Predicting Bias Words through Feature Contributions

Sophia Walton

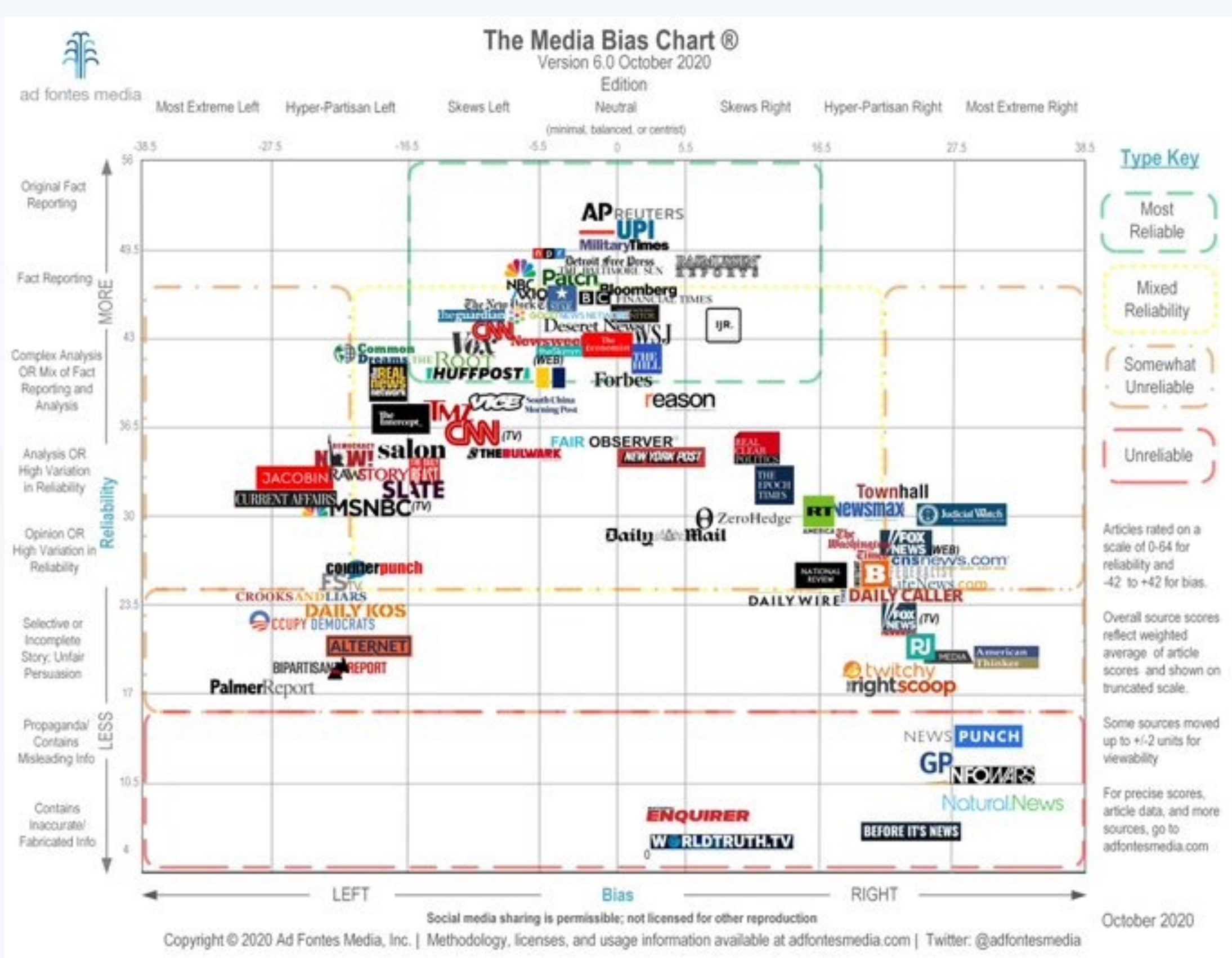
University of Virginia Interpretable Machine Learning

INTRODUCTION

What is Media Bias?

In today's world, media bias has become a huge problem that permeates society. This can have far-reaching consequences. News is not always news, and we must identify what is real news.

When reading news, it is important for people to recognize the biased positions that our media presents. This can help reduce the consequences of media bias on society.



(Aquino)

How does this affect Machine Learning?

