Bias as a feature for generative AI detection

Comparing stylometric and bias-related features in AI detection models

Universität Zürich
Computational Forensic Linguistics
Prof. Dr. Gerold Schneider
Ahmet Yavuz Uluslu
Fall Semester 2025
Andrea Eva Scheck

11-734-969
Kamorstr. 1., 9000 St. Gallen
andreaeva.scheck@uzh.ch
4th semester

DRAFT

1 Introduction

Studies show that large language models exhibit systematic social biases in their writing, often reproducing and amplifying stereotypes present in their training data regarding gender, racial and ethnic background, religion, national or geographic origin, age, disability or sexual orientation (Gallegos et al., 2024). LLMs additionally show content and style bias (Bang et al., 2024): they use generic language, hedge their statements, use less polarized vocabulary and remain more unsentimental and polite than human writers. In terms of political stance, LLMs are trained to display neutral attitudes. For example, in October 2025, OpenAI claimed that the latest models (GPT-5 instant and GPT-5 thinking) demonstrate a 30 % reduction in political bias by expressing less personal political opinions or politically one-sided responses (OpenAI, 2025).

In short, LLMs don't simply replicate human language, they replicate it with a characteristic set of content patterns. All of these bias patterns could therefore function as markers to flag AI generated text when compared to authentic human writing. In my project, I therefore want to test **if bias-related feature training improves human-vs-AI text classification** beyond a non-bias-based baseline.

1.1 Related work

TO BE COMPLETED

2 Methods

To extract and systematically test different features on AI and human generated text, three linear classifiers were trained using logistic regression.

2.1 Baseline model

The baseline model was constructed exclusively of **handcrafted, low-dimensional stylometric features** intended to capture surface-level linguistic regularities rather than semantic content.

— Using spaCy sentence and token segmentation, **sentence and word length statistics** were computed while excluding punctuation tokens.

- The **stopword rate** was computed as the proportion of tokens marked as stopwords by spaCy's built-in English stopword list, relative to the total number of non-punctuation tokens in the text.
- The relative **frequency of punctuation marks** (, . ; : ? ! ' " ( ) - …) was computed from the raw text. For each punctuation mark, the count was normalized by the total number of characters and scaled per 1,000 characters to make texts of different lengths comparable.
- The **ratio of uppercase characters, title case, digits and whitespaces** were computed to capture orthographic patterns in the raw text.
- Lexical diversity was measured as **Type-Token ratio** by extracting lowercased alphabetic tokens.
- The **frequencies of Universal POS tags** were computed for the following categories: NOUN, VERB, ADJ, ADV, PRON, ADP, DET, CCONJ, SCONJ, NUM, AUX, INTJ, PART, PROPN.
- The proportion of tokens which were part of any **5-gram sequence repetition** was captured
- The **Jaccard similarity between the top-200 most frequent unigrams** in the first and second halves of each text was measured.

Each feature was developed and its relevance tested on the training data from the Subtask 1 of the PAN 2025 Voight-Kampff challenge (see Section 3.1). No information from the validation split was used during feature development or model fitting.

2.1.1 Features intentionally discarded for the baseline model

During early development, character- and word-level n-gram features were also explored using TF-IDF representations implemented via scikit-learn's TfidfVectorizer. Character n-grams were extracted from lowercased text, while word n-grams were extracted after removing numeric characters and applying a simple token pattern.

These TF-IDF features were ultimately excluded from the baseline model. The resulting feature space was extremely high-dimensional relative to the size of the dataset, producing a sparse representation that showed substantially higher performance on the training split than on the validation split. This discrepancy strongly suggested overfitting, with the classifier likely capturing surface-level artefacts specific to the training data.

- ⊠“Training performance was much higher than validation performance” (ideally give the numbers if you still have them).

2.2 Bias model

A second feature set was designed to capture potential content-level, stylistic and rhetorical biases that may differ between human- and AI-generated texts. These features were not intended to encode topic content directly, but rather higher-level tendencies related to sentiment, stance, identity references and argumentative style.

