Detecting biases in fake news detection
Project PoC for NLP Course, Winter 2024

Dawid Płudowski, Anotni Zajko, Mikołaj Roguski, Piotr Robak
Warsaw University of Technology
{dawid.pludowski, antoni.zajko, mikolaj.roguski, piotr.robak}.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Automated fake news detection is a topic of great importance for modern society and for which NLP techniques have great potential to be applied. While researchers propose multiple well-performing fake news detectors each year, the question arises as to whether these models are not biased towards certain, specific entities, like persons and organizations. While much research about bias towards a general group of people was done, none of them explains how to detect a bias towards smaller, specific subgroups that do not occur widely in the training datasets. In this work, we propose a framework to assess biases in the models that operate on tokens – namely transformers. For this purpose, we will use custom-parametrized eXplainable AI (XAI) techniques to detect the importance of chosen named entities on the final prediction of models, as well as counterfactual techniques to present how swapping between certain persons or organizations in the sentence may lead to model's reasoning to change its decision. Finally, we suggest mitigation measures to prevent this harmful behaviour.

1 Introduction

In this article, we verify whether the fake news detectors based on neural networks are biased toward certain entities. As much news on the Internet is focused on certain persons (e.g., politicians, influencers, celebrities), there is a significant risk that many state-of-the-art models for fake news detection may focus their decision processes mainly on the particular sub-group of named entities or even on separate persons, which may lead to undesired model behaviour that cannot be detected by evaluation on a test dataset. Such analysis can be realized manually with the help of explainability AI (XAI) techniques, e.g., attributions methods like shapley values [1] or feature ablation [2]. However, this kind of assessment tends to require a tremendous workload of human beings and is prone to errors. Moreover, the quality check of this process is, in fact, as demanding as the process itself. Thus, there is a great need for a fully automated framework to verify the magnitude of the bias that the specific model has on the specific persons. In this work, we approach this task using counterfactual-based methods that not only produce easy-to-interpret bias scores but also a series of semantic examples of bias in sentences by simple replacement of tokens.

We believe that our work is of special significance when explaining a model of not nearly perfect performance. In these cases, the complex and black-box nature of NLP-oriented models, e.g., transformers, often obscures the rather simple and straightforward nature of the reasoning method that is focused on the overrepresentation of the specific examples in the training dataset [3].

We address this problem twofold: first, we detect how important named entities (NE) are to state-of-the-art models. For this purpose, we prepare NE-aware modification of already existing XAI methods and counterfactual-inspired swapping methods to present how changing only the names of persons or organizations may lead to a flip of the predicted label. Next, we verify the impact of anonymizing some of the NEs on the state-of-the-art models' performance. Our results suggest that the magnitude of the bias hidden in the training data can have a great impact on the fine-tuned model's reasoning, but this can be easily mitigated using simple preprocessing techniques.

The rest of this article is structured as follows: first, we describe the current state of the knowledge in the area of fake news detection and using

XAI in NLP in Section 2. Next, we explain the methodology of our work, together with the listing of the used artifacts in Section 3. Based on it, we go to the results of our work, described in Section 4. Finally, we discuss and summarize the outcome of the article in Sections 5 and 6. Additionally, we discuss a contribution of each team member and our replies to the reviewers in Sections 7 and 8.

2 Related works

In recent years, several works about detecting fake news in an automated manner were proposed [4, 5]. In particular, NLP-based approaches occurred to be the most successful in this field [6, 7, 8]. Among all of the articles on this topic, the ones based on deep learning transformers highlight the high performance of the transformers [9, 10], yet the risk of using "black-box" approaches in such a human-oriented task poses a challenge of explaining the models' reasoning. Models like RoBerTa [11] embed the words into latent space in which some sensitive words can be used in an unpredicted, harmful way. Some authors highlight the fact that this can be done twofold [12]: the transformers may be already pre-trained with the bias, or the bias can be introduced on the stage of fine-tuning. In reality, both sources of bias are relevant in real-life scenarios, but the distinction between them is of great importance as we can directly mitigate only the latter one effectively [13].

