Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

POLAR: A Per-User Association Test in Embedding Space

Pedro Bento, Arthur Buzelin, Arthur Chagas, Yan Aquino, Victoria Estanislau, Samira Malaquias, Pedro Robles Dutenhefner, Gisele L. Pappa, Virgilio Almeida, Wagner Meira Jr.

Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
{pedro.bento, arthurbuzelin, arthurchagas, yanaquino, victoria.estanislau, samiramalaquias, virgilio, glpappa, meira}@dcc.ufmg.br, pedroroblesduten@ufmg.br

Abstract

Most intrinsic association probes operate at the word, sentence, or corpus level, obscuring author-level variation. We present *POLAR* (Per-user On-axis Lexical Association Report), a per-user lexical association test that runs in the embedding space of a lightly adapted masked language model. Authors are represented by private deterministic tokens; POLAR projects these vectors onto curated lexical axes and reports standardized effects with permutation p-values and Benjamini–Hochberg control. On a balanced bot–human Twitter benchmark, POLAR cleanly separates LLM-driven bots from organic accounts; on an extremist forum, it quantifies strong alignment with slur lexicons and reveals rightward drift over time. The method is modular to new attribute sets and provides concise, per-author diagnostics for computational social science.

Code — https://github.com/pedroaugtb/POLAR

Introduction

Measuring social associations in language technologies is often approached through intrinsic probes that quantify how strongly words relate to semantic categories in embedding spaces. Among such probes, association tests inspired by psychological instruments have become widely used to study how distributional representations encode stereotypes and domain semantics across tasks and models (Sun et al. 2019; Gonen and Goldberg 2019; Kurita et al. 2019). These tests provide a compact, model-agnostic signal, by projecting embeddings onto interpretable lexical axes and ask whether observed differences are larger than expected under exchangeable labelings of the attribute sets.

At the same time, *user embeddings* have surged in prominence across machine learning subfields. In recommender systems, neural collaborative filtering represents each person with a learned vector that captures stable preferences (He et al. 2017). In networked settings, node-embedding methods such as DeepWalk, node2vec, and GraphSAGE map users to points in a geometry that preserves social neighborhoods and behavioral regularities (Perozzi, Al-Rfou, and Skiena 2014; Grover and Leskovec 2016; Hamilton, Ying, and Leskovec 2017). In dialogue, persona-conditioned models leverage profile representations to steer generation style and content (Zhang et al. 2018). Across these lines, a shared intuition emerges: low-dimensional user vectors act as condensed, reusable summaries of a person's linguistic habits, social position, and topical affinities.

Despite the ubiquity of intrinsic association tests and user-level embeddings, there is still no methodology that directly applies association testing to individual user vectors. Existing intrinsic probes overwhelmingly target words, sentences, or a single shared model space (Sun et al. 2019; Kurita et al. 2019), so author-level associations are typically accessed only indirectly, via task-specific classifiers or by aggregating users into group summaries. This is limiting for computational social science: aggregation can obscure within-group heterogeneity, and in extreme cases even reverse or cancel effects that are consistently present at the individual level (the ecological fallacy) (Robinson 1950). For instance, a corpus might exhibit a strong average association between "immigrants" and "crime" while actually mixing users who consistently use neutral or humanizing language with others who repeatedly deploy slurs; a single aggregate score cannot distinguish these trajectories.

Concretely, common "aggregate-probe" strategies collapse many authors into one representation, for example, computing a corpus-level WEAT/SEAT score in a shared embedding space, averaging sentence-level association scores across all posts, or pooling representations such as mean [CLS] over the dataset. These summaries answer a different question: they characterize the average association of the corpus, not the distribution of associations across authors (Wakefield and Lyons 2010). POLAR instead treats each author as the unit of analysis, producing per-user effect sizes $s(u; \mathcal{A}, \mathcal{B})$ that retain heterogeneity and can reveal whether an association is widespread, concentrated in a minority, or shifting over a user's posting history.

This gap motivates a method that treats each user vector as a first-class object for calibrated association testing. We propose *POLAR*, a per-user lexical association test that operates in the same embedding space as the language model. Concretely, we hash each author to a private token, lightly adapt a masked language model so that the token's vector summarizes that author's linguistic context, and then com-

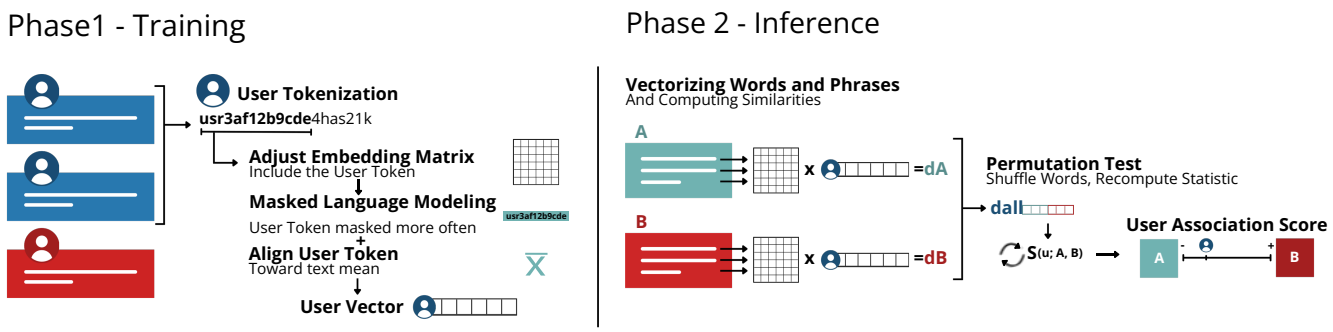

Figure 1: Overview of *POLAR*, which first learns user vectors via masked language modeling with injected tokens (Phase 1), then computes association scores through cosine similarity and permutation testing against lexical attribute sets (Phase 2).

pute standardized effect sizes by projecting the user vector onto curated semantic axes (e.g., *X* vs. *Y*) with Monte Carlo permutation p-values and false-discovery control. Because user and lexical embeddings cohabit the same space, associations are immediately interpretable in both sign and magnitude, revealing not only *which side* of a semantic contrast a user falls on, but also *by how much*, without requiring labels or external supervision. POLAR thus enables: (i) user-token adaptation in pretrained language models, (ii) effect-size inference over lexical axes, and (iii) author-level diagnostic applications, bridging a key methodological gap in current probing paradigms.

