Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 179-190 recibido 31-03-2023 revisado 07-05-2023 aceptado 20-05-2023

Identifying Media Bias beyond Words: Using Automatic Identification of Persuasive Techniques for Media Bias Detection

Identificación del sesgo en los medios más allá de las palabras: uso de la identificación automática de técnicas persuasivas para la detección del sesgo mediático

Francisco-Javier Rodrigo-Ginés[1], **Jorge Carrillo-de-Albornoz**[1,2], **Laura Plaza**[1,2]
[1]NLP & IR UNED, 28040 Madrid, Spain
[2]RMIT University, VIC 3000, Australia
frodrigo@invi.uned.es, {jcalbornoz, lplaza}@lsi.uned.es

Abstract: Detecting media bias is a challenging task due to the complexity and ambiguity of language. Current approaches are limited in their ability to generalise across regions and styles of journalism. This paper proposes a new approach that focusses on detecting rhetorical linguistic techniques rather than just analysing words or contextual representations. We compare three different systems based on different techniques for identifying media bias, including a lexical-based system, a language transformers-based system, and a cascade transformers system that relies on persuasive techniques detection. We have evaluated these systems using a Ukraine crisis news dataset and splitting it by according to the country to generate training and test sets, i.e. different sets for each country. The results of the cascade system outperforms by at least a 6% the other approaches in identifying media bias when evaluating with different countries setup. Our results suggest that models capable of detecting rhetorical and persuasive linguistic techniques are necessary to generalise media bias effectively.
Keywords: Natural Language Processing, Disinformation, Media bias detection.

Resumen: Detectar sesgo mediático es una tarea desafiante debido a la ambigüedad del lenguaje. Los enfoques actuales tienen dificultades para generalizar entre regiones y estilos periodísticos. Proponemos un enfoque centrado en la detección de técnicas lingüísticas en lugar de analizar palabras o representaciones contextuales. Comparamos tres sistemas diferentes basados en diferentes técnicas para identificar el sesgo de los medios: un sistema basado en léxico, un sistema basado en transformers y un sistema de transformers en cascada capaz de detectar técnicas persuasivas. Hemos evaluado estos sistemas utilizando un conjunto de datos de noticias de la guerra de Ucrania. Los resultados del sistema en cascada superan en al menos un 6% a los demás enfoques a la hora de identificar el sesgo de los medios de diferentes países. Nuestros resultados sugieren que los modelos capaces de detectar técnicas lingüísticas retóricas y persuasivas son necesarios para generalizar la detección de sesgo de los medios de manera efectiva.
Palabras clave: Procesamiento de Lenguaje Natural, Desinformación, Detección de sesgo mediático.

1 Introduction

The detection of bias in the media has been an active research area in recent years. The importance of detecting media bias cannot be overstated, especially in today's world, where the media has a significant influence on public opinion and political decision-making. Moreover, it is important to note that biases in media reporting can vary in their intentionality, ranging from conscious and

ISSN 1135-5948 DOI 10.26342/2023-71-14 ©2023 Sociedad Española para el Procesamiento del Lenguaje Natural

deliberate to unconscious and unintentional. However, the impact of media bias on public opinion and discourse cannot be overlooked, which makes it imperative to address this issue and promote media literacy.

The task of automatic detection of bias in news articles is a challenging task due to the complexity and ambiguity of the language (Aggarwal et al., 2020). Early research in this area focused on hand-developed lexical and linguistic techniques, along with sentiment analysis and topic modeling, to detect bias at the document level (Lin, Bagrow, and Lazer, 2011). Nonetheless, lexical-based models applied to media bias detection may be good for detecting word choice/labeling bias, but may be not enough for detecting other forms of media bias such as persuasion or rhetoric. More recently, deep learning approaches such as recurrent neural networks have been used to detect bias in news articles (Baly et al., 2020). However, despite the increasing number of proposed methods, there is still a lack of understanding as to how these models actually learn to detect bias.

This problem was highlighted in (Cremisini, Aguilar, and Finlayson, 2019) experiments. In 2019, they presented the Ukraine crisis news dataset. In their research, they found that the behaviour of the implemented classification models drops drastically when trained only with news from Russia and evaluated with news from Ukraine. This fact arises the question as to whetther their models were learning the regional journalistic style instead of generalizing the detection of media bias.

Following this idea, we hypothesize that current models that are based solely on words or contextual representations are not capable of generalizing media bias effectively. While these models may be effective at capturing certain regional journalistic styles, they fail to identify and account for more subtle rhetorical linguistic techniques that may be employed to convey media bias. As a result, there is a need for more precise forms that can detect these techniques and facilitate more generalizable models.