Similarly, Large Language Models (LLM) like GPT [14] may be biased [15] towards certain responses, even if guards and prompt engineering are applied [16, 17]. This fact is even more alarming than the bias introduced in transformers, as the complexity of the LLMs makes fine-tuning them almost impossible for the average user. Moreover, some malicious techniques may be used as a source of the bias, for example, prompt engineering [18].

This risk should be evaluated and the bias should be explained. The current state of the XAI techniques for NLP, however, is still premature [19]. While many methods were invented and even more were re-implemented for attention-based models [20], it seems that the community has not agreed on the universality of any of them, contrary to, e.g., using Shapley values [1] on tabular data or diffusion models to create counterfactuals in computer vision [21]. This creates many risks, as there is no straightforward and simple way to secure the black-box models in NLP. However, it also creates an opportunity to experiment with different methods in this field. In particular, counterfactual-based methods are of great value [22], as they provide the most intuitive way of explanation, answering the "what if?" question [23] which can be used to fast and robust manual quality check of the explanation.

To the extent of our knowledge, no NEs analysis for fake news detection has been published in the literature to this moment. In fact, several articles about general bias exist [15, 17], however, they do not provide a framework to detect the bias towards specific entities. To fill this gap, we propose our analysis supported by XAI tools.

3 Methodology

Our research is based on the 5 artifacts provided in Table 1. Below, we discuss each of these elements.

Table 1: Table containing major elements of our methodology. In order of datasets used, models we chose, mechanism of named entity extraction and eXplainable Artificial Intelligence methods.

Artifact	Elements
Data	ISOT, CoAID, LIAR
Models	RoBERTa, ERNIE
NER	spaCy
XAI	Attribution maps, Counterfactuals

3.1 Data

As our data, we use three datasets about fake news detection: LIAR [24], CoAID [25], and ISOT [26]. The statistics from them are summarized in Table 2.

LIAR dataset contains the tweets of American politicians, often referring to other politicians. Each tweet is labelled with the level of the truth in the text (e.g., truth, partially truth, not truth). Because we would like to focus here on the binary classification, the levels of truth were grouped into two levels used for the training and the evaluation phase – "truth" and "not truth". The data also contains a lot of additional features, such as describing the political views of the speaker, their state of origin, the topic of news, and many more. While many works use this additional information for feature engineering purposes, we decided to drop them from the data so the model learns only from the raw text. We believe this approach is closer to the realistic case of training in production-ready fake news detection when the volume of data makes this additional labelling cumbersome. The dataset has become popular in the realm of fake news detection, with nearly $2,000$ citations. What is important from our perspective, it contains a lot of NEs related to persons.

The CoAID dataset lists articles and news about the COVID-19 pandemic. The news is labelled as either containing true information or misinformation. In the case of this dataset, the number of NEs related to persons and institutions is rare compared to LIAR, so we expect this particular dataset will create a smaller bias in the model.

ISOT fake news dataset is a set of articles from several sources, gathered and published by the University of Victoria[1]. Although it has not been published at any conference, it has a significant number of downloads on the Kaggle platform, over 10 thousand. We find no information about the quality of this data, so the results obtained with it should be treated as secondary compared to the previous two.

We acknowledge that there are multiple data sources that could be additionally used in our research. We did not decide to use them mostly because of the size of the experiments (over 24h GPU time so far). What is more, we would like to use in our analysis only data that is either well-cited or well-downloaded in Kaggle.

