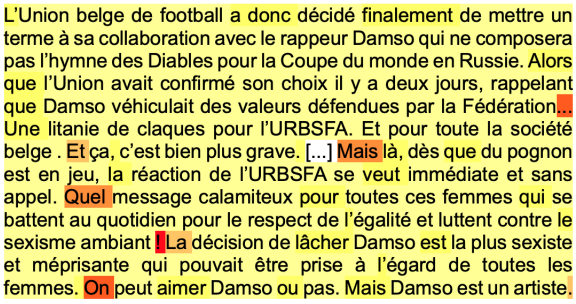

Figure 2: LRP attention map of an extract of a *news* article published in 2018 by *L'Avenir*, classified as *opinion* by CamemBERT.

2.2. News vs. Opinion Classification

Traditionally, the distinction between the *news* (e.g., straight news, reportage) and opinion genres (e.g., editorials, op-eds) in journalism rests on the professional ideal of objectivity (Schudson, 2001). Although widely contested concept in field of journalism studies (Muñoz-Torres, 2012), objectivity continues to function as a structuring principle in journalistic practice and audience expectations. Much like the subjective *opinion* genre, in manuals of journalistic writing, news reporting is considered as a distinct and codified genre, with its specific stylistic, rhetorical, and structural norms, found for example in journalism textbooks (Koren, 2004).

From a computational perspective, *news* vs. *opinion* text classification is often framed as a preliminary subtask of subjectivity detection within sentiment analysis (Ravi and Ravi, 2015). The goal of this subjectivity classification task is to determine whether an article belongs to one or another of these macro-genres of journalism. This document-level task can be achieved using traditional NLP approaches relying on lexicons of subjective terms

and manually crafted features, often using syntactic, lexical, or discourse-level indicators (Krüger et al., 2017; Alhindi et al., 2020). Features such as evaluative adjectives, modal verbs, or discourse markers can help distinguish between neutral reporting and opinionated commentary (Benamara et al., 2017; Todirascu, 2019).

In recent years, however, these traditional methods have been largely overtaken by deep learning approaches, especially those based on transformer architectures like BERT and RoBERTa (Otter et al., 2020). Fine-tuning these models on annotated datasets enables them to capture complex patterns of subjectivity and stance, yielding impressive performance for sentiment analysis and text classification tasks (Batra et al., 2021; Joseph et al., 2022). LLMs have set new benchmarks for subjective text classification in journalistic corpora (Deping et al., 2021), in English and in less-resourced languages. As stated in section 2.1, this boost in performance comes at the cost of reduced explainability, which is crucial for ethically sensitive tasks like journalistic genre classification.

To address this, the present study proposes methods for uncovering the linguistic and structural patterns captured by transformer models fine-tuned to distinguish between *news* and *opinion*, which can be applied to other text classification tasks.

3. Methods

3.1. Data

We use a corpus of 10,000 press articles published online between 2013 and 2023 by 4 Belgian media (*RTBF Actus*, *Le Soir*, *La Libre Belgique* and *L'Avenir*) and 4 Canadian French (Quebec) media (*La Presse*, *Le Devoir*, *Le Journal de Montréal* and *Le Soleil*). The corpus contains 1,250 articles of each media, and it is completely balanced between *news* and *opinion* articles (5,000 of each). Articles were collected using scraping tools and were classified based on the genre labels provided by the media websites. We selected the final articles by applying the LDA-based topic modeling method described in Escouflaire et al. (2024a), which allows for a balanced corpus in terms of topic distribution among the two genres. The corpus is divided in a training set (8,000 articles), a development set (1,000) and a test set (1,000), all containing balanced amounts of *news* and *opinion* articles.

3.2. Classification Model and Explainability Method

We use CamemBERT (Martin et al., 2020), a transformer model based on the RoBERTa architecture (Liu et al., 2019) and pre-trained on French data.

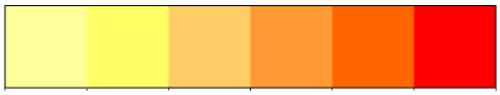

Figure 3: Shades of color used to highlight tokens following various degrees of relevance in the LRP attention maps, for the *opinion* class.

