Pluralistic Recommendation in News

Reference Contact - Paolo Ferragina (`paolo.ferragina@unipi.it`)
Lorenzo Bellomo, Virginia Morini, Dino Pedreschi, Giulio Rossetti

Contents

1 Introduction

News sources have a strong tendency to polarize their readers through over-personalized recommender systems, as those ones tend to suggest content already coming from the users' belief chamber. Accordingly, biased recommendation is one of the most crucial problems of our society. While such an issue has been widely investigated in the USA domain, as far as we know, no work has been performed in the European domain, which is characterized by a more multifaceted media landscape both from a political, linguistic and geographic point of view.

Given this scenario, the original goal of the microproject was to design and implement a prototype of an AI system able to predict the bias of European news and then use it to build a bias-free (thus, pluralistic) recommender system for European news sources.

The first step we envisioned was the one of building a European dataset with the following goals:

1. **Diversity in political view**: This was ensured by picking news sources with a clear political label on Eurotopics[1].

2. **Several Media from all the European countries**. We manually selected media that comply with the previous constraint coming from all around Europe.

3. **Possibly large**. We extracted this dataset from CommonCrawl-News[2], which is by far the largest internet archive for news all around the world.

After gathering this dataset, the next step was the one of building a political leaning classifier. We decided to employ BERT [1], as it obtained state-of-the-art performance on several tasks for multiple languages. We performed BERT classification on this dataset with the end goal of obtaining a political leaning classifier. The performances we obtained were great when predicting known sources but severely degraded when performing predictions on unknown ones. This prompted us to perform Model Explainability. Results are described in details in 4. In summary, our classifier is not predicting the political leaning, but it is merely predicting the source that wrote the news.

Before proceeding to the recommendation task, we had to find a solution to this issue.

We identified three main reasons to explain why our classifier attempt failed:

- *Lack of a manually annotated political leaning dataset for the European Field*. The only dataset we were able to find is from Baly et al [2] and contains only USA articles.

- *Lack of topics*. Our dataset has sport, economic, political articles all together, and labeled indistinctly. We would like to perform classification only on politically relevant articles. This is of course impossible when there are no topics in the dataset

- *Difficulty of the task*. Associating a political leaning to a well written piece of text is considered as a hard problem, even for humans. For this reason, this kind of tasks are usually performed by human experts. This problem is further caused by the fact that news tend to use a *"smoother* language" than the one used in social media. This makes it even harder to identify hidden subjectivity.

Given the issue in classification, we decided to shift the focus of the project to building another dataset, with the same goals in mind we had for the first one, but with a way to extract topics. We need topics because we need a way to filter out articles that do not carry political bias, such as those dealing with sports and gossip. The end goal is to re-perform news-bias classification on a fully "political" subset of articles and then perform *Explainability* in order to find the most distinctive words between leanings.

The two datasets are presented in section 2.

[1] `https://www.eurotopics.net/en/` European news aggregator published by the German federal government agency Bundeszentrale für politische Bildung.

[2] `https://commoncrawl.org/2016/10/news-dataset-available/`

Relevant Links It must be noted that, regarding the release of the two datasets, we have not yet received clearance from Eurotopics. Hereby, the two datasets must not be shared outside of the Humane-AI network. Here are provided all the links to the code and the datasets:

- `https://github.com/LorenzoBellomo/BiasClassification` - Repository containing all the code regarding the political leaning classifiers, together with the explainability part.
- `https://github.com/LorenzoBellomo/EU-NewsDataset` - Repository containing all the code regarding the two EU datasets, comprehending data cleaning pipelines, topic extraction, and all the topic modeling tests.
- `https://drive.google.com/file/d/1Qq2khT7lM-5_oHSNJhbK_-EATNdOSY-n/view?usp=sharing` - Link to the NO-TOPIC dataset.
- `https://drive.google.com/file/d/1KGy-FcLulACK_Fa3Abd9Xr4qaurBnu2S/view?usp=sharing` - Link to the dataset with topics.

2 Datasets

In this Section, we present the two datasets used in this project. Both of them contain European news articles. The first one has no annotated topics for the articles, while the second one does. The first one is described in section 2.1, while the second is shown in section 2.2.

2.1 *EU News Articles* Dataset without Topics

This dataset is composed of over 15 million European news articles coming from 192 media outlets belonging to 27 European countries with the addition of the United Kingdom. Further, articles range in a time period from 2016 to 2021 and are written in their original languages, for a total of 24 different languages included in the dataset. Each article (i.e., title and textual content) was extracted from the **Common Crawl Corpus**[3], which contains petabytes of raw web page data collected since 2008.

The dataset is released without any text pre-processing, but we performed some for the classification task. We further apply a text pre-processing pipeline in order to clean the raw news articles. In particular, we remove escape characters, URLs, mentions, and hashtags.
Moreover, we enrich our dataset with several media metadata (e.g., frequency of publication, distribution area). Among them, we obtain the political orientation of each media outlet that a team of Eurotopics experts (i.e., journalists from all over Europe) manually annotated. As shown from the label distribution across different countries in Figure 1, each media was labeled with one out of six leanings: Left-wing, Centre-left, Liberal, Liberal-conservative, Conservative, Right-wing.

We leverage this dataset to train a text classifier able to detect the political bias of news pieces. However, since the news articles dataset does not contain topic annotation at the article level, we assign a political label to each article following a Distant supervision scheme, i.e., we give each news piece the label of the media they belong to. Despite this being the only automatic way to accomplish the task, such a technique comes with disadvantages, such as the possibility of assigning a political label to an article that do not has political connotation (e.g., sport or entertainment articles) and, in the end, affect the quality of the training set.

2.2 *EU News Articles* Dataset with Topics

Following the issue described on the *topic-less* dataset, we decided to perform another round of Common Crawl-News scraping, aiming to also obtain the URLs of articles. We use these URLs to extract topics. The dataset is formed by over 4 million articles, distributed on the 6 political leanings. They span the time range 2016 - 2021, and they contain articles from the same 28 countries described above. The articles come from 175 different media outlets.

Figure 2 shows two plots. One displays the distribution of political leanings for each of the 28 countries, while the second one provides the distribution of detected topics.