3.2 Models

For our study, we use two pre-trained transformer models, which we fine-tuned to each dataset separately. Each training is performed three times with different seeds. All of the models are taken into consideration when comparing the performance of the models in the classification task. For each model, the run with the best validation accuracy was selected for XAI tasks.
The first model we used is RoBERTa, while the second one is ERNIE [27]. RoBERTa is a BERT encoder with a special training technique that is proven to provide state-of-the-art results in fake news detection [28]. On the other hand, ERNIE is a model that is trained on external knowledge, such as knowledge graphs or linguistic information, and it has been proven to be suitable for challenging semantic tasks. What is even more important from the perspective of our research is that ERNIE is trained using a more careful selection of data sources and was primarily developed to support the Chinese language. As our data are focused on news created by the Western world, especially the United States of America, its bias towards specific persons from American politics may be less highlighted in the pre-trained version. We narrowed our research to these models because they yield state-of-the-art performance and are purely transformers, which is crucial for the XAI methods we plan to apply.

Table 2: Table containing basic statistics about datasets. From the top: number of observations, average observation text length, the average number of NEs in an observation, average ratio of NEs to text length (in tokens), and ratio of fake and factual news.

Dataset	CoAID	ISOT	LIAR
Observations	5457	44954	12796
Avg. text len.	66.5	80.1	107.1
Avg. # NE	0.668	1.15	0.78
# NE / Text len	0.058	0.076	0.037
Fake / True	0.17	0.48	0.47

We decided to exclude from our research large language models (LLM) due to the multiple obstacles to performing the evaluation of equal quality as on the transformers. We listed some of them to support our decision: refusing to give an answer, nondeterministic behavior, the need for extensive prompt engineering, lack of access to the most popular ones, computing and financial thresholds, and the inability to mitigate their potential weaknesses using fine-tuning.

3.3 Named entity recognition

For the recognition of named entities (NEs), we use spaCy [29] package, which contains models performing this task. We decided to use their most capable model for English language – `en_core_web_lg`. During manual verification,

[1] `https://onlineacademiccommunity.uvic.ca/isot/datasets/`

we found out that it is capable to recognize as a person tokens containing surnames of most popular USA politics. While choosing the correct NER model is crucial for the success of our analysis, we want to analyze transformers, not NEs, and thus, we do not decide to test multiple of them.

3.4 XAI methods

As explainability methods, we put emphasis on the two main branches of XAI: attribution maps [30] and counterfactuals [31].

3.4.1 Attribution maps

There are many techniques for the task of assigning attribution for NLP models [32]. For example, in [33], authors propose attention rollout as an attention method specifically designed for attention-based models. Another gradient-based method was proposed in [34], where authors show how attention can be used to detect the interaction between elements of the input. While the mentioned methods present great value for our research, they suffer from the requirement of access to the model's weights, which is unrealistic in the black-box scenario, which we would like to address in this work. Thus, we decided to keep our research simple and stick to the feature ablation method [2], which is model-independent and does not require significant computation time.

3.4.2 Counterfactuals

Using the outcomes from the previous attribution-based analysis, we try to construct counterfactuals to fool the model [23]. In particular, we base our method on swapping important persons among the dataset to trick a model into changing its decision. As an example, we construct a counterfactual observation that will change "Person X said [objective fact]" into "Person Y said [objective fact]". This task can formalized in the following form:

Definition 1. *Let f be a classification function, G be a named entities group, x be an observation and x_i be i-th token. We consider $\hat{x}$ as a good counterfactual of x if $f(x) \neq f(\hat{x})$ and $\{i : x_i \neq \hat{x}_i\}$ is of minimal size and contains only indices of tokens from group G.*

Thus, the difference minimization is done in Hamming distance [35] with constraint to a subgroup of tokens. In our research, we treat G as a group of all person-related tokens. For algorithmic simplicity, we aim to construct counterfactuals that have a difference only on a single token.

We construct this type of counterfactual explanation by using the following algorithm.