Figure 4: Shades of color used to highlight tokens following various degrees of relevance in the LRP attention maps, for the *news* class.

We fine-tune the case-insensitive base version of CamemBERT (110 million parameters) on our training corpus of 8,000 articles for our binary text classification task, with the hyperparameters used by Bogaert et al. (2024) for the same task: learning rate of 2×10^{-5}, batch size of 4, and 2 epochs.

We selected CamemBERT over other French or multilingual transformer models due to its compatibility with the explainability method we use: Layer-wise Relevance Propagation (LRP). LRP is a local explanation method that attributes a model's prediction to individual input features (in our experiments, text tokens) by propagating relevance scores backward from the output layer through the network. Initially introduced by Bach et al. (2015) and later adapted for text classification by Arras et al. (2017), LRP works by redistributing relevance scores across layers while maintaining conservation principles. We chose to use this explanation method because it provides a relevance value for each token, indicating its contribution to the model's final prediction. Methods like LRP, which rely on backpropagation, provide transparent insights into a model's internal decision-making process and can be visualized as attention maps for easier human interpretation. Backpropagation remains among the most reliable and interpretable methods for generating token-level explanations of deep learning models' predictions with strong faithfulness (Lyu et al., 2024).

For our experiments, we use Chefer et al. (2021)'s implementation of LRP to generate attention-based explanations of predictions made by our fine-tuned CamemBERT models for the 1,000 articles in the test set.

To make the model's explanations accessible and interpretable to human readers, we present them using the attention map format (Reif et al., 2019). While attention maps offer primarily localized, token-level insights and may not fully capture

higher-level features such as long-range dependencies or syntactic structures, we argue that their visual clarity makes them particularly useful for discourse analysis. In these visualizations, tokens are color-coded according to their relevance in the model's prediction: for *opinion* predictions, we use a gradient ranging from red (most relevant) to orange and yellow (least relevant), while for *news* predictions, the gradient spans from blue (most relevant) to teal and green (least relevant), as detailed in Figures 3 and 4. The LRP method distributes the model's attention across all tokens in the predicted text, meaning that every token is highlighted (even those receiving very low attention scores). Tokens with the highest relative attention are rendered in the darkest shades of color, making them visually prominent. To enhance the readability of the attention maps, spaces between tokens (which CamemBERT does not consider in its textual modeling) are shaded with the same color as the token that directly precedes them.

4. Results and Discussion

In this section, we present the results of our case study in two phases. First, we demonstrate how the qualitative analysis of a random sample of attention maps allows us to identify regularities in the model's explanations. Then, we show that the complementary analysis of the tokens receiving the most attention in the model's explanations of our test set confirms and expands on the results of the qualitative analysis.

4.1. Qualitative Analysis

We randomly selected a balanced sub-sample of 250 articles from the test set (125 *opinion* and 125 *news* articles) and qualitatively examined, using a discourse analysis framework, the attention maps generated using the LRP method from the predictions made by CamemBERT for these 250 texts. In this section, we describe the patterns that we identified through the analysis of this sub-sample. To illustrate our findings, we present a selection of six attention maps, two of which represent explanations of extracts of texts that were misclassified by the model. The model reached a prediction accuracy of 93.2% when classifying the 1,000 articles of the test set.

Across the sub-sample, we observed two broad types of attention map configurations: focused maps, where high attention is concentrated on a small number of tokens, and diffuse maps, where attention is spread more evenly across the text. Importantly, these configurations did not appear to correlate with the genre of the text (i.e., *opinion* vs. *news*). For example, Figure 2 shows a focused attention pattern in an extract classified *opinion*, where the model's attention is sharply directed toward expressive punctuation marks, pronouns and connectives. Figure 5 is also an example of a focused map, but in a *news* article.

