An Integrated Approach for Political Bias Prediction and Explanation Based on Discursive Structure

Nicolas Devatine, Philippe Muller, Chloé Braud

An Integrated Approach for Political Bias Prediction and Explanation Based on Discursive Structure

Nicolas Devatine[1], **Philippe Muller**[1,3], **Chloé Braud**[2,3]
[1]IRIT, University of Toulouse
[2]IRIT, CNRS
[3]Artificial and Natural Intelligence Toulouse Institute (ANITI)

firstname.lastname@irit.fr

Abstract

One crucial aspect of democracy is fair information sharing. While it is hard to prevent biases in news, they should be identified for better transparency. We propose an approach to automatically characterize biases that takes into account structural differences and that is efficient for long texts. This yields new ways to provide explanations for a textual classifier, going beyond mere lexical cues. We show that: (i) the use of discourse-based structure-aware document representations compare well to local, computationally heavy, or domain-specific models on classification tasks that deal with textual bias (ii) our approach based on different levels of granularity allows for the generation of better explanations of model decisions, both at the lexical and structural level, while addressing the challenge posed by long texts.

1 Introduction

In an expanding information-based society, where public opinion is influenced by a plurality of sources and discourses, there is growing concern about fair information sharing. Biased speech, slanted presentation of events are inevitable, whether intentional or not, but must be transparent to ensure a more democratic public space. This has motivated substantial work on text classification to identify political orientation, what stances are supported by a text, or to characterize misleading or fake information (Hamborg et al., 2019). It is also important that such methods can provide justifications to their decisions, both to understand what linguistic expressions are characteristic of certain positions, and also to provide some transparency in the analysis itself. Explainability of supervised models is now a large subfield addressing this concern, with methods providing justifications, mostly in the form of relevant tokens in the case of textual tasks, e.g. (Kusner et al., 2015).

In this work, we contribute to both these lines of research by proposing an integrated approach for predicting and explaining political biases, where the structure of the document can inform the proposed bias characterization, as opposed to current approaches only relying on lexical, local cues. Indeed, by focusing on local formulation, existing research (Da San Martino et al., 2020; Field et al., 2018) ignores that political expression also relies on argumentation, i.e. the way information is presented. Example 1 is segmented into Elementary Discourse Units (EDUs), the minimal spans of text to be linked by discourse relations as described e.g. in the Rhetorical Structure Theory (Mann and Thompson, 1988). The discourse structure built upon these segments represents how information is conveyed in a right-leaning text about climate and can inform on how the information is presented (why the climate is not a problem, what opposing argument the writer wants to highlight), and also to detect the most important spans of texts.

Example 1. [*There's nothing abnormal about the weather this January,*]$_1$ [*it's just part of the Earth's natural climate patterns.*]$_2$ [*The mainstream media is just pushing the idea of climate change*]$_3$ *[to push their own agenda.*]$_4$

To the best of our knowledge, we are the first to investigate discourse-based information for bias characterization, and we do so through: (i) a segmentation of the texts based on discourse units rather than sentences, (ii) experiments on discourse connectives that can be seen as shallow markers of the structure, (iii) and crucially, a model based on latent structures, as a proxy for discourse structures, that can help the prediction and provide a different sort of input for explainability methods.

Furthermore, while recent progress on text classification has been largely due to the wide-spread use of pretrained language models, fine-tuned on specific tasks, they remain limited in terms of input size (i.e. 512 sub-tokens in general) and can-

Findings of the Association for Computational Linguistics: ACL 2023, pages 11196–11211
July 9-14, 2023

not easily deal with phenomena that relate elements far apart. Long texts are also problematic for many explanation methods. Our proposed approach addresses this limitation on both sides. The code is available at: https://github.com/neops9/news_political_bias.

Our work makes the following contributions:
- we propose a model to predict political bias of news articles, with unrestricted input length, using latent structured representations on EDUs;
- we propose improvements to perturbation-based explanation methods, using different levels of granularity (i.e. words, sentences, EDUs, or structures);
- we evaluate experimentally our propositions for both the prediction and the explanation of bias.

2 Related work

The prediction of the political orientation in texts has long been of interest in political science (Scheufele and Tewksbury, 2007), and has generated growing interest in NLP, either for classification at document level, e.g. detecting extreme standpoints (Kiesel et al., 2019) or more general left/center/right orientation in news (Kulkarni et al., 2018; Baly et al., 2020; Li and Goldwasser, 2021), but also at a finer-grain local level, locating specific framing (Card et al., 2015; Field et al., 2018), or various linguistic devices such as "propaganda techniques", as in the SemEval 2020 task (Da San Martino et al., 2020). For a more general view, see the survey in (Hamborg et al., 2019). Recently, Liu et al. (2022) have developed a language model over RoBERTa (Liu et al., 2019b), fine-tuned on a large corpus of news to address both stance and ideology prediction, by incorporating new "ideology-driven" pre-training objectives, with very good results. In contrast, we develop a generic approach that could be applied as is to new classification tasks.

Aside from approaches whose objective is just prediction of an orientation, some studies aim at characterizing bias, and rely on lexical statistics or surface cues (Gentzkow et al., 2019; Potthast et al., 2018). In contrast, we want to investigate other factors as well, at a more structural level, mainly document-level organization aka discourse structure. Automated discourse analysis is the subject of a rich body of work but current parsers still have rather low performance and weak generalization. This is why we took inspiration from Liu and Lapata (2018), who use structural dependencies over sentences that are induced while encoding the document to feed downstream supervised models. Their results indicate that the learned representations achieve competitive performance on a range of tasks while arguably being meaningful. This approach is effective for summarization with the learned structures, while less complex than relying on rhetorical relations, capturing consistent information (Liu et al., 2019a; Isonuma et al., 2019; Balachandran et al., 2021). Similar results were found for fake news classification (Karimi and Tang, 2019). Our model relies on these approaches, but adds a finer-grain level of analysis relying on Elementary Discourse Units.