To tackle this, we have systematic analyzed the literature, both from the journalism and computer science perspective, and identified different forms of media bias commonly found in news. We categorize these forms into two types: depending on the intention bias and depending on the context. Intention bias refers to a journalist's deliberate attempt to influence the reader's opinion, while context bias refers to the bias that can arise from the way a story is presented or from the journalist's choice of sources. We have identified 17 forms of media bias, which include bias through word choice or labeling. This can manifest in the form of using pejorative language to negatively depict certain groups or employing loaded terms to frame a story with a particular narrative. For instance, a news article might describe a protest as a 'riot,' indicating a negative bias against the event. Another prevalent form of bias involves persuasive techniques such as appeals to emotion, authority, or groupthink. In the case of appeals to authority, a news article may quote a high-ranking government official to substantiate a particular perspective. For instance, an article might assert, "According to the Secretary of State..." to lend credibility to a certain point of view.

In light of these considerations, the present study aims to delve into the detection of media bias by proposing a novel approach. Instead of solely relying on lexical analysis or contextual representation, our approach focuses on detecting rhetorical linguistic techniques.

We compare three different systems based on different techniques for identifying media bias. The first system was a lexical-based system, which relied solely on identifying certain words or phrases that were indicative of media bias. The second system was a language transformers-based system. Finally, the third system was a cascade transformers system that relies on persuasive techniques detection using the SemEval'23 task 3 dataset.

We evaluated these three systems using the Ukraine crisis news dataset and compared their performance on different subsets for training and test that includes: all news, news only from Ukraine, and all other except news from Ukraine. Additionally, we used LIME and SHAP explainability techniques to mask and remove words from the texts to determine whether the lexical-based systems were capable of identifying bias in such conditions comparing to our cascade model.

Based on our results, we found that the cascade system, which was capable

of detecting subtle linguistic patterns and techniques, outperformed the other two approaches in terms of identifying media bias in the Ukraine news dataset when training and testing the model with news from different countries. This finding supports our hypothesis that more precise forms that detect rhetorical linguistic techniques are necessary to generalize media bias effectively.

The paper is structured as follows. Section 2 presents the background to the problem and related work. Section 3 describes the dataset and preprocessing pipeline, and the models used. Section 4 presents how we have evaluated the models, the experimental setup, the explainable AI techniques used, and finally the results of the experiments. To conclude the paper, Section 5 discusses the results and presents the conclusions.

2 Background

In recent years, researchers have taken different approaches to generalizing media bias. The methods can be divided into two categories: non-neural network models and neural network models. Non-neural network models are mainly based on statistical learning or machine learning and require handcrafted features, such as linguistic (Hube and Fetahu, 2018) or reported speech features (Lazaridou and Krestel, 2016). Neural network models, on the other hand, can automatically learn feature representations from text and have been shown to outperform traditional methods. In particular, RNNs (Rashkin et al., 2017) and transformers (Baly et al., 2020) are the most commonly used neural networks for media bias detection.

Additionally, some researchers have explored other methods such as stakeholder mining (Ogawa, Ma, and Yoshikawa, 2011), community detection (Patricia Aires, G. Nakamura, and F. Nakamura, 2019), and information theory (Aires, Freire, and da Silva, 2020) approaches. Overall, the most common approach to detecting media bias is to use supervised machine learning models to classify news articles as biased or unbiased. However, these models lack interpretability and transparency and there is a need for more specific labels so that the models can detect more specific forms of bias.

Some authors are already working in introducing fine-grained bias labels (Piskorski et al., 2023). Instead of simply classifying news articles as biased or unbiased, models are now being trained to detect specific forms of bias, such as bias by word choice/labeling, appeals to emotion, authority or groupthink, red herring, and loaded language. This allows for a more nuanced understanding of media bias and can help in developing targeted strategies to counteract specific forms of bias.

In summary, media bias detection research has witnessed the exploration of diverse methodologies, ranging from traditional statistical models to state-of-the-art neural network models. The challenge lies in not only improving the performance of bias detection models but also enhancing their interpretability and expanding the range of detectable biases.

3 Methodology

In this section, we describe the datasets used, the preprocessing pipeline and the models developed.

3.1 Datasets description

In order to train our systems, we have used two different datasets: the Ukraine crisis news dataset, created by (Cremisini, Aguilar, and Finlayson, 2019), and a persuasive techiques multilabel dataset that we will called the SemEval'23 task 3 dataset developed by (Piskorski et al., 2023).

The Ukraine crisis news dataset includes 4,538 articles in English related to the 2014 Crimea crisis from 227 news sources in 43 countries. The articles have been manually classified as either pro-Russian, pro-Western, or Neutral, and also aligned with a master timeline of 17 major events. This dataset is a multiclass dataset, as the goal of the task is to classify the articles in one of the 3 classes (pro-Russian, pro-Western, and Neutral).