[3] `https://commoncrawl.org/2016/10/news-dataset-available/`

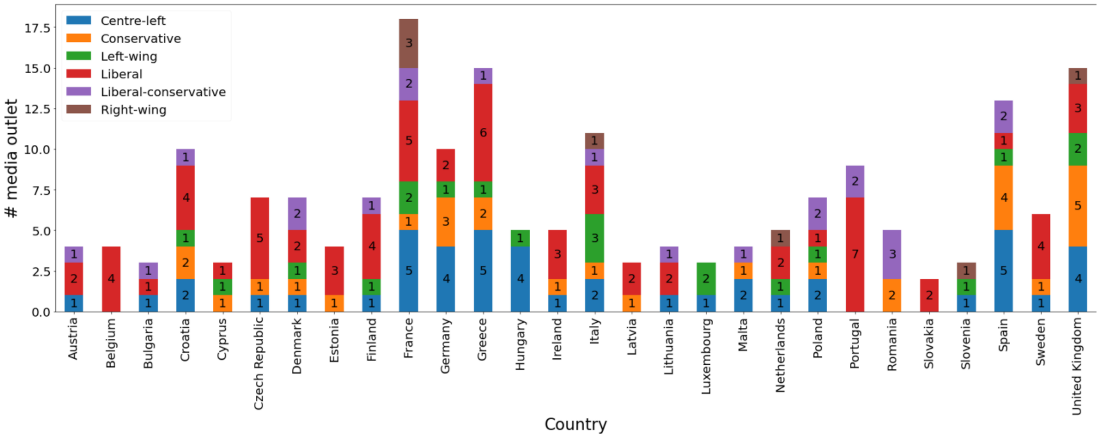


Figure 1: For each European country, the number of media outlets coloured with respect to their political orientation.

The protocol for extracting the topics is described in the following paragraph.

Topic Extraction To extract topics, we started with one key observation. Many of the URLs present in our dataset comply with the following pattern:

```
https://news_source.com/topic/article-name
```

We decided to split the path of the URLs, and we obtained the most common words that consistently appeared *"between the slashes"*.

On this set of words, we performed multilingual topic modeling with BERTopic[4] [3]. This step created 10 preliminary topics identified by words with similar meanings. We then manually checked those words to obtain a set that was fully satisfactory to us.

Firstly we removed the two "noise" topics and the topic which contained just variations of the word *news*. We then renamed the other topics to be easier-to-read (i.e. *"9_politique_politik_elections_politica"* to just *politics*). Lastly, we checked the list of most common words and all the words already present in the topics and moved them in order to better fit with the topic name itself.

This left us with 7 topics, and a set of words. If we find any of these words inside an article URL *between two slashes*, we assign the corresponding topic to that article.

Given an URL, and calling t_i the i-th word between slashes of the path part of said URL, the protocol proposed to assign a topic to an article is:

1. We split the topic part of the URL on the " / "
2. We check if any of the words present in our topic dictionary are present and exactly matches between the slashes (in the examples above, if t_i is equal to any of those words)
3. If any matching word is present, the corresponding topic is assigned to the article
4. if more than one word matches we pick the first one (t_1 has priority over t_2)
5. The topic coronavirus is the one with lowest priority, so it is assigned only if it is the only topic that matches on said URL.

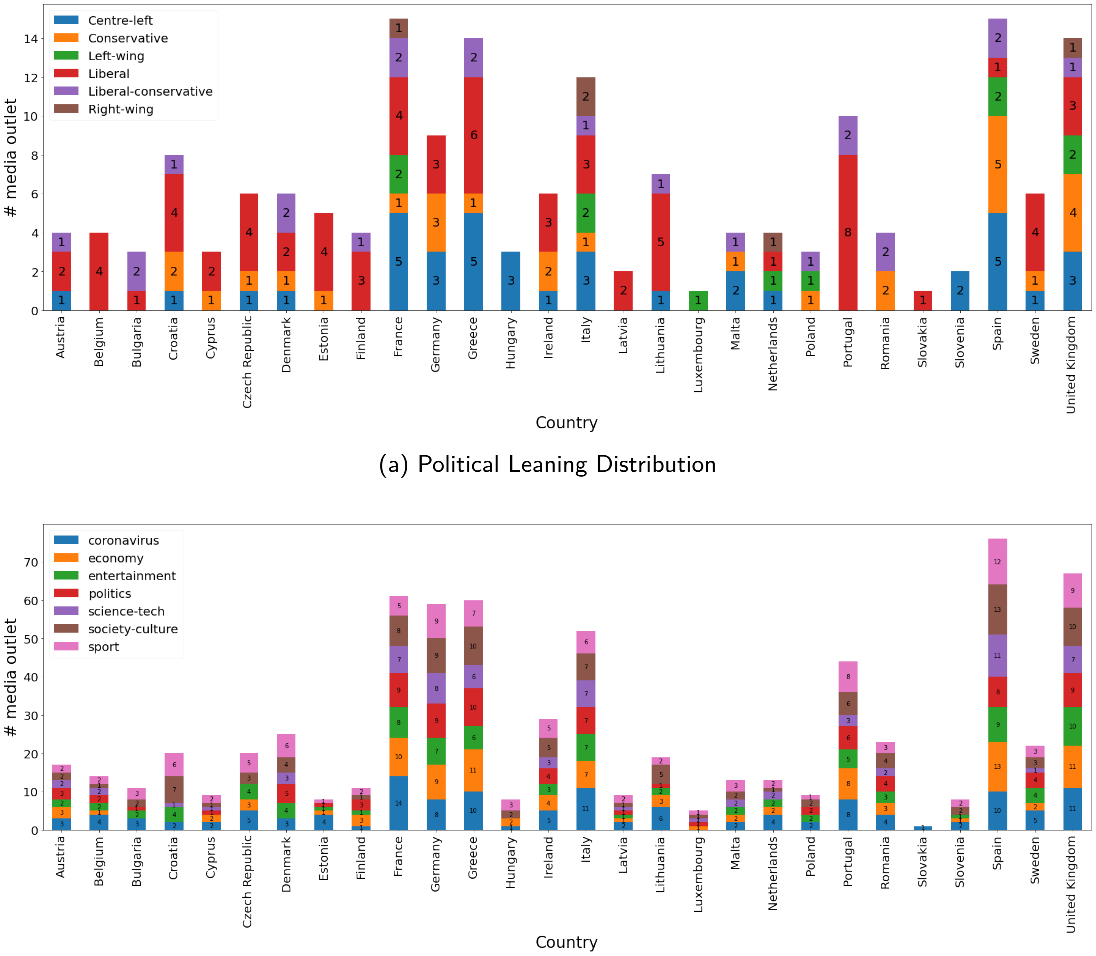
Having run this protocol on our articles, we found the following topic-sizes in our dataset.