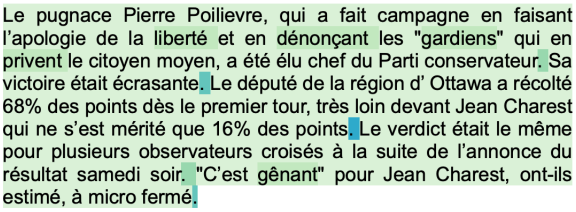

Figure 5: LRP attention map of an extract of a *news* article published in 2022 by *Le Devoir*, classified as *news* by CamemBERT.

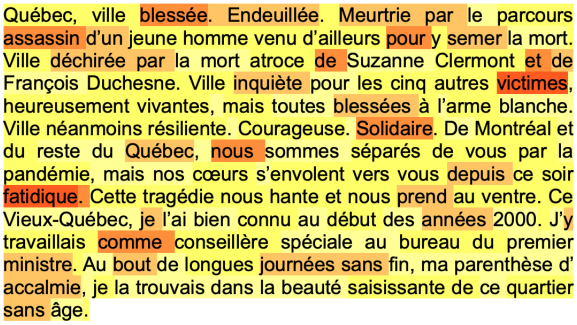

Figure 6: LRP attention map of an extract of an *opinion* article published in 2020 by *Le Journal de Montréal*, classified as *opinion* by CamemBERT.

In many of the attention maps examined, we observed that CamemBERT often directs a significant portion of its attention to punctuation marks, particularly separators (commas, colons, semicolons) and final punctuation (periods, question marks, exclamation marks). While this is somewhat expected for expressive or argumentative punctuation (e.g., *?* and *!* as markers of opinion), it is more surprising in cases where attention is overwhelmingly focused on neutral punctuation like periods and commas, with minimal attention given to the surrounding lexical tokens. This pattern was especially frequent in explanations for texts classified as *news*. Previous studies (Clark et al., 2019; Kovaleva et al., 2019) have documented similar behavior in transformer models and suggest that highly frequent tokens, such as the period, may receive undue attention due to their overrepresentation in training corpora. Another plausible explanation is that these punctuation tokens serve a structural function in the model's internal representations, acting as anchors or aggregators for segment-level information. In the context of our classification task, we propose that the model may interpret the consistent use of neutral punctuation as an implicit marker of *news* style,

contrasting with the expressive or rhetorical punctuation more typical of opinion pieces. Thus, while the model's attention to such tokens may seem superficial, it may in fact reflect a learned sensitivity to genre-specific stylistic conventions. However, because of the opacity of models like CamemBERT, confirmation of this hypothesis remains beyond the scope of our current investigation.

Through our qualitative analysis, *opinion* markers most strongly highlighted by the model appear diverse and reflect a sensitivity to various linguistic features commonly associated with subjective and argumentative discourse. In Figure 2, the model focuses its attention on expressive punctuation (*!* and *...*), exclamative *quel* ("what"), indefinite pronoun *on* ("one"), and logical connectives *et* ("and") and *mais* ("but"). Then, in Figure 6, attention is concentrated on adjectives (*blessée*, *solidaire*, *fatidique*) and axiological nouns, with a strong positive or negative connotation (*assassin*, *victimes*), as well as on first person pronouns (*nous*, *je*). These patterns, frequently observed across the *opinion* texts in the sub-sample, suggest that CamemBERT leverages lexical and stylistic cues that are strongly correlated with the expression of opinion in journalistic writing.

On the other side, different attention patterns appear to be more characteristic of texts classified as *news*. In Figure 7, the model assigns high attention to explicit mentions of information sources (*universités*, *chercheurs*), which reflects a typical requirement of journalistic objectivity. Similarly, Figure 1 shows attention focused on quotation marks and reporting verbs, such as *a annoncé* ("announced") and *indique* ("indicates"), which are conventional markers of direct or indirect speech in news discourse. This example shows that CamemBERT is able to capture the presence of reported speech in a sentence, even without the presence of quotation marks, as almost no attention is given to tokens belonging to the quoted portions of the text. The model also often highlights temporal expressions, such as days of the week (e.g. *mardi* in Figure 7), which function as non-deictic temporal anchors situating the narrative within a specific context. These temporal markers are emblematic of so-called "hot news" articles, which report on recent events with immediate relevance, as opposed to "cold news" that engages in deeper analysis or commentary (Tuchman, 1973; Pilmis and Matthews, 2014). The model's attention maps highlighting these different elements suggests that it has captured the formal structure and evidentiary grounding of news reporting, in contrast with subjective writing.