The news annotated as pro-Russia includes the following topics: Crimea coming home; Russia welcomes Crimea; Crimea's accession to Russia; Russia welcomes Crimea; Admission of Crimea into Russia; Ukraine took over Crimea; Crimea wants to go back to its roots in Russia; Referendum website hit by cyber-attack; The U.S. plans to supply weapons to Ukraine. The news annotated as pro-Western covers Russia stealing land from a sovereign nation;

Russian Separatists; Annexation by Moscow; Russia stages coup; Russia took over Crimea; Russia does not fear the West; Crimea has been isolated by Russia; Putin admits Russian actions to take over Crimea; Putin refuses to rule out intervention in Donetsk. And the news annotated as neutral includes: Mention frames from both sides equally, reporting facts, or offer explanation for both pro-Russian and pro-Western frames. State factual information without any emotional, political or ideological charge.

We noted a significant dataset imbalance, with a considerably larger number of articles related to Russia compared to Ukraine and the Neutral category. Specifically, there were 3372 articles related to Russia, 908 articles related to Ukraine, and 258 articles classified as Neutral. The dataset imbalance poses challenges for training and evaluating media bias detection models. The unequal representation of bias categories can lead to model biases, where the model may prioritize or perform better on the majority class while struggling to accurately detect bias in the minority classes. This issue hinders the models' ability to generalize effectively across all bias categories and may result in skewed performance metrics.

In the other hand, the SemEval'23 task 3 dataset(Piskorski et al., 2023) focuses on the detection of genre, framing and persuasion techniques in online news articles in a multilingual setup. The data presented in the task is unique in its kind as it is both multilabel and multilingual, and it also covers complementary dimensions of what makes text persuasive, namely style and framing. The task covers multiple languages, including English, French, Spanish, or Italian. For the development of our systems we have used the subset in English for the detection of persuasive techniques.

3.2 Data preprocessing pipeline

The data preprocessing pipeline consists of the following two steps:

1. Stopwords removal: Stopwords were removed using the NLTK stopwords list. Only done for the Logistic Regression models.

2. Media outlets names removal: The names of 83 media outlets have been removed from the texts in order to reduce bias from the model.

3.3 Models description

We implemented three different models: (1) a lexical-based Logistic Regression model; (2) a transformer-based model; and (3) a multilabel cascade transformer-based model capable of identifying different forms of persuasive techniques.

3.3.1 Logistic Regression

We chose to include a logistic regression (LR) model as one of our implemented models due to its simplicity, and widespread usage in various text classification tasks, including media bias detection (Chen et al., 2020).

In our implementation, we utilized the scikit-learn library to train and evaluate the logistic regression model. To represent the text data, we employed the TF-IDF (Term Frequency-Inverse Document Frequency) scheme, which assigns weights to words based on their frequency in the document and across the entire dataset. TF-IDF helps capture the importance of words within a document while downweighting terms that appear frequently across multiple documents.

No extensive hyperparameter tuning was performed for the logistic regression model, and we relied on the default parameter values provided by scikit-learn. This decision was made to establish a baseline performance for media bias detection and to compare the effectiveness of more complex models against this simple yet widely-used approach.

Apart from serving as a baseline model for media bias detection, we also utilized the logistic regression model for explainable AI techniques. By leveraging the interpretability of logistic regression, we were able to extract representative terms and features that contributed to the model's bias classification. These representative terms played a crucial role in subsequent mask and removal evaluations, allowing us to identify the impact of specific words on the bias detection process.

3.3.2 Fine-tuned transformer-based model

We have implemented a DistilBERT-based model for media bias detection. We have used the HuggingFace library (Wolf et al., 2020) for fine-tuning the model. DistilBERT

is a smaller, faster and cheaper version of BERT, that has been shown to perform as well as BERT in many downstream NLP tasks (Sanh et al., 2019). The model is fine-tuned in the Ukraine crisis news dataset. This model is a more robust state of the art baseline representative of contextual representation techniques.

It is worth noting that, unlike the logistic regression model, the transformer-based model does not rely on manually designed features or explicit rule-based systems. Instead, it learns representations directly from the text, allowing it to capture complex patterns and dependencies between words.

3.3.3 Cascade transformer-based model

The system developed for multi-label detection of persuasion techniques is based on a cascade transformer-based model that incorporates two trained models to carry out cascading inference (Enomoro and Eda, 2021). This model architecture leverages the power of transfer learning from pre-trained transformer models, which have been shown to outperform traditional machine learning approaches in natural language processing tasks.

In this model, we trained two separate DistilBERT models, which were fine-tuned on different datasets. The first model was fine-tuned on the SemEval'23 task 3 dataset, while the second model was fine-tuned on the Ukraine crisis news dataset. This was done to ensure that the models were able to capture a wide range of language patterns and persuasion techniques.