The last aspect of our approach is the use of explainable methods to characterize bias. We propose an integrated approach where a classification model is used with methods to explain its decision, thus providing cues about the way bias is present and detected in texts. Numerous explainability methods have been proposed in recent years, most of which are amenable to being used on text classification tasks. Almost all of them are *local* i.e. provide information about the role of separate parts of the input for a given instance only, e.g. input tokens most relevant to a model's prediction for textual tasks. These methods can be either black box methods, operating only on predictions of the models (Castro et al., 2009; Ribeiro et al., 2016), or can observe the impact of the input on some of its internal parameters (Simonyan et al., 2014; Sundararajan et al., 2017). We extend the use of such methods to take into account structural elements. Although some studies have recently investigated how structural / discourse information is encoded in pretrained languages models (Wu et al., 2020; Huber and Carenini, 2022), to the best of our knowledge, we are the first to explore textual explainable methods not relying only on surface form information. This is crucial for long texts, as methods such as LIME (Ribeiro et al., 2016) that rely on sampling word perturbations can become expensive for high token counts.

3 Integrated bias detection and characterization

Our approach is based on a model that predicts a bias while inducing a structure over documents, and explanation methods that could either take as inputs simply the tokens, the EDUs, the sentences, or that could be based on the induced structures, see Figure 1. In this section, we describe our model

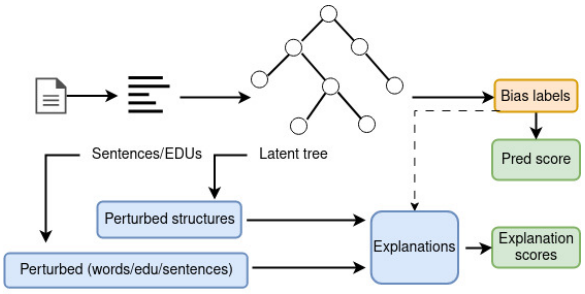

Figure 1: Overview of the approach: a supervised classification model relies on latent structures over textual units, and a module provides perturbation-based explanations, relying on various levels of analysis: words, sentences, EDUs, or latent trees.

for predicting bias, on which we rely to produce structure-based explanations.

3.1 Base Bias Prediction model

In Liu and Lapata (2018), the sentences are composed of sequences of static word embeddings that are fed to a bi-LSTM to obtain hidden representations used to compute the sentence representations, that are then passed through another bi-LSTM to compute the document representation. At both levels, representations are built using the structured attention mechanism allowing for learning sentence dependencies, constrained to form a non-projective dependency tree. Finally, a 2-layer perceptron predicts the distribution over class labels. Note that LSTMs do not have limitations on the input size.

We modify the model to include the improvements proposed by Ferracane et al. (2019). In particular: (i) we remove the document-level bi-LSTM, (ii) for the pooling operation, we aggregate over units using a weighted sum based on root scores, instead of a max pooling, (iii) we perform several additional levels of percolation to embed information from the children's children of the tree, and not only direct children. On top of that, we skip the sentence-level structured attention, as it adds an unnecessary level of composition that was found to have a negative empirical impact on the results.

3.2 Improvements

We make two additional important modifications to the classification model, one generic (replace the base unit of the latent structure), the other specific to the task considered.

Segmentation The learning of a latent structure is supposed to leverage argumentative processes that can reflect the author's political orientation. We thus changed the base textual units from sentences to more discourse-oriented ones, as given by a discourse segmenter. Discourse segmentation is the first stage of discourse parsing, identifying text spans called Elementary Discourse Units that will be linked by discourse relations. We chose to use an existing segmenter (Kamaladdini Ezzabady et al., 2021)[1] as it showed good performance on the latest segmentation shared task (Zeldes et al., 2021), while being the only one from that campaign not needing features other than tokens.

Adversarial Adaptation Media source of an article can be easily determined using some specific lexical cues, such as the media name. Since most articles from a media share the same political label, a model could exploit these features, that wouldn't generalize to other news sources. It is difficult to remove these cues via preprocessing, as they can be various and source-specific. Baly et al. (2020) suggest two approaches: adversarial adaptation (AA) (Ganin et al., 2016), and triplet loss pre-training (Schroff et al., 2015), and chose the latter based on preliminary results, while we found AA more promising. AA involves incorporating a media classifier in the model's architecture and maximizing its loss using a gradient reversal layer, resulting in a model that is discriminative for the main task yet independent of the media source.

4 Lexical and Structural Perturbation-Based Explanations

Among the numerous existing methods for interpreting a model's decision, we chose to focus on so-called black box approaches, only relying on a model output predictions, and not its internal representations, for more generality. However, the most popular black box approaches, LIME (Kusner et al., 2015), Anchor (Ribeiro et al., 2018) or Shap (Lundberg and Lee, 2017) rely on lexical features when applied to textual tasks, looking for relevant subsets of features or using perturbations by removing/switching words in the input which makes them computationally expensive for high token counts, or forces approximation via sampling, which still has to be representative enough to be useful. Of these methods we chose to only consider LIME, which is intrinsically based on sampling and has been shown by Atanasova et al. (2020) to have the

[1] https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut

best or near-best performance on their metrics, and thus present a good compromise.

