National University of Computer and Emerging Sciences

Bias Detection System for Pakistani News

Project Proposal

Team

Ahmed Nadeem……L21-5393

Supervised by

Ms. Anosha Khan

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

February 2025

1. Introduction

The rise of digital journalism has led to an increase in news content being shared across various platforms. However, with this surge in information, concerns regarding media bias have also grown. Biased reporting can influence public opinion, shape narratives, and affect social and political discourse. This project aims to develop a Bias Detection System for Pakistani News, leveraging Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system will analyze news articles, detect potential biases, and classify them accordingly. By providing a quantitative measure of bias, this system aims to promote media transparency and encourage objective journalism.

2. Goals and Objectives

Goal 1: Develop a system that automatically detects and quantifies bias in Pakistani news articles.

- Objective: Implement sentiment analysis and entity recognition to evaluate potential biases in news content.

Goal 2: Create a labeled dataset of Pakistani news articles with bias scores.

- Objective: Scrape and preprocess news articles from various sources and assign bias labels based on sentiment and contextual analysis.

Goal 3: Build a machine learning model that classifies news articles based on their bias level.

- Objective: Train and evaluate ML models using labeled data to enhance detection accuracy.

3. Research Gap and Research Questions

Existing bias detection models are often trained on Western media sources, limiting their applicability to Pakistani news, which has unique contextual nuances. The absence of a dedicated **bias-labeled dataset for Pakistani news** creates a gap in automated bias detection systems.

Research Questions:

1. How can NLP techniques be effectively utilized to detect bias in Pakistani news articles?
2. What are the key linguistic indicators of media bias in the Pakistani news industry?
3. How does sentiment polarity correlate with bias in news reporting?

4. Scope of the Project

- **Key Functionalities:**
  - Automated scraping and preprocessing of Pakistani news articles.
  - Sentiment analysis and Named Entity Recognition (NER) to assess bias.
  - Development of a bias classification model using machine learning.
  - Visualization of bias scores in an interactive dashboard.
- **Boundaries and Limitations:**
  - Focus on English news sources from Pakistan.
  - Bias detection is based on textual analysis and may not capture nuanced biases like framing or omission.
  - The accuracy of the system depends on dataset quality and annotation reliability.

5. Methodology (Dataset/Approach)

Dataset:

- **Source:** Web scraping from popular Pakistani news websites such as **Dawn, The Express Tribune, The Nation, and ARY News**.
- **Preprocessing:** Tokenization, stopword removal, lemmatization, and feature extraction.
- **Annotation:** Bias scores assigned based on sentiment polarity and contextual analysis.

Approach:

1. **Data Collection:** Scrape 50,000+ news articles using Selenium.
2. **Data Processing:** Perform Named Entity Recognition (NER) and Sentiment Analysis using **spaCy and TextBlob**.
3. **Feature Engineering:** Extract linguistic features indicative of bias.
4. **Model Training:** Train ML models such as **Naive Bayes, SVM, and Transformer-based models (BERT, RoBERTa)** for bias classification.
5. **Evaluation:** Use accuracy, F1-score, and AUC-ROC to assess model performance.

6. Expected Outcomes

- A **functional bias detection system** that classifies news articles as **biased or neutral** with a **quantitative bias score**.
- A **publicly available dataset** of Pakistani news articles labeled with bias scores.
- Insights into how sentiment polarity and entity bias affect news reporting.
- A potential tool for **fact-checkers, journalists, and media analysts** to assess news credibility.

7. Initial Study

The project builds upon prior research in NLP-based bias detection. Key references include:

References

[1] Y. Baly et al., "Predicting Factuality of Reporting and Bias of News Media Sources," *Empirical Methods in NLP*, 2018.
[2] T. Horne et al., "Evaluating Models for Fake News Detection," *Proceedings of the 15th International AAAI Conference on Web and Social Media*, 2019.
[3] L. Van der Lee et al., "Analyzing Textual and Contextual Features for News Media Bias Detection," *Proceedings of NLP in Journalism*, 2020