- Sentence-level **sentiment** was computed using the VADER sentiment analyzer.
- A lexicon-based **emotional tone measure** was implemented using lexicons of positive and negative opinion words. For each document, the relative frequency of positive and negative words was computed, along with a simple polarity score derived from their difference. This feature conceptually overlaps with sentiment analysis but relies on transparent lexical counts rather than a pretrained model.
- A lexicon of **hedging and modality markers** was used to estimate how cautiously or tentatively claims are phrased. The lexicon included modal verbs (e.g. may, might), softening adverbs (e.g. perhaps, presumably), epistemic verbs (e.g. seem, suppose) and multi-word phrases (e.g. in my opinion, there is a chance).
- A lexicon of **subjectivity markers** was used to compute the rate of explicit personal perspective indicators (e.g. first-person opinion markers). This feature partially overlaps with sentiment and hedging and was treated cautiously in the analysis.
- A lexicon of **profanity** was used to compute the rate of toxic, profane and insulting terms per document.
- Two complementary lexical features were implemented to capture **rhetorical strength**: A lexicon of categorical statements as expressed in absolute or extreme expressions (e.g. always, never, everyone) and a lexicon of forceful argumentative or strongly assertive terms (e.g. prove, must, it is certain).

Again, each feature was developed and its relevance tested on the training data from the Subtask 1 of the PAN 2025 Voight-Kampff challenge. As a result, several bias-oriented features were explored but ultimately excluded from the final model.

2.2.1 Features discarded for the bias model

- A set of **emojis and emoticons** was explored, but in the PAN dataset, no emojis were observed, making this feature effectively constant.
- **Politeness and impoliteness markers** were extremely sparse, leading to unstable estimates.
- A **pretrained political stance classifier** was applied to each document to estimate probabilities for left, center and right orientations. Its outputs were sparse and unstable across the PAN dataset's heterogeneous genres, showed minimal class separation and did not improve validation performance, likely due to domain mismatch with non-political texts.
- A measured ratio of **we vs. they terms** was unstable due to very small numerators and denominators, with variance exceeding the mean.
- **Semantic or embedding-based bias representations** (e.g. Word2Vec similarity) were considered but rejected due to the short length of each text.

For consistency, these features were also excluded from subsequent experiments on my own custom built migration dataset.

2.3 Merged model

The merged model combined all baseline stylometric features (Section 2.1) and bias-oriented features (Section 2.2) into a single feature representation. For each document, the full baseline feature vector and the full bias feature vector were concatenated into a single feature matrix. Feature extraction, scaling and classification followed the same procedure as for the individual models.

3 Data

3.1 PAN 2025 Voight-Kampff dataset

The primary dataset was retrieved from Subtask 1 of the PAN 2025 Voight-Kampff challenge, which is about binary authorship attribution. The dataset contained texts from multiple genres:

- fiction: 16,534 texts
- news: 5,436 texts
- essays: 5,326 texts

The AI-generated portion included outputs from numerous models, such as GPT-3.5, GPT-4 variants, Gemini, LLaMA, DeepSeek, Mistral, Falcon, Qwen and others. Overall the label counts were moderately imbalanced towards AI:

- AI-generated (label = 1): 16,918 texts
- Human-written (label = 0): 10,378 texts

Because of the imbalance toward AI-generated texts in the dataset, a majority-class baseline always predicting the most frequent label (AI) achieves an accuracy of 0.644 on the validation set. All reported model performances on the pan data set should therefore be interpreted relative to this baseline.

The dataset was provided with a predefined train–validation split. The training set contains 23,707 texts (approximately 87%) and the validation set contains 3,589 texts (approximately 13%). Label distributions were similar across splits:

- Training set: 61.6% AI, 38.4% human
- Validation set: 64.4% AI, 35.6% human

3.2 Custom migration-focused dataset

To evaluate model behavior under tighter topical and genre control, I constructed an additional dataset focused on a single politically controversial topic: immigration.