LIME works by learning a simple model around an instance, which approximates the prediction of the model in the "neighborhood" of the instance. The neighborhood of an instance is sampled by slightly perturbing the input with respect to some features, words in the case of textual models, yielding a set of (perturbed) instances. Then a simple linear model is fitted on these instances to match the model predictions, with a weight given to the instances according to their distance from the original instance. The parameters of the simple model then yield importance scores for the input features, and the best ones are chosen as an "explanation" of the decision on the original instance.

Despite its usefulness, LIME has some known limitations, regarding the cost of the sampling process (Molnar, 2022, section 9.2.5) or the robustness of the explanations (Alvarez-Melis and Jaakkola, 2018). The main issue is that the quality of the explanations highly depends on the amount of generated perturbed samples, to be representative of the model's behavior, and to avoid spurious or not robust explanations. For texts, where features are words, this can mean a high computational cost, especially for long documents, since the number of possible perturbations of a text grows exponentially with its size. We thus propose four strategies to reduce this cost and still produce relevant explanations, by focusing on different levels of granularity.

Token-level explanations The first level still operates at the token level, removing tokens randomly, but focusing on specific words. We consider three subcases: (1) ignoring functional words, less likely to be relevant to a classification decision, while being very frequent; or (2) sampling only with respect to some specific classes of tokens: (2a) named entities extracted with spaCy,[2] and (2b) discourse connectives (Webber et al., 2019), using the extended list of markers[3] proposed by Sileo et al. (2019), that could act as shallow indicators of argumentative structures.

EDU/Sentence-level The second level moves away from word-based explanations to focus on a higher granularity: either sentences, preprocessed using Stanza (Qi et al., 2020), or EDUs to take into account the general organization of the document. EDUs are supposed to be the atomic level of structure analysis, and thus more coherent in terms of size and content than full sentences. The process for generating explanations is then very similar to word-based ones: instead of perturbing a document by removing a random set of words, we remove a random set of EDUs. An EDU-based explanation then consists of a subset of the most impactful EDUs for the model. This also reduces drastically the perturbation space, making it more feasible and reliable to sample.

Two-level explanations Using a higher level of granularity may provide less detailed explanations, we thus propose to combine the previous level of analysis, EDU-based, with the classical word-based approach, restricted to the selected EDUs. In practice, we define a hyperparameter k, apply the first stage of explanation, and then generate word-level perturbations only for words present in the k most impactful EDUs of the explanation.

Structure-Level Explanations Finally, we propose to generate explanations directly at the level of the structure learned by the model, still using the LIME method. Here, we will perturb the entire structure extracted *via* the latent model for a given example (see Section 3.1). We chose to rely on perturbations that remove a subset of head-dependent relations in the original tree, i.e. a pair of segments. An explanation of the structure is then the subset of the most impactful relations in the tree.

By combining all levels of explanation presented, we can generate an enhanced explanation covering multiple aspects of the data (see Figure 2).

5 Explanation evaluation metrics

Evaluating the explanations is an important challenge, and common practices mostly depend on costly human judgments. Here we rely on the diagnostic properties proposed by Atanasova et al. (2020) in the context of text classification. We discarded two measures that cannot be computed: the *agreement with human rationales* measure, since we do not have access to human annotations for the explanation of political datasets, and the *rationale consistency* measure, since it is meant to compare an explanation method across different models. We consider that a document is composed of a set of features, and that our explanation method generates a saliency score for each of them.

[2] https://spacy.io/

[3] https://github.com/sileod/Discovery/blob/master/data/markers_list.txt

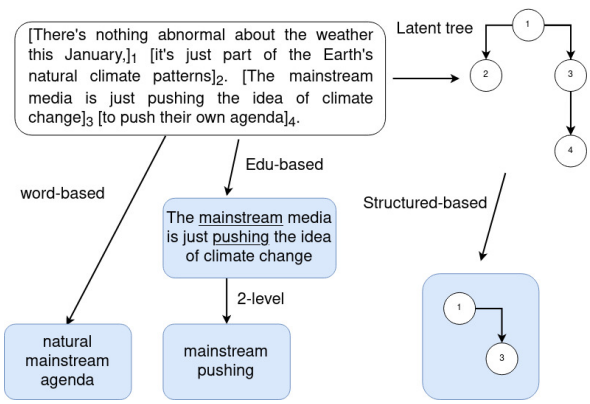

Figure 2: Fabricated examples of generated explanations (blue), according to which part of the input is perturbed to generate the LIME approximation around an instance. Structure-based explanations need the structure produced by the model. Numbers in the structure refers to EDUs.

	#BERT Tokens	#EDUs	#Sent.
Allsides	1257 ± 863	58 ± 44	32 ± 25
C-POLITICS	1008 ± 1106	100 ± 112	20 ± 24
HP	780 ± 691	81 ± 74	25 ± 24

Table 1: Mean and standard deviation for various levels of each dataset: subtokens, EDUs, sentences.

Confidence Indication (CI) When generating an explanation, the feature scores for each possible class can be computed. It is then expected that the feature scores for the predicted class will be significantly higher than those of the other classes. If not, this should indicate that the model is not highly confident in its prediction, and the probability of the predicted class should be low. We can then measure a confidence indication score as the predictive power of the explanation for the confidence of the model. Predicted confidence is computed from the distance between saliency scores of the different classes and then compared to actual confidence by using the Mean Absolute Error (MAE).