Algorithm 1. *Having the text observation $x = (x_1, \ldots, x_k)$, the named entity mask $m_G \in [0, 1]^k$, and the list of the most important I_{min} (negatively) and I_{max} (positively) person entities in the dataset, the counterfactual is crafted by replacing token x_i if $m_i = 1$ with the randomly selected named entities from I_{min} (if the original prediction is positive) or I_{max} (if it is negative).*

Intuitively, the counterfactuals are easy to create if the replaced NE has significant attribution with the same sign to the model's prediction (negative attribution for predictions close to 0 and positive for close to 1). However, this algorithm does not guarantee success, and in particular, attributions do not reflect the full model's reasoning, so using them to select the most promising choices may be misleading. In our experiments, we set a size of both I_{max} and I_{min} to 10. If the observation x contains more than one person token (formally, $\#\{i = 1 : i \in m_G\}$), we choose one to be replaced randomly.

3.5 Evaluation methods

During our research, we examine the values of the following metrics, which allow us to conclude whether a certain model is biased on a certain dataset:

- model performance measured in *accuracy* – the traditional performance of the classification,
- ratio between average importance of person tokens and rest of the tokens – metric of ability of generating counterfactuals,
- number of persons that easily change the prediction of the model – the number of entities affected by bias,
- number of observations for which the counterfactual can be crafted – true measure of bias.

Among all of them, we will consider the last one as the most important in our research.

4 Results

In this section, we describe our research results. First, we provide details about the training of the transformers. Next, we show the results of

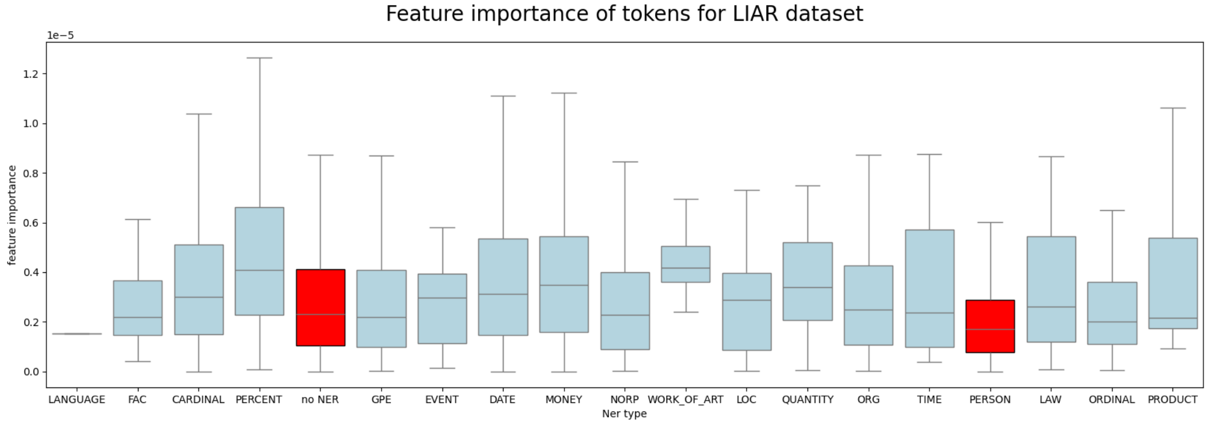

Figure 1: Feature importance of each NE group for LIAR dataset, according to RoBerTA as a classifier. We emphasize the results for “PERSON” tokens and no-entity tokens using the red colour.

attribution-based analysis. Finally, we present the results of the counterfactual generation process, along with their randomly sampled examples.

4.1 Transformers

In Table 3, we provide metrics of the fine-tuned models on all datasets. We performed two types of training – firstly, we trained on a basic version of datasets, and then we trained on the version with all tokens corresponding to persons replaced with ”John” so the models can still recognize person-related entities but cannot distinguish between them. This simple yet effective technique is our proposal for mitigation of bias in transformers. In most cases, masking a person’s NE results in better scores, up to even 10% for RoBerTa on the ISOT task. For datasets on which the score is worsened after masking, the performance degradation is not greater than 1%. However, employing masking for the ERNIE model, makes the results more unstable (see standard deviation in Table 3. While the magnitude of training’s instability is not high, further investigation should be conducted to verify this phenomenon.