3.2.1 Human corpus collection

Human-written texts were drawn from the SFU Opinion and Comments Corpus (SOCC). This large corpus consists of editorials, opinion columns and op-eds published in *The Globe and Mail* between January 2012 and December 2016.

This source was chosen for several reasons: the texts are explicitly opinionated and argumentative, they were all written by professional journalists before thewidespread public use of large language models and they originate from a single publication, reducing stylistic and institutional variance.

As a first step, the over 10,300 articles from the corpus were searched and only those containing at least two migration-related terms from a predefined lexicon (e.g., *immigration, migrant, refugee, asylum, border, deportation, undocumented, visa, integration*) were selected. This yielded **466 candidate articles**.

To further filter migration-focused texts, a zero-shot topic classification step was applied using a BART-based natural language inference model (*facebook/bart-large-mnli*). Articles were classified into four coarse topical categories, including a migration-specific label. Based on content relevance and argumentative focus, a final human corpus of **300 migration-focused human opinion texts** was built.

3.2.2 AI-generated corpus design

To create comparable AI-generated texts, I generated opinion-style articles on the same topic using OpenAI GPT (gpt-4o-mini), Google Gemini models (accessed via Google's legacy SDK) and DeepSeek models (accessed via an OpenAI-compatible API).

To match the genre and communicative intent of the human texts, prompts explicitly requested newspaper editorials or opinion pieces written for a general audience and focusing on immigration. To reduce prompt-specific artifacts, a small set of prompt variants was used (e.g., *"Write a newspaper opinion piece about immigration"*, *"Write an editorial on immigration for a general newspaper audience"*) and prompts were randomly sampled across generations.

Text length was treated as a potential confound and therefore controlled explicitly, creating a character-based length distribution equal to the length distribution of human texts.

The complete dataset was then split into training and test sets using an 87% / 13% split as seen in the PAN dataset, applying stratification to preserve balance across human vs. each AI model and text length. The final data set contained **600 texts (300 human, 300 AI), 522 for training, 78 for testing**.

4 Results

4.1 Baseline model performance

4.1.1 PAN dataset

On the PAN 2025 Voight-Kampff Subtask 1 validation set, the baseline model achieved strong overall performance using only lightweight stylometric features.

The classifier reached an **accuracy of 0.952**, with balanced precision and recall across classes (Table X). Performance was slightly higher for AI-generated texts than for human-written texts, reflected in a higher recall for the AI class (0.972 vs. 0.915).

Class	Precision	Recall	F1-score	Support
Human	0.948	0.915	0.931	1,277
AI	0.954	0.972	0.963	2,312
Accuracy			**0.952**	3,589

Out of 3,589 validation texts, **172 instances were misclassified**. False positives (human classified as AI) were more frequent than false negatives (AI classified as human), consistent with the dataset's moderate class imbalance toward AI-generated texts.

Coefficient analysis of the trained logistic regression model reveals systematic stylistic differences captured by the baseline features. Features with the strongest positive weights toward the AI class include higher pronoun and determiner frequencies, increased type–token ratio, longer average word length and higher repetition and self-similarity scores. In contrast, human-written texts are characterized by higher stopword rates, greater sentence-length variability, higher rates of proper nouns and increased use of punctuation such as semicolons and hyphens.

4.1.2 Migration dataset

When evaluated on the custom migration-focused dataset, the same baseline model achieved **perfect classification performance**, correctly classifying all 78 test instances (38 human, 40 AI), with no misclassifications observed.

Class	Precision	Recall	F1-score	Support
Human	1.000	1.000	1.000	38
AI	1.000	1.000	1.000	40
Accuracy			**1.000**	78

Since these results were so strong, I wanted to rule out errors. First, I removed all explicit length-based features from the baseline feature set, suspecting them of overfitting; performance remained at 100%. Second, the model was trained only on texts generated by GPT and Gemini and evaluated on previously unseen DeepSeek-generated texts; again, classification accuracy remained perfect. These additional checks suggest that the **observed performance is not attributable to trivial explanations** like length leakage, train–test contamination, or generator-specific memorization.

