UNIVERSITAS
GADJAH MADA
Analysis on Ethnic Bias in Indonesian Toxicity and Polarization Model Based on IndoDiscourse Dataset
Muhammad Fata Akyas, Drs. Medi, M.Kom.; Rifki Afina Putri, S.T., M.S., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

ABSTRACT

Analysis Of Ethnic Bias in Indonesian Toxicity and Polarization Model Based on IndoDiscourse Dataset

By
Muhammad Fata Akyas
20/454539/PA/19570

Social media texts containing toxicity and polarizing ideologies pose significant psychological risks to readers, making the detection of hate speech a critical focus in global and local Natural Language Processing (NLP) research. The IndoDiscourse project provides a foundational dataset for researching Indonesian toxicity and polarization. However, to optimize their IndoBERTweet baseline model, the original researchers utilized data aggregation methods, specifically majority voting and the discarding of ambiguous cases. This approach risks dropping specific annotator perspectives during training, potentially embedding ethnic bias into the model.

This research aims to analyze the extent of annotator ethnic bias present in the IndoDiscourse baseline model. Through a comprehensive evaluation involving data annotation agreement (Gwet's AC1 and Cohen's Kappa), ethnic test-set performance comparisons, and demographic-conditioned modeling, this study demonstrates varying performance metrics across different ethnic groups. While most groups show similar scores to the baseline, specific minority subsets, Chinese annotator group for instance, exhibited reduced Macro F1 scores and a dependency on explicit demographic tagging for performance improvement in both toxicity and polarization set, indicating a divergence from the mainstream model understanding. Furthermore, the results prove that a low annotation agreement score does not inherently translate to worse model performance, indicating that annotator bias and model predictive bias are not strictly correlated.

Keywords: Natural language processing, ethnic bias analysis, annotation agreement, demographic conditioning.

UNIVERSITAS
GADJAH MADA
Analysis on Ethnic Bias in Indonesian Toxicity and Polarization Model Based on IndoDiscourse Dataset
Muhammad Fata Akyas, Drs. Medi, M.Kom.; Rifki Afina Putri, S.T., M.S., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

INTISARI

Analisis Bias Etnis dalam Model Toksisitas dan Polarisasi Indonesia Berdasarkan Dataset IndoDiscourse

Oleh
Muhammad Fata Akyas
20/454539/PA/19570

Teks-teks di media sosial yang mengandung konten yang toksik dan berpolarisasi pada ideologi menimbulkan risiko psikologis yang signifikan bagi pembaca, sehingga deteksi ujaran kebencian menjadi salah satu fokus utama dalam penelitian Pemrosesan Bahasa Alami (NLP) baik di tingkat global maupun lokal. Proyek IndoDiscourse menyediakan dataset dasar untuk meneliti konten bertoksisitas dan mempolarisasi di Indonesia. Namun, untuk mengoptimalkan model dasar IndoBERTweet mereka, para peneliti sebelumnya menggunakan metode agregasi data, khususnya pemungutan suara mayoritas dan penyisihan kasus-kasus yang ambigu. Pendekatan ini berisiko mengabaikan perspektif-perspektif tertentu dari para anotator selama pelatihan, yang berpotensi menanamkan bias etnis ke dalam model.

Penelitian ini bertujuan untuk menganalisis tingkat bias etnis yang ditemukan dalam model dasar IndoDiscourse. Melalui evaluasi komprehensif yang mencakup persetujuan anotasi data (AC1 Gwet dan Cohen's Kappa), perbandingan kinerja pada kumpulan data uji berdasarkan etnis, serta pemodelan yang disesuaikan dengan kondisi demografis, penelitian ini menunjukkan variasi metrik performa pada kelompok etnis yang berbeda. Meskipun sebagian besar kelompok menunjukkan skor yang serupa dengan model dasar, subkelompok minoritas tertentu, seperti kelompok anotator Tionghoa, menunjukkan skor Macro F1 yang lebih rendah dan ketergantungan pada penandaan demografis untuk peningkatan metrik pada kedua kumpulan data toksisitas dan polarisasi, mengindikasikan adanya penyimpangan dari pemahaman model utama. Selain itu, hasil penelitian ini menunjukkan bahwa skor persetujuan anotasi yang rendah tidak langsung mengindikasikan performa model yang lebih buruk, yang artinya bias anotator dan bias prediktif model tidak berkorelasi secara signifikan.

Kata kunci: Pemrosesan bahasa alami, analisis bias etnis, persetujuan anotasi, pengkondisian demografis.