We demonstrate that POLAR surfaces author-level structure that aggregate probes miss. On a balanced bot–human Twitter corpus, POLAR recovers clear register differences across multiple axes without using labels at training time. On an extremist forum, POLAR quantifies strong positive alignment with slur lexicons and reveals rightward drift as users continue posting – an interpretable, label-free indicator of progressive radicalization. Together, these findings show that (i) compact user vectors capture stable lexical preferences; (ii) association testing can be cleanly adapted to the author level; and (iii) the resulting scores provide actionable, statistically grounded diagnostics for sociotechnical analysis.

Related Work

Here we span three strands: (i) user representations and compact conditioning that learn concise per-user vectors; (ii) intrinsic association probes (e.g., WEAT/SEAT) applied to words, sentences, or shared spaces; atnd (iii) author-identified datasets on bots, toxicity, and political discourse. These threads motivate our contribution of applying calibrated association testing directly to individual user embeddings.

User Representations in NLP and CSS

User embeddings are now a standard primitive for personalization and social analysis, learned from text, interactions, or both (Pan and Ding 2019a; Kumar, Zhang, and Leskovec 2019; Cécillon et al. 2021; Pougué-Biyong et al. 2023; Cinus et al. 2025). Dense user vectors have long supported personalization and social analysis, with early work learning multiview embeddings from text, networks, and profile signals to predict behavior and demographics (Benton, Arora, and Dredze 2016). In dialogue, persona-based models encode speaker embeddings to capture style and audience effects across interactions (Li et al. 2016). Beyond prediction, user-level vectors illuminate audience effects and individual differences in community settings (Sepahpour-Fard et al. 2023). These works established that users can be represented as stable vectors, making it possible to analyze individual differences and audience effects in online communities.

User Profiling and Bot Detection with Embeddings

A substantial literature builds user representations from social media activity for profiling and detection tasks, including bot identification. For example, Alekseev and Nikolenko (2017) study word-embedding-based user profiling in online social networks, and Heidari, Jones, and Uzuner (2020) use deep contextualized embeddings for text-based user profiling to detect social bots on Twitter. More broadly, user embeddings have been leveraged to separate bots from organic accounts using learned representations of language and behavior. POLAR is complementary: instead of training a task-specific classifier as the primary object, it provides per-user association effect sizes with permutation p-values along interpretable lexical axes, which can then optionally be used as features for downstream models.

Compact Conditioning for Personalization

Recent work compresses user histories into low-dimensional vectors or soft prompts that efficiently steer language models without concatenating long input traces (Doddapaneni et al. 2024; Liu et al. 2025). Prompt tuning, prefix-tuning, adapters, and low-rank adaptation all demonstrate that a small, efficient per-user state can steer large language models (Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Houlsby et al. 2019; Hu et al. 2021). These methods demonstrate the practicality of compact per-user state for conditioning and retrieval.

Intrinsic Bias and Association Tests

The Word Embedding Association Test (WEAT) introduced cosine-based association testing with permutation significance for static embeddings (Caliskan, Bryson, and Narayanan 2017). Subsequent variants extend to sentence encoders and highlight design sensitivities (May et al. 2019; Goldfarb-Tarrant et al. 2020). Prior work has noted that intrinsic bias scores can be unstable or only loosely connected to downstream harms (Blodgett et al. 2020). These analyses, however, remain largely aggregate, targeting words, sentences, or model-wide shared spaces and treating the corpus or model as a single unit of analysis. As a result, they miss heterogeneity across individual authors and cannot distinguish whether a stereotype is uniformly encoded across users or concentrated in a subset of accounts. These concerns motivate complementary approaches that refine how such probes are applied, for instance by shifting attention from aggregate to individual-level embeddings.

In parallel, user-level embeddings are widely used for personalization and prediction in social media, including mental-health assessment (Amir et al. 2017), lifestyle and motivation modeling (Islam and Goldwasser 2021), and unsupervised stance detection in polarized settings (Rashed et al. 2021), with broader surveys reviewing social-media-based user embedding methods (Pan and Ding 2019b). These approaches typically learn low-dimensional user representations from posting histories or social graphs and then feed them into downstream classifiers or clustering pipelines, without quantifying calibrated associations with specific lexical axes. POLAR bridges this gap by treating individual user vectors as first-class targets of association testing, yielding author-level effect sizes and p-values along interpretable semantic frames rather than corpus-level surrogates.

Bots, Toxicity, and Political Discourse

The fox8-23 benchmark documents LLM-powered bot accounts alongside humans (Yang and Menczer 2024), while hate-speech corpora from extremist forums such as Stormfront (de Gibert et al. 2018) provide author-identified data on hostile language. Beyond NLP, social science studies have documented how exposure and community norms shape polarization and radicalization (Bail et al. 2018; Matias 2019). Hate/toxicity corpora with author identifiers provide a complementary lens on hostile language and community norms, while political-discourse datasets capture stance and polarization in the wild. Together they stress-test whether POLAR captures domain-general signals or domain-specific structure.

Despite advances in user embeddings, compact personalization, and intrinsic bias probes, no prior work has systematically adapted association testing to the level of individual users. Existing approaches remain focused on words, sentences, or aggregated representations, overlooking how associations may vary across individual authors. To our knowledge, this leaves an actionable gap: a method that treats user embeddings as first-class targets of calibrated association testing, yielding interpretable, statistically grounded, per-user scores rather than aggregate surrogates.

Methodology

We introduce *POLAR*: a per-user lexical association test run *directly* in the embedding space of a masked-language model (MLM). The pipeline has two phases. First, we obtain a compact vector for each user by inserting a private, deterministic user token into their posts and lightly adapting the MLM so that the token vector summarizes that user's typical linguistic context. Second, holding the model fixed, we compare each user vector to curated attribute word/phrase sets (e.g., *gun safety* vs. *gun rights*) and compute standardized, permutation-tested effect sizes with false-discovery control, as illustrated in Figure 1.