When biased news becomes a part of the training data of a model, the biases that exist in the media can be perpetuated to the predictions that a model makes.

But what if we can identify bias in a text as simply biased or factual? This could allow us to account for at least one aspect of the bias problem in machine learning.

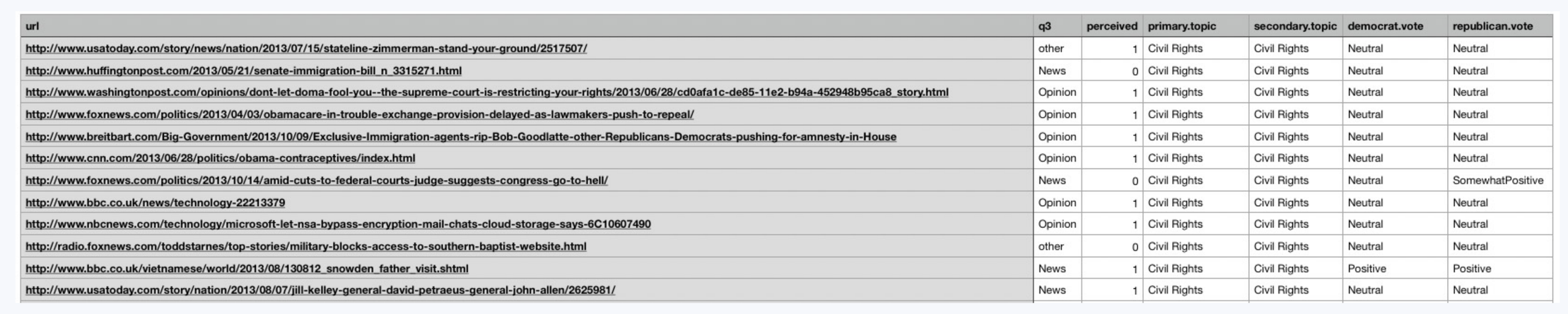
OBJECTIVES

1. Use a neural network to identify biased news articles and texts.
2. Measure the quality of predictions in a biclass classifier.
3. Determine the contributions of each feature to the prediction.
4. Learn whether or not the model is able to recognize the same bias words as human evaluators and can be effective at detecting bias.

METHODS

DATASETS AND PREPROCESSING

1. Media Bias Dataset (Ceren Budak et. al 2016)



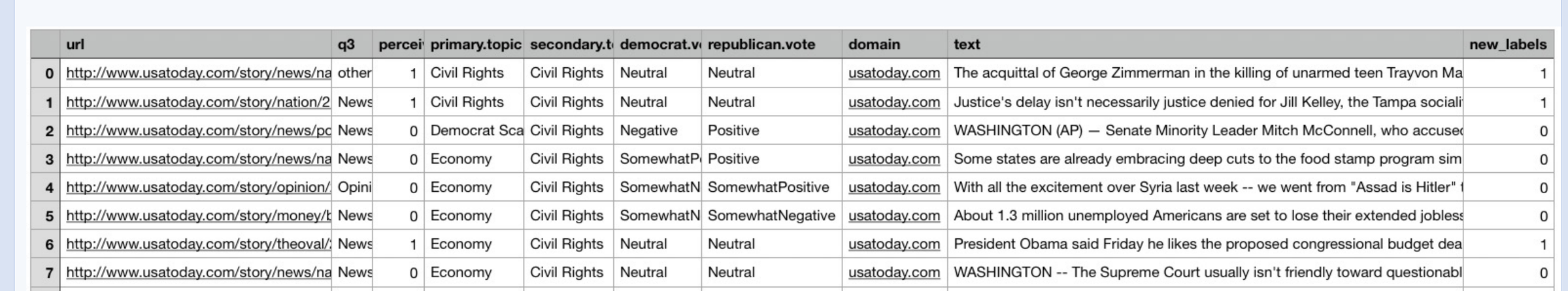
The original dataset

A collection of ~21000 URLs from 16 news sites

↓ Data cleaning (remove error codes and video data)