Faithfulness Faithfulness is an indication that features selected in an explanation were actually useful for the model to make a prediction. It is measured by the drop in the model's performance when a percentage of the most salient features in the explanation are masked. Starting from 0%, 10%, up to 100%, we obtain the performance of the model for different thresholds. From these scores, the faithfulness is then measured by computing the area under the threshold-performance curve (AUC-TP).

Dataset Consistency (DC) DC measures if an explanation is consistent across instances of a dataset. Two instances similar in their features should receive similar explanations. Similarity between instances is obtained by comparing their activation maps, and similarity between explanations is the difference between their saliency scores. The consistency score is then the Spearman's correlation ρ between the two similarity scores. The overall dataset consistency is the average obtained for all the sampled instance pairs.

6 Datasets

We evaluate the effectiveness of our approaches on three English-language datasets[4] which contain annotations of political leaning (bias) of long news articles, and thus particularly relevant to the context of this study. Lengths of documents are shown in Table 1: *Allsides* and *C-POLITICS* present the longest texts (additional statistics in Appendix A).

Allsides This media-based news articles dataset proposed by Baly et al. (2020)[5] contains $30,246$ articles with 3-class annotations: *left, center, right*. Media present at training time are excluded from evaluation. The articles were crawled from Allsides[6] which is a platform that offers an analysis of the political leanings of various English-language media at the article level. An article is labeled by the political positioning of its media.

Hyperpartisan (HP) A binary classification task (Kiesel et al., 2019) of predicting whether a given news article is hyperpartisan or not (takes an extreme left-wing or right-wing standpoint), task 4 of SemEval-2019. We considered the dataset containing $1,273$ manually annotated articles.

C-POLITICS We built on the large-scale news articles dataset POLITICS[7] (Liu et al., 2022). It comes with an aligned version containing $1,060,512$ clusters of articles aligned on the same story from 11 media. We propose a reduced version of this dataset meeting three desirable constraints: class balance, temporal framing and media-agnostic. We kept only articles published

[4]Distributed under Apache License 2.0, CC BY 4.0 and CC BY-NC-SA 4.0, for *Allsides*, *Hyperpartisan* and *POLITICS* respectively.

[5]https://github.com/ramybaly/Article-Bias-Prediction

[6]https://www.allsides.com

[7]https://github.com/launchnlp/POLITICS

between 2020 and 2021 (annotation stability), excluding the possibility of a media appearing in several splits (train, validation, test) and forcing to have at least one article of each label per cluster (homogeneity). We evaluate on the 3-ways classification task of predicting the political leaning (left, center, right). We ended up with a dataset containing 37,365 articles for 12,455 clusters. An article is labeled by the political positioning of its media. This will be made available upon acceptance.

7 Experimental Settings

Baselines For *Allsides* and *Hyperpartisan*, we compare to the results obtained by the authors of the datasets, and the winners of the task (HP). We also compare to three additional transformer-based baselines on the three tasks, for which we fine-tuned a classification model (on a single run): (1) RoBERTa-base (Liu et al., 2019b) (2) Longformer-4096 (Beltagy et al., 2020), a language model designed to handle very long sequences of text, up to 4096 tokens (3) POLITICS (Liu et al., 2022), a state-of-the-art language model built over RoBERTa-base for political ideology prediction, pretrained on more than 3.6M news articles (see above). RoBERTa and POLITICS are fine-tuned on the whole input using a sliding window of size 512 and an overlap of size 64; we built on Liu et al. (2022)'s implementation[8]. All baselines and proposed models have similar numbers of parameters (cf. the appendix). For the explanations, we compare to the original version of LIME for text classification, which is based on words perturbation, and a random explanation on the whole input.

Settings For the classification model, we built on Ferracane et al. (2019)'s implementation,[9] itself based on Liu and Lapata (2018)'s. We adapted the code according to the modifications and additions proposed in our approach, as detailed in Section 3.1. Hyperparameters were set using grid search and are the same for all tasks (Table 8 in Appendix B). We used pretrained 300D GloVe vectors (Pennington et al., 2014). For the AA training, since the training set may contain many media sources with a long tail distribution, we only consider the 10 most frequent sources. Hyperparameters for the fine-tuning of RoBERTa, POLITICS and Longformer are given in Appendix B. 2-level explanations are generated using the 10 most impactful EDUs.

[8] https://github.com/launchnlp/POLITICS/

[9] https://github.com/elisaF/structured/

Evaluation We evaluate two versions of the classification model: segmentation into sentences, or into EDUs (on a single run). We report accuracy as it is the standard measure in previous work on these tasks. We built on the LIME python package[10] to implement our methods (Section 4). We generate and evaluate explanations on 100 documents from the test set for 1,000 and 10,000 perturbed samples and compute a score for each feature. Explanations are generated for our trained classification model with EDU segmentation (Section 3.1). The confidence interval for the evaluation of the explanations is only given for the baseline (LIME Words) for 10 generations. Since each of the proposed improvements has a reduced perturbation space relative to the baseline, which is the impact factor of the variance, and to avoid a disproportionate computational cost, we consider that the confidence interval will be at worst equal or better, and therefore we do not give it for all experiments.

8 Results

Results obtained for the different classification tasks are given in Table 2. As expected, the fine-tuning of the pre-trained and specialized model POLITICS obtains the best results on all tasks. Followed closely by Longformer with an average of -3.45 points, which shows the interest of keeping the whole document as input. Regarding our structured approaches, we can note that despite lower scores compared to POLITICS and Longformer, the EDU-based version performs better than RoBERTa on corpora with the longest text lengths (i.e. *Allsides* $+1.76$ points, *C-POLITICS* $+4.37$ points). The segmentation into EDUs significantly improves the results on all tasks compared to the segmentation into sentences ($+4.59$ points on average), showing the importance of the fine-grain discourse approach. Putting these results in perspective, our approach is more generic than POLITICS, as it does not require heavy and domain-specific pre-training, and much lighter than Longformer (w.r.t. computational cost).