Attribute-Set Construction and Documentation

Attribute pairs $(\mathcal{A}, \mathcal{B})$ define the semantic axes probed by POLAR, so their construction is part of the method rather than an afterthought. We build these sets by combining resources from prior association-test work (e.g., WEAT/SEAT-style lists) with domain knowledge and platform-specific lexicons (such as bot-disclaimer phrases or slur/jargon lists used in hate-speech research), and we apply simple sanity checks to keep the resulting axes interpretable and stable. In practice, we balance list sizes when possible, favor high-frequency lexical items and short phrases to reduce tokenization drop-out, remove near-duplicates and overly ambiguous items, and record per-(user, pair) coverage statistics indicating how many attributes survive tokenization.

Data Construction and User Tokens

Let $\mathcal{D} = \{(u_i, x_i)\}_{i=1}^{N}$ denote posts x_i authored by users u_i. Each user u is mapped to a deterministic token

$$t_u = \texttt{usr} \| \text{SHA1}(u)_{[:10]},$$

added to the tokenizer vocabulary and *prepended* to every post: $\tilde{x}_i = (t_{u_i} \| x_i)$. We lowercase, apply wordpiece tokenization (BERT-base-uncased by default), discard users with fewer than two posts, and optionally cap per-epoch posts per user to curb dominance by prolific accounts.

Training sketch Starting from `bert-base-uncased`, we resize the input embedding matrix $E \in \mathbb{R}^{V' \times d}$ to include the user tokens. Batches are balanced across users. We optimize the standard masked-LM loss with user-aware masking (higher mask probability on the user token) and a small alignment term that nudges the user vector $E[t_u]$ toward the average representation of the post's non-user tokens. A short freeze–then–unfreeze schedule stabilizes user token learning; all hyperparameters are documented in the Appendix.

POLAR in the Learned Space

After training, each user u is represented by the row-normalized embedding

$$\hat{\mathbf{e}}_u = \frac{E[t_u]}{\|E[t_u]\|_2} \in \mathbb{R}^d.$$

Attributes are specified as two finite sets $(\mathcal{A}, \mathcal{B})$ of lexical items, each item being either a unigram or a phrase. We map

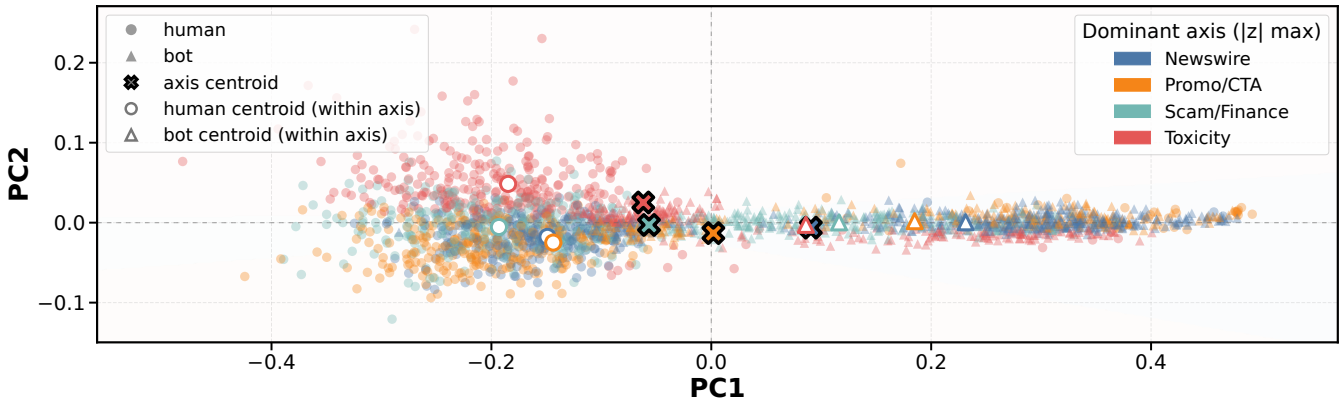

Figure 2: PCA projection of users in the 4D POLAR-axis score space. Marker shape encodes the gold label (bot vs. human), while color indicates the dominant axis (largest $|s|$ for that user). Centroids (black $\times$) summarize average positions; PCA is used for visualization, while classification is performed in the original axis space (Tables 1–2).

any item w to an embedding by averaging its wordpiece vectors under the *trained* tokenizer and then ℓ_2-normalizing; denote this map by ϕ : vocab $\to \mathbb{R}^d$. Stack the resulting unit vectors into matrices

$$A = \begin{bmatrix} \phi(a_1)^\top \\ \cdots \\ \phi(a_m)^\top \end{bmatrix} \in \mathbb{R}^{m\times d}, \qquad B = \begin{bmatrix} \phi(b_1)^\top \\ \cdots \\ \phi(b_n)^\top \end{bmatrix} \in \mathbb{R}^{n\times d}.$$

Statistic. POLAR follows the geometry of the original WEAT statistic, adapted to the user-embedding setting. For a given user u, we measure cosine similarities

$$d_A = A\,\hat{\mathbf{e}}_u \in \mathbb{R}^m,$$
$$d_B = B\,\hat{\mathbf{e}}_u \in \mathbb{R}^n,$$
$$d_{\text{all}} = \begin{bmatrix} d_A \\ d_B \end{bmatrix} \in \mathbb{R}^{m+n}.$$

and define the standardized effect size

$$s(u;\mathcal{A},\mathcal{B}) = \frac{\text{mean}(d_A) - \text{mean}(d_B)}{\text{sd}(d_{\text{all}})}. \qquad (1)$$

The numerator can be written $(\bar{\mathbf{a}} - \bar{\mathbf{b}}) \cdot \hat{\mathbf{e}}_u$, where $\bar{\mathbf{a}} = \frac{1}{m}\sum_{i=1}^m \phi(a_i)$ and $\bar{\mathbf{b}} = \frac{1}{n}\sum_{j=1}^n \phi(b_j)$; thus s is a projection of the user vector onto the semantic axis $\bar{\mathbf{a}} - \bar{\mathbf{b}}$ rescaled by the empirical dispersion of all similarities for that user–pair. Standardization yields a unitless, scale-stable quantity that is comparable across pairs with different (m, n) and internal variability.

