IDENTIFIKASI BIAS SOSIAL EKONOMI DALAM MODEL BAHASA AI INDONESIA MELALUI *ETHICAL PROBING*

Fadilah Zahra Dwi Kinanti*[1)], Ari Maulida Aprilia[2)], Aldila Rachma Aulia[3)], Anis Nadhirotul Mustafida[4)], Dicky Anggriawan Nugroho[5)]

1. Informatika, Universitas Islam Negeri K.H Abdurrahman Wahid Pekalongan email: fadilah.zahra.dwi.kinanti24084@mhs.uingusdur.ac.id
2. Informatika, Universitas Islam Negeri K.H Abdurrahman Wahid Pekalongan email: ari.maulida.aprilia24068@mhs.uingusdur.ac.id
3. Informatika, Universitas Islam Negeri K.H Abdurrahman Wahid Pekalongan email: aldila.rachma.aulia24066@mhs.uingusdur.ac.id
4. Informatika, Universitas Islam Negeri K.H Abdurrahman Wahid Pekalongan email: anis.nadhirotul.mustafida24043@mhs.uingusdur.ac.id
5. Informatika, Universitas Islam Negeri K.H Abdurrahman Wahid Pekalongan email: dicky.anggriawannugroho@uingusdur.ac.id

Abstract

This study evaluates socioeconomic bias in three large language models (LLMs) that support Indonesian Nusantara, IndoGPT, and SEA-LION using an ethical probing approach. A total of 100 short narrative prompts (4–11 words) were compiled to represent issues of poverty, informal employment, access to education, and regional contexts. Each model output was analyzed using five key indicators: emotional valence, stereotypes, narrative themes, framing, and deontic indicators. The results show that all three models tend to produce neutral responses, especially SEA-LION, which has the highest proportion of neutral responses. However, stereotypes still appeared at almost the same level across all models, indicating that a neutral tone does not guarantee bias-free output. IndoGPT showed the highest use of normative language, while Nusantara more often displayed structural framing and empathetic nuances. In contrast, SEA-LION was the most stable in maintaining neutrality without eliminating implicit stereotypical tendencies. These findings confirm that socioeconomic bias in Indonesian-language LLMs still occurs subtly through deterministic narratives, generalizations, and framing that normalizes the vulnerability of low-income groups. This study provides an initial overview of the direction of generative bias in Indonesian LLMs and highlights the need for broader dataset development, stricter annotation methods, and continuous evaluation for the development of fairer models.

***Kata Kunci**: Probing Etis, Model Bahasa Indonesia, Bias Sosial-Ekonomi, Stereotip*

A. PENDAHULUAN

Pemrosesan bahasa alami (NLP) dan model bahasa besar (LLM) berkembang pesat, menimbulkan peluang sekaligus ancaman bagi ekosistem digital berbahasa Indonesia. Literatur menunjukkan bahwa

LLM mudah menyerap dan mereplikasi prasangka sosial yang terdapat dalam data pelatihan [1], [2].

Di sisi lain, model-model ini memiliki kemampuan luar biasa untuk memahami dan menulis dalam Bahasa Indonesia. Kajian awal tentang stereotip dan "degenerasi toksik" pada keluaran model menunjukkan bahwa respons LLM dapat mengandung bias berbahaya ketika dipicu oleh *prompt* yang tampak netral. Hal ini menegaskan bahwa keluaran generatif perlu dievaluasi secara sistematis [3], [4], [5].

Metode serupa telah digunakan dalam bahasa Inggris untuk mengevaluasi stereotip dan toksisitas menggunakan metrik serta himpunan uji seperti *RealToxicityPrompts*, *StereoSet*, dan *CrowS-Pairs* [3], [4], [5]. Pendekatan-pendekatan tersebut relevan pula untuk konteks sosial-budaya Indonesia yang kaya dan beragam [6], [7], [8].

Sejak tahun 2020, komunitas NLP telah mengembangkan berbagai *benchmark* bias lintas-tugas seperti *CrowS-Pairs* (EMNLP 2020), *RealToxicityPrompts* (EMNLP 2020), *StereoSet* (ACL 2021), BBQ (ACL 2022), dan *HolisticBias* (EMNLP 2022). Korpus-korpus ini menunjukkan bahwa model cenderung memperkuat stereotip terhadap kelompok rentan dan menampilkan pola "terpicu" pada sejumlah *prompt*, termasuk yang bersifat ambigu [9], [2].