Coefficient analysis on the migration dataset shows a broadly similar pattern of influential features as in the PAN dataset, though with smaller absolute weights due to the reduced dataset size. AI predictions were primarily associated with higher noun, pronoun and determiner frequencies, as well as increased word-length variability, while human texts were more strongly associated with proper nouns, sentence-length variability and casing-related features.

4.2 Bias model performance

4.2.1 PAN dataset

On the PAN 2025 Voight-Kampff Subtask 1 validation set, the bias-focused model achieved moderate overall performance using only handcrafted bias-related features. The classifier reached an accuracy of **0.843**, substantially lower than the baseline stylometric model but still clearly above the majority-class baseline.

Performance was again higher for AI-generated texts than for human-written texts, with a recall of **0.891** for AI compared to **0.757** for human texts.

Class	Precision	Recall	F1-score	Support
Human	0.793	0.757	0.775	1,277
AI	0.869	0.891	0.880	2,312
Accuracy			**0.843**	3,589

Out of 3,589 validation texts, **562 instances were misclassified**. Again, false positives (human texts classified as AI) were more frequent than false negatives, indicating that the bias model tends to over-predict the AI class under uncertainty.

Coefficient analysis indicates that **sentiment- and emotion-related features dominate the bias model's decisions**. Strong positive weights toward the AI class were observed for negative and positive emotional tone rates, as well as mean sentiment score. In contrast, human-written texts were more strongly associated with higher rates of identity-related terms (gender, race/ethnicity, nationality), profanity and assertive language.

Overall, the bias model captures differences in **emotional polarity, evaluative language and identity references**, but lacks the fine-grained stylistic sensitivity of the baseline model, resulting in substantially lower performance on the heterogeneous PAN dataset.

4.2.2 Migration dataset

When evaluated on the custom migration-focused dataset, the bias model achieved **high classification performance**, reaching an accuracy of **0.962** on the held-out test set. Only **3 out of 78** test texts were misclassified.

Class	Precision	Recall	F1-score	Support
Human	0.973	0.947	0.960	38
AI	0.951	0.975	0.963	40
Accuracy			**0.962**	78

The confusion matrix shows a small number of errors in both directions (two human texts classified as AI and one AI text classified as human), indicating balanced behavior across classes despite the small dataset size.

Feature-level analysis reveals a pattern consistent with the PAN results but with **stronger and more concentrated weights**. AI predictions were again driven primarily by higher positive and negative emotional tone rates and higher mean sentiment scores. Human-written texts were more strongly associated with higher rates of identity-related terms, profanity and assertive language.

Compared to the PAN dataset, the bias features appear substantially more effective in the migration setting, likely reflecting the **topic's inherently evaluative and polarized nature**, which amplifies emotional and identity-related signals captured by the bias feature set.

4.3 Merged model performance

4.3.1 PAN dataset

The merged model which combined all baseline stylometric features with the bias-oriented lexical and affective features into a single classifier reached an accuracy of **0.962,** achieving the highest overall performance among all evaluated models.

Class	Precision	Recall	F1-score	Support
Human	0.960	0.931	0.945	1,277
AI	0.963	0.978	0.970	2,312
Accuracy			0.962	3,589

Out of 3,589 validation texts, **138** instances were misclassified. Compared to the baseline model (172 errors) and the bias model (562 errors), the merged model substantially reduces misclassification, particularly false positives for the human class.

Coefficient analysis shows that the merged model is dominated by baseline stylometric features, but that bias features contribute complementary signal. The strongest weights pushing predictions toward the AI class are primarily baseline features, such as higher pronoun and determiner frequencies, higher type–token ratio, longer average word length and increased repetition and self-similarity. However, several bias features also appear among the top contributors, most notably negative and positive emotional word rates and mean sentiment score.

Conversely, predictions toward the human class are again driven mainly by baseline features such as higher stopword rates, greater sentence-length variability, increased use of proper nouns and specific punctuation patterns. Bias features such as profanity rate, polarity score, categorical language and nationality-related identity terms also contribute negatively, reinforcing human predictions.