Null, p-value, and exchangeability. The null hypothesis states that user u exhibits no differential association with $\mathcal{A}$ versus $\mathcal{B}$. Under this null, labels on the $m+n$ entries of d_{all} are exchangeable. We estimate a two-sided p-value by Monte Carlo permutation with M random re-labelings (default M=2000). Writing s_{obs} for the observed statistic and $s^{(k)}$ for the statistic at permutation k, the estimate

$$\hat{p} = \frac{1 + \sum_{k=1}^M \mathbb{I}\big(|s^{(k)}| \geq |s_{\text{obs}}|\big)}{1+M}$$

uses add-one smoothing to avoid zero p-values under finite M and to bound the Monte Carlo error. This test requires no distributional assumptions beyond exchangeability and remains valid for small lists as long as $m, n > 0$.

Multiplicity Control. Because each attribute pair is tested across many users, we control the expected false discovery proportion *per pair* via the Benjamini–Hochberg procedure at level α=0.05, applied to the vector of permutation p-values $\{\hat{p}(u;\mathcal{A},\mathcal{B})\}_u$. Reporting both the raw statistic s and the FDR-adjusted decision enables effect-size interpretation alongside inferential guarantees.

Interpretability and Benefits. Two properties make POLAR particularly useful for computational social science. First, user vectors and lexical items occupy the same embedding geometry; the sign of s encodes direction along a named semantic axis, while its magnitude captures graded alignment. Analysts may therefore read results as *which side* of a frame a user aligns with, and by how much, without training auxiliary classifiers. Second, inference is modular: adding or revising attribute sets does not alter the estimation or testing machinery, and phrase-level attributes are handled natively via subword averaging. In practice, we accompany per-user scores with basic diagnostics: realized coverage (how many attributes survived tokenization for a pair and user), a separability sanity check via $\cos(\bar{\mathbf{a}}, \bar{\mathbf{b}})$, and explicit flagging of degenerate cases where $\text{sd}(d_{\text{all}}) = 0$.

Evaluation Protocol and Baselines

We evaluate discriminative performance on the fox8–23 bot–human benchmark using a one-layer logistic regression model trained on four distinct, label-agnostic feature sets. These include: (i) POLAR axes – per-user association scores $s(u;\mathcal{A},\mathcal{B})$ computed with labels held out; (ii) Mean CLS – the mean pooled [CLS] embedding of a user's posts, reduced via PCA to 64 dimensions; (iii) Aggregated Sentence WEAT – post-level $s(x)$ scores using the same attribute sets as POLAR, averaged across posts per user; and (iv) Detox/Botspeak proxies – average toxicity scores and simple behavior

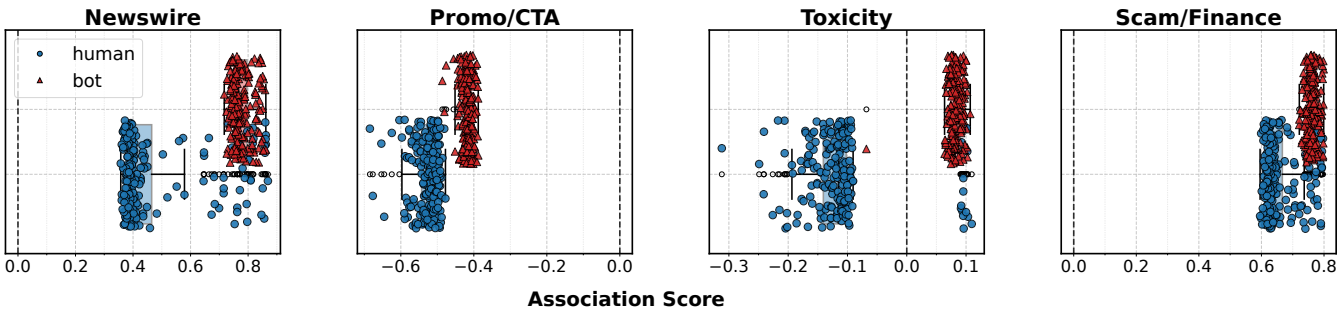

Figure 3: Box-and-swarm plots of per-user *POLAR* scores $s(u; \mathcal{A}, \mathcal{B})$ (y-axis; Eq. 1) for humans and bots across four lexical axes. Points are individual users; boxplots summarize median and interquartile range.

indicators such as URL, hashtag, mention, and call-to-action rates.

To avoid label leakage, we apply stratified 5-fold cross-validation. For (i), we select the top-k axes within each training fold by ranking them via univariate AUROC. Logistic regression is then trained on those selected axes and evaluated on the held-out fold. We report out-of-fold (OOF) metrics, including AUROC with 95% confidence intervals from 2,000 bootstrap resamples, as well as PR–AUC and Brier score. Paired bootstrap tests on OOF predictions assess whether increasing the number of axes (e.g., $k = 2, 3, 4$) significantly improves performance over $k = 1$. All POLAR computations – including axis statistics and permutations – remain fully unsupervised; only the logistic regression classifier is trained with labels.

Alongside discrimination, we report expected calibration error (ECE), computed on out-of-fold (OOF) probabilities with 10 equal-width bins unless stated otherwise. ECE summarises the average gap between predicted confidence and empirical accuracy across bins; lower is better. We include ECE to show that the probabilities derived from single-axis POLAR scores can be used as risk-aware signals, not just for ranking.

Reproducibility

We fix stochastic seeds, store paths and hyperparameters in `meta.json`, and save the exact tokenizer (including user tokens). Unless noted otherwise, defaults are: sequence length 128, batch size 128, learning rate 5×10^{-5}, 4 epochs, warm-up ratio 0.03, gradient clip 1.0, user-mask probability 0.30, MLM mask 0.15, alignment weight 0.2, and M=2000 permutations. Mixed precision (bf16 when available, else fp16) is used automatically. Full training details are provided in the Appendix; the inference described above is unchanged by those choices.

Datasets

We evaluate POLAR on two corpora with user identifiers but distinct sociolinguistic regimes: (i) a balanced Twitter benchmark of LLM-powered bots and humans, used as a "vanilla" test of whether per-user lexical associations capture broad register differences; and (ii) a white-supremacist forum corpus with sentence-level hate labels, used to probe users' stance and association with sensitive targets and policy frames.

