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, on which NLP techniques show great impact. While researchers propose multiple well-performing fake news detectors each year, the question arises as to whether these models are not biased towards certain entities, like persons, and organizations. In this work, we propose a framework to assess biases in the models that operate on tokens – transformers and LLMs. 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.

1 Introduction

In this research, we would like to verify whether the fake news detectors based on neural networks are biased toward certain entities. As many news on the Internet are focused on certain persons (e.g., politicians, influencers, celebrities), and institutions (e.g., government, companies), 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, which may lead to undesired model behavior that cannot be detected by evaluation on a test dataset. For example, if the model is trained on tweets of American politicians, there is a risk that the detection process will be biased towards the texts that refer to the Republican or Democrat politicians, depending on a multitute of factors, including the source of the training data, political views of model creators and many other, very hard to predict reasons. This is especially probable for tasks that are hard (e.g. when performance metrics of SoTA models like F1 do not reach even the 0.8 threshold) where the sophisticated process of the training of the model may collapse to mapping certain tokens as indicators of fake news.

We would like to address this potential problem twofold: first, we would like to detect how important named entities (NER) are to state-of-the-art models. For this purpose, we prepare NER-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 would like to verify the impact of anonymizing some of the NERs on the state-of-the-art models' performance. If possible, measures to mitigate the bias will be proposed.

2 Significance and related works

In recent years, several works about detecting fake news in an automated manner were proposed [1, 2]. In particular, NLP-based approaches occurred to be the most successful in this field [3, 4, 5]. Among all of the articles on this topic, the ones based on deep learning transformers highlight the high performance of the transformers [6, 7], 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 [8] embed the words into latent space in which some sensitive words can be used in an unpredicted, harmful way. Similarly, Large Language Models (LLM) like GPT [9] may be biased [10] towards certain responses, even if guards and prompt engineering are applied [11, 12].

To the extent of our knowledge, no NERs analysis for fake news detection has been published in the literature to this moment. To fill this gap, we propose our analysis supported by explainable AI (XAI) tools.

3 Concept and work plan

To perform the proposed analysis, we are planning to perform the following steps:

1. We will download data and perform simple data engineering on it. It will mainly consist of the extraction of relevant columns, binarization of labels, fixing data format, and train-test split. In the case of CoAID and ISOT datasets, we will only take into account the titles of articles.

2. Next, we will train two fake news detectors. Both of them will be fine-tuned transformers. For each dataset, fine-tuning will be performed separately.

3. Having such models, we will perform their validation on test sets to validate whether they are acceptable fake news detectors.

4. Finally, we will apply XAI methods that will aim to detect biases towards specific named entities.

4 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 datasets used, models we chose, mechanism of named entty extraction and eXplainable Artificial Intelligence methods.

Artifact	Elements
Data	ISOT, CoAID, LIAR
Models	RoBerTa, KnowBert
NER	spaCy
XAI	Attribution maps, Counterfactuals, Thesis evaluation

4.1 Data

As our data, we use three datasets about fake news detection: LIAR [13], CoAID [14], and ISOT [15]. 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). This dataset become popular in the realm of fake news detection, with nearly 2000 citations. What is important from our perspective it contains a lot of NER related to persons.

The CoAID dataset lists articles and news about the COVID pandemic. The news is labeled as either containing true information or misinformation. In the case of this dataset, the number of NER related to persons and institutions is rare compared to LIAR so we expect this particular dataset will create a smaller bias to 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 is not published at any conference, it has a significant number of downloads on the Kaggle platform over 10 thousand.

