Bias Beyond Demographics: Fairness-Aware Evaluation in Fake News Classification

Michelle Lynn George (Elle Lynn)
CS 5891 – Responsible AI
12 April 2025

Overview

This report builds upon earlier work in bias detection by expanding the analysis of fake news classification models to include fairness-aware machine learning techniques. While prior efforts focused on surface-level prediction patterns, this phase explores how stylistic and linguistic factors — rather than traditional demographics — may contribute to biased outcomes. In the absence of age, race, or gender data, fairness was reframed in terms of how models interpret language: tone, structure, and emotional charge embedded in news headlines. This approach aligns with the goals of Assignments 6 and 7 by integrating both bias detection and fairness-aware evaluation strategies.

Datasets Used

The primary dataset was constructed by merging Fake.csv and True.csv, forming a labeled corpus of news headlines. A cleaned version was stored in WELFake_Dataset.csv. Feature engineering added surrogate indicators for potential bias, such as political keywords, clickbait formatting, and sentiment

tone. (See Figure 1) These features enabled fairness auditing that does not rely on demographic attributes, utilizing both statistical and visual metrics.

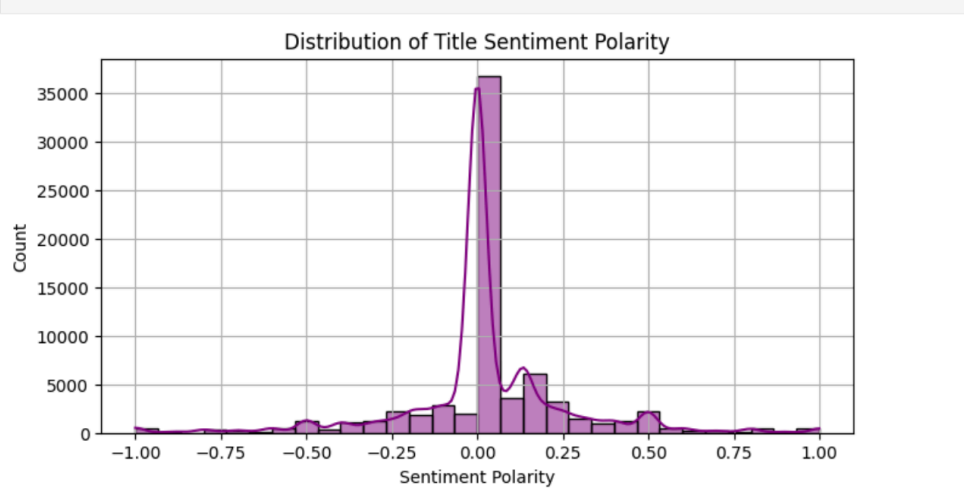


Figure 1: Sentiment Polarity
Distribution of sentiment polarity scores for real vs. fake news headlines, revealing tone-based distinctions that may influence model predictions.

Methodology

The core of this project involved engineering features based on how the model might interpret — or misinterpret — the tone or style of a headline. These included binary flags for clickbait-style headlines (is_clickbait), sentiment scores using TextBlob (title_sentiment), and structural metrics such as title length, caps ratio, and punctuation count (see Figure 2).

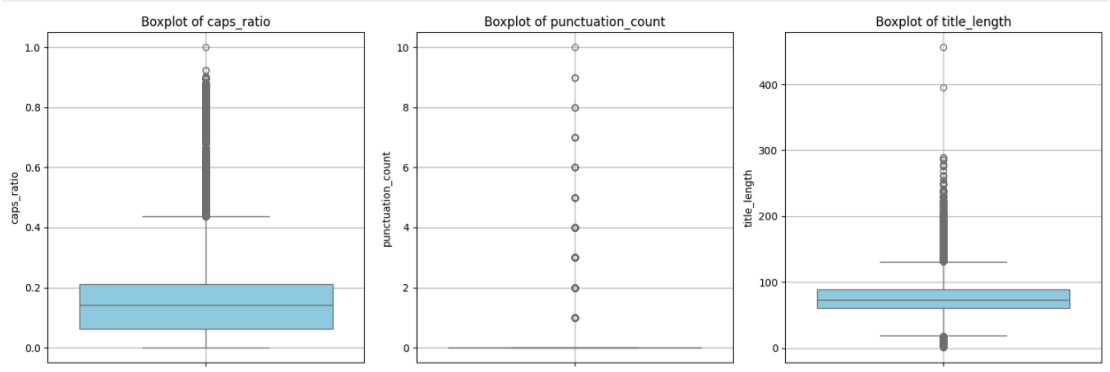
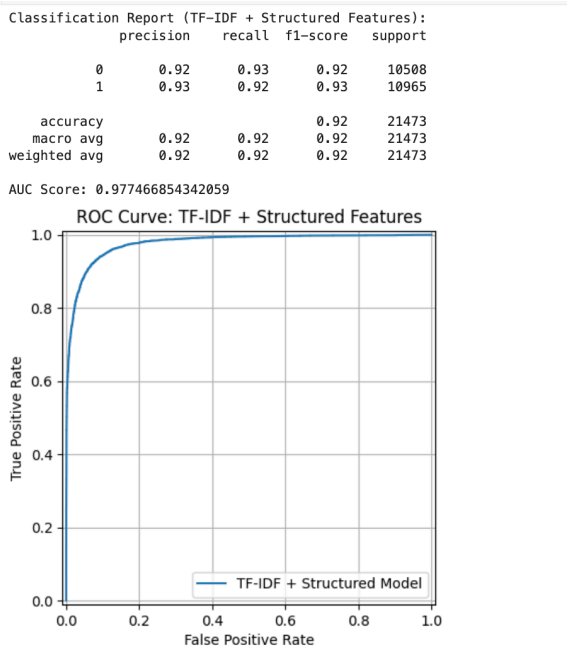