Bot–Human Twitter Benchmark (fox8–23)

The fox8–23 collection comprises recent tweets from 1,140 accounts identified as AI-powered social bots and 1,140 human accounts, yielding a balanced bot–human sample at the *user* level. Accounts were discovered and validated in a study of an LLM-driven botnet; the release provides tweet timelines for both groups.[1]

What we test with POLAR. This corpus serves as a sanity check that POLAR detects coarse stylistic divergences without supervision. We instantiate attribute pairs that contrast LLM-style "bot speak" against ordinary conversational language – e.g., apology/safety disclaimers and instructional scaffolding (*"I apologize", "as an AI…", "let's break this down"*) versus colloquial, experiential, and interactional markers (*"I think/feel", "lol", "gonna", "in my experience"*). The hypothesis is that bots align more strongly with the former and humans with the latter; POLAR operationalizes this as a per-user standardized effect $s(u)$ along each lexical axis with permutation p-values and BH–FDR control.

Stormfront White-Supremacist Forum

The Stormfront dataset contains 10,568 English sentences extracted from posts published between 2002 and 2017 on a white-supremacist forum. Sentences were annotated at sentence granularity (HATE vs. NOHATE, plus auxiliary labels) under detailed guidelines; crucially for our purposes, the release includes per-sentence *user identifiers*, post positions, and subforum metadata, enabling reconstruction of user histories while shielding the original source via placeholder IDs.

What we test with POLAR. Here we use POLAR to quantify each user's lexical alignment with sensitive topics and targets. Attribute pairs trace ideologically loaded frames and target-group semantics, contrasting derogatory

[1]We treat any dataset-provided account identifier as a user key for tokenization; raw screen names are never emitted by our pipeline.

or securitizing lexicons (*e.g.*, outgroup slurs; "invasion", "criminal aliens") against neutral or humanizing counterparts (*e.g.*, group self-references; "immigrants", "refugees", "equal rights"). Additional pairs cover policy frames frequently debated in such forums (immigration enforcement vs. reform, religious intolerance vs. freedom, gendered derogation vs. gender equality). POLAR yields per-user effect sizes and significance flags, surfacing heterogeneity within the community (users with strong positive/negative alignment on different axes) beyond aggregate hate/no-hate rates.

Preprocessing and inclusion. Across both corpora we lowercase, tokenize with the trained wordpiece model, and retain users with at least two posts (or sentences) to ensure stable vectors; each user is mapped to a deterministic hashed token and raw text is not required during testing.

Results

This section quantifies how POLAR surfaces meaningful linguistic variation at the author level. We first use the fox8–23 bot–human benchmark to verify that our method detects coarse stylistic divergences when the underlying language models differ sharply (LLM-generated vs. organic tweets). We then turn to Stormfront to probe ideological alignment and radicalisation. Unless noted, figures show standardized effect sizes $s(u; \mathcal{A}, \mathcal{B})$ (Eq. 1) after Benjamini–Hochberg correction, with each dot representing a single user.

Bot–Human Baseline

Figure 2 projects each user into the first two principal components of the four-dimensional POLAR score space.[2] Users with similar association profiles concentrate near category centroids (black $\times$), suggesting that frames such as Promo/CTA or Toxicity capture stable register preferences. Clusters are elongated and overlapping: many accounts mix registers, and a non-trivial share of humans lie near bot-dominated regions along PC_1, reflecting topic drift and short timelines.

To quantify separability, we train out-of-fold (OOF) logistic regressions on POLAR scores in two complementary ways: (i) one score at a time to assess the informativeness of individual axes (Table 1), and (ii) a compact top-k model that selects axes within each training fold to test whether combining them helps (Table 2). On the single-axis setting (users N=2,277), Scam/Finance vs. Daily Life and Newswire vs. Personal Experience each exceed AUROC 0.95, with strong PR–AUC and good probability quality (low Brier; moderate ECE), while Promo/CTA vs. Hedges remains informative and Toxicity vs. Civility is clearly weaker – consistent with style and topical framing, not generic toxicity, driving the bot signal. The attribute anchor sets themselves are well separated in embedding space (cosine between positive/negative centroids 0.74–0.86 across axes), supporting construct validity of the frames.

A complementary view of raw score distributions appears in Figure 3: four box-and-beeswarm panels (one per axis) where humans (blue circles) and bots (red triangles) form two clouds with medians differing by more than one pooled standard deviation on every axis, and whiskers that rarely overlap – making the separation visually apparent.

[2]Colours encode the user's *dominant* category, i.e., the category where $|s|$ is largest.

Axis	AUROC	PR–AUC	Brier	ECE
SCAM/FINANCE	**0.961**	0.973	0.061	0.073
NEWSWIRE	**0.950**	0.962	0.087	0.108
PROMO/CTA	0.915	0.930	0.116	0.068
TOXICITY	0.868	0.878	0.145	0.054

Table 1: Single-axis OOF performance on fox8–23 (one LR on each *POLAR* score). Best two AUROC in bold.

Features (axes)	AUROC [95% CI]	PR–AUC	Brier
POLAR (k=1)	0.961 [0.952, 0.970]	0.911	0.060
POLAR (k=2)	0.960 [0.951, 0.969]	0.913	0.061
POLAR (k=3)	0.961 [0.952, 0.970]	0.918	0.060
POLAR (k=4)	0.961 [0.952, 0.970]	0.918	0.060

Table 2: OOF performance on fox8–23 using a 1-layer logistic regression fed only with *POLAR* axis scores. AUROC CIs from $2{,}000$ bootstrap resamples.

Finally, the top-k OOF model (Table 2) shows that adding axes beyond k=1 yields negligible differences: AUROC/PR–AUC remain essentially unchanged and calibration hardly improves. Paired bootstrap comparisons versus k=1 are not significant after Benjamini–Hochberg correction. Overall calibration is reasonable (ECE $\leq$ 0.108), with Scam/Finance offering the best Brier (0.061).

Note the k=1 model reselects the top axis within each training fold by AUROC; consequently, its fold-varying choice can yield lower aggregate PR–AUC than the best fixed single-axis model in Table 1.