- *society-culture*: $517\,522$

[4] https://github.com/MaartenGr/BERTopic

- *coronavirus*: $56\,822$
- *economy*: $689\,568$
- *politics*: $644\,706$
- *sport*: $1\,344\,183$
- *entertainment*: $775\,758$
- *science-tech*: $247\,783$

We simply dropped every article which did not have a topic and built the dataset with all the remaining articles.



(a) Political Leaning Distribution

(b) Topic Distribution

Figure 2: For each European country, the number of media outlets coloured with respect to *(a)* their political orientation and *(b)* the topic of their articles.

3 Political Bias Classifier

The task of detecting and predicting different kinds of bias (e.g., social, cultural) in written texts is typically formulated as a text-classification problem, where the textual content of a biased instance is encoded into a vector representation that is used to train a classifier to predict one of C classes.

Of course, when dealing with textual data, it is of utmost importance to take into account both the suitable type of word representation and the proper type of classifier. Since traditional word representation (i.e., bag-of-words model) encode words as discrete symbols not directly comparable to others [4], they are not fully able to model semantic relations between words. Instead, word embeddings like Word2vec [5], BERT Embeddings [1], Glove [6] mapping words to a continuously valued low dimensional space, can capture their semantic and syntactic features. Also, their structure makes them suitable to be deployed with Deep Learning models, fruitfully used to address Natural Language Processing-related (NLP) classification tasks. Among the powerful available NLP classifiers (e.g., Recurrent Neural Networks like LSTM [7]), recently, in the literature have been introduced the so-called Transformer models that, differently from the previous ones, can process each word in a sentence simultaneously via the attention mechanism [8]. In particular, autoencoding transformer models such as Bidirectional Encoder Representations from Transformers (BERT) [1] and the many BERT-based models spawning from it (e.g., RoBERTa [9], DistilBERT [10]), has proven that leveraging a bidirectional multi-head self-attention scheme yields state-of-the-art performances when dealing with sentence-level classification.

Following such rationale, we decide to rely on BERT-based models to predict the **political bias** in the *EU News Articles Dataset without Topics*. In the following paragraphs, we will explain in detail the experimental setup and the model evaluation step.

3.1 Experimental Setup

In the data splitting step, to train and test the model, we split it into three different sets: i) balanced *Training set* composed of 200k news articles for each political label; ii) *Test set known media* that consist of articles whose media was used for training the model (1k articles for each label); iii) *Test set unknown media* that consist of articles whose media was not used for training the model (1k articles for each label). We define these two different test sets to make a preliminary investigation on the distant supervision issue and thus to assess if BERT is probably learning to model the source of the article instead of predicting his political bias. During model selection, since we have a multilingual, multi-class dataset, we leverage the pre-trained model BERT-BASE-MULTILINGUAL-CASED presented by Devlin et al. [11] and publicly released by Hugging Face[5]. As for the previous model, we varied the length of the input $[64, 128, 256, 512]$ and the learning rate $[2e^{-5}, 3e^{-5}, 5e^{-5}]$. In addition, we took into account the type of text preprocessing, i.e., with or without punctuation. We obtain the best results on the validation set, leaving punctuation, 512-token input, and a learning rate of $2e^{-5}$, reaching an average accuracy of 93.1%.

3.2 Model Evaluation

During the model evaluation step, we assess BERT performances both on the *Test set known media* and *Test set unknown media*. The model reaches a high accuracy value (i.e., 92.8%) on the test set composed of articles whose media were used during training. At the same time, it performs really poorly (i.e., 32.5%) on the test set consisting of articles whose media were not seen in the training step. As previously mentioned, such a difference in results could be attributable to the fact that the model is learning the media source rather than the political bias due to the lack of topic annotation (e.g., sports or entertainment news pieces could be not politically aligned) and the technique of distant supervision. For such reasoning, we decide to further investigate such an aspect with respect to the *Test set known media* by leveraging the explainability methods introduced below.

4 Model Explainability

In this Section, we investigate the behavior of the BERT-based model that we used to predict the political bias of the *EU News Articles Dataset without Topics*, via explainability methods. In particular, our goal is to assess if the classifier is actually predicting the source (i.e., media) of each news piece instead of capturing their political bias. Below, we briefly give an overview of the explanations algorithms used to interpret model outputs, review results obtained in a global and local setting and draw conclusions.

[5] `https://huggingface.co/bert-base-multilingual-cased`

4.1 Explanation Methods

Following recent surveys on Explainable AI [12, 13, 14, 15, 16, 17, 18], we briefly define the field to which the explainers we use in this contribution belong, i.e., post-hoc explainability methods. This branch pertains to the black-box explanation methods. The aim is to build explanations for a black-box model, i.e., a model that is not interpretable or transparent regarding the automatic decision process due to the complexity of its internal dynamics. Post-hoc strategies can be *global* if they target explaining the whole model, or *local* if they aim to explain a specific decision for a particular record. The validity of the local explanation is therefore dependent on the particular instance chosen, and often the findings by themselves are not generalizable to describe the overall model logic. In addition, the explanation technique can be *model-agnostic* if they are independent with respect to the type of black-box to be inspected (e.g., tree ensemble, neural networks, etc.) or *model-specific* if they involve strategies that work only with specific types of models.

In the following, we briefly present the explanation techniques we chose to use. Specifically, Integrated Gradients and LIME are used both locally and globally, as described in Section 4.2.1.

4.1.1 Integrated Gradients

Integrated Gradients (IG) [19] is a post-hoc, model-specific explainability method for deep neural networks that attributes a model's prediction to its input features. In other words, it is able to compute how much a given input feature is important for the output prediction. Differently from mostly attribution methods [20, 21], IG satisfies both the attribution axioms *Sensitivity* (i.e., relevant features have not-zero attributions) and *Implementation variance* (i.e. the attributions for two functionally equivalents models are identical). Indeed, IG aggregates the gradients of the input by interpolating in small steps along the straight line between a baseline and the input. Accordingly, a large positive or negative IG score indicates that the feature strongly increases or decreases the model output, while a score close to zero indicates that the feature is not relevant with respect to the output prediction. IG can be applied to any differentiable model and thus handle different kinds of data like images, texts, or tabular ones. Further, it is used for a wide range of tasks like understanding feature importance by extracting rules from the network, debugging deep learning models performance, and identifying data skew.

4.1.2 LIME

LIME[6] [22] is among the most widely adopted local post-hoc model-agnostic approaches [13]. It returns an explanation by auditing the black box on randomly generated neighbors of the instance under examination. The explanation consists of a set of k *features importance* obtained as coefficients of a linear regression model locally approximating the behavior of the black box on the synthetic neighborhood. Positive values indicate positive contributions to the class, while negative values indicate negative contributions. High values indicate a higher contribution to the classification outcome, while values close to zero indicate negligible contribution.

