Political Bias in Neural Retrieval: Model-Agnostic Measurement on Canadian Social Media

Anonymous Authors

Motivation. Representational bias inside embedding spaces is a *structural* property: geometric distortions that persist across tasks and prompts. Because the same vector geometry underlies ranking, classification, clustering, and generation, intrinsic bias can reappear unpredictably at deployment [1, 2]. We study *model-agnostic, no-retraining* ways to measure and ultimately mitigate such bias in real systems.

Setup. We evaluate widely used embedding models (BGE-M3, Contriever, E5, E5-Large, Gemini text-embedding, GTE, MiniLM, MiniLM-12, MPNet) alongside a sparse baseline (BM25) on a balanced political corpus built from large-scale Canadian social media posts (790K) authored by elected officials (four groups: Conservative, Liberal, NDP, Others (like Green Party, Bloc Québécois etc.). Queries reflect real information needs (e.g., short, Google-Trends-like terms). All evaluations are run on the same balanced pool to decouple model behavior from corpus volume.

Metrics (model-agnostic). We combine three complementary measures: (i) *Polarization Entropy (PE)*: concentration of one affiliation in top-k; (ii) α-*nDCG*[3]: early-rank viewpoint diversity (rewarding novel affiliations high in the list); (iii) *Normalized KL divergence (nDKL)*: deviation from an ideal uniform exposure. Together, they capture concentration, diversity-at-the-top, and global deviation.

Key findings. We observe systematic, statistically reliable differences across models: some dense retrievers surface disproportionately Conservative content, others skew Liberal on the same balanced pool; early ranks show the largest disparities. A compact MiniLM variant tends to exhibit lower polarization (more balanced exposure), while several contrastively trained models show higher nDKL on politically charged queries. Category-wise analysis (e.g., Law & Government vs. Sports) indicates context sensitivity: fairness can break down on ideologically loaded topics even when overall averages look reasonable. These results confirm that intrinsic geometry matters for political exposure and that auditing must focus on early ranks, where user attention is highest.

Implications. Our measurements motivate deployable, *post-hoc* filters that (a) damp explicit clustering when a single affiliation dominates high ranks and (b) apply lighter directional corrections when bias appears as subtle shifts. Because the approach is model-agnostic and requires no retraining, it is feasible for closed/proprietary APIs and production IR/RAG stacks.

Limitations & Next Steps. Balanced-party pools reduce volume confounds but do not eliminate topic-prevalence differences across parties. Future work: (1) intrinsic-space correction operators with utility guarantees (e.g., preserving retrieval fidelity), (2) extension to non-political axes (e.g., gender) where bias appears as small directional drifts rather than separable clusters, and (3) longitudinal audits on live systems.

References

[1] Bolukbasi et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? *Neural Information Processing Systems (NIPS)*.

[2] Blodgett et al. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Association for Computational Linguistics*.

[3] Clarke et al. (2008). Novelty and Diversity in IR Evaluation (α-nDCG). *SIGIR*.