4.3.2 Migration dataset

The merged model was not evaluated on the custom migration dataset, since the baseline model already achieved perfect classification accuracy (100%). Under these conditions, adding bias features couldn't possibly improve measurable performance or provide additional insight.

5 Discussion

5.1 Misclassifications

5.1.1 Overlapping misclassifications

Across all three models, **66 texts were consistently misclassified** (18 AI texts, 48 human texts). An error analysis of these cases shows that the baseline and bias models are **not failing for opposing or competing reasons**. Instead, they fail **for the same underlying reason** and therefore do not "contradict each other" on these texts.

These instances are best characterized as **globally ambiguous texts**. They do not exhibit strong signals in either feature space: **Stylometric features** show no pronounced deviation toward AI-like or human-like patterns; values largely cluster around the training mean. **Bias-related features** likewise lack extreme values that would clearly indicate human stance-taking or AI-like neutrality.

As a result, neither model produces strong, confident evidence in favor of one class.

Crucially, these are **not conflicted texts**, where different feature groups point in opposite directions. Rather, they are cases of **signal absence**: both feature spaces are weakly informative.

This explains why the merged model does not recover these errors. Feature stacking is effective only when complementary signals exist across feature sets. Consequently, merging baseline and bias features does not improve classification performance, because the underlying ambiguity persists across all representations.

5.1.2 Baseline model misclassifications

The baseline model produced **172 misclassifications** on the PAN dataset. Of these, **66 errors (38%)** were shared with all models, while **45 errors (26%)** were **specific to the baseline model**. The remaining errors were shared with either the bias or merged model. Overall, the baseline model misclassified **64 AI texts as human** and **108 human texts as AI**.

All texts in the baseline-specific misclassifications exhibit unusually strong and internally consistent stylistic profiles, deviating sharply from typical examples of their true class across multiple features. In stylistic space, they more closely resemble the opposing class than their true label, making the error systematic rather than accidental.

Importantly, these errors are not associated with weak or noisy feature representations. On the contrary, they are among the **most stylometrically distinctive texts** relative to the training distribution. These findings indicate that the baseline model fails when **stylometric cues are actively misleading rather than absent**. Specifically, it struggled with **AI-generated news texts**, which presumably closely imitate the structural properties of human journalism. Similarly, **highly polished human texts** that exhibit unusually regular structure and reduced idiosyncratic variation also posed a challenge.

5.1.3 Bias model misclassifications

The bias model produced **562 misclassifications** on the PAN dataset. Of these, **66 errors (12%)** were shared across all models, while **472 errors (84%)** were **specific to the bias model**. The remaining errors were shared with either the baseline or merged model. Overall, the bias model misclassified **252 AI texts as human** and **310 human texts as AI**.

Bias-specific misclassifications exhibited a markedly different profile from baseline-specific errors. They all showed low to moderate assertive rates across most texts, with limited variance (in particular for sentiment scores) and very low identity-term rates across all identity categories.

In contrast to baseline errors, these misclassifications do not arise from unusually marked feature patterns. Instead, they involve texts whose content-level signals remain close to the model's expectations for both classes. The bias model fails primarily on content-neutral or weakly opinionated texts, where sentiment, identity and stance-related cues are sparse or absent. Rather than presenting conflicting evidence, these texts simply offer too little signal for reliable discrimination.

This includes **straight news reporting**, **neutral essays**, **factual summaries** and **fiction without a clear moral or ideological stance**.

5.1.4 Merged model misclassifications

The merged model produced **138 misclassifications** in total. Of these, **66 errors (48%)** were shared across all models, while only **9 errors (7%)** were **specific to the merged model**. The remaining errors were shared with either the baseline or bias model. Overall, the merged model misclassified **50 AI texts as human** and **88 human texts as AI**.

