Project Topic Description

Group Project Members: Michael Mistarz, Nikita Sharma, Salma Aly, Chris Matthews, Jamia Rusell
Team Lead: Jamia Russell

Problem Question/Statement: How accurately can machine learning models classify political bias in news articles and social media communities using language patterns and content from Reddit and major news outlets?

We are planning to designate conservative and liberal news articles and channels using posts from subreddits identified as either right-leaning/conservative and left-leaning/liberal. Using that data we can gauge different news articles from similar dates to see if the program can adequately consolidate different news articles into either bucket classified as liberal or conservative news. The model will take language used within news articles to identify similarities in language across reddit posts that are upvoted and consistently used within subreddit groups. Including a probability scale, we can identify the scope of news articles and what direction they lean on the political spectrum. We have two proposed methodologies to complete this project as seen below.

Methodology 1:

Develop a machine learning model to classify political messaging into left, center, and right-leaning categories based on content extracted from social media communities such as Reddit. The model will then be used to test and evaluate political bias in media articles from major news outlets.

Methodology 2:

Develop a machine learning model to classify political messaging into left, center, and right-leaning categories based on content extracted from media articles from major news outlets. The model will then be used to test and evaluate social media communities such as Reddit.

We hope to receive feedback on which method we should use to develop the model. **Which model would be most effective in detecting and classifying language on the political spectrum?** We believe training the subreddit posts language may be more representative of either party identity. These communities often include members who fall on the farther end of each side of the spectrum with keywords/buzzwords indicative of party identity. Comparing media and news channels - which aim to be unbiased- we can designate where the messaging lands on the spectrum. On the other hand, using an existing vetted scale of political party identification documented in new sources as training data we can determine subreddit group party affiliation and identity similarly.

Research Reports:

Analysing Bias in Political News

Abstract:

Although of paramount importance to all societies, the fact that media can be biased is a troubling thought to many people. The problem, however, is by no means easy to solve, given its high subjectivity, thereby leading to a number of different ap- proaches by researchers. In this work, we addressed media bias according to a tripartite model whereby news can suffer from a combination of selective coverage of issues (Se- lection Bias), disproportionate attention given to specific subjects (Coverage Bias), and the favouring of one side in a dispute (Statement Bias). To do so, we approached the problem within an outlier detection framework, defining bias as a noticeable deviation from some mainstream behaviour. Results show that, in following this methodology, one can not only identify bias in specific outlets, but also determine how that bias comes about, how strong it is, and the way it interacts with other dimensions, thereby rendering a more complete picture of the phenomenon under inspection.

Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception

Abstract:

The pervasive spread of misinformation and disinformation in social media underscores the critical importance of detecting media bias. While robust Large Language Models (LLMs) have emerged as foundational tools for bias pre- diction, concerns about inherent biases within these models persist. In this work, we inves- tigate the presence and nature of bias within LLMs and its consequential impact on media bias detection. Departing from conventional approaches that focus solely on bias detection in media content, we delve into biases within the LLM systems themselves. Through metic- ulous examination, we probe whether LLMs exhibit biases, particularly in political bias pre- diction and text continuation tasks. Addition- ally, we explore bias across diverse topics, aim- ing to uncover nuanced variations in bias ex- pression within the LLM framework. Impor- tantly, we propose debiasing strategies, includ- ing prompt engineering and model fine-tuning. Extensive analysis of bias tendencies across different LLMs sheds light on the broader land- scape of bias propagation in language models. This study advances our understanding of LLM bias, offering critical insights into its implica- tions for bias detection tasks and paving the way for more robust and equitable AI systems.

Discovering and Categorising Language Biases in Reddit*

Abstract:

We present a data-driven approach using word embeddings to discover and categorize language biases on the discus- sion platform Reddit. As spaces for isolated user communities, platforms such as Reddit are increasingly connected to issues of racism, sexism and other forms of discrimina- tion. Hence, there is a need to monitor the language of these groups. One of the most promising AI approaches to trace linguistic biases in large textual datasets involves word em- beddings, which transform text into high-dimensional dense vectors and capture semantic relations between words. Yet, previous studies require predefined sets of potential biases to study, e.g., whether gender is more or less associated with particular types of jobs.

This makes these approaches un- fit to deal with smaller and community-centric datasets such as those on Reddit, which contain smaller vocabularies and slang, as well as biases that may be particular to that community. This paper proposes a data-driven approach to auto- matically discover language biases encoded in the vocabulary of online discourse communities on Reddit. In our approach, protected attributes are connected to evaluative words found in the data, which are then categorised through a semantic analysis system. We verify the effectiveness of our method by comparing the biases we discover in the Google News dataset with those found in previous literature. We then successfully discover gender bias, religion bias, and ethnic bias in differ- ent Reddit communities. We conclude by discussing potential application scenarios and limitations of this data-driven bias discovery method.

References

De Arruda, Gabriel, Norton Roman, and Ana Monteiro. "Analysing Bias in Political News." J.UCS (Annual Print and CD-ROM Archive Ed.) 26, no. 2 (2020): 173–99. https://doi.org/10.3897/jucs.2020.011

Ferrer, Xavier, Tom van Nuenen, Jose M Such, and Natalia Criado. "Discovering and Categorising Language Biases in Reddit." arXiv (Cornell University), 2020. https://doi.org/10.48550/arxiv.2008.02754.

Lin, Luyang, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. "Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception." arXiv (Cornell University), 2024. https://doi.org/10.48550/arxiv.2403.14896.

Similar projects:
https://surajkarak.github.io/projects/NLP-Reddit-Political-Bias/

Research Reports

1. https://ojs.aaai.org/index.php/ICWSM/article/view/18048/1785
2. Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception

News Ratings:
Media Bias | AllSides

Subreddits arranged from most liberal (left) to most conservative... | Download Scientific Diagram (researchgate.net)