This distinction is particularly salient in Figure 8, an *opinion* article in which CamemBERT exhibits a nuanced attention pattern similar to that seen in news texts: in the sections containing indirect speech, attention is focused on reporting verbs

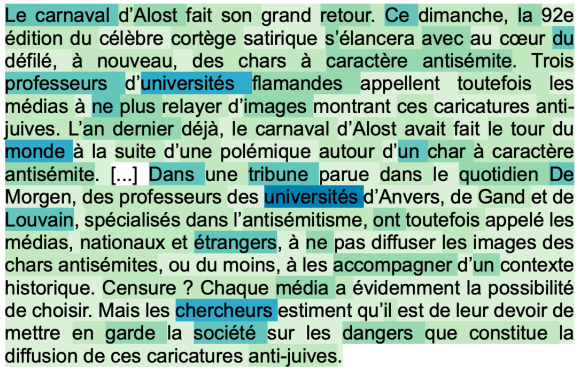

Figure 7: LRP attention map of an extract of a *news* article published in 2020 by *RTBF*, classified as *news* by CamemBERT.

rather than on the reported content itself. Yet despite the presence of reported speech and attributions, the model correctly classifies the article as opinion. A closer look reveals that, unlike in Figure 1, where the citation verbs are neutral ("announce", "indicate"), the verbs used in Figure 8 carry a clear evaluative stance. The repeated use of the verb *prétend* ("pretend") suggests the journalist's skepticism toward the reported statements, while the phrase *tente de répliquer* ("tries to respond"), in which the first verb is also given attention, implies an unsuccessful argumentative effort. These subtle lexical connotations appear to be captured by the model, indicating that beyond merely identifying quotation structures, CamemBERT may have learned to interpret their pragmatic and rhetorical functions. Additionally, in this example and in others found throughout the sub-sample, the model assigns notable attention to elements relative to the stylistic and argumentative structure of the article, such as the anaphoric repetition of the words *faux* and *encore faux* ("false again"), systematically highlighted in the attention map in Figure 8. These attention patterns suggest that CamemBERT is not only sensitive to overt markers of *opinion* but also to the deeper architecture of the text and to evaluative nuances embedded in journalistic discourse.

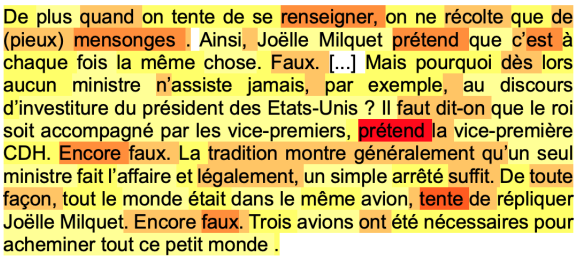

Figure 8: LRP attention map of an extract of an *opinion* article published in 2013 by *RTBF*, classified as *opinion* by CamemBERT.

4.2. Token Attention Distribution

To find out which tokens are the most important overall in guiding CamemBERT towards one genre or another, we measure the average attention attributed to each token occurring at least 50 times in the 1,000 attention maps generated by applying the LRP method to the predictions of the fine-tuned CamemBERT model on all 1,000 articles of the test set (Zhou et al., 2022). We then rank all tokens based on their average attention, in descending order. Then, we analyze the ranked list of tokens in order to identify linguistic patterns which can be converted into features. We restrict our analysis to the fifty tokens which receive the most attention for each class (*opinion* and *news*). The two resulting lists are presented in table 1.