4.2 Results

This Section reports the results of the experiments carried out to test our hypotheses. We focus the analysis on BERT-based classifiers as black boxes, adopting IG and LIME as explainers.

4.2.1 Global Explanations

In the following paragraphs, we discuss the results obtained by scaling from local to global explanations in an attempt to explain the whole model.

A very simple way to accomplish this task is to obtain local predictions for a large number of items and then average the scores assigned to each feature across all the local explanations to produce a global one. Accordingly, for each record in the dataset, we store the local explanation, which consists of a key, i.e., the word present in the phrase, and a value, i.e., the feature importance. Then we average the

[6] `https://github.com/marcotcr/lime`

Explainer	Class	Pattern	% Records with pattern	Δ Accuracy
IG	*Centre-left*	'BL', '15th', 'EF'	88.3	0.010
	Conservative	'Traffic', 'Hot', 'SO'	88.3	0.033
	Left-wing	'Ike', 'nav', 'Laguna'	39.5	0.165
	Liberal	'Cap', 'Kyle', 'Mario'	16.3	0.102
	Liberal-conservative	'×', 'Caro', '©'	4	0
	Right-wing	'Uruguay', 'Twee', 'Notice'	33.8	0.061
LIME	*Centre-left*	'JuntsxSí', 'RSA', 'flamingo'	14.9	0
	Conservative	'Pressemeldung', 'Maardu', 'Polizeimeldung'	0.2	0
	Left-wing	'Parijs', 'Insta', 'Aurelia'	10.1	0
	Liberal	'ROZZANO', 'SASSARI', 'WIEN'	1.4	0.214
	Liberal-conservative	'lente', 'Continental', 'dispensados'	3.3	0
	Right-wing	'chilling', 'dioxide', '394'	0.4	0

Table 1: For each class are shown the three words that obtain the higher global scores by different explanation methods, the percentage of the records that contain these words and the delta accuracy obtained by removing them from the records.

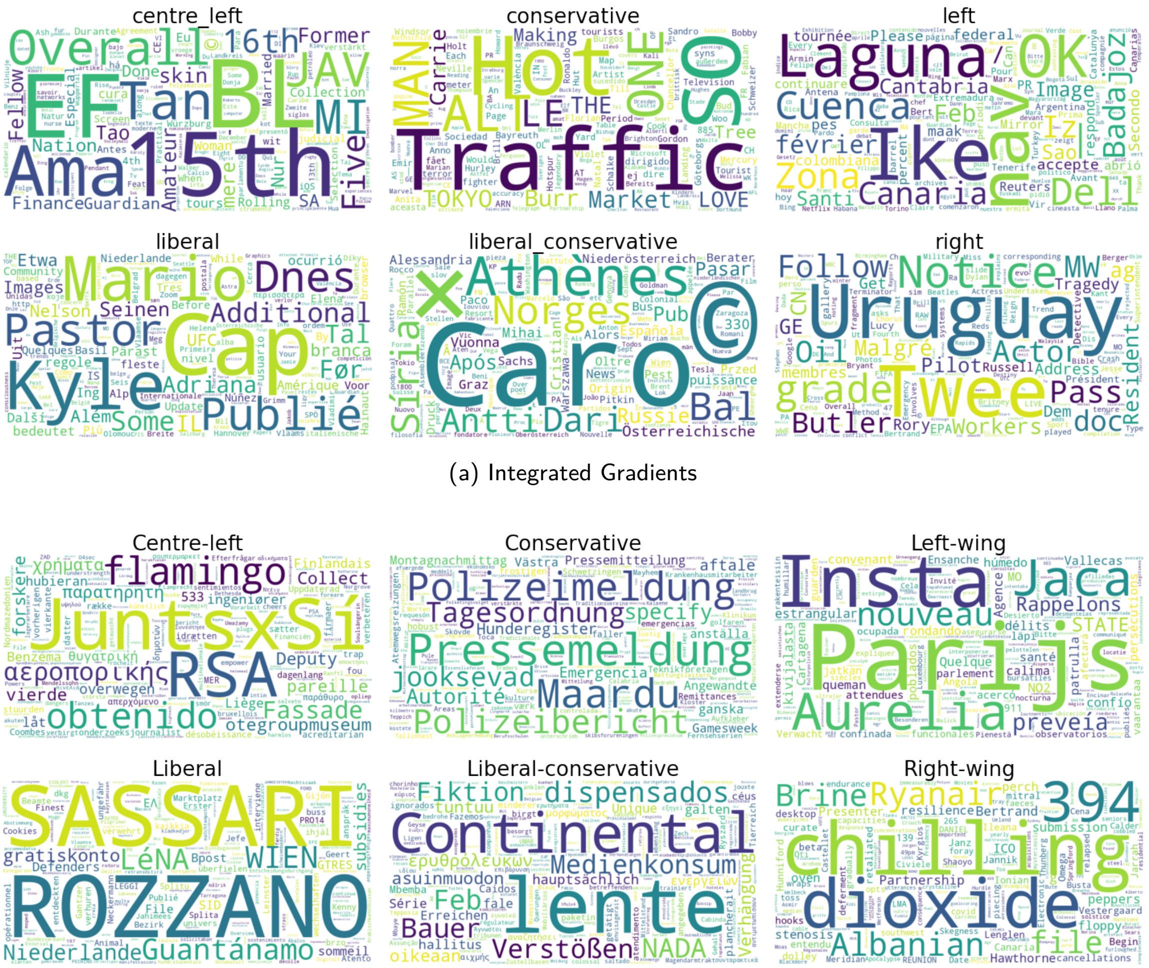
obtained scores for each word. This process is repeated for each class predicted by the model in such a way to find what are the words that led the model to output a specific class. To validate the explanation, we divide the dataset by predicted class, and in each instance, we remove the top three words with the highest importance, computed and retrieved globally. Next, we recompute the model predictions on each subset of the predicted class and report the difference in accuracy (Δ Accuracy), i.e., with and without the highest-importance words, to quantify how their absence impacts classification and performance.

