Data Science Canvas

Project: Bias Detection in Indian News Media

Team: Girish S N, Bharath Karanth A, Saravanakumar R, Anmol Gupta

Problem Statement

Business Case & Value Added

News media outlets are often accused of biased reporting across various categories. With the rapid growth of digital news, readers are exposed to vast and diverse information at their fingertips, often with subtle biases.

There is a need for a solution that can analyze news articles for bias – offering transparent, multi-dimensional, and interpretable results.

This project seeks to provide researchers, fact-checkers, and the public with actionable insights into media bias in Indian journalism.

Data Landscape

There was no publicly available dataset for Indian news articles with enough features that could have been used to meet our objectives.

Hence, we created our own dataset through web scraping of online Indian news articles.

Model Selection

Various bias detection methods were employed.

Keyword-based lexicon matching for different bias categories (gender, religion, caste, region, etc.).

Sentiment analysis using VADER and TextBlob.

Emotion detection via lexicons.

Discourse analysis (active/passive voice, agency).

Implicit association tests inspired regex patterns.

Semantic similarity using TF-IDF and BERT embeddings.

Topic classification (LDA, K-means) to contextualize bias across topics.

Model Requirements

Models must work with multilingual content and the Indian news context.

Bias scoring requires proper feature engineering including careful preprocessing.

Models should combine interpretable lexicon features with more complex semantic and ML ensemble scores for robustness.

The overall bias score is a weighted aggregation across multiple bias categories, implying the model must output category-wise scores reliably.

The system requires consistent updates as it depends on evolving news data, so resilience in scraping and model retraining is crucial.

Skills

Expertise in Python programming.

Experience in web scraping and MongoDB database management.

Understanding of ensemble modeling and semantic embeddings.

Familiarity with bias detection concepts and Indian socio-political contexts for meaningful feature engineering and interpretation.

Visualization skills to create insightful charts for bias trends.

Software & Libraries

Python was utilized as the main software language.

Key libraries included *pandas* and *numpy* for data handling, *nltk* for NLP preprocessing, *requests* and *beautifulsoup4* for web scraping, *pymongo* for database connectivity, *python-dateutil* for date handling, and *matplotlib* and *seaborn* for plotting and visualization.

MongoDB served as the primary database for storing scraped and processed data.

A Docker setup was provided for running MongoDB and Mongo Express.

Execution & Evaluation

Model Evaluation

Sentiment and emotion scores also provide validation layers to distinguish tone differences.

Explainability is addressed by lexicon-based features and clear bias type classifications.

Data Storytelling

Target users might require clear visualizations showing bias trends over time and among various media sources.

Charts were generated to highlight key narratives and support exploratory analysis.

Emphasis was also done on providing context around events and media in the summary insights, to make the graphs meaningful and for useful insights to be extracted.

Data Collection & Preparation

Data Selection & Cleansing

Initially scraped raw data included only a csv of URLs and timestamps.

Pre-processing and cleaning of the raw data was done, so that it can be further used for training and analysis.

Data Integration

Raw data, scraped from different online news websites and across different languages, was ultimately integrated into a single comprehensive dataset. Python scripts were used to scrape as well as integrate the data.

The total number of articles in the final dataset was 3,96,739.

Data Collection

As a first step, raw data was collected through the web scraping of online Indian news articles.

The non-English articles (Hindi, Kannada, Tamil) were translated to English.

Extensive feature extraction techniques were applied to eventually create a robust and comprehensive dataset.

Creation of the dataset itself was one of the two major objective of our project.

Explorative Data Analysis

The data includes articles scraped from multiple media outlets and languages, with preprocessing to normalize text.

Outliers and structure are considered through statistical metrics and by aggregating bias scores with 5-year moving averages to smooth trends.

Calendar and event metadata were integrated to relate bias trends to major social and political events, helping to identify patterns and anomalies.