The tokens receiving the most attention in texts classified as *news* reveal a strong alignment with the observations made in section 4.1, as well as with lexical and structural conventions of objective journalism. Among the most salient tokens in the *news* list of Table 1 are verbs of attribution and citation such as *explique* ("explains"), *indiqué* ("indicated") and *déclaré* ("declared"), which are used to convey reported speech from external sources and frequently appear in the present or past perfect tense. These are often accompanied by auxiliary verbs like *a* or *a-t-il*, used in compound tense constructions that lend a neutral, factual tone to the narration. The list also includes nouns that refer to information sources or institutional entities, such as *données* ("data"), *informations*, *enquête* ("investigation"), *bureau*, and *presse*, which recall the evidentiary focus typical of news discourse. Prepositions such as *selon* ("according to") also figure prominently, as they serve to attribute claims or facts to those specific sources or actors, thereby reinforcing the article's impression of objectivity. Temporal markers are also highlighted: on the one hand, absolute temporal markers, e.g. *vendredi* ("Friday") and *décembre*, on the other hand, deictic temporal markers, e.g. *hier* ("yesterday"), *nuit* ("night") and *matin* ("morning"). The presence of these markers in the list suggests the importance of temporal anchoring in *news* texts, which was seemingly encoded by CamemBERT. The abbreviation *h* for *heure* ("hour"), which also appears in the list, is another direct time reference, appearing in expressions such as *à 3h du matin* ("at 3 in the morning"). Surprisingly, some modal adverbs like *notamment* ("especially") and *également* ("also") are strongly weighted towards *news*. It can be argued that unlike other adverbs, which are more likely to be markers of the author's subjectivity, these adverbs are often used for illustrating or enumerating items or events. Notably absent from the *news* top-ranking tokens are adjectives, pronouns, or determiners, confirming that CamemBERT avoids focusing on overt markers of subjectivity.

In contrast, the tokens most strongly associated with the *opinion* class accordingly reflect a more subjective and evaluative mode of expression. First-person pronouns and determiners such as *je* ("I"), *nous* ("we"), *mon* and *mes* ("my") are among the most heavily weighted, clearly indicating the author's presence and personal involvement in the discourse. The indefinite pronoun *on* ("we"/"one"), frequently used to mark distanciation and in informal writing, also ranks highly, showing that its contribution to a more interpretative tone is captured by CamemBERT. Modal adverbs like *bien* ("well"), *mal* ("badly"), *seulement* ("only"), *presque* ("almost") or *trop* ("too much") function as intensifiers or attenuators, usually displaying evaluative stance or opinionated commentary. The presence in the list of the deictic adverb *maintenant* implicates a strong temporal immediacy in discourse, while that of the vague spatial marker *partout* suggests a spatial approximation which would not be acceptable for factual *news* texts. Both of these dimensions appear to be important for CamemBERT's classification of text genres. Argumentative markers are also prominent: conjunctions such as *parce (que)* ("because"), *si* ("if"), *comme* ("like"/"as"), and *puis* ("then"), are important cues for the model's predictions of the *opinion* class. Finally, many abstract nouns found in the *opinion* list of Table 1, such as *manière* ("manner"), *position*, *idée* ("idea"), *liberté* ("freedom"), and *message* point to a reflective and conceptual discourse, which is opposed to concrete event reporting, and may be therefore leveraged by CamemBERT in its classification.

Then, a contrastive analysis of tenses and moods of the verbs found in the lists of Table 1 further reinforces the linguistic divergence between the two classes. For CamemBERT's attention when predicting *news*, the most prominent verb forms include the present and the past perfect, typically used to report facts or statements made by others in a neutral tone. The future tense (*sera*, *seront*) is also present, often employed to project expected developments in factual reporting. The imperfect appears once in the list (*avaient*), probably as part of the common compound verbal structure of the past perfect ("plus-que-parfait"), anchoring events within a coherent chronological framework. In *opinion* articles, by contrast, verbs are less frequently given attention, suggesting a more nominal or adjectival style captured by CamemBERT. When present, verb forms tend to include the simple present to express opinions or modality, e.g., *semble* ("seems"), possibility, e.g. *pu* ("been able to") or obligation, e.g., *faut* ("must"), and the conditional mood with *serait* ("would be") which can be used to introduce hypotheses, soften claims, or create rhetorical distance, reinforcing the typical