These results validate POLAR on a setting where linguistic differences are stark: without supervision, it exposes author-level effect sizes aligned with intuitive category semantics and yields diagnostic plots that make the separation visible.

White-Supremacist Forum

Figure 4 summarises POLAR scores for seven semantic axes in the Stormfront corpus (N=2,082). All slur-based categories – Racial, Gender, LGBTQ, and Incel jargon – cluster on the positive side, with means of 1.33, 1.32, 0.71, and 1.24, respectively; Violence shows a moderate positive mean (0.26), and Sentiment is negative (-0.79). The corresponding anchor centroids are also strongly separated (cosine 0.78–0.87 across axes), indicating coherent attribute sets. Thus, in a community with an overt ideological leaning, POLAR pinpoints that bias numerically and visually, expliciting that the users embeddings sit closer to hateful or derogatory lexicons than to their neutral counterparts across every sensitive dimension we probe.

Because the forum lacks user-level gold labels, we report descriptive statistics for each axis: the mean standardized effect s across users and the fraction significant after Benjamini–Hochberg at α=0.05. Slur/jargon axes show large positive means with near-universal significance,

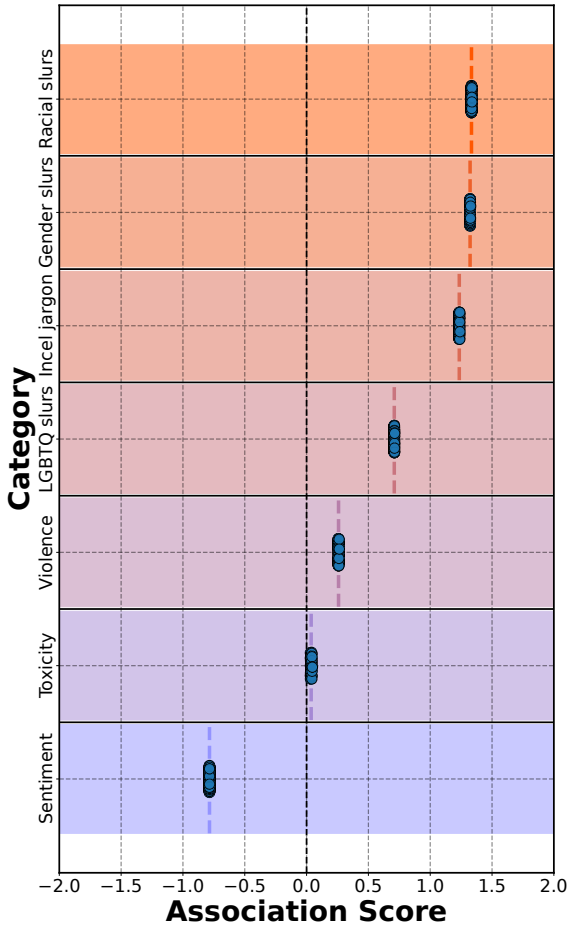

Figure 4: Stormfront per-user *POLAR* associations on sensitive targets and policy frames.

while generic Toxicity and Violence exhibit small or non-significant shifts (Table 3), indicating that *domain-specific lexicons* – not blanket hostility – drive the embedding signal here.

Figure 5 illustrates how selected users' POLAR scores change as their posting histories grow. For each highlighted author, we recompute a cumulative association score after the t-th post, denoted $s_t(u; \mathcal{A}, \mathcal{B})$, using all posts observed up to index t. Each panel plots these cumulative scores on the x-axis (labeled "WEAT score" for historical continuity with prior association-test literature), while the y-axis is an uninformative jitter used only to reduce overlap and make individual steps visible. Colored markers correspond to successive posting steps (numbers indicate t), and connecting lines are drawn solely to guide the eye through the temporal order.

The highlighted trajectories drift steadily rightward across all four axes, indicating that continued participation is associated with increasing alignment between user embeddings and the slur lexicons. This pattern is observational (not causal) but consistent with progressive radicalization signals accumulating in users' text histories. Because POLAR is lightweight, label-free, and recomputable after each message, these rightward drifts offer an interpretable early-warning indicator – even for short or sparsely annotated histories.

Axis	Mean s	Sig. BH
Racial Slurs vs. Neutral	1.332	1.00
Gender Slurs vs. Respect	1.320	1.00
Incel Jargon vs. Relationships	1.232	1.00
LGBTQ Slurs vs. Neutral	0.707	1.00
Violence vs. Peace	0.258	0.00
Toxicity vs. Civility	0.033	0.00
Sentiment (neg. vs. pos.)	−0.784	1.00

Table 3: Stormfront: axis-level descriptives. *Sig. BH* is the share of users significant after BH–FDR at α=0.05 for that axis.

Conclusion

This study introduced *POLAR*, the first *per-author* lexical association test that operates directly in the embedding space of a lightly adapted masked-language model. By hashing user identifiers into private tokens and fine-tuning only those token vectors, we generate compact, interpretable representations that remain geometrically compatible with ordinary wordpieces. Standardized effect sizes computed between each user vector and curated attribute sets then provide a scale-stable, author-level measure of stance or bias – complete with permutation p-values and false-discovery control.

Empirical results across two contrasting corpora underscore the method's versatility. On the fox8–23 bot–human benchmark, POLAR cleanly separates LLM-driven bots from organic accounts along multiple stylistic axes without ever observing class labels. In the Stormfront forum, the same statistic exposes strong, uneven alignment with hateful or derogatory lexicons, and cumulative trajectories signal progressive radicalization over time. Taken together, these findings show that user embeddings capture stable lexical preferences, that these preferences can be probed with lightweight, modular tests, and that the resulting scores offer immediate, interpretable diagnostics for computational social science.

Limitations and Future Work

Scope of corpora. Our evaluation spans Twitter and a single extremist forum; generalising to other platforms (e.g., long-form blogs, chat applications) or multilingual settings remains future work.

Short histories and silent users. POLAR requires at least two posts per author; users with very sparse histories are discarded. Alternative smoothing or hierarchical pooling strategies could extend coverage.