Table 3 presents the evaluation metrics for each of the proposed LIME alternatives. We observe that in general, except for discourse markers and named entities, the two-level explanation performs better, obtaining strong evaluation scores for all the proposed metrics. The use of a higher level of

[10] https://github.com/marcotcr/lime

granularity (sentences, EDUs) improves the quality of the explanations compared to the baseline; note that between EDUs and sentences, the finer segmentation into EDUs is the most accurate, showing the effectiveness of discourse-based approaches. The higher CI score for EDUs shows that it is the appropriate level of granularity with respect to the impact of their content on the model decision, it is also the level of segmentation on which the model has been trained. Similarly, reducing the perturbation space by targeting classes of words generates better quality explanations, in particular for named entities, which are particularly informative for the model as already shown in the literature (Li and Goldwasser, 2021). Regarding the explanation of the structure, although the scores obtained are in the low range, we can state that they represent relevant information for the decision of the model as compared to baselines. In general, the two-level explanation seems to be the best compromise between explanation quality, computational cost, and level of detail, while the LIME baseline (words) suffers from a high perturbation space.

As we are reducing the sampling space in our approaches, we also made comparisons on the number of samples used to generate the explanation for these metrics, between $1,000$ and $10,000$ samples. We notice that the scores obtained by most of our approaches on $1,000$ samples remain better than those of the baseline for $10,000$ samples. This shows that it is possible to generate good explanations, and often of better quality, with a number of samples 10 times smaller, which is a major improvement over the computational cost.

Model	Allsides	C-POLITICS	HP
Literature			
Baly et al. (2020)	51.41*	-	-
Jiang et al. (2019)	-	-	82.2*
Fine-tuned PLMs			
RoBERTa	52.63	49.24	80.41
Longformer-4096	56.11	55.07	85.23
POLITICS	**60.44**	**60.52**	**85.82**
Structure-based models			
Structured Attention/Sent	48.76	48.57	75.63
Structured Attention/EDU	54.39	53.61	78.73

Table 2: Accuracy% (test set). $*$ indicates results not reproduced, taken from the original papers. Note that POLITICS is based on RoBERTa, and already specifically fine-tuned on political texts before our own fine-tuning.

Explainability technique	CI MAE ↓	F AUC-TP ↓	DC ρ ↑
Random explanation	0.053	47.45	0.010
base LIME (words)	0.036	45.78	−0.003
EDUs	**0.029**	38.80	0.075
Sentences	0.034	37.90	0.014
Structure	0.038	36.00	0.065
2-level EDUs+Words	0.034	36.40	0.131
Words w/o Stopwords	0.031	44.80	0.045
Discourse Markers	0.032	43.14	0.119
Named Entities	0.033	**35.25**	**0.176**

Table 3: Confidence Indication (CI), Faithfulness (F) and Dataset Consistency (DC) scores for the different versions of LIME described in Section 4, on the *Allsides* dataset. For each document, $10,000$ perturbed samples are generated. For "LIME Words", the standard deviation is ± 0.002 for Confidence Indication, ± 2.2 for Faithfulness, and the estimated p-value for the correlation of Dataset Consistency is 0.002.

9 Analysis of explanations

By looking at the explanations generated for the different levels of granularity and properties targeted, we can gain some insights about the model's decisions. An important property that must be fulfilled by the explanation is its comprehensibility by a human in order to characterize biases. We propose a qualitative analysis of the explanations and a comparison of the various approaches, both at the lexical and structural level.

Table 4 shows the most recurrent and impactful words in the explanations, as given by the aggregated saliency scores of the 100 generated explanations, for each class for the *Allsides* task, depending on the method of explanation. Similar results are reported for *Hyperpartisan* and *C-POLITICS* in Table 11 and 12 of the Appendix C. Overall, the words that emerge seem consistent with the classes, and it is relatively straightforward to understand the possible biases that characterize them. Regarding the differences between word-based explanation approaches, we observe that two-level explanations yields more relevant information and specific lexical cues (e.g. *environmental, transgender, scientists, archbishops*), which confirms the interest of a first pass through an adapted level of granularity in order to target the most interesting parts of the text. Explanations based on discourse markers or named entities show overlap with the other methods, indicating consistency between approaches. EDU-based explanations are more comprehensive and self-sufficient, while covering information con-

Explainability technique	Left	Center	Right
LIME Words	obama, pacific, brass, mccain, barack, after, percent, donald, aids, with	trump, donald, continued, washington, said, ginsburg, iran, options, this, china	scalise, garnering, heard, that, anti-muslim, only, fired, president, media, surveillance
EDUs	"when mainstream columnists start using words like aristocracy and kleptocracy"	"according to the american psychiatric association, not all transgender individuals suffer from gender dysphoria."	"because Stossel had done the shovel work (*cough*) of introducing fundamental concepts and breaking in nerds."
2-level EDUs+Words	media, percent, barack, columnist, worse, contrarian, sundays, interested, nationwide, watching	trump, twitter, dysphoria, manafort, donald, gender, environmental, transgender, scientists, ginsburg	stossel, scalise, president, cohen, sentamu, disgusting, nobody, media, archbishops, garnering
Discourse Markers	absolutely, surely, lately, only, maybe	then, perhaps, already, frequently, still	here, though, however, obviously, naturally
Named Entities	Barack Obama, David Pecker, John Mccain, Preet Bharara, Hillary Clinton	Donald Trump, Paul Manafort, Bader Ginsburg, Christopher Wray, Mark Zuckerberg	Steve Scalise, John Sentamu, John Stossel, Jerry Falwell, Michael Cohen

Table 4: Prototype explanations by class (Allsides), ordered from most to least impactful, as given by the highest saliency scores of the explanations.

tained in word-based explanations. This seems to make it an appropriate compromise between human readability and computational cost. Furthermore, there does not seem to be any particular trend in the relative position of the most impactful EDUs in the text, which confirms the interest of keeping the entire document (Figures 6, 7 and 8 of Appendix C).