The metrics we obtained can be considered good enough to perform further analysis (according to, e.g., results from [24]). Except for the ISOT dataset, both models achieve similar results, which suggests their comprehension of the task is on a similar level. This makes potential differences in their bias more highlighted, as one could simply choose a less biased one without loss of performance.

4.2 Explanations

In the explanation analysis, we experimented with attribution methods to capture the dependency of the model on specific names or surnames. Not surprisingly, considering the results shown in Table 3, the NEs related to the persons are not more important compared to other tokens. This is shown in the Figure 1. However, for some observations, a person entity may be among the most important tokens in a way that may be harmful to the particular person. For these observations, the impact of the specific persons may be significant and creates an opportunity to easily fool the model to change its prediction. We show this phenomenon in Figure 2.

Next, we try to craft counterfactuals by swapping names between persons to show that the model is biased toward one of them. the results are presented in Figure 3 and 4. The persons listed in Figure 4 can be successfully inserted into the sentence to fool the model to change its prediction. Thus, we conclude they may be really negatively affected by using this kind of fake news prediction in real-life production systems.

This part of the research was successful, mostly in the LIAR dataset. For the CoAID dataset, the presence of the person NE was too sparse to have any major impact, and any permutation of them did not create a difference significant enough to flip the model prediction. The ISOT dataset proved to be the most problematic. Although many counterfactuals were created during the evaluation phase (over 40% success rate in this task), we observed that the model’s behaviour tends to be much more random, and thus, we con-

Table 3: Comparison of accuracies of RoBERTa and ERNIE fine-tuned on datasets with and without persons. We report averages and standard deviations.

Dataset	**Accuracy**	
	RoBerTa	ERNIE
LIAR	0.667 +/- 0.013	0.669 +/- 0.009
LIAR without persons	0.675 +/- 0.011	0.666 +/- 0.033
COAID	0.979 +/- 0.001	0.979 +/- 0.000
COAID without persons	0.982 +/- 0.001	0.971 +/- 0.008
ISOT	0.841 +/- 0.000	0.983 +/- 0.000
ISOT without persons	0.935 +/- 0.000	0.984 +/- 0.001

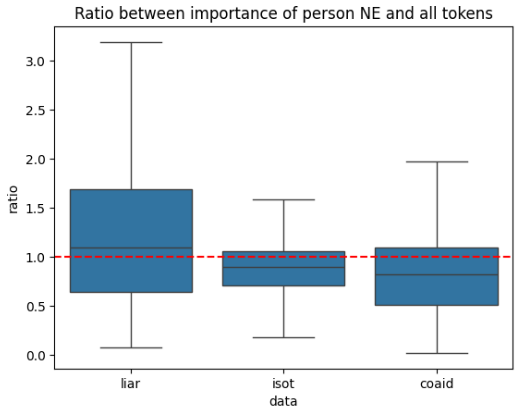

Figure 2: Ratio of importance between person NEs and other tokens. For the LIAR dataset, in over half of observations, person NEs are more important than the rest part of the sentence. For all datasets, in over 25% observations, the person NEs are more important, suggesting models' reliance on these tokens.

sider this phenomenon not as a bias but rather a lack of complex reasoning of the model. While investigating the cause of such behaviour is out of the scope of our work, we would like to highlight that models trained on this dataset tend to return only two unique values of probability (presented in Table 4, which may suggest some overtraining or even data leakage between observations in the original data.

Looking at the Figures, it is clear that the model favours specific persons, depending on the distribution in the dataset. Examples of such bias, based on LIAR dataset, are presented below.

Example 1. *Mitt **Romney** drove to Canada with the family dog Seamus strapped to the roof of the car. – 8% probability of fake news.*

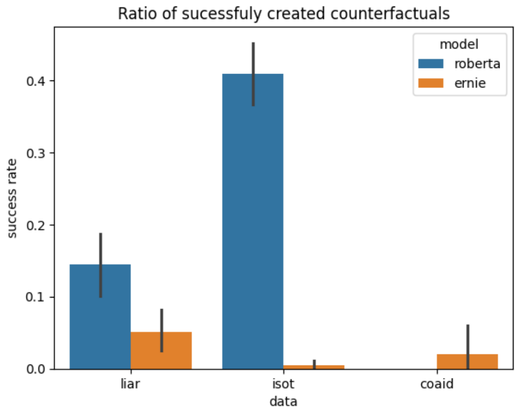

Figure 3: Ratio of success in creating counterfactuals on different datasets. The high success rate of RoBerTa on the ISOT dataset comes from overfitting.