Table 2: Table containing basic statistics about datasets. From the top: number of observations, average observation text length, average number of ners in an observation, average ratio of NERs 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. # NER	0.668	1.15	0.78
# Ner / Text len	0.058	0.076	0.037
Fake / True	0.17	0.48	0.47

4.2 Models

For our study, we will use two pre-trained BERT models, which we fine-tune to each dataset separately. The first one is RoBerTa while the second one is KnowBert [16]. RoBerTa is a BERT encoder with a special training technique that is proven to provide SOTA results in fake news detection [17]. On the other side, KnowBert is a BERT model that is trained on the knowledge base built upon Wikipedia, which enriches it with news verification capabilities. We narrow our research

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

to these models because they yield state-of-the-art performance and are purely transformer which is crucial for the XAI methods we plan to apply. If possible, from the perspective of using raw tokens, we also try to use Gemini to compare the LLM approach to the transformers.

4.3 NER

For the recognition of named entities, we will use spaCy [18] package, which consists of models performing this task. For the detection of political biases, we will possibly replace these named entities with simple placeholders denoting entities' names which means that all politicians will be consistently called as, e.g., a person.

4.4 XAI methods

As explainability methods, we plan to put emphasis on the two main branches of XAI: attribution maps [19] and counterfactuals [20].

4.4.1 Attribution maps

Our preliminary research showed that some fine-tuned transformers treat NERs associated with persons as twice as important compared to the rest of the tokens. Moreover, collapsing all person-related tokens to only one yields a dramatic decrease in the performance of transformer models. This leads us to the more detailed research in which we:

- attempt to tell if some famous persons are treated as especially important from the perspective of the model's inner reasoning using feature ablation methods [21],
- verify if any other NERs, such as organizations, names of countries etc. have more importance compared to others.

There are many techniques for the task of assigning attribution for NLP models [22]. For example, in [23] authors propose attention rollout as an attention method specifically designed for attention-based models. Another gradient-based method was proposed in [24], 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 weight which is unrealistic in the black-box scenario (LLM case). Thus, we decided to keep our research simple and stick to the feature ablation method [21] which is model-independent.

4.4.2 Counterfactuals

Using the outcomes from the previous attribution-based analysis, we will try to construct counterfactuals to fool the model [25]. In particular, we will base our method on swapping important persons among the dataset to trick a model into changing its decision. As an example, we may 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 [26] with constraint to a subgroup of tokens.

4.5 Thesis evaluation methods

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

- model performance measured in *accuracy*,
- ratio between average importance of person tokens and rest of the tokens,
- number of outlier-like important persons,
- ability to create the counterfactuals based on swapping person tokens between observations.

Among all of them, we will consider the last one as a final measure of bias in the model.

5 Results

In this section, we shortly describe our preliminary research results, presented as a proof of concept (PoC). First, we provide details about the training of the transformers. Next, we show sample examples of explanations we want to expand on in the next weeks. Finally, we discuss what is still missing to finish the project.

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

Dataset	Accuracy
LIAR	0.655 +/- 0.006
LIAR without persons	0.664 +/- 0.015
COAID	0.979 +/- 0.005
COAID without persons	0.982 +/- 0.002
ISOT	0.841 +/- 0.225
ISOT without persons	0.935 +/- 0.049

5.1 Transformers

As for now we focused on the analysis of one model – RoBERTa [8]. In Table 3 we provide metrics of fine-tuned model 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 are not able to distinguish between them. Each training was performed three times with different seeds.

It can be seen that the performance of the model did not degrade after masking of persons. This means that there is a possibility of easy removal of model biases towards specific persons without significant degradations of the performance.

5.2 Explanation

As 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 NERs related to the persons are not more important compared to other tokens. This is shown in the Figure 1. This, however, does not mean that for some observations persons entity may be among the most important tokens in a way that may be harmful for the particular persons.