~16000 URLS from 13 news sites

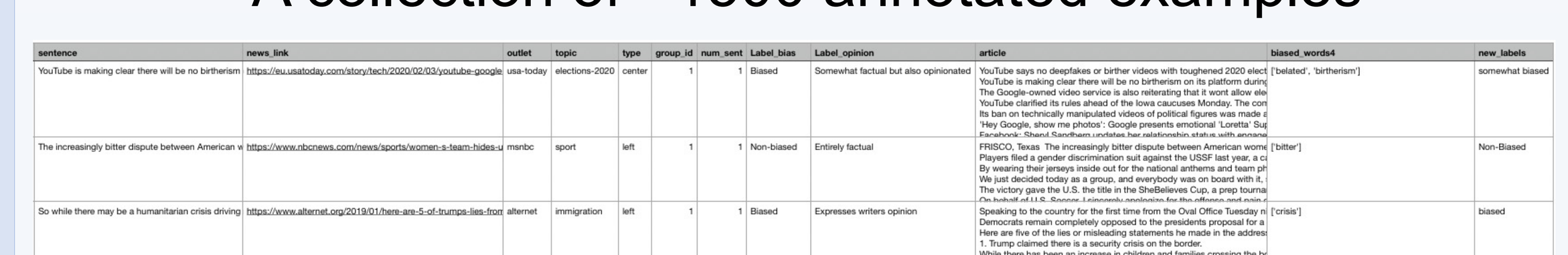
↓ Extract text with Beautiful Soup



The final dataset

2. MBIC Dataset (Spinde et. al. 2021)

A collection of ~1500 annotated examples



Data Cleaning

- New Labels Assignment: (0 – unbiased, 1- biased)

MBIC	Label_bias column with these numerical values
Media Bias	|perceived|

MODELS

1. A Random Forest model with 60 trees
   1. Vectorization using TFID Vectorizer
2. A Convolutional Neural Network
   1. Tokenization using SpaCy English
   2. Vectorization using TorchText
   3. 10 Epochs with Batch Size 16

Data Split:

Data Split	Number of Examples
Training Set – Media Bias	~13000 (Vocab Size = 57511)
Validation Set – Media Bias	~1500
Test Set – Media Bias	~1500
Out of Domain Test Set - MBIC	~1500

STEPS:

1. Training and Validation of the Models
2. Testing of the Models both on the in domain and out of domain test sets
3. Use post-hoc explanation methods for the models to find feature contribution scores to the model's prediction

EXPLANATION METHODS

- Found top 6 examples from each text with two explanation methods:

SAMPLING SHAPLEY

$$\phi_i = \sum_{m=1, S_m \in F \setminus \{i\}}^{M} \frac{1}{M} [f_{\hat{y}}(x_{S_m \cup \{i\}}) - f_{\hat{y}}(x_{S_m})].$$

(Chen)

- Used M=50 samples for estimation
- Allows for more faithful and complex model understanding

LIME

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

(Molnar)

- cosine similarity distance and exponential kernel
- Allows for quicker, less complex explanations of a relatively simple model architecture

RESULTS

Model	Validation Accuracy	ID test accuracy	OOD test accuracy
Random Forest	0.6781	0.586	0.6875
CNN	0.5072	0.526	N/A

Vocabulary Size = 57511

- Due to the low accuracy, LIME and SHAP were only performed on the Random Forest model

Explanation Example:

Explanation method	Annotated text
Human Annotation	YouTube is making clear there will be no "**birtherism**" on its platform during this year's U.S. presidential election – a **belate**d response to a type of conspiracy theory more prevalent in the 2012 race.
Sampling Shapley	YouTube is making clear there will be no "birtherism" on its platform during this **year**'s U.S. presidential election – **a** belated response to a **type of** conspiracy **theory** more prevalent in the 2012 race.
LIME	YouTube is making clear **there will** be no "birtherism" on its platform during **this year**'s U.S. presidential election – a belated response to a type of conspiracy **theory** more prevalent in the 2012 race.

Top 5 token weights:

LIME Example	Sampling Shapley Example
['will', -0.0292)	('theory', 0.054295)
('theory', -0.0223642),	('year's', 0.0082490)
('this', 0.01984)	('a', 0.0),
('year', 0.01553)	('type', -0.00260),
('there', 0.011122)	('of', -0.0120503323)