The first model is used to identify if the given text contains any of the media bias forms identified, including appeal to authority, appeal to groupthink/popularity, red herring, and loaded language. If no media bias form is detected, the prediction is set to "neutral". If any media bias form is detected, the second model is used to generate the final predictions.

The first model is multilabel, which means that it can predict multiple persuasion techniques simultaneously. We have used a threshold of 0.20 to decide if a persuasion technique is detected, based on optimization during the training phase. This means that if the predicted probability for a given persuasion technique is greater than or equal to 0.20, it is considered to be present in the text.

To identify the most effective threshold value, we employed the softmax function within the context of a multilabel classification problem. This determination process unfolded in two distinct stages:

1. In the first stage, we conducted a series of experiments with *macro thresholds* spanning from 0.1 to 0.9. By calculating the F1 score and flat accuracy for each macro threshold, we were able to ascertain the *best macro threshold*—the threshold that produced the maximum F1 score.
2. In the following stage, we computed *micro thresholds*. These were determined by augmenting the *best macro threshold* value with increments ranging from 0.01 to 0.09. For each of these micro thresholds, we calculated the corresponding F1 score and flat accuracy.

Ultimately, we selected the threshold that yielded the highest F1 score as the best threshold.

Overall, the cascade transformer-based model we have developed is a robust and effective approach for multi-label detection of persuasion techniques. The models were trained using the Adam optimizer with a learning rate of 5e-5 and a batch size of 32, which is a commonly used setting in transformer-based models.

In Figure 1, we provide an overview of the model architecture and the cascading inference process, which shows the complexity of the approach taking advantage of the persuassion dataset for a preliminary filtering step. The use of pre-trained transformer models and cascading inference represents a state-of-the-art approach to natural language processing and holds great promise for future research in this field.

4 Evaluation

In this section, we present the evaluation of the performance of the three different approaches for automatic detection of media bias. We evaluate the three systems on the Ukraine crisis news dataset, which contains news articles from various countries, languages, and media outlets. The dataset

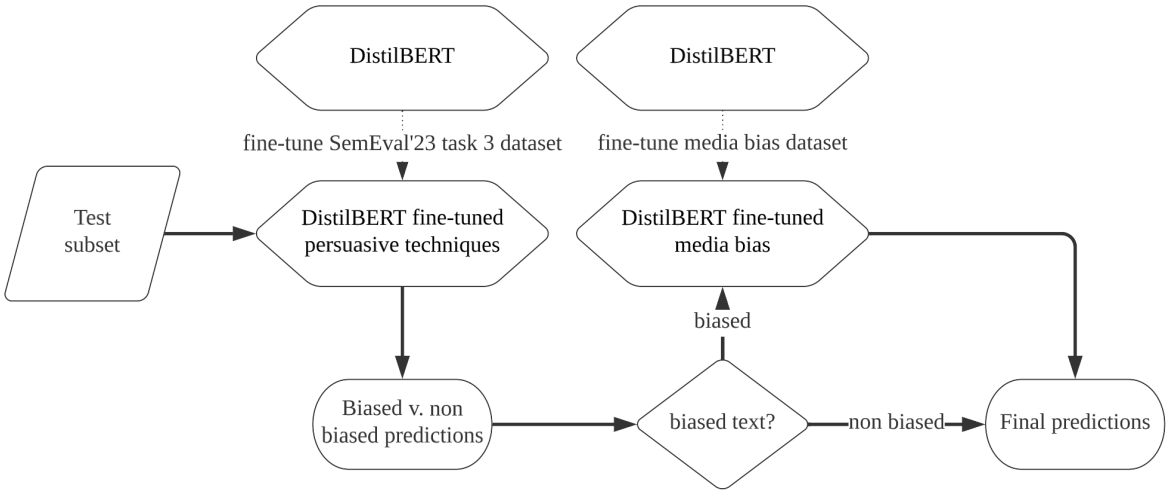

Figure 1: Cascade transformer-based model overview.

is annotated with three labels: pro-Russian, pro-Western, and neutral. We randomly split the dataset into training and testing sets, where 80% of the dataset was used for training and the remaining 20% for testing.

Since the dataset also includes information about which country each news item is published in, we have used that information to evaluate our models in different contexts. The evaluations setup are the following: (1) models trained and evaluated with news from multiple countries; (2) models trained and evaluated with news from any country except for Ukraine; (3) models trained and evaluated with news only from Ukraine; (4) models trained and evaluated with news from any country except for Ukraine and evaluated with news from Ukraine; (5) models trained with news only from Ukraine and evaluated with news from any country except for Ukraine.

Also, as we mentioned in the introduction, we believe that lexical-based models applied to media bias detection may be good for detecting word choice/labeling bias, but may be not enough for detecting other forms of media bias such as persuasion or rhetoric. Therefore, in order to determine how specific terms and phrase may affect the predictions of the models we have developed a lexicon of words that induce word choice/labeling bias in the given context. The idea of these experiemnts is to measure how models behaveis when masking and deleting such words in the texts.