Figure 2: Structural Feature Boxplots

Boxplots showing variation in headline structure (caps ratio, punctuation count, and title length), used to examine stylistic indicators that may introduce bias.

A logistic regression model was trained using TF-IDF embeddings (see figure 3) of the title text combined with these engineered features. The model's predictions were then evaluated using fairness metrics to identify performance discrepancies across stylistically different headline groups.



Classification Report (TF-IDF + Structured Features):

	precision	recall	f1-score	support
0	0.92	0.93	0.92	10508
1	0.93	0.92	0.93	10965
accuracy			0.92	21473
macro avg	0.92	0.92	0.92	21473
weighted avg	0.92	0.92	0.92	21473

AUC Score: 0.977466854342059

Figure 3: TF-IDF Visualization

TF-IDF vector representation of headline text, forming the embedding foundation for classification and fairness evaluation.

Results

The combined TF-IDF + structured model performed very well overall, achieving an AUC of approximately 0.977. However, fairness metrics revealed notable differences in how the model handled clickbait-style headlines.

Statistical Parity Difference (SPD) showed a gap of ~0.46 between groups, suggesting the model was more likely to label clickbait headlines as fake. Equal Opportunity Difference (EOD), (see figure 4) which compares true positive rates, also revealed a gap of ~0.08 favoring clickbait headlines.

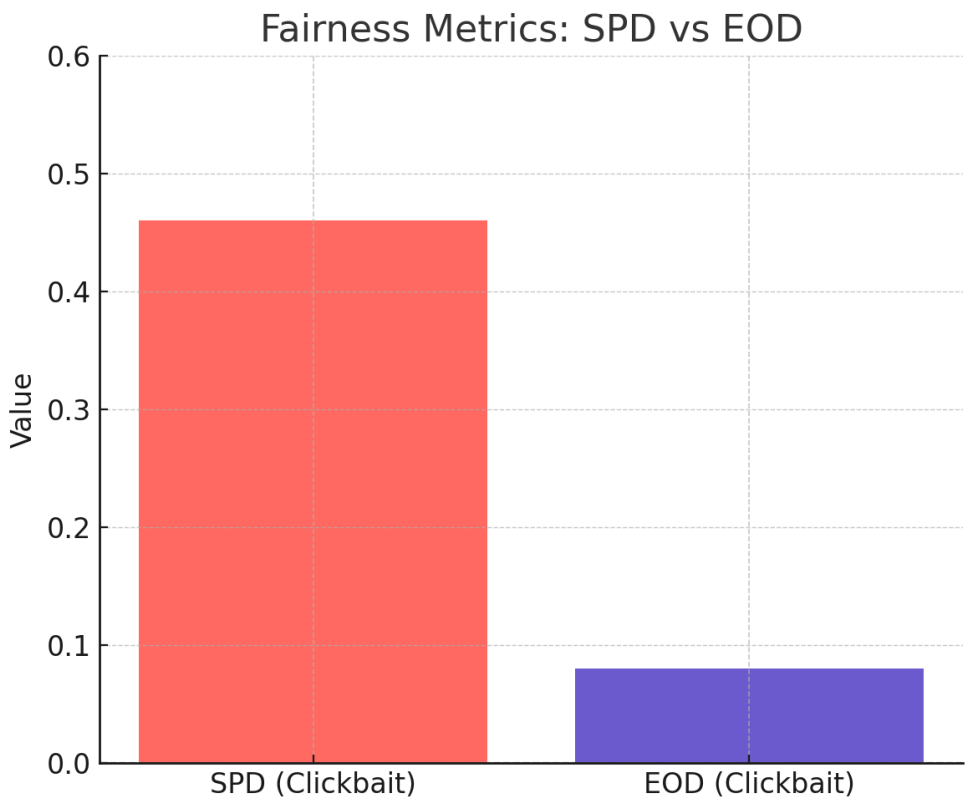


Figure 4: Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD)
Fairness metric comparison showing measurable disparity between clickbait and non-clickbait headline groups in terms of both classification likelihood (SPD) and true positive rates (EOD).