News		Opinion	
explique (explains)	*indiqué* (indicated)	*parce* (because)	*passe* (goes/happens)
mardi (Tuesday)	*vendredi* (Friday)	*message* (message)	*seulement* (only)
déclaré (declared)	*a-t-il* (did he)	*maintenant* (now)	*nom* (name)
selon (according to)	*jeudi* (Thursday)	*manière* (manner)	*partout* (everywhere)
lundi (Monday)	*mercredi* (Wednesday)	*parti* (party)	*je* (I)
annoncé (announced)	*samedi* (Saturday)	*mon* (my)	*comme* (like/as)
hier (yesterday)	*dimanche* (Sunday)	*droits* (rights)	*sait* (knows)
Mme (Mrs.)	*également* (also)	*vie* (life)	*si* (if)
h (h)	*toutefois* (however)	*sans* (without)	*cause* (cause)
seront (will be)	*juin* (June)	*tant* (so much)	*risque* (risk)
matin (morning)	*données* (data)	*position* (position)	*idée* (idea)
a (has)	*annonce* (announcement)	*mes* (my)	*français* (French)
décembre (December)	*notamment* (especially)	*nous* (we)	*grands* (great)
soir (evening)	*informations* (information)	*pu* (been able to)	*école* (school)
sera (will be)	*presse* (press)	*presque* (almost)	*économique* (economic)
bureau (office)	*enquête* (investigation)	*faut* (must)	*bien* (well)
dit (says)	*avaient* (had)	*N-VA* (N-VA)	*gouvernement* (government)
site (site)	*nuit* (night)	*celui* (that one)	*trop* (too much)
ancien (former/ancient)	*reçu* (received)	*serait* (would be)	*jeu* (game)
après (after)	*entreprise* (company)	*on* (one/we)	*surtout* (especially)
partir (to leave)	*juillet* (July)	*question* (question)	*même* (same/even)
suite (following)	*lors* (during)	*liberté* (freedom)	*semble* (seems)
direction (direction)	*parmi* (among)	*Ottawa* (Ottawa)	*mal* (badly)
appel (call)	*secteur* (sector)	*droit* (right/law)	*pouvoir* (power)
groupe (group)	*janvier* (January)	*puis* (then)	*plutôt* (rather)

Table 1: Top-50 tokens with the highest average relevance values in attention maps derived from CamemBERT predictions of the test set, grouped by genre and with their English translation.

reflexivity of opinionated press discourse.

Finally, some thematic observations drawn from the high-attention tokens in Table 1 also point to potential biases in the corpus used for the experiment. Tokens frequently highlighted in *opinion* texts include political references, such as *parti* ("party"), *gouvernement* ("government"), *N-VA* (name of the Belgian right-wing party), and *Ottawa* (city where the Canadian government is located), suggesting a prevalence of politically oriented opinion articles. On the other hand, tokens associated by CamemBERT with the *news* class often relate to the economic or industrial domain, including *entreprise* ("company"), *groupe* ("group"), and secteur ("sector"). While the dataset was constructed to be as balanced as possible in terms of genre, as described in section 3.1, the presence of these tokens in Table 1 may indicate a slight thematic imbalance that could have influenced how CamemBERT learned to differentiate the two classes. The model might be partially relying on topic-specific lexical cues rather than purely stylistic or linguistic features.

5. Conclusions

This paper has introduced an interdisciplinary approach to transformer model explainability by applying discourse analysis to transformer-based attention explanations. Through a case study on French text classification, we demonstrated that human interpretive insight, grounded in linguistic theory, can significantly enrich our understanding of what attention maps reveal about model behavior.

Our study faces some limitations. The interpretive human analysis of the attention maps is partially subjective and dependent on linguistic expertise, which may reduce reproducibility. The results of our case study, although the corpus is representative of Belgian and Canadian French journalistic writing, are restricted to two genres (news and opinion), one language (French), one model (CamemBERT) and one explanation method (LRP), limiting the generalizability of our findings. Further research should investigate to what extent the results would differ when using other models and explanation methods, and how they would apply to other classification tasks and languages. Additionally, LRP and attention-based methods do not offer a full picture of model internals, especially when multiple heads and layers are involved. While attention-based backpropagation may not be the best method for transformer model explainability (Bibal et al., 2022), it still remains one of the most faithful ones available today (Lyu et al., 2024). While this article is focused on attention-based explanations, we previously explored human-based and feature-based classification and explanation for the same task and on similar data (Escouflaire et al., 2024b; Bogaert et al., 2024).