By comparing the results between the different classes (left, center, right), and without entering into political considerations, we can establish a first diagnosis of the biases that characterize them. From the word-based explanations, we observe a shift in the lexical fields between classes (*pacific, aids, percent – transgender, environmental, scientists – fired, surveillance, archbishops*), which indicates a bias in topics covered and in the way information is conveyed. Articles from the right class seem to favor negative-sounding terms, while the pitch used is more neutral for the center and left classes. We can also note the over-representation of public and political figures in the explanations, which is distinguished between each class by the political leaning and the social category of the people being mentioned. In particular, we notice that articles from the right are almost exclusively mentioning personalities from their side, with the specificity of recurrently referring to religious figures (e.g. *John Sentamu, Jerry Falwell*). While the profiles are more diversified for the left and center classes, giving a lot of attention to right-wing personalities. About discourse markers, three trends can be identified from each of the classes. The left class seems to prefer markers of certainty or uncertainty (e.g. *absolutely, maybe*). The center class focuses on markers indicating time or frequency (e.g. *then, already, frequently*). Finally, the right class favors markers that indicate contrast or emphasis (e.g. *though, however, obviously, naturally*).

For the analysis of the structure and its explanation, we compare various statistics following Ferracane et al. (2019). Average height of trees (6.36), average proportion of leaf nodes (0.87) and the average normalized arc length (0.35) are equivalent between classes, although the right-wing class have slightly more shallow trees. Regarding the explanations, the most impactful relationships are mainly located in the first levels of the tree, close to the root, independently of the class. Although the explanation by perturbing the tree relations is not the most intuitive at first sight, it allows for a new level of abstraction by providing an understanding of the model's decisions with respect to the induced structure, which combined with other methods of analysis, can reveal additional biases.

10 Conclusion

We propose an integrated approach to both predict and analyze political bias in news articles, taking into account discourse elements. We show that structured attention over EDUs yields significant improvement at different levels over existing approaches, or comparable results, if lower, with respect to data- or computation-hungrier models. We

also proposed new variants for perturbation-based explanation methods when dealing with long texts, both at the lexical and structural level, that would not be possible with the other models. We demonstrate the effectiveness of our system by evaluating it on a series of diagnostic properties, and propose a qualitative analysis and comparison of the various approaches for the characterization of political bias.

Limitations

We reused data collected by previous work in the literature. Collecting news articles is susceptible to various sampling biases, related to the sources collected, the topics covered, and the time span of the collection, which influences what appears in the articles. In addition, labels given to articles are actually the political orientation of their source in the case of the Allsides and POLITICS datasets, which is obviously likely to induce errors. They rely on expertise provided respectively by the Allsides[11] and Ad Fontes[12] websites. The exact methods are undisclosed, but such labeling has necessarily a subjective aspect, oversimplifying predefined political categories, and can evolve in time. This affects classification reliability when applied to different sources, different times, different topics. This is on top of any specific elements related to the language (English) and cultural background of the sources (predominantly U.S.-based sources). This study is not intended to provide an accurate tool for predicting the political orientation of a text, but to provide analyses of the linguistic expression of bias, as seen through a supervised model.

Ethical considerations

Studying the political orientation of various media is already the objective of various institutions (Allsides, Ad Fontes, Media Bias/Fact Check). It depends on many factors, and a reliable automatic identification is still out of reach of current models, as can be seen from existing experimental results, and some of the limitations underlined above. These models should thus not be used for something other than research purposes, or supporting human analysis. This is one of the reasons why we develop an explainable approach to bias prediction, but these also have their own limitations, and shouldn't be used either as a strong indication of bias in one way or another without careful human examination.

[11] https://www.allsides.com/media-bias/media-bias-rating-methods

[12] https://adfontesmedia.com/how-ad-fontes-ranks-news-sources/

Acknowledgements

Nicolas Devatine's work is supported by the SLANT project (ANR-19-CE23-0022). This work was partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program. This work is also partially supported by the AnDiaMO project (ANR-21-CE23-0020). Chloé Braud and Philippe Muller are part of the programme DesCartes and are also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. We thank Farah Benamara for her helpful comments and suggestions on an earlier version of the paper.

References

David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the robustness of interpretability methods. *CoRR*, abs/1806.08049.

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

Vidhisha Balachandran, Artidoro Pagnoni, Jay Yoon Lee, Dheeraj Rajagopal, Jaime Carbonell, and Yulia Tsvetkov. 2021. StructSum: Summarization via structured representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2575–2585, Online. Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames

corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the shapley value based on sampling. *Computers Operations Research*, 36(5):1726–1730. Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653, Florence, Italy. Association for Computational Linguistics.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *Int. J. Digit. Libr.*, 20(4):391–415.

Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394, Seattle, United States. Association for Computational Linguistics.

Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy. Association for Computational Linguistics.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. 2019. Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using ELMo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Morteza Kamaladdini Ezzabady, Philippe Muller, and Chloé Braud. 2021. Multi-lingual discourse segmentation and connective identification: MELODI at disrpt2021. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 22–32, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 957–966.

Chang Li and Dan Goldwasser. 2021. Mean: Multi-head entity aware attention networkfor political perspective detection in news media. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom Censorship, Disinformation, and Propaganda (NLP4IF)*.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019a. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Dietram A. Scheufele and David Tewksbury. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of Communication*, 57(1):9–20.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Dataset Statistics

Statistics about the datasets are reported in Tables 5, 6 and 7. The distributions of the number of tokens per dataset (Figures 3, 4 and 5) show that *Hyperpartisan* has overall shorter news articles compared to *Allsides* and *C-POLITICS*.

	Left	Center	Right	Total
Train	9,618	6,683	7,189	23,490
Valid.	98	618	1,640	2,356
Test	599	299	402	1,300

Table 5: Statistics about the *Allsides* dataset.

	Left	Center	Right	Total
Train	8,543	8,543	8,543	25,629
Valid.	890	890	890	2,670
Test	3,022	3,022	3,022	9,066

Table 6: Statistics about the *C-POLITICS* dataset.

	Non-HP	HP	Total
Train	407	238	645
Test	314	314	628

Table 7: Statistics about the *Hyperpartisan* (HP) dataset.

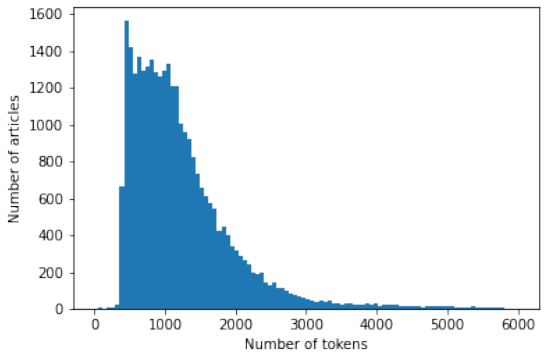

Figure 3: Distribution of the number of (BERT) tokens per article for the *Allsides* dataset.

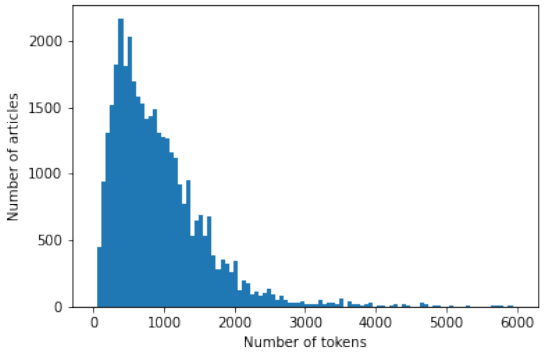

Figure 4: Distribution of the number of (BERT) tokens per article for the *C-POLITICS* dataset.

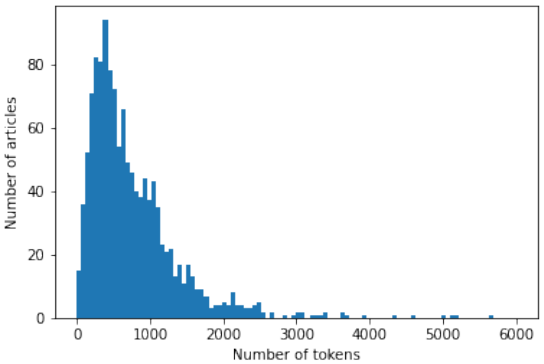

Figure 5: Distribution of the number of (BERT) tokens per article for the *Hyperpartisan* dataset.

B Settings

RoBERTa and POLITICS are initialized using the hyperparameters given in Table 9, Table 10 is for Longformer. The classification model we propose (Structured Attention/EDU) contains about 120M parameters, RoBERTa and POLITICS contain about 125M parameters, and it is about 148M for Longformer. Training is done on an Nvidia GeForce GTX 1080 Ti GPU card.

Hyperparameter	
# Epochs	10
Learning Rate	0.01
Batch size	8
Loss Function	Cross Entropy
Optimizer	Adagrad
Weight Decay	0.01
Bi-LSTM Hidden Dim.	200
2-layer Perceptron Dim.	200
Classifier Dropout	0.5
Adversarial Adaptation λ	0.7

Table 8: Hyperparameters used for training the latent structured attention model (see Section 3.1).

Hyperparameter	
# Epochs	15
Learning Rate	$1e-4$
Batch size	4
Loss Function	Cross Entropy
Optimizer	AdamW
Weight Decay	0.01
Classifier # Layers	2
Classifier Hidden Dim.	768
Classifier Dropout	0.1
Sliding window size	512
Sliding window overlap	64

Table 9: Hyperparameters used to fine-tune RoBERTa and POLITICS.

Hyperparameter	
# Epochs	10
Learning Rate	$2e-5$
Max Input Length	4096
Batch size (via gradient accumulation)	4
Loss Function	Cross Entropy
Optimizer	AdamW
Weight Decay	0.01
Classifier # Layers	2
Classifier Hidden Dim.	768
Classifier Dropout	0.1

Table 10: Hyperparameters used to fine-tune Longformer.

C Explanations

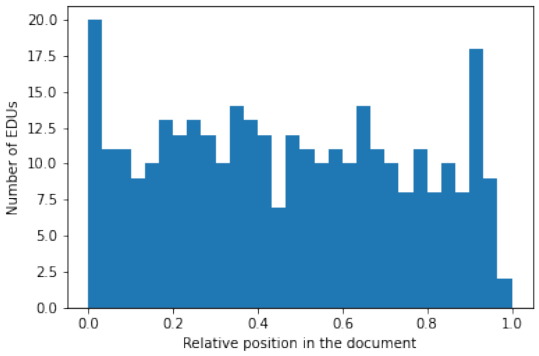

Figure 6: Distribution of relative positions of the most impactful EDUs for the left class (*Allsides*).