All the results refer to the *Test set known media* that consists of articles whose media was used for training the model (1k articles for each label). In Figure 3, thanks to the WordClouds visualization, we show the results of the local-to-global process explained in Section 4.2.1. At a first glance, we can notice that the words with higher importance for each class do not seem to have a political connotation. Indeed, among them we have several city names (e.g, *Sassari*, *Rozzano*, *Parijs* for LIME and *Uruguay*, *Athènes*, *Canaria* for IG), references to social networks (e.g., *Insta* for LIME, *Tweet* for IG), or other terms not particularly related to politics (e.g., *chilling, Traffic, Notice*). However, although definitely fewer, we find two words with high importance that are strongly related to the predicted political class (i.e., *Mario* in the IG Liberal class that refers to Mario Draghi's work at BCE and *JuntsxSì* in the LIME Centre-left class that is a parliamentary alliance focused on achieving the independence of Catalonia from Spain).
To assess the differences in the model performance (e.g., Δ Accuracy), we remove the three words with the highest importance from their respective classes. Looking at the results reported in Table 1, it seems evident that LIME and IG outputs are pretty different. Indeed, while IG words are frequent in the articles (made exception for Liberal-conservative) and their removal leads to a Δ Accuracy up to 16.5%, LIME words are definitely less frequent in the dataset, and we observe a change in performance equal to 21.4% only for the Liberal class.
As a preliminary observation, we are able to state that, as expected, the lack of topic annotation and thus the presence in the training set of articles not politically-related and annotated with distant supervision introduce a huge quantity of noise in the dataset, affecting the capacity of the model to detect politically biased terms correctly.

4.2.2 Local Explanations

From the global explanations, we notice that the distant supervision technique used to annotate the dataset introduces significant noise affecting the training data quality and thus the model outputs. Accordingly, in this Section, our goal is to assess if, as a consequence of the annotation, the black-box model is actually predicting the source of each article (i.e., stylistic patterns of the media they belong to) rather than their political bias. To analyze this aspect, we adopt an inverse strategy: *i)* for each media contained in the *Test set known media* we compute local explanations for a sample of articles; *ii)* we look for specific words or sentences that are relevant for the outputs; *iii)* we remove words/sentences from the articles and compute Δ Accuracy.

By looking at LIME and IG local explanations for each media outlet article, we are able to detect



(a) Integrated Gradients

(b) LIME

Figure 3: For each class is shown a WordCloud representing the words that obtained the higher global scores by different explanation methods for the *EU News Articles* dataset.

recurrent patterns (i.e., words or sentences) that are specific to the media outlet. Unfortunately, most of these patterns are not semantically related to their target class and thus should be regarded as noise. In particular, we identify both patterns that are common in more than one media (like the media name in the article or the date at the beginning of the article) and specific to the media ones. Below, we describe some of these patterns as well as the respective variation in performance:

1. *Media name in upper case*: in the text preprocessing pipeline, we removed the exact matches of media outlets' names (case sensitive) because many of them are also common words used in the articles (e.g., La Repubblica). For this reason, we notice, thanks to explanations techniques, that references to the outlet media name in upper case (e.g., LE MONDE, DER STANDARD, DEMOKRAATTI) are still present and have a high impact on the model output. Indeed, by looking at Figures 4 and 5, we can clearly see both from IG and LIME explanations that the model exploits such pattern to discriminate among different political classes. Accordingly, the Δ Accuracy obtained by removing these patterns is equal to 16.8% for IG and to 17.5% for LIME.

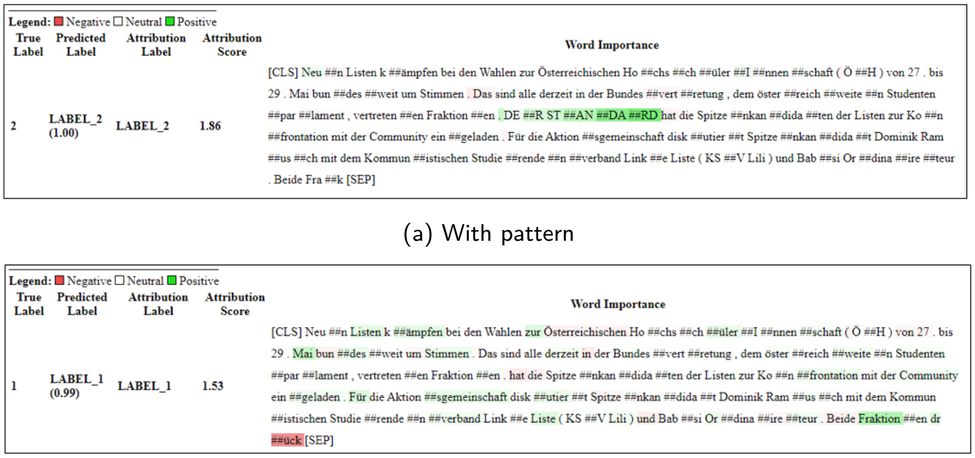
2. *Date at the beginning of an article*: as in the previous example, IG and LIME local explanations highlighted the presence in several articles of the date at the beginning of them. Such a pattern, even if the date changes every day, seems to slightly affect the model predictions, obtaining a Δ

in Accuracy of 1.2% for IG and of 1.3% for LIME.

3. *Daily Express - GETTY*: several articles belonging to Daily Express media outlet contain a cover image which is usually provided by the agency Getty Images, Inc.[7] For this reason, many articles begin with the *GETTY* pattern in order to credit the actual image owner. As shown in the supplementary material Figure 8 and 9 such a word obtains very high feature importance for both explainers. Accordingly, by removing its occurrences, we observe a decrease in accuracy equal to 23.0% for IG and 28.3% for LIME.

4. *Daily Mirror - Subscribe*: similarly, every article of the Daily Mirror starts with this sort of advertising *'Get money updates directly to your inbox + Subscribe Thank you for subscribing! Could not subscribe, try again later Invalid Email'*. Thus, this sentence becomes a pivotal feature in the classification process, as pinpointed by both IG and LIME. If we remove it, we experience a drop in performance of 62.0% for the former and of 62.1% for the latter.

5. *Diário de Notícias - Pub*: several articles belonging to Diário de Notícias start with the word *Pub* that, as shown in the supplementary material Figure 6 and 7, drove the classifier to use such a pattern to discriminate among different political orientation as in the case of the *Media name in upper case*. Here, we have a Δ in the accuracy of 22.7% for IG and of 24.6% for LIME.



