Domain background : The project's domain background is in Media bias. Media bias exists and shapes the views of their readers subconsciously. Bias are exhibited via language choices.

Problem statement : Currently, there is not enough research on the intersection of computer science and social sciences in identifying bias in the media [1] . This project will build and deploy a classification model that can suggest if an article text is leaning towards the Democrats or the Republicans. The development of AI-based bias detector tools is becoming popular in recent years [2]. There are different approaches to identifying potential bias in text. For example, a past project [3] made use of averaged feature values of 141 features used in another project to detect fake news, while another [4] made use of news outlet classification. This project will utilize NLP techniques on actual article text.

Datasets and inputs : The dataset is obtained from https://deepblue.lib.umich.edu/data/concern/data_sets/8w32r569d?locale=en and consists of 21004 rows of Article URL links and their corresponding positive, neutral, or negative ratings.

I obtained the article text from the article URL links via web scraping. After removing entries with 404 errors (articles removed), the dataset now consists of 18758 entries , with the following non-exclusive ratings: 12314 Neutral-Democrat , 3012 SomewhatNegative98-Democrat, 2080 SomewhatPositive-Democrat, 990 Negative-Democrat, 362 Positive-Democrat , 13512 Neutral-Republican, 2861

SomewhatNegative-Republican, 1215 SomewhatPositive-Republican, 986 Negative-Republican, and 184 Positive-Republican.

Solution statement : If the model has a high accuracy rate, it will be able to identify if an article is potentially biased towards any political party.

Benchmark model : Support Vector Machine Model [3] to identify bias in articles based on Article Body feature with averaged F1 and Accuracy score: 36.63 and 41.74

Evaluation metrics : F1 and Accuracy score

Project design :

1. Further process the dataset to remove entries with positive/negative for both republican and democrat.
2. Data cleaning
3. Create training and testing sets of equal number of neutral/positive/negative articles
4. Tokenization and create countvectorizer dictionary for training set
5. Create training job and train the model on a Classification Model.
6. Select best classification model.
7. Deploy model to a simple web app.

[1] Hamborg, F., Donnay, K. & Gipp, B. Automated identification of media bias in news articles: an interdisciplinary literature review. *Int J Digit Libr* **20,** 391–415 (2019). https://doi.org/10.1007/s00799-018-0261-y

[2] https://www.forbes.com/sites/simonchandler/2020/03/17/this-website-is-using-ai-to-combat-political-bias/#6ef847446f4c

[3] Baly, Ramy & Karadzhov, Georgi & Alexandrov, Dimitar & Glass, James & Nakov, Preslav. (2018). Predicting Factuality of Reporting and Bias of News Media Sources. 10.18653/v1/D18-1389.

[4] https://towardsdatascience.com/media-bias-detection-using-deep-learning-libraries-in-python-44efef4918d1