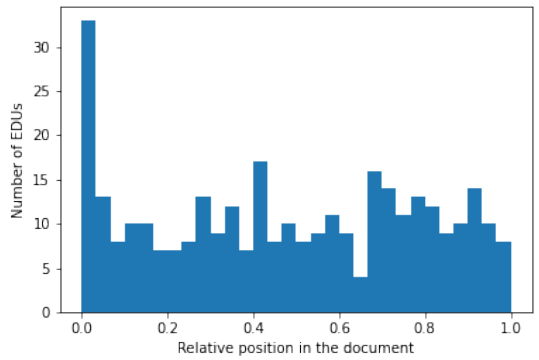

Figure 7: Distribution of relative positions of the most impactful EDUs for the center class (*Allsides*).

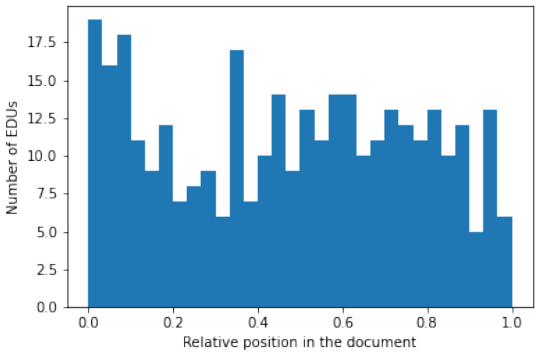

Figure 8: Distribution of relative positions of the most impactful EDUs for the right class (*Allsides*).

Explainability technique	Non-hyperpartisan	Hyperpartisan
LIME Words	reported, lewandowski, according, donald, could, news, corey, hustler, unaired, police	trump, reveals, discomfiting, reputation, controversial, hillary, politicians, immigrants, criminals, guns
EDUs	"if the 14,000 hours of unaired 'apprentice' tapes are released."	"it is an evil, oppressive ideology with governmental, judicial, educational, militaristic, and societal aspects to it"
2-level EDUs+Words	said, facebook, reported, news, tweeted, lewandowski, donald, weinstein, instagram, media	tyranny, racist, chargeable, abiding, trump, treasonous, shameful, clintons, deserved, reveals
Words w/o Stopwords	weinstein, lewandowski, said, news, facebook, texas, reported, president, twitter, police	trump, hillary, tyranny, abiding, racist, obama, treasonous, reputation, shameful, melania
Discourse Markers	first, then, eventually, this, recently	then, perhaps, here, again, only
Named Entities	Harvey Weinstein, Nikki Haley, Allie Clifton, Corey Lewandowski, Jake Tapper	Donald Trump, Chrissy Teigen, Hillary Clinton, Mike Pence, Barack Obama

Table 11: Prototype explanations by class (Hyperpartisan), ordered from most to least impactful, as given by the highest saliency scores of the explanations.

Explainability technique	Left	Center	Right
LIME Words	disparaging, trump, melania, pitfalls, honors, attacking, authorities, explain, which, surprising	bemoaned, reason, president, irrational, true, accomplishments, republicans, stadium, reeves, participated	president, sweeping, spokesman, chinese, surrounding, doom, lashed, caucuses, nevada, virus
EDUs	"but trump complied,"	"whom republicans have criticized throughout the impeachment process."	"that democrats only increased the support for late-term abortion and abortion on demand."
2-level EDUs+Words	contributed, e.g., repeats, replies, stance, explains, nonsense, refusing, disparaging, unhelpful	bemoaned, referencing, said, frequent, abusing, quoting, criticized, impeachment, unlike, legal	america, warn, boom, president, boycott, political, democrats, ideological, lockdown, wuhan
Words w/o Stopwords	trump, click, contributed, e.g., explains, stance, attempted, nonsense, refusing, concerned	bemoaned, quoting, berkovitz, heralded, political, accomplishments, frequent, impeachment, coronavirus, legal	america, china, president, democrats, political, chinese, warn, wuhan, boom, boycott
Discourse Markers	honestly, increasingly, evidently, then, surprisingly	also, however, absolutely, obviously, then	meantime, rather, this, also, together
Named Entities	Donald Trump, Deb Riechmann, Tom Barrett, Joe Biden, Kamala Harris	Tobe Berkovitz, Devin Brosnan, Bernie Sanders, Hunter Biden, Bill Stepien	Pete Buttigieg, Donald Trump, Steve Mnuchin, Robert Unanue, Marsha Blackburn

Table 12: Prototype explanations by class (C-POLITICS), ordered from most to least impactful, as given by the highest saliency scores of the explanations.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Discussed in section "Limitations".

☑ A2. Did you discuss any potential risks of your work?
Discussed in section "Ethical considerations".

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1 (Introduction).

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 6 (Dataset).

☑ B1. Did you cite the creators of artifacts you used?
Section 6 (Dataset).

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 6 (Dataset).

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 6 (Dataset).

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 6 (Dataset) and Appendix A.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 6 (Dataset) and Appendix A.

C ☑ Did you run computational experiments?

Section 7 (Experimental Settings).

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 7 (Experimental Settings) and Appendix B.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 7 (Experimental Settings) and Appendix B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 8 (Results).

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 4, 5 and 7.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.