Our main contributions are threefold. First, we propose the integration of discourse analysis as a

complementary method for interpreting attention-based explanations, especially in text classification tasks where textual genre and style are relevant. Second, we show that transformer models like CamemBERT capture discourse-level and pragmatic features, such as subjectivity markers, reported speech and discourse structure, in addition to surface-level lexical cues. Third, we illustrate the limitations of purely visual or token-level explanation techniques, arguing that their plausibility is enhanced when examined with linguistic or contextual knowledge.

6. Ethics Statement

This research is primarily focused on model explainability, a domain with significant ethical implications. By providing a discourse-analytic framework for interpreting LLM predictions, our work directly contributes to the ethical requirements of transparency and accountability in text classification. The fine-tuned CamemBERT model is used for classifying journalistic genres (news vs. opinion). While this task is not inherently high-risk, a lack of transparency could mask learned biases. Our method, by revealing the specific linguistic cues the model relies on, allows researchers and practitioners to identify and mitigate unintended bias. For example, if the model disproportionately relies on specific regional language features or stylistic habits learned from the training data, our method makes this reliance visible. The corpus of 10,000 French press articles used in this article was collected using scraping tools. The articles were classified based on publicly available genre labels provided by the media websites. Our code is publicly available on Github (the link will be attached in the final version of this paper).

7. Bibliographical References

Francisca Adoma Acheampong, 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.

Josh Achiam, 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*.

Chirag Agarwal, 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*.

Tariq Alhindi, 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.

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

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Himanshu Batra, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. Bert-based sentiment analysis: A software engineering perspective. In *Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27–30, 2021, Proceedings, Part I 32*, pages 138–148. Springer.

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

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900.

Jérémie Bogaert, Marie-Catherine de Marneffe, Antonin Descampe, Louis Escouflaire, Cédrick Fairon, and François-Xavier Standaert. 2024. Sensibilité des explications à l'aléa des grands modèles de langage: le cas de la classification de textes journalistiques [sensitivity of explanations to the randomness of large language models: a case study on journalistic text classification]. *Traitement Automatique des Langues*, 64(3):15–40.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,

Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Kevin Clark, 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*.

Lin Deping, Wang Hongjuan, Liu Mengyang, and Li Pei. 2021. News text classification based on bidirectional encoder representation from transformers. In *2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)*, pages 137–140. IEEE.

Louis Escouflaire, Antonin Descampe, and Cédrick Fairon. 2024a. Automated text classification of opinion vs. news french press articles. a comparison of transformer and feature-based approaches. *Language & Communication*, 99:129–140.

Louis Escouflaire, Antonin Descampe, and Cédrick Fairon. 2024b. 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).

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

Jyothis Joseph, S Vineetha, and NV Sobhana. 2022. A survey on deep learning based sentiment analysis. *Materials Today: Proceedings*, 58:456–460.

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

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Katarina R Krüger, 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.

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

Yinhan Liu, 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*.

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

Louis Martin, 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.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Juan Ramón Muñoz-Torres. 2012. Truth and objectivity in journalism: Anatomy of an endless misunderstanding. *Journalism studies*, 13(4):566–582.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Olivier Pilmis and Toby Matthews. 2014. Producing in urgent situations. *Revue française de sociologie*, 55(1):101–126.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. *Advances in Neural Information Processing Systems*, 32.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

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

Cansu Sen, 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, Online. Association for Computational Linguistics.

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

Gaye Tuchman. 1973. Making news by doing work: Routinizing the unexpected. *American journal of Sociology*, 79(1):110–131.

Miles Turpin, 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yilun Zhou, Marco Túlio Ribeiro, and Julie Shah. 2022. Exsum: From local explanations to model understanding. In *NAACL-HLT*, pages 5359–5378. Association for Computational Linguistics.