Attribute design bias. Choice of attribute sets steers interpretation. Although our pairs are drawn from prior literature and domain expertise, they may omit salient frames

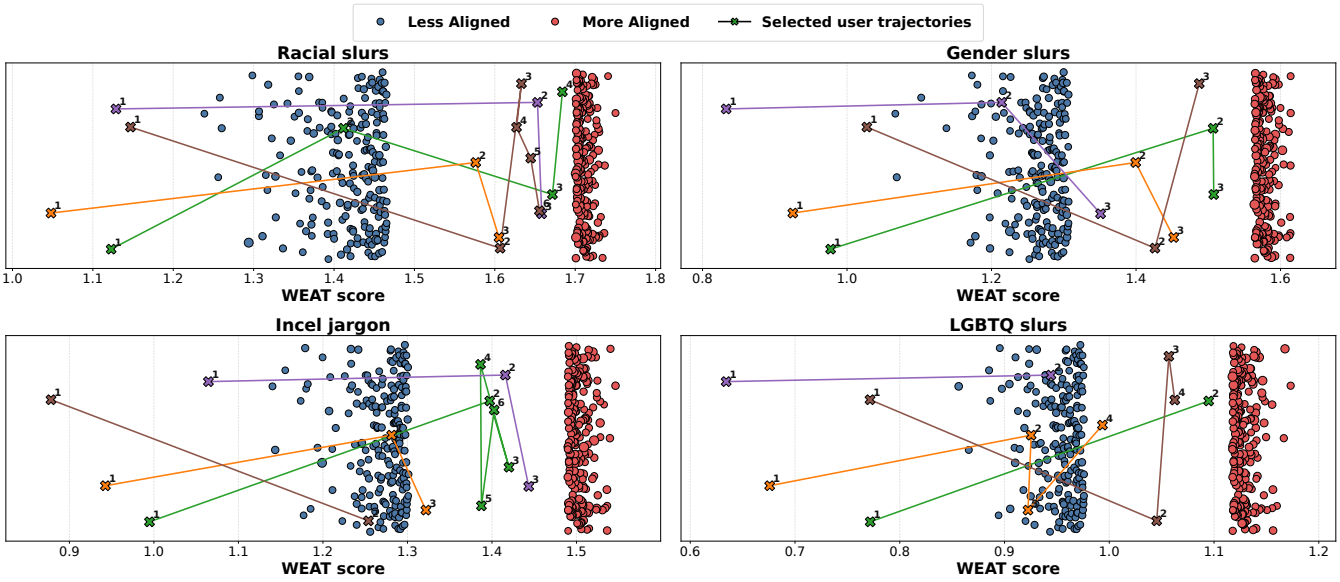

Figure 5: Stormfront dynamics: per-user scatterplots mark least- (blue) and most-aligned (red) accounts on four slur axes; coloured trajectories track selected users with the fastest drift toward alignment.

or inadvertently encode researcher bias. Crowdsourced or community-generated attributes are a promising mitigation.

Ethical Statement

We use only public, researcher-released datasets (fox8–23; Stormfront) under their licenses – no new collection or user interaction. Analyses rely on pseudonymous IDs and irreversible hashed user tokens; we release no account mappings or additional personal data. Sensitive lexicons are used solely for measurement; quotes are minimized with content warnings; results are aggregate only. Code is research-only and explicitly forbids surveillance/profiling uses. This observational public-data study did not require human-subjects review; we follow ICWSM ethics guidelines and discuss risks and limitations.

Acknowledgments

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

Alekseev, A.; and Nikolenko, S. 2017. Word embeddings for user profiling in online social networks. *Computación y Sistemas*, 21(2): 203–226.

Amir, S.; Coppersmith, G.; Carvalho, P.; Silva, M. J.; and Wallace, B. C. 2017. Quantifying Mental Health from Social Media with Neural User Embeddings. arXiv:1705.00335.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. B. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Benton, A.; Arora, R.; and Dredze, M. 2016. Learning Multiview Embeddings of Twitter Users. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 14–19. Berlin, Germany: Association for Computational Linguistics.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Cécillon, N.; Labatut, V.; Dufour, R.; and Linares, G. 2021. Graph embeddings for abusive language detection. *SN Computer Science*, 2(1): 37.

Cinus, F.; Monti, C.; Bajardi, P.; and Morales, G. D. F. 2025. On the Inference of Sociodemographics on Reddit. *arXiv preprint arXiv:2502.05049*.

de Gibert, O.; Perez, N.; García-Pablos, A.; and Cuadros, M. 2018. Hate Speech Dataset from a White Supremacy Forum. In Fišer, D.; Huang, R.; Prabhakaran, V.; Voigt, R.; Waseem, Z.; and Wernimont, J., eds., *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, 11–20. Brussels, Belgium: Association for Computational Linguistics.

Doddapaneni, S.; Sayana, K.; Jash, A.; Sodhi, S.; and Kuzmin, D. 2024. User Embedding Model for Personalized Language Prompting. In Deshpande, A.; Hwang, E.; Murahari, V.; Park, J. S.; Yang, D.; Sabharwal, A.; Narasimhan, K.; and Kalyan, A., eds., *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE*

2024), 124–131. St. Julians, Malta: Association for Computational Linguistics.

Goldfarb-Tarrant, S.; Marchant, R.; Sánchez, R. M.; Pandya, M.; and Lopez, A. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In Axelrod, A.; Yang, D.; Cunha, R.; Shaikh, S.; and Waseem, Z., eds., *Proceedings of the 2019 Workshop on Widening NLP*, 60–63. Florence, Italy: Association for Computational Linguistics.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 855–864. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 173–182. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450349130.

Heidari, M.; Jones, J. H.; and Uzuner, O. 2020. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. In *2020 International Conference on Data Mining Workshops (ICDMW)*, 480–487. IEEE.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Islam, T.; and Goldwasser, D. 2021. Analysis of Twitter Users' Lifestyle Choices using Joint Embedding Model. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1): 242–253.

Kumar, S.; Zhang, X.; and Leskovec, J. 2019. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1269–1278. ACM.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. In Costa-jussà, M. R.; Hardmeier, C.; Radford, W.; and Webster, K., eds., *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. Florence, Italy: Association for Computational Linguistics.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003. Berlin, Germany: Association for Computational Linguistics.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597. Online: Association for Computational Linguistics.