Example 2. *Mitt **Obama** drove to Canada with the family dog Seamus strapped to the roof of the car. – 79% probability of fake news.*

Fake news is mostly intended to describe someone in a negative context. Thus, the examples above show that the model is in strong favour of Barack Obama while it is offensive towards Mitt Romney. Here, it is obvious that the fake part of the news is "dog [...] strapped to the roof of the

Table 4: Probability values returned by both RoBerTa models on the test part of ISOT dataset.

	Value	Ratio
ISOT	0.000347	37%
	0.999907	63%

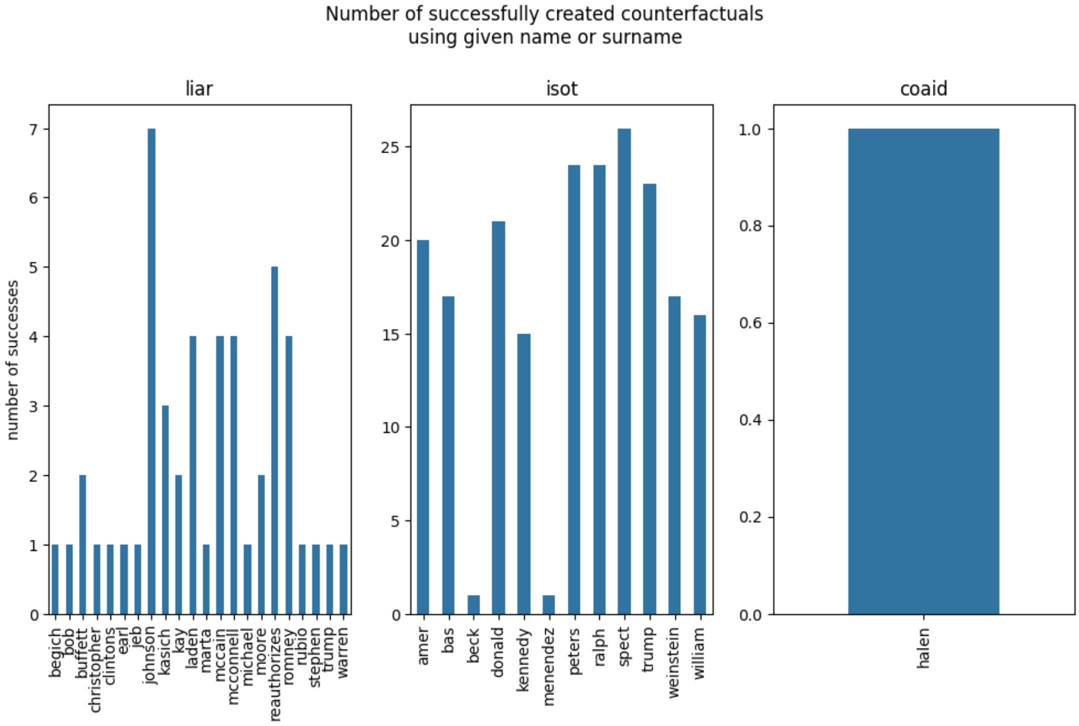

Figure 4: Persons's names and surnames that were successfully used to create counterfactuals. Models used in our research are biased towards these persons the most.

car" and we would like our model to not discriminate just because of the name of the person.

A similar, less obvious example is provided below. Ad-hoc analysis of the model's behaviour suggests that it classifies any news with negative sentiment as a fake if "Barack Obama" is present in its content.