We have used LIME and SHAP explainability techniques to understand the impact of specific words on the models' decision-making, and used them to challenging new set-up evaluations. To do this, we trained the Logistic Regression model on the Ukraine crisis news dataset, and applied LIME and SHAP over that models, creating the mentioned lexicons so we could mask/remove certain words to study the behavior of the models.

4.1 Detecting bias by word choice using explainable AI (XAI) techniques

In order to study how automatic techniques detect bias by word choice, we have used two explainable AI (XAI) techniques, namely LIME (Ribeiro, Singh, and Guestrin, 2016) and SHAP (Lundberg and Lee, 2017). These techniques have been used in combination to obtain a lexicon of words that influence the media bias classification models. These techniques allow us to uncover the patterns in the model's predictions and to identify the words and terms that are most predictive of bias.

4.1.1 LIME

Local Interpretable Model-Agnostic Explanations (LIME) algorithms (Ribeiro, Singh, and Guestrin, 2016) are used to explain the predictions of a given model. LIME can be used to find the features (words, entities, etc.) that are most important in the model's predictions and to uncover the

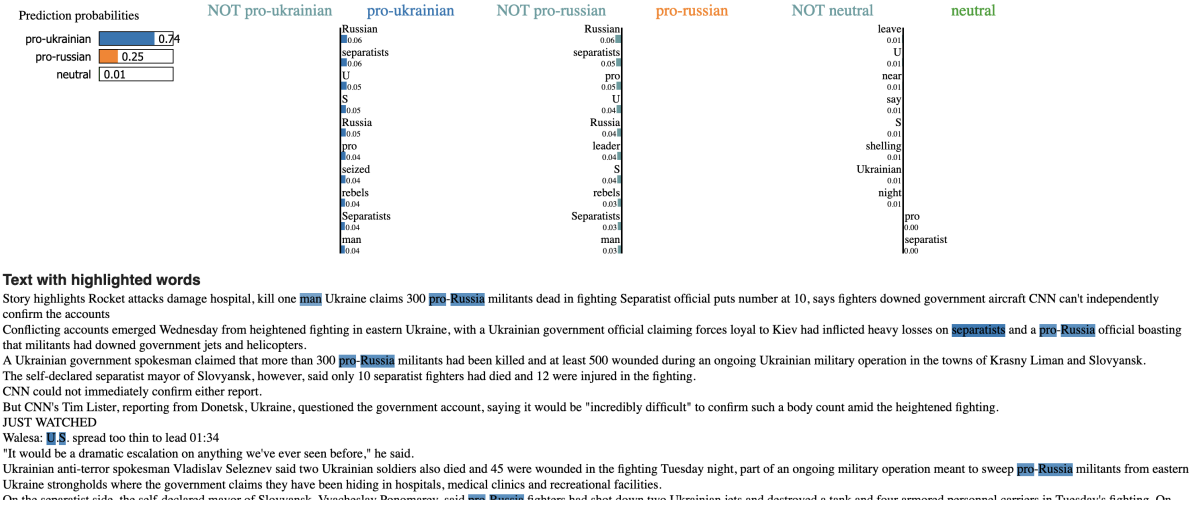

Figure 2: Example of LIME output.

patterns in the model's predictions. In Figure 2 we show an example of a LIME output. Please note, that terms highlighted in Figure 2 do not have to be included in our lexicons.

LIME works by perturbing the input data (in our case, the article embeddings) and measuring the resulting change in the model's predictions. This allows us to identify the words and phrases that are most important for the model's predictions. For example, if a certain word or phrase is consistently associated with a particular class of the model's prediction (in our case, bias), then it is likely to be an important feature for the model.

4.1.2 SHAP

The SHapley Additive exPlanations (SHAP) algorithm (Lundberg and Lee, 2017) is an explainable AI technique that can be used to explain the predictions of a model. SHAP assigns each feature a score that indicates how important it is in the model's predictions.

This algorithm works by combining the features of a given input and measuring the resulting change in the model's predictions. As the LIME technique, we are able of identify the features that are most important for the model's predictions.

4.1.3 Building the word choice lexicon

Using the scores given by LIME and SHAP for each feature (word) in the models, we created three lexicons of words that may induce word choice/labeling bias for each given context. In order to mark a word as a word choice biased one, we have manually reviewed each word taking into account the annotation guidelines provided by (Cremisini, Aguilar, and Finlayson, 2019).

The resultants lexicons includes some of the following words: separatist, rebel, independence, supporters, annexed, protester, trade, demonstrator, activist, withdrawal, deliveries, opposition, pro, unsupported, intervention, agreement, defeat, disruptions, disruption, offensive, exercises, militant, proclaimed, humanitarian, conflict, soviet, symbolic, border, rights, overturned, terrorist, aggressor, insurgent, freedom, fighter, proxy, victim, provocation, and evil.