Next, we try to craft counterfactuals by swapping names between persons to show that the model is biased toward one of them. This part of the research was successful and by our examples, it is clear that the model favours specific persons, depending on the distribution in the dataset. Examples of such bias 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.*

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 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.*

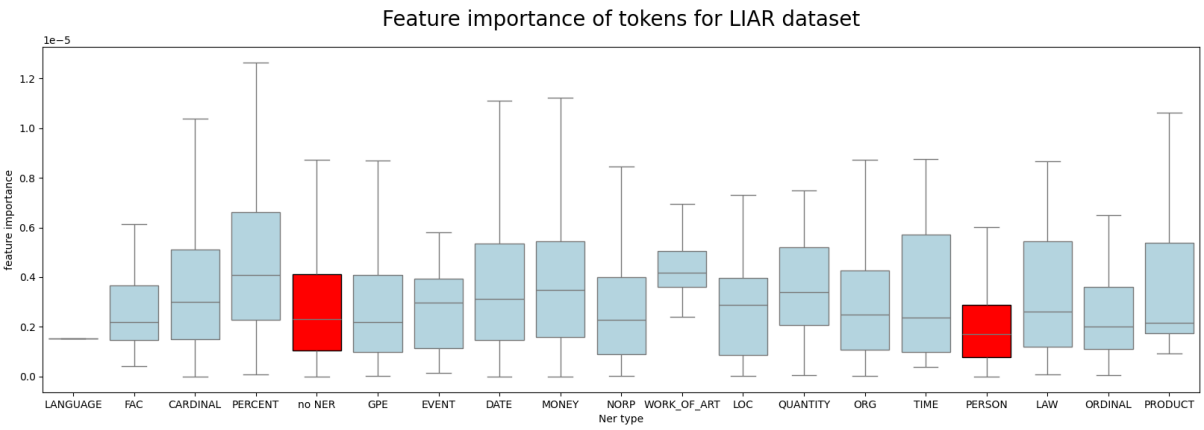

Figure 1: Feature importance of each NER.

For now, our methodology for crafting such examples is as follows:

Definition 2 *Having the text observation $x = (x_1, \ldots, x_k)$, the named entity mask $m \in [0, 1]^k$, and the list of the most important I_{min} (negatively) and I_{max} (positively) person entities in the training 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 label is positive) or I_{max} (if it is negative).*

The definition above provides a good starting point for the automation of creating counterfactuals, yet it is still under development.

5.3 Missing parts

For now, our experiments were based on only one model – RoBerTa. While fine-tuning this model leads to state-of-the-art results, we still want to compare it with KnowBert and possibly an instance of the LLM. Next, we would like to extend our analysis of the importance of person-related NERs. Current analysis may be seen as unintuitive, as the average importance of person NERs presented in Figure 1 is not outstanding, yet we are able to perform successive counterfactual generation. We would like to fill this gap with different importance aggregation techniques to explain this phenomenon more clearly. Finally, our explanatory results are created semi-automatically and only on one dataset. At the same time, we aim to automate this procedure which requires us to craft an appropriate definition of what should be captured by counterfactuals to provide only the most important information to the analysis.

6 Discussion

Our preliminary results reveal the bias hidden in state-of-the-art solutions for fake news detection. While analysing our results, one needs to remember that we selected 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 NERs that may be considered undesired behaviour. Our work provides a sample framework to analyse these scenarios.

Moreover, we showed that simply masking NERs with meaningful, yet repetitive phrases does not decrease the model's performance while making the risk of bias less severe. 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.

The open question is the reason for the bias in the model. In the examples from Section 5, 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%. Our further research will try to answer this question.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] Mohammad Hadi Goldani, S. Momtazi, and R. Safabakhsh. Detecting fake news with capsule neural networks. *ArXiv*, abs/2002.01030, 2020.

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

[7] 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.

[8] 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.

[9] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[10] 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.

[11] 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.

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

[13] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection, 2017.

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

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

[16] Matthew E. Peters, Mark Neumann, Robert L Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *EMNLP*, 2019.

[17] 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.

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

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

[20] 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.

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

[22] Pouya Pezeshkpour, Sarthak Jain, Byron C Wallace, and Sameer Singh. An empirical comparison of instance attribution methods for nlp. *arXiv preprint arXiv:2104.04128*, 2021.

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

[24] 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.

[25] 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.

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