Example 3. *Toomey and **Trump** will ban abortion and punish women who have them. – 7% probability of fake news.*

Example 4. *Toomey and **Obama** will ban abortion and punish women who have them. – 68% probability of fake news.*

5 Discussion

Our results confirm the hypothesis of the bias towards NEs in state-of-the-art solutions for fake news detection. It is aligned with the results obtained in the literature (see Section 2). While analysing our results, one needs to remember that we succeeded on showing the bias on datasets containing a lot of tokens related to persons. We did so to highlight this specific kind of bias. However, different types of datasets may suffer from bias toward other NEs that may be considered undesired behaviour. Our work provides a sample framework to analyse these scenarios.

Moreover, we showed that simply masking NEs with meaningful yet repetitive phrases does not decrease the model's performance while making the risk of bias less severe. The presented results suggest also that the ERNIE model is less prone to be biased towards specific persons. This fact should be further investigated, especially in the context of the quality of their training dataset. We believe that the outcome of our work will inspire the researchers to craft a fair and unbiased model in the domain of natural language processing, which is crucial to not harm people by using AI in production systems.

6 Conclusions

To summarize, in our work, we achieve the following things:

1. the bias towards specific persons was recognized in the widely-used model for fake news classification,

2. the magnitude of the bias was measured and plotted,

3. persons treated unfairly by the model were identified and listed,

4. the mitigation measure was proposed and successfully applied.

We believe our work is of of the the first that applies the counterfactual techniques to access biases in the black-box model in NLP on a large scale.

Impact and ethical concerns. Our work aims to decrease the ethical risks that are part of the using AI-powered solutions in the public services. However, it might be used to mislead the owners of the AI solution that it is fully ethical and secure to be released on the production. Thus, we would like to highlight here that one should assess the bias in their model in multiple independent ways and by no means, should they really solely on our framework.

Future works. We acknowledge that we left several experiments and extensions to this article as future work. Here, we discuss only the bias towards person NEs, which is often seen as a group that should be treated with particular security in AI-oriented pipelines. However, a similar evaluation should be done for other groups of NEs, such as organizations, locations or products. In our work, we do not distinguish between names, surnames, and titles (e.g., sir, Mr.) which decreases the plausibility of the generated counterfactuals (in particular, semantically incorrect sentences may appear in which name is used as a surname, e.g., “Barack Donald did ...”. To fix this, the named entities should be captured with better algorithms to match name, surname and title, even if they do not appear one by one in the sentence.

Another big step in this research is to extend it to evaluate LLM, both open-source and commercial. To do so, the mitigation measure other than processing the fine-tuning data needs to be researched. For both transformers and LLMs, the bias introduced in the foundational model and additional context (training data for the former and prompt for the latter) should be distinguished and measured.

The open question is the reason for the bias in the model. In the examples from Section 4, it was presented that the presence of “Obama” increases the chances that the model classifies the news as fake. We verified that 70% of news about Barack Obama in training data contains fake information. On the contrary, the ratio of fake information about Donald Trump of Mitt Romney is only 30% which suggest that the models’ reasoning may tend to be much more simplistic than we expect from their complexity. Our further research will try to answer this question.

7 Rebuttal

During our work we received reviews from two teams of our peers. Both reviews proposed some interesting improvements, and highlighted the strengths and weaknesses of our approach, with mostly positive feedback. Of the suggestions we received most were what we had already planned for the final milestone such as:

- evaluation of different models,
- full automation of counterfactual generation,
- evaluation of whether masking named entities would improve training outcomes.

All of these suggestions were applied as per the original plan.

The other suggestions we received were:

- To extend the scope of the project by including other types of named entities – considering other named entities would go far beyond our planned scope of work.
- Include other model evaluation metrics – however as the underlying task was classification with a sufficiently balanced distribution of positive and negative samples we deem accuracy a sufficient and best fit metric of evaluation.