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
2	LABEL_2 (1.00)	LABEL_2	1.86	[CLS] Neu ##n Listen k ##ämpfen bei den Wahlen zur Österreichischen Ho ##chs ##ch ##üler ##I ##nnen ##schaft (Ö ##H) von 27 . bis 29 . Mai bun ##des ##weit um Stimmen . Das sind alle derzeit in der Bundes ##vert ##retung , dem öster ##reich ##weite ##n Studenten ##par ##lament , vertreten ##en Fraktion ##en . DE ##R ST ##AN ##DA ##RD hat die Spitze ##nkan ##dida ##ten der Listen zur Ko ##n ##frontation mit der Community ein ##geladen . Für die Aktion ##sgemeinschaft disk ##utier ##t Spitze ##nkan ##dida ##t Dominik Ram ##us ##ch mit dem Kommun ##istischen Studie ##rende ##n ##verband Link ##e Liste (KS ##V Lili) und Bab ##si Or ##dina ##ire ##teur . Beide Fra ##k [SEP]

(a) With pattern

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (0.99)	LABEL_1	1.53	[CLS] Neu ##n Listen k ##ämpfen bei den Wahlen zur Österreichischen Ho ##chs ##ch ##üler ##I ##nnen ##schaft (Ö ##H) von 27 . bis 29 . Mai bun ##des ##weit um Stimmen . Das sind alle derzeit in der Bundes ##vert ##retung , dem öster ##reich ##weite ##n Studenten ##par ##lament , vertreten ##en Fraktion ##en . hat die Spitze ##nkan ##dida ##ten der Listen zur Ko ##n ##frontation mit der Community ein ##geladen . Für die Aktion ##sgemeinschaft disk ##utier ##t Spitze ##nkan ##dida ##t Dominik Ram ##us ##ch mit dem Kommun ##istischen Studie ##rende ##n ##verband Link ##e Liste (KS ##V Lili) und Bab ##si Or ##dina ##ire ##teur . Beide Fraktion ##en dr ##ück [SEP]

(b) Without pattern

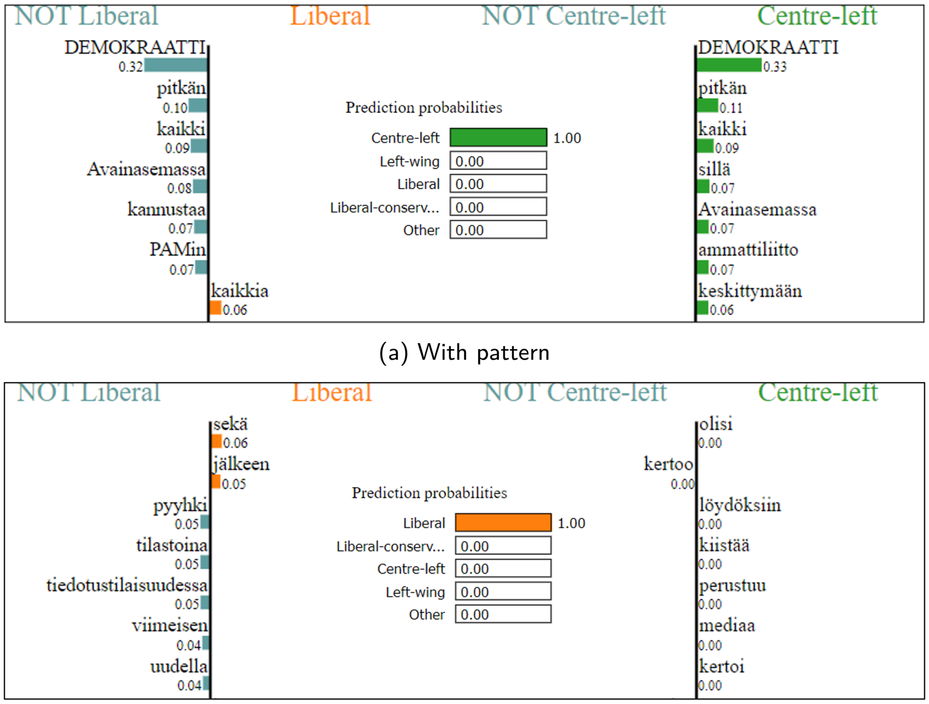
Figure 4: IG explanations - *Media* pattern

4.3 Discussion

In the following, we discuss the retrieved insights obtained via global and local explanations on the *EU News Articles Dataset without Topics*.

As a first consideration, we can notice that the results retrieved via global explanations are extremely uninformative (e.g., *"lente"*, *"chilling"*, *"394"*, *"x"*, or *"©"*). This initial evidence emphasizes the impact of a human manually curated dataset versus an automatically crawled one. The data selection process is therefore highly relevant. Indeed, the automatic labeling process assumes as a general rule that all of the news articles produced by a media outlet should be aligned to its political leaning. As previously stated, this assumption is clearly simplistic since articles could also be non-political related (e.g., lifestyle, sports news). For this reason, the explanations returned by LIME and IG (both globally and locally) suggest that the classifier relies on a more spurious correlation between lexical items and labels in the classification process. Distant-supervision introduces noise that independently affects all of the training samples. For this reason, statistical information inferred from co-occurrences in the training samples

[7] `http://gettyimages.com/`



(a) With pattern

(b) Without pattern

Figure 5: LIME explanations - *Media* pattern

is way less significant, leading to a classifier more prone to false-positive errors (all the uncorrelated words of not-political articles are now evidence of political bias), dimming the informativeness of truly high-correlated words (e.g.,*"Polizeimeldung", "Traffic", "Mario"*).

In conclusion, distant supervised labels are prone to inject noise signal arising from *wrong or simplistic assumptions* taken in this process and thus, in this setting, lead the model to predict the news source instead of their political bias. For such reasoning, in the attempt to mitigate this effect, we decide to define a new dataset, described in Section 2.2, that contains topic information for each news piece. As a short-term direction, we plan to i) rely on this new dataset to perform the task of predicting political bias by selecting only news pieces that talk about political issues and ii) update the text preprocessing pipeline according to explainers' suggestions.

5 Future Works

In this Section, we show some potential directions of work both for the short and the long term.

5.1 Source Bias Analysis

Having extracted a massive news article dataset, we could extract all the URLs present in the texts. URLs can give us crucial information, like, for instance, which news article cites another as a source. Having extracted URLs, we can build a graph where nodes are news sources. An edge between source A and source B is added if an article from source A contains an URL pointing to source B.

This can lead to a set of different analyses:

- Do we observe political segregation in the citation pattern between left and right in Europe?
- Do we observe an increase, a decrease, or a stable situation with respect to the 6 years of our articles?
- Does this segregation have different behaviors in different countries?