Liu, L.; Liu, S.; Yuan, Y.; Zhang, Y.; Yan, B.; Zeng, Z.; Wang, Z.; Liu, J.; Wang, D.; Su, W.; Wang, P.; Xu, J.; and Zheng, B. 2025. UQABench: Evaluating User Embedding for Prompting LLMs in Personalized Question Answering. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, 5652–5661. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714542.

Matias, J. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116: 201813486.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. *CoRR*, abs/1903.10561.

Pan, S.; and Ding, T. 2019a. Social media-based user embedding: A literature review. *arXiv preprint arXiv:1907.00725*.

Pan, S.; and Ding, T. 2019b. Social Media-based User Embedding: A Literature Review. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6318–6324. International Joint Conferences on Artificial Intelligence Organization.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. DeepWalk: online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '14, 701–710. ACM.

Pougué-Biyong, J.; Gupta, A.; Haghighi, A.; and El-Kishky, A. 2023. Learning stance embeddings from signed social graphs. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 177–185.

Rashed, A.; Kutlu, M.; Darwish, K.; Elsayed, T.; and Bayrak, C. 2021. Embeddings-Based Clustering for Target

Specific Stances: The Case of a Polarized Turkey. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1): 537–548.

Robinson, W. S. 1950. Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, 15(3): 351–357.

Sepahpour-Fard, M.; Quayle, M.; Schuld, M.; and Yasseri, T. 2023. Using word embeddings to analyse audience effects and individual differences in parenting Subreddits. *EPJ Data Science*, 12(1): 38.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. Florence, Italy: Association for Computational Linguistics.

Wakefield, J.; and Lyons, H. 2010. Spatial aggregation and the ecological fallacy. *Handbook of spatial statistics*, 20103158: 541–558.

Yang, K.-C.; and Menczer, F. 2024. Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, 4.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213. Melbourne, Australia: Association for Computational Linguistics.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes.
   (f) Have you related your theoretical results to the existing literature in social science? Yes.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA.

(d) Did you discuss how data is stored, shared, and deidentified? NA.

Appendix

Artifacts & File Schema

We release one run folder per corpus:

- **fox8–23 (bots vs. humans)**: `out_bot_human_bert_space/`
- **Stormfront (white-supremacist forum)**: `out_hate_bert_space/`

Each contains:

- `model/`: HuggingFace tokenizer and embeddings (includes user tokens).
- `users.csv`: columns `user_id`, `n_posts`; optional `token`, `label_majority`, `targets`.
- `meta.json`: run metadata (e.g., `usr_prefix`, `min_posts_per_user`, seed, training flags).
- `per_user_scores.csv` (long format): `user_id`, `pair`, `s`, `p_perm`, `n_posts`, `n_pos_attr`, `n_neg_attr`, `label_majority`, `targets`, `signif_bh_fdr_0.05`.

Attribute Sets

We embed exact word/phrase lists directly from the trained vocabulary and publish them with the code/artifacts. Examples (only to aid figure reading):

fox8–23 (register axes). *Newswire vs. Personal Experience* (e.g., "court, officials" *vs.* "tbh, lol"); *Promo/CTA vs. Hedges* ("offer, discount" *vs.* "I think, not sure"); *Toxicity vs. Civility* ("idiot, garbage" *vs.* "thank you, empathy"); *Scam/-Finance vs. Daily Life* ("airdrop, seed phrase" *vs.* "movie night, weekend").

Stormfront (sensitive axes). *Sentiment*; *Toxicity vs. Civility*; *Racial slurs vs. Neutral*; *Gender slurs vs. Respect*; *LGBTQ slurs vs. Neutral*; *Violence vs. Peace*; *Incel jargon vs. Relationships*.

Sensitive lexicons are used strictly for measurement on public datasets; outputs are anonymized at analysis time.

Diagnostics & Edge Cases

- **Coverage**: per (user, pair) we record how many attribute items survive tokenization; low coverage is flagged in `per_user_scores.csv`.
- **Axis sanity**: we log $\cos(\bar{\mathbf{a}}, \bar{\mathbf{b}})$ (centroid similarity) to detect poorly separated pairs.
- **Degeneracy**: if $\mathrm{sd}([d_A; d_B]) = 0$ or one side is empty post-tokenization, we set `s`/`p_perm`=`NaN`; these remain non-significant after FDR.
- **Missing user tokens**: users whose tokens are absent from the saved vocabulary are skipped and counted at inference time.

Compute & Scaling

Let U be users, d the embedding dimension, and K total attributes per pair. Inference costs $O(UKd)$ for dot products; each permutation replicate operates in $O(K)$ (relabel only). With M=2000 and modest K (tens), runtime grows linearly with U and the number of tested pairs.

Training & Hyperparameters

This section centralizes training settings to avoid cluttering Methodology.

Base LM & user tokens. `bert-base-uncased` (HuggingFace). One deterministic private token per user, format `usr<sha1[:10]>`, *prepended* to every post. Vocabulary resized once to include all user tokens. Users with <2 posts are excluded.

Objective. Masked-LM cross-entropy with mask prob. 0.15. Additional user-token mask prob. 0.30. Alignment term that nudges the user token toward the mean of non-user token embeddings in the post; weight 0.2.

Batching. Batches are approximately balanced across users. Optional per-epoch cap of posts per user (default: off). Values are recorded in `meta.json`.

Optimizer & schedule. AdamW (betas $[0.9, 0.999]$, $\epsilon = 10^{-8}$, weight decay 0.01). Learning rate 5×10^{-5} with linear warm-up (ratio 0.03) then linear decay. Gradient clip 1.0.

Run settings. Sequence length 128; batch size 128; epochs 4; gradient accumulation 1; dropout as in BERT-base defaults. Mixed precision: bf16 when available, else fp16. Random seed 123 (NumPy/PyTorch). No early stopping; final checkpoint is used.

Stability. Short freeze–then–unfreeze schedule for user-token rows during early steps (stabilizes token learning without changing downstream inference).

Compute. Single-GPU training; inference is CPU/GPU agnostic.

Provenance. Runs write `meta.json` (hyperparameters, `usr_prefix`, `min_posts_per_user`) and save the tokenizer with user tokens under `model/`.