8 Contribution

Here, we present the contribution of each coauthor of this paper. As we worked with internal review system, most coding and writing task have assigned two persons – one for the actual work and the other for the quality review. Please note that theoretical aspects of this work were discussed by all team member equally, regardless of the final decisions and implementations. The contributions are presented in Table 5.

References

[1] Eyal Winter. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

Table 5: A contribution of each team member.

Task	Owner	Reviewer
Problem definition	Dawid Płudowski	
Project management	Dawid Płudowski	Antoni Zajko
Literature review – XAI	Piotr Robak	Dawid Płudowski
Literature review – models and data	Anotni Zajko	Piotr Robak
Model training and evaluation	Anotni Zajko	Mikołaj Roguski
Attribution implementation	Dawid Płudowski	Piotr Robak
NE-attribution mapping implementation	Piotr Robak	Dawid Płudowski
Counterfactual implementation	Mikołaj Roguski	Dawid Płudowski
Clean code and scripts	Mikołaj Roguski	Piotr Robak
Results analysis	Dawid Płudowski	Antoni Zajko
Infrastructure (cloud computing etc.)	Mikołaj Roguski	Antoni Zajko
Review session	all	

[2] Luke Merrick. Randomized ablation feature importance. *arXiv preprint arXiv:1910.00174*, 2019.

[3] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184, 2021.

[4] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 729–736, New York, NY, USA, 2013. Association for Computing Machinery.

[5] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40, September 2020.

[6] Humberto Fernandes Villela, Fábio Corrêa, Jurema Ribeiro, Air Rabelo, and Darlinton Carvalho. Fake news detection: a systematic literature review of machine learning algorithms and datasets. *Journal on Interactive Systems*, 14:47–58, 03 2023.

[7] Fang Ma and Guoxian Tan. Nlp in fake news detection. pages 71–83, 2021.

[8] Mohammad Hadi Goldani, Saeedeh Momtazi, and Reza Safabakhsh. Detecting fake news with capsule neural networks. *Applied Soft Computing*, 101:106991, 2021.

[9] Tianle Li, Yushi Sun, Shang ling Hsu, Yanjia Li, and R. C. Wong. Fake news detection with heterogeneous transformer. *ArXiv*, abs/2205.03100, 2022.

[10] U. S. S. Varshini, R. P. Sree, M. Srinivas, and R. Subramanyam. Rdgt-gan: Robust distribution generalization of transformers for covid-19 fake news detection. *IEEE Transactions on Computational Social Systems*, 2023.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, 2021.

[13] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *Association for Computational Linguistics (ACL 2019)*, 2019.

[14] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[15] Erik Weber, Jérôme Rutinowski, Niklas Jost, and Markus Pauly. Is gpt-4 less politically biased than gpt-3.5? a renewed investigation of chatgpt's political biases, 2024.

[16] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *ArXiv*, abs/2305.13860, 2023.

[17] Raluca Alexandra Fulgu and Valerio Capraro. Surprising gender biases in gpt. *Computers in Human Behavior Reports*, page 100533, 2024.

[18] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023.

[19] Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. Challenges in applying explainability methods to improve the fairness of nlp models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, 2022.

[20] Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batarseh. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6:1225093, 2023.

[21] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022.

[22] Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. A survey on natural language counterfactual generation. *arXiv preprint arXiv:2407.03993*, 2024.

[23] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[24] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

[25] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset, 2020.

[26] Matthew Iceland. How good are sota fake news detectors, 2023.

[27] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019.

[28] A. Kitanovski, M. Toshevska, and G. Mirceva. Distilbert and roberta models for identification of fake news. In *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, pages 1102–1106, 2023.

[29] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[30] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[31] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[32] Pouya Pezeshkpour, Sarthak Jain, Byron Wallace, and Sameer Singh. An empirical comparison of instance attribution methods for nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, 2021.

[33] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[34] Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878, 2021.

[35] Abraham Bookstein, Vladimir A Kulyukin, and Timo Raita. Generalized hamming distance. *Information Retrieval*, 5:353–375, 2002.