- Do we see a shift in citation pattern after a disruptive historical event (i.e., a government change)?
- Do we see a significant difference when dealing with all citations, rather than only the political ones?
- Do we see some form of "selection bias" from different leanings with regards to some non-political event (for example, a book release)?

5.2 Entity Framing

Different news sources describe the same event in slightly different manners, be it because of editorial constraints or because of the style of the authors. This concept is called *framing*, and it is mostly studied in academia when the differences arise from opposing political views. This study has been widely performed in single countries (mostly USA), but as far as we know, no cross-country attempt has been performed.

We can find a set of *politically-relevant* events across Europe, translate all the articles dealing with that topic, and try to match sentences in a sort of "parallel view". This can aid us in highlighting recurring differences in the way some topics are dealt with. This can range from the different choice of words to some entities getting completely hidden in the narrative.

5.3 Recommender System

The next step we envision is the one of attempting again on build the recommender system. This requires us to perform classification on the political articles rather than on all of them as we did up until this point.

The recommender system has to be built in such a way that it minimizes bias. Bias is not only defined by the political leaning of an article but also by other factors such as geography, time of the event, and topic described. The recommender must be able to maximize content diversity but avoid falling into the pitfall of the backfire effect.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, 2019.

[2] R. Baly, G. Da San Martino, J. Glass, and P. Nakov, “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)*, pp. 4982–4991, Association for Computational Linguistics, 2020.

[3] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” *arXiv preprint arXiv:2203.05794*, 2022.

[4] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” *Information*, vol. 10, no. 4, p. 150, 2019.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[6] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” *arXiv preprint arXiv:1910.01108*, 2019.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Comput. Surv.*, vol. 51, no. 5, pp. 93:1–93:42, 2019.

[13] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, “Benchmarking and survey of explanation methods for black box models,” *CoRR*, vol. abs/2102.13076, 2021.

[14] A. A. Freitas, “Comprehensible classification models: a position paper,” *SIGKDD Explor.*, vol. 15, no. 1, pp. 1–10, 2013.

[15] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.

[16] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, L. Pappalardo, S. Ruggieri, and F. Turini, “Open the black box data-driven explanation of black box decision systems,” *CoRR*, vol. abs/1806.09936, 2018.

[17] L. Longo, R. Goebel, F. Lécué, P. Kieseberg, and A. Holzinger, “Explainable artificial intelligence: Concepts, applications, research challenges and visions,” in *CD-MAKE*, vol. 12279 of *Lecture Notes in Computer Science*, pp. 1–16, Springer, 2020.

[18] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. Müller, “Toward interpretable machine learning: Transparent deep neural networks and beyond,” *CoRR*, vol. abs/2003.07631, 2020.

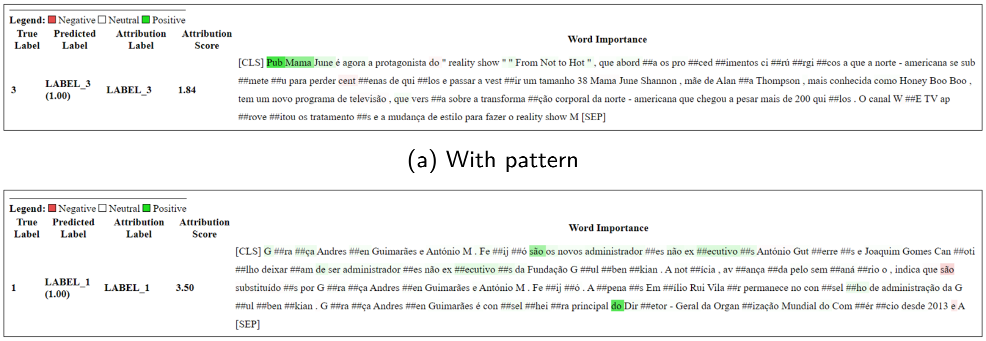
[19] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[20] S. Vashishth, S. Upadhyay, G. S. Tomar, and M. Faruqui, “Attention interpretability across nlp tasks,” *arXiv preprint arXiv:1909.11218*, 2019.

[21] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, “Layer-wise relevance propagation for neural networks with local renormalization layers,” in *International Conference on Artificial Neural Networks*, pp. 63–71, Springer, 2016.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, “"why should I trust you?": Explaining the predictions of any classifier,” in *KDD*, pp. 1135–1144, ACM, 2016.

Supplementary Material



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
3	LABEL_3 (1.00)	LABEL_3	1.84	[CLS] Pub Mama June é agora a protagonista do " reality show " " From Not to Hot " , que abord ##a os pro ##ced ##imentos ci ##rú ##rgi ##cos a que a norte - americana se sub ##mete ##u para perder cent ##enas de qui ##los e passar a vest ##ir um tamanho 38 Mama June Shannon , mãe de Alan ##a Thompson , mais conhecida como Honey Boo Boo , tem um novo programa de televisão , que vers ##a sobre a transforma ##ção corporal da norte - americana que chegou a pesar mais de 200 qui ##los . O canal W ##E TV ap ##rove ##itou os tratamento ##s e a mudança de estilo para fazer o reality show M [SEP]