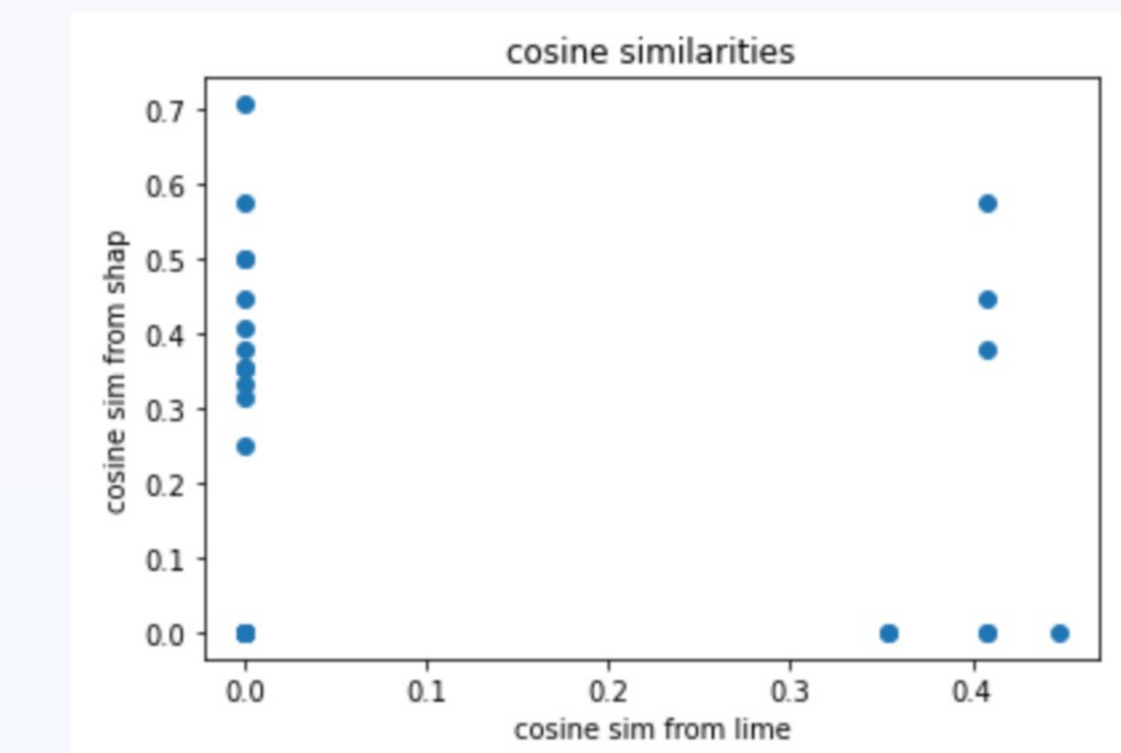
Performance Comparison

- Cosine Similarity Scores:

LIME predictions: 0.002021228

Shapley Values: 0.00435207

Between LIME and Shapley: 0.375514



CONCLUSION

- Based on these findings, bias cannot be effectively predicted by the model in either a CNN or a Random Forest Model
- The annotations that the model made are not similar to those made by humans
- Due to the limitations of the model predictions, it is hard to identify if the lack of similarity between human annotations and explanations is due to their ineffectiveness or due to the model itself

LIMITATIONS

- Dataset size: Both the datasets used were very small and not large enough to provide a quality training corpus for the model
- Vocab Size: The vocab size was incredibly large compared to the dataset size - High proportion of Unknown Tokens
- Model complexity: Only a Simple CNN and a Random Forest Model were used. These are not the most effective

FUTURE GOALS

- Use a larger training corpus that would improve the model accuracy
- Perform the same experiment on other types of biased data that contain less complicated unknown words
- Increase the model complexity to help further improve these aspects

REFERENCES

Aquino, Christine. "Military Times Recognized as One of the Most Reliable Media Sources in Latest Media Bias Chart." *Sightline*, 19 Oct. 2020, https://sightlinemediagroup.com/military-times-recognized-as-one-of-the-most-reliable-media-sources-in-latest-media-bias-chart/.

Chen, Hanjie. "Homework 3: Shap and Integrated Gradients", Interpretable Machine Learning, 20 March 2021.

Ceren Budak, Sharad Goel, Justin M. Rao, "Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis", Public Opinion Quarterly, Volume 80, Issue S1, 2016, Pages 250–271, https://doi.org/10.1093/poq/nfw007

Molnar, Christoph. "Interpretable Machine Learning." *9.2 Local Surrogate (LIME)*, 29 Mar. 2022, https://christophm.github.io/interpretable-ml-book/lime.html.

T. Spinde, L. Rudnitckaia, K. Sinha, F. Hamborg, B. Gipp, K. Donnay "MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics". In: Proceedings of the iConference 2021.