Grouped ROC curves (see figure 5) confirmed these trends visually, showing slightly higher recall performance for the clickbait group.

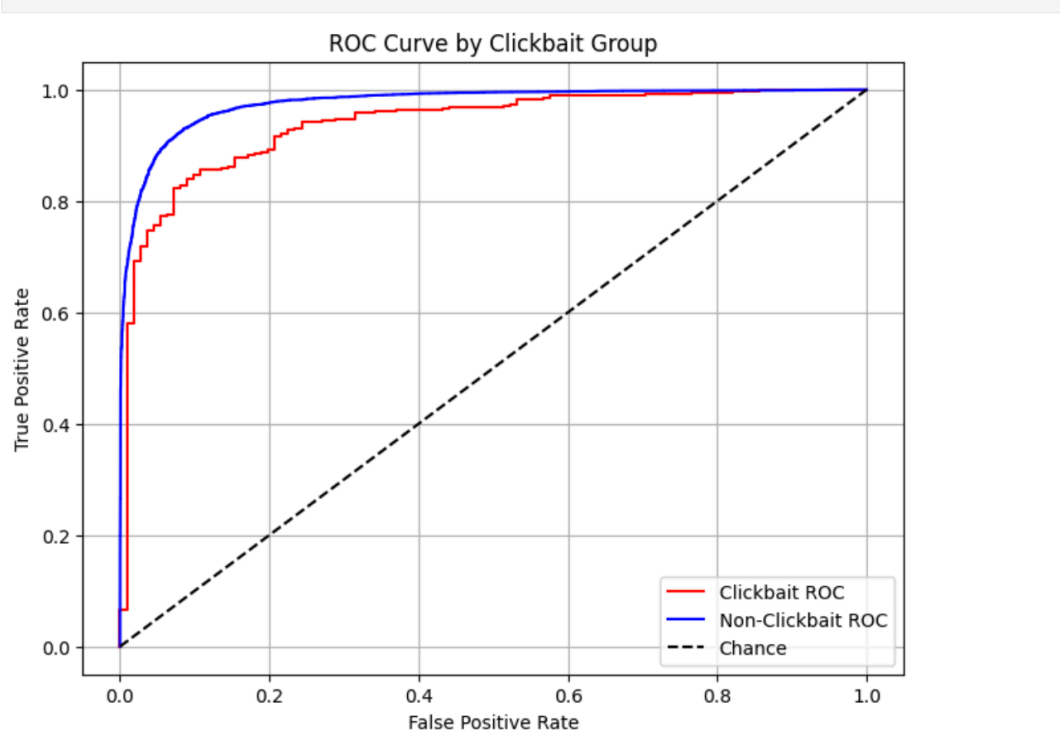


Figure 5: Grouped ROC Curves
ROC curves stratified by headline type, highlighting differences in recall performance and model behavior across stylistic groups.

Reflection

Fairness auditing in NLP does not require access to sensitive personal data; it can be effectively applied to the ways in which a model interprets language itself. Phrasing, formatting, and emotional tone in headlines can lead to unintended associations and biased outcomes in classification models.

This perspective expands the traditional understanding of fairness by highlighting that bias may arise not only from who is represented, but also from how information is framed. Evaluating bias through linguistic structure provides

a deeper view into model accountability and interpretability in real-world applications.

Future Work

Future directions for this work include exploring mitigation techniques such as Reject Option Classification using AIF360, focusing on low-confidence prediction regions. Incorporating sentence embeddings (e.g., BERT) may enhance the model's ability to understand semantic nuance and detect subtler forms of bias. Calibration curves can provide insights into confidence disparities among stylistic groups. If demographic metadata is available, it could facilitate intersectional fairness analyses that incorporate both linguistic and identity-based dimensions.

Conclusion

Assignments 6 and 7 form a cohesive fairness audit: Assignment 6 initiated detection and group comparisons, while Assignment 7 implemented formal fairness metrics and model evaluation. This work validates that bias can emerge in many forms — not just from who is represented, but from how they are represented. In the realm of fake news classification, fairness requires attention to voice, format, and emotional framing — and this project takes a meaningful step toward that goal.