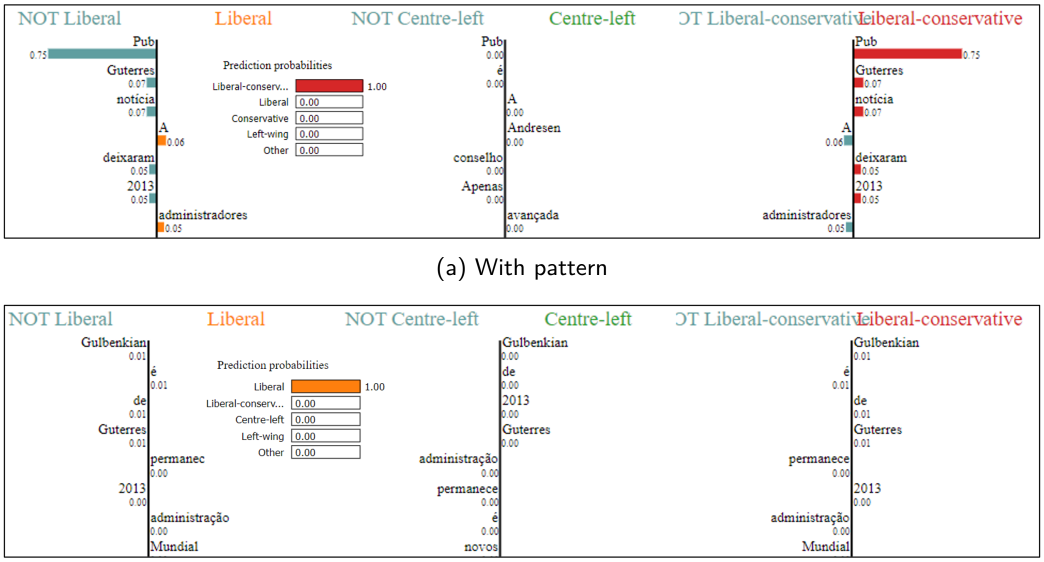
(a) With pattern

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
1	LABEL_1 (1.00)	LABEL_1	3.50	[CLS] G ##ra ##ça Andres ##en Guimarães e António M . Fe ##ij ##ó são os novos administrador ##es não ex ##ecutivo ##s António Gut ##erre ##s e Joaquim Gomes Can ##oti ##lho deixar ##am de ser administrador ##es não ex ##ecutivo ##s da Fundação G ##ul ##ben ##kian . A not ##ícia , av ##ança ##da pelo sem ##aná ##rio o , indica que são substituído ##s por G ##ra ##ça Andres ##en Guimarães e António M . Fe ##ij ##ó . A ##pena ##s Em ##ílio Rui Vila ##r permanece no con ##sel ##ho de administração da G ##ul ##ben ##kian . G ##ra ##ça Andres ##en Guimarães é con ##sel ##hei ##ra principal do Dir ##etor - Geral da Organ ##ização Mundial do Com ##ér ##cio desde 2013 e A [SEP]

(b) Without pattern

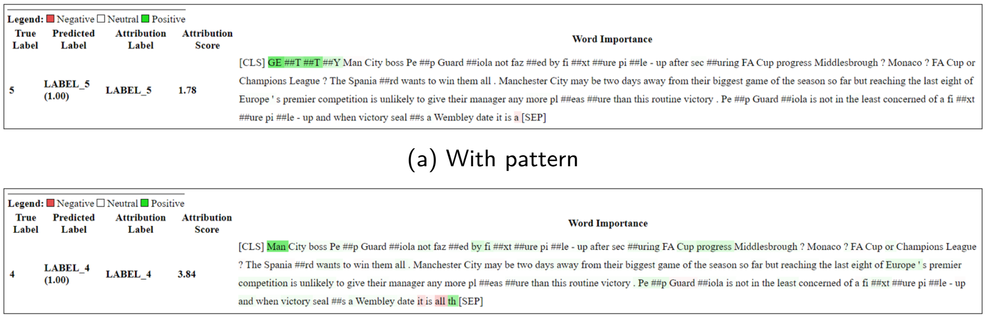
Figure 6: IG explanations - *Pub* pattern



(a) With pattern

(b) Without pattern

Figure 7: LIME explanations - *Pub* pattern



Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
5	LABEL_5 (1.00)	LABEL_5	1.78	[CLS] GE ##T ##T ##Y Man City boss Pe ##p Guard ##iola not faz ##ed by fi ##xt ##ure pi ##le - up after sec ##uring FA Cup progress Middlesbrough ? Monaco ? FA Cup or Champions League ? The Spania ##rd wants to win them all . Manchester City may be two days away from their biggest game of the season so far but reaching the last eight of Europe ' s premier competition is unlikely to give their manager any more pl ##eas ##ure than this routine victory . Pe ##p Guard ##iola is not in the least concerned of a fi ##xt ##ure pi ##le - up and when victory seal ##s a Wembley date it is a [SEP]

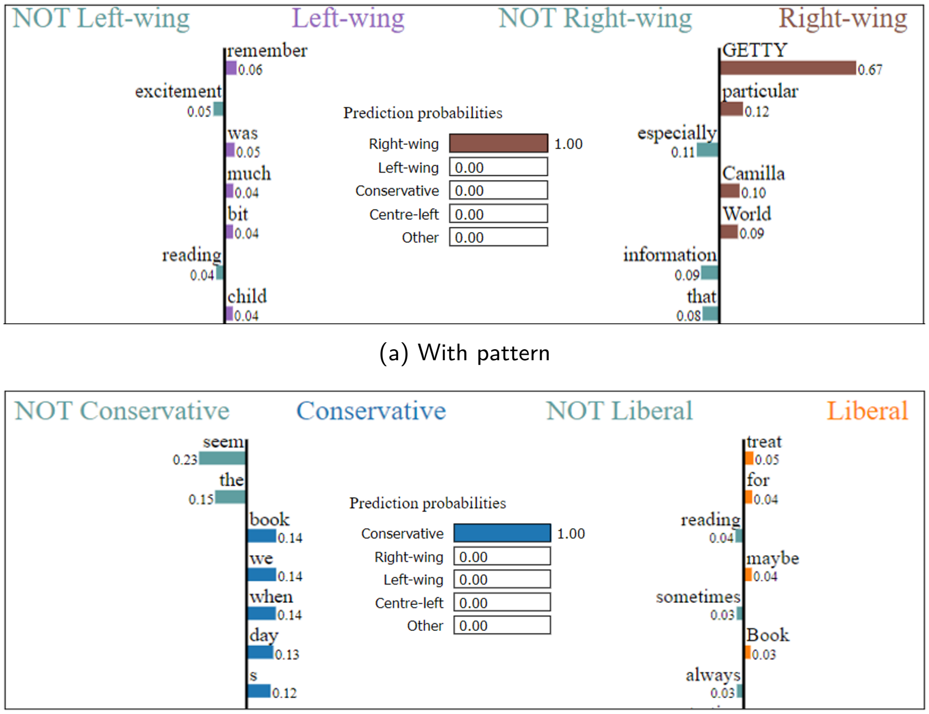
(a) With pattern

Legend: 🟥 Negative □ Neutral 🟩 Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
4	LABEL_4 (1.00)	LABEL_4	3.84	[CLS] Man City boss Pe ##p Guard ##iola not faz ##ed by fi ##xt ##ure pi ##le - up after sec ##uring FA Cup progress Middlesbrough ? Monaco ? FA Cup or Champions League ? The Spania ##rd wants to win them all . Manchester City may be two days away from their biggest game of the season so far but reaching the last eight of Europe ' s premier competition is unlikely to give their manager any more pl ##eas ##ure than this routine victory . Pe ##p Guard ##iola is not in the least concerned of a fi ##xt ##ure pi ##le - up and when victory seal ##s a Wembley date it is all th [SEP]

(b) Without pattern

Figure 8: IG explanations - *Getty* pattern



(a) With pattern

(b) Without pattern

Figure 9: LIME explanations - *Getty* pattern