Donald, Temuulen, Isaia

Detecting Political Bias and Misinformation in News and Social Media Using NLP

Problem Statement

In the midst of political discourse on educational access and defunding research comes the risk of the spread of misinformation across online platforms. Discourse typically exists in the form of tweets and social media discussions, which are often influenced by bias - whether left-leaning or right-leaning; bias can result in a selective news feed and cultivates a limited presentation of information. It is all the more important to be able to detect misinformation. We can see current examples of fact-checking mechanisms such as X's Community Notes, a crowdsourced fact-checking system designed to provide context on potentially misleading posts (Epstein et al., 2023). Community Notes allow users to flag posts and provide factual annotations to misleading content, making it a very valuable tool for training models to detect misinformation and bias. Additionally, recent research shows that incorporating contrastive learning frameworks (i.e., SimCSE) is useful for detecting political bias in news articles. This project will build on these findings by incorporating natural language processing (NLP) algorithms to develop a machine learning model that detects political bias in news sources and online discussions on social media. Beyond detection, the project will explore the relationships and insights between political bias and misinformation, contributing to a deeper understanding of how these elements influence public discourse.

Datasets:

Sentiment analysis models often use the Hugging Face platform to find relevant datasets that contain labeled political bias data (. Examples of datasets we can use are:

- News Sentiment Data (sweatSmile/news-sentiment-data)
- Twitter Community Notes Dataset (from Twitter API)
- Political Bias Dataset (cajcodes/political-bias)
- Political Stance Dataset (strombergnlp/polstance)
- Reddit - reddit comments
- News - Political Bias Dataset

Machine Learning Models

Firstly, we plan to use an unsupervised k-means clustering technique in order to find bias patterns. This will allow us to group together political bias from different sources to visualize the size of the clusters on specific political stances and/or how separated the data is across the political spectrum. We plan to implement an NLP model like a TensorFlow Deep Neural Network, a basic deep learning model. This model will incorporate text representation techniques such as Bag of Words (BOW) representation to classify political bias and text, and Term Frequency-Inverse Document Frequency (TF-IDF) technique, which is to classify text

based on word importance and refines bias classification. Lastly, we plan to refine the model using pre-trained transformer models such as BERT or RoBERTa. How this will work is that it processes all words in a sentence at the same time and using a mechanism called self-attention that analyzes the relationship between the words, as opposed to sequential text processing which is typical of traditional models. Additionally, instead of models reading text from left to right, BERT and RoBERTa read text bidirectionally, which means they analyze a word by considering its relationship with the words before and after it. For our project, we can use these models on a labeled dataset leveraging Twitter's Community Notes and the aforementioned datasets to classify tweets and news articles into categories like political stance (left or right), and factual information (factual or misinformation).

Evaluation Metrics

We use accuracy and ROC-AUC score as measures of model performance for the label-based models. For unsupervised evaluation, we use k-means clustering since k-means do not have pre-defined class labels, but we will look at the distribution of the political sources within clusters. We also plan to use Spearman Correlation for SimCSE (Contrastive Learning Model) which calculates the average similarity between test sentences and sources, which predicts the most likely news source based on the highest correlation.

References

Epstein, R., Zollo, F., & Del Vicario, M. (2023). The role of crowdsourced fact-checking: Analyzing Twitter's Community Notes in combating misinformation. Journal of Computational Social Science, 5(2), 89-104.

Nadeem, M. U., & Raza, S. (2023). Detecting Bias in News Articles using NLP Models. Stanford CS224N Project Report. Retrieved from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1224/reports/custom_116661041.pdf