Most of these terms can induce bias in news outlets by presenting a particular viewpoint of the conflict as well as creating a certain narrative of the parties involved. For example, using the term "separatists" can imply that the people fighting against the Ukrainian government are separate from Ukraine, which could lead to a pro-Russia narrative. Similarly, using the term "rebels" can suggest that those fighting against the Ukrainian government are doing so out of a desire to overthrow it, potentially leading to a more pro-Ukraine perspective. Finally, using the term "terrorists" to describe the opponents of the Ukrainian government could lead to a very negative portrayal of those parties, which could lead to a negative bias

against them.

Our main hypotheses are that lexical-based techniques are useful for detecting word choice/labeling media bias, but they are not capable of generalizing other forms of media bias. To evaluate our hypotheses, we compared the performance of the proposed models against the whole Ukraine crisis news dataset, as well as excluding and masking them making. The way we masked the words from the lexicon was replacing the words for the `[MASK]` token, as it has been done in the DistilBERT training pipeline.

4.2 Results

In this section, we present the results of our experiments on media bias detection using different models and evaluation approaches. The results are based on the analysis of various metrics and comparisons between the models implemented, including logistic regression and transformer-based models.

Table 1 provides an overview of the performance of the models. It is evident that the transformer-based models consistently outperform the logistic regression models in terms of bias detection. Specifically, the DistilBERT fine-tuned model, when trained and tested with news articles from the same countries, achieves the best results. This finding aligns with our expectations and highlights the effectiveness of fine-tuned transformer models in capturing bias patterns within specific contexts.

On the other hand, if we evaluate the results obtained by testing the models in contexts other than the training one, the cascade model obtains much better results. In the case of the models trained with news published in Ukraine, and tested with news from outside Ukraine, this improvement is 6%. In the opposite case (trained with news published outside Ukraine, and tested with news published in Ukraine), the improvement reaches up to 90% (0.18 macro F1-score on fine-tuned DistilBERT v. 0.33 macro F1-score on cascade model).

These results confirms that our cascade model trained with persuasive techniques greatly improves the results in the evaluation set-up in which the models are trained with texts from countries other than those with which they are tested. As (Cremisini, Aguilar, and Finlayson, 2019) suggested, it seems that both words and contextual representation based models do not generalize media bias, but rather are learning local journalistic styles. In that sense, our cascade model, which is trained with persuasive techniques, is able to better detect media bias in different countries.

Also interesting is the comparison of the models across the different evaluations, *whole dataset* vs *masked one* vs *deleted*, where we can see that the Logistic regression model highly drops in efficiency when removing the lexicon with biased words, comparing with the two transformers approaches. When evaluating the impact of the word choice lexicon to mask and remove the most predictive words, we can see that the performance of the lexical-based model is worse (around a 10%) when removing the words from the lexicon, but similar when masking it. Interestingly, the masked version did not perform worse than the non-masked version, which might seem counterintuitive. We believe that this unexpected finding could be attributed to the effect of the masking process. When certain terms are masked, the model receives a new hint or cue that indicates the presence of bias. Regarding the DistilBERT approach and the proposed cascade approach, it can be seen that the results of our approach is consistently across these evaluations in all setup, while the differences of the DistilBERT highly varies of the training/test subsets.

Finally, we believe that the proposed methodology is a promising approach for detecting media bias. We have demonstrated that explainable AI techniques can be used to identify the words and phrases that are most influential in the model's predictions, and that cascade transformer-based models are capable of detecting more subtle forms of media bias, such as persuasion or rhetoric, generalizing the detection of the media bias and avoiding the words bias problem.

5 Discussion and conclusion

In this paper, we proposed a novel approach to media bias detection that focuses on detecting rhetorical linguistic techniques rather than just analyzing words or contextual representation. We compared three different systems based on different techniques for identifying media bias, and evaluated them using the Ukraine crisis news dataset (Cremisini, Aguilar, and Finlayson,

Models trained and tested with news from all countries. Best results highlighted in bold

Method	Train	Test	Prec.	Recall	F1
Logistic Regression	All	All	0.56	0.54	0.56
Fine-tuned DistilBERT	**All**	**All**	**0.78**	**0.77**	**0.78**
Cascade model	All	All	0.63	0.62	0.63
Logistic Regression w. masked lexicon	All	All	0.60	0.54	0.56
Fine-tuned DistilBERT w. masked lexicon	**All**	**All**	**0.82**	**0.83**	**0.82**
Cascade model w. masked lexicon	All	All	0.66	0.62	0.65
Logistic Regression w. deleted lexicon	All	All	0.51	0.47	0.47
Fine-tuned DistilBERT w. deleted lexicon	**All**	**All**	**0.76**	**0.74**	**0.75**
Cascade model w. deleted lexicon	All	All	0.61	0.66	0.62