The merged model benefits from combining the two complementary perspectives of the baseline and bias model, which substantially reduces brittleness relative to single-view approach. This is most clearly illustrated by a small but informative subset of 11 texts that were misclassified by both the baseline and bias models but correctly classified by the merged model. Notably, 8 of these 11 texts (73%) are long-form news explainers characterized by structured, factual and neutral reporting.

5.2 Differences between the datasets

The difference in performance between the PAN dataset and my own migration dataset are best explained by structural properties of the datasets rather than model behavior.

The PAN dataset spans a wide range of topics, genres and writing conditions, including news, fiction, essays and multiple styles of human authorship. This diversity leads to substantial stylistic overlap between human and AI-generated texts. As a result, stylometric features such as sentence length, POS balance, lexical richness and punctuation regularity do not consistently separate the two classes. Even highly regular writing can plausibly originate from either humans or LLMs, which limits separability and results in a performance ceiling of approximately 95% for the baseline model.

In contrast, the migration dataset is highly constrained. All texts address a single topic and belong to a single genre (opinion/editorial), with tightly controlled prompting conditions. This restriction concentrates the stylistic signal: human texts exhibit consistent author-specific variation, while AI-generated texts display uniform regularities associated with prompt-following and generation smoothing. Under these conditions, surface-level stylometric features become highly discriminative, allowing perfect separation, even when evaluated on an unseen generator.

Importantly, the migration dataset is also small and was compiled with a limited set of prompts. Under such circumstances, a model can exploit stable editorial conventions and prompt-regularities, without true generalizability. This result is therefore explicitly **in-domain** and does not imply robustness across domains or genres. The observed performance should therefore be interpreted as an upper bound under idealized conditions, rather than evidence of real-world deployment readiness.

6 Conclusion

Given the strong standalone performance of the baseline model shown in this study, one might reasonably ask whether exploring bias features is necessary at all. At present, stylometric features remain highly effective, often outperforming content-based signals by a wide margin. However, this advantage is unlikely to persist indefinitely. As large language models continue to improve and as human writers increasingly adapt to AI-mediated writing norms, purely stylistic differences are likely to erode. Moreover, models can already be prompted to produce text that deliberately mimics human stylistic variation in order to evade form-based detection.

This study shows that stylometric and content-based bias features fail in systematically different ways. Baseline-only misclassifications arise when structural cues are misleading: texts are well-formed, stylistically regular and genre-appropriate, even though their content provides little discriminatory signal. In contrast, bias-only misclassifications arise when content cues are absent: texts are informational, neutral and low in stance or identity signaling, even though their structure is unremarkable and human-like.

This complementarity explains why the merged model substantially reduces bias-specific errors. Stylometric features compensate precisely in cases where content-based signals collapse, while bias features provide coverage where stylistic cues are genre-confounded or misleading. Crucially, the merged model does not simply succeed by stacking more features. Instead, it **recovers cases where neither stylistic deviation nor bias cues alone are sufficient**, but where their interaction exposes inconsistencies between surface regularity and semantic intent. These structural–semantic mismatches are characteristic of advanced AI-generated text and are only detectable when both perspectives are evaluated jointly.

However, several limitations must be acknowledged. First, bias features are inherently correlated with topic and genre. For example, news articles about technical infrastructure must be expected to exhibit lower bias and stance signals than opinion pieces on contentious social issues, regardless of whether they are written by humans or AI. As a result, content-based features inevitably reflect dataset composition in addition to differences between human and AI authorship.

Second, many features commonly associated with AI-generated text, such as politeness, positivity, or certain punctuation patterns, are also part of legitimate human writing styles. Humans may naturally write in a highly regular, polished manner or adopt stylistic conventions that resemble those of language models. This makes false positives, particularly for human texts, a persistent risk and highlights the importance of maintaining high human recall in any detection system.

7 Sources

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097–1179. https://aclanthology.org/2024.cl-3.8/

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. https://aclanthology.org/2024.acl-long.600.pdf

OpenAI (9.10.2025). Defining and evaluating political bias in LLMs. https://openai.com/index/defining-and-evaluating-political-bias-in-llms/ (last accessed 16.10.2025)