Models trained and tested with news from all countries except for Ukraine:

Method	Train	Test	Prec.	Recall	F1
Logistic Regression	A-U	A-U	0.61	0.51	0.54
Fine-tuned DistilBERT	**A-U**	**A-U**	**0.63**	**0.65**	**0.63**
Cascade model	A-U	A-U	0.49	0.57	0.52
Logistic Regression w. masked lexicon	A-U	A-U	0.61	0.51	0.53
Fine-tuned DistilBERT w. masked lexicon	**A-U**	**A-U**	**0.57**	**0.65**	**0.60**
Cascade model w. masked lexicon	A-U	A-U	0.61	0.54	0.55
Logistic Regression w. deleted lexicon	A-U	A-U	0.53	0.49	0.48
Fine-tuned DistilBERT w. deleted lexicon	**A-U**	**A-U**	**0.60**	**0.61**	**0.61**
Cascade model w. deleted lexicon	A-U	A-U	0.50	0.53	0.51

Models trained and tested with news from Ukraine:

Method	Train	Test	Prec.	Recall	F1
Logistic Regression	U	U	0.75	0.53	0.40
Fine-tuned DistilBERT	**U**	**U**	**0.79**	**0.69**	**0.67**
Cascade model	U	U	0.57	0.52	0.53
Logistic Regression w. masked lexicon	U	U	0.74	0.54	0.39
Fine-tuned DistilBERT w. masked lexicon	**U**	**U**	**0.77**	**0.66**	**0.65**
Cascade model w. masked lexicon	U	U	0.61	0.57	0.58
Logistic Regression w. deleted lexicon	U	U	0.63	0.47	0.34
Fine-tuned DistilBERT w. deleted lexicon	**U**	**U**	**0.63**	**0.61**	**0.59**
Cascade model w. deleted lexicon	U	U	0.51	0.52	0.51

Models trained with news from all countries except for Ukraine and tested with news from Ukraine:

Method	Train	Test	Prec.	Recall	F1
Logistic Regression	A-U	U	0.27	0.02	0.04
Fine-tuned DistilBERT	A-U	U	0.21	0.17	0.18
Cascade model	**A-U**	**U**	**0.42**	**0.28**	**0.33**
Logistic Regression w. masked lexicon	A-U	U	0.26	0.02	0.05
Fine-tuned DistilBERT w. masked lexicon	A-U	U	0.16	0.15	0.15
Cascade model w. masked lexicon	**A-U**	**U**	**0.36**	**0.27**	**0.32**
Logistic Regression w. deleted lexicon	A-U	U	0.24	0.03	0.05
Fine-tuned DistilBERT w. deleted lexicon	A-U	U	0.20	0.19	0.19
Cascade model w. deleted lexicon	**A-U**	**U**	**0.37**	**0.25**	**0.33**

Models trained with news from Ukraine and tested with news from all countries except for Ukraine:

Method	Train	Test	Prec.	Recall	F1
Logistic Regression	U	A-U	0.27	0.40	0.20
Fine-tuned DistilBERT	U	A-U	0.34	0.37	0.37
Cascade model	**U**	**A-U**	**0.42**	**0.51**	**0.39**
Logistic Regression w. masked lexicon	U	A-U	0.24	0.41	0.21
Fine-tuned DistilBERT w. masked lexicon	U	A-U	0.35	0.37	0.36
Cascade model w. masked lexicon	**U**	**A-U**	**0.39**	**0.48**	**0.36**
Logistic Regression w. deleted lexicon	U	A-U	0.23	0.36	0.17
Fine-tuned DistilBERT w. deleted lexicon	U	A-U	0.32	0.31	0.32
Cascade model w. deleted lexicon	**U**	**A-U**	**0.36**	**0.35**	**0.37**

Table 1: Model performance, measured in macro precision, macro recall, and macro F1 score, evaluated depending on training and testing subsets. The best results are highlighted in bold.

2019). The results showed that the proposed cascade system, which was capable of detecting subtle linguistic patterns and techniques, outperformed the other two approaches in terms of identifying media bias in the Ukraine crisis news dataset when training and testing the model with news from different countries.

Our results also indicate that classical lexical-based techniques are useful for detecting word choice/labeling media bias, but they are not as capable of generalizing other forms of media bias as models based on transformers. Furthermore, introducing methods that detect specific forms of bias improves even more state of the art performance models. This is an important finding, as it suggests that lexical-based models may not be suitable for detecting subtle forms of media bias, such as persuasion or rhetoric. It is therefore important to develop more sophisticated models that are capable of detecting more subtle forms of media bias.

Our proposed methodology is a promising approach for detecting media bias. Also, using the explainable AI techniques used to obtain the lexicon have allowed us to uncover the patterns in the model's predictions and to identify the words and terms that are most predictive of bias in this given context, helping us in the evaluation process. In addition, the cascade transformer-based model has allowed us to detect specific forms of media bias, such as appeal to authority, appeal to groupthink/popularity, red herring, and loaded language.

In conclusion, our experiments show that cascade transformer-based models are better suited for detecting media bias than lexical-based models. Additionally, our experiments suggest that these models are better able to detect more subtle forms of bias, such as persuasion or rhetoric. We believe our approach holds promise for detecting media bias in different contexts, and could be further improved and adapted to detect other types of bias.

In future work, we plan to extend our methodology to other domains and contexts, as well as to explore how our approach can be applied to media bias detection in other languages. In addition, we plan to further improve the models by introducing additional bias lexicons and incorporating additional explanatory AI techniques, such as counterfactual explanation (Hsieh, Moreira, and Ouyang, 2021).

Furthermore, we recognize the importance of considering media bias detection in different languages. Language-specific nuances and cultural contexts play a significant role in shaping bias, and studying media bias detection in other languages will help broaden the applicability and impact of our research. By incorporating multilingual datasets and language-specific models, we can capture and analyze bias patterns unique to different linguistic contexts.

We will also expand the application of our methodology to other domains, such as social media content analysis, political speech analysis, and more. We aim to investigate media bias detection in different languages to make our approach more universally applicable. This would involve incorporating additional bias lexicons such as the Hyperpartisan news detection (Kiesel et al., 2019) from SemEval 2019, and conducting multilingual media bias analysis, with the intent of identifying and understanding regional and cultural nuances in media bias.

To ensure the transparency and interpretability of our models, we also intend to leverage advanced explanatory AI techniques like counterfactual explanation (Hsieh, Moreira, and Ouyang, 2021). This will provide insights into the decision-making process of the models and will help users understand why a particular piece of content was flagged as biased.

Through these ongoing efforts, we aim to advance the field of media bias detection, contribute to more comprehensive and interpretable models, and provide resources that promote media literacy and critical thinking.

Acknowledgments

This work was supported by the Spanish Ministry of Science and Innovation under the research project FAIRTRANSNLP-DIAGNÓSTICO: Midiendo y cuantificando el sesgo y la justicia en sistemas de PLN (PID2021-124361OB-C32). This work has been also funded by the Ministry of Universities and the European Union through the EuropeaNextGenerationUE funds and the "Plan de Recuperación,

Transformación y Resiliencia".

Also, we would like to thank Andres Cremisini, Daniela Aguilar, and Mark A. Finlayson for providing the media bias detection dataset.

References

Aggarwal, S., T. Sinha, Y. Kukreti, and S. Shikhar. 2020. Media bias detection and bias short term impact assessment. *Array*, 6:100025.

Aires, V. P., J. Freire, and A. S. da Silva. 2020. An information theory approach to detect media bias in news websites.

Baly, R., G. D. S. Martino, J. Glass, and P. Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*.

Chen, W.-F., K. Al-Khatib, B. Stein, and H. Wachsmuth. 2020. Detecting media bias in news articles using gaussian bias distributions. *arXiv preprint arXiv:2010.10649*.

Cremisini, A., D. Aguilar, and M. A. Finlayson. 2019. A challenging dataset for bias detection: the case of the crisis in the ukraine. In *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, pages 173–183. Springer.

Enomoro, S. and T. Eda. 2021. Learning to cascade: Confidence calibration for improving the accuracy and computational cost of cascade inference systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7331–7339.

Hsieh, C., C. Moreira, and C. Ouyang. 2021. Dice4el: interpreting process predictions using a milestone-aware counterfactual approach. In *2021 3rd International Conference on Process Mining (ICPM)*, pages 88–95. IEEE.

Hube, C. and B. Fetahu. 2018. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786.

Kiesel, J., M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. 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, June. Association for Computational Linguistics.

Lazaridou, K. and R. Krestel. 2016. Identifying political bias in news articles. *Bulletin of the IEEE TCDL*, 12.

Lin, Y.-R., J. Bagrow, and D. Lazer. 2011. More voices than ever? quantifying media bias in networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 193–200.

Lundberg, S. M. and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ogawa, T., Q. Ma, and M. Yoshikawa. 2011. News bias analysis based on stakeholder mining. *IEICE transactions on information and systems*, 94(3):578–586.

Patricia Aires, V., F. G. Nakamura, and E. F. Nakamura. 2019. A link-based approach to detect media bias in news websites. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 742–745.

Piskorski, J., N. Stefanovitch, G. Da San Martino, and P. Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada, July.

Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Ribeiro, M. T., S. Singh, and C. 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, pages 1135–1144.

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

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.