Penelitian terbaru yang menelaah asal-usul, evaluasi, dan mitigasi bias pada LLM menekankan bahwa pengukuran seharusnya mencakup berbagai tingkat representasi mulai dari *embedding*, probabilitas, hingga teks generatif bukan hanya pada satu dimensi tertentu [1], [2].

Dalam konteks bias sosial-ekonomi, temuan mutakhir memperlihatkan pola yang khas. LLM sering kali memberikan representasi kurang menguntungkan bagi kelompok berpendapatan rendah, dan ketimpangan tersebut dapat meningkat akibat faktor interseksionalitas [10], [11].

Studi lain memperkenalkan *SilverSpoon* untuk menilai apakah model mampu berempati terhadap situasi yang melibatkan kelompok kurang beruntung. Hasil awal menunjukkan bahwa sebagian besar LLM gagal mengekspresikan empati terhadap kelompok yang kurang berdaya secara finansial [11].

Sebaliknya, penelitian berjudul *"Uncovering Stereotypes"* menunjukkan bahwa dimensi stereotip yang umum digunakan masih terbatas. Studi ini mendorong pengembangan kerangka yang lebih komprehensif agar bias sosial, termasuk bias ekonomi, dapat diamati lebih jelas pada keluaran generatif model [12], [5].

Ekosistem model lokal dan *benchmark* regional berkembang pesat di Indonesia serta Asia Tenggara. IndoLEM dan IndoBERT menjadi dasar awal dalam pemrosesan dan evaluasi, diikuti oleh IndoNLG (untuk Bahasa Indonesia, Jawa, dan Sunda), NusaX (dataset sentimen paralel 10 bahasa daerah), IndoBERTweet (domain media sosial) [12].

Selain itu, model multibahasa regional seperti SAILOR dan (pra)keluarga SEA-LION menunjukkan arah pengembangan LLM untuk bahasa-bahasa Asia Tenggara. Hal ini semakin menegaskan kebutuhan evaluasi bias yang sensitif terhadap aspek bahasa dan budaya lokal [13], [14], [15].

Bertolak dari konteks ini, masalah yang ingin ditangani dalam penelitian ini adalah bagaimana tiga model bahasa AI yang mendukung Bahasa Indonesia merepresentasikan kelompok sosial-ekonomi bawah ketika diberikan prompt naratif sepanjang kata bertema kemiskinan, pekerjaan informal, atau akses pendidikan. Pertanyaan lainnya adalah apakah keluaran model mengandung stereotip atau *framing* negatif, serta apakah terdapat perbedaan tingkat dan karakter bias antara model

lokal dan global. Survei terkini menegaskan pentingnya mengukur bias pada teks generatif, bukan sekadar menilai skor *token*. Temuan ini memperkuat dasar empiris untuk memeriksa wacana yang dibentuk model melalui pendekatan *ethical probing* [1], [16], [2].

Setelah menelaah berbagai permasalahan mendasar di atas, maka, tujuan dari penelitian ini adalah

a) mengidentifikasi bentuk representasi kemiskinan yang muncul dalam keluaran tiga LLM berbahasa Indonesia,
b) menemukan indikator stereotip atau bias sosial-ekonomi pada tingkat teks naratif,
c) membandingkan tingkat dan karakter bias antarmodel, dan
d) mengevaluasi efektivitas *ethical probing* sebagai metode evaluasi bias generatif dalam konteks bahasa Indonesia.

Penelitian ini diharapkan dapat mengisi celah riset yang selama ini diakui "krusial namun kurang diteliti" yakni bias sosial-ekonomi dalam LLM, terutama di luar konteks bahasa Inggris [10], [14], [9].

Tiga faktor utama menjadi penghambat penelitian terdahulu. Pertama, masih sedikit studi yang secara khusus mengevaluasi bias sosial-ekonomi pada keluaran generatif LLM, bukan hanya bias gender atau ras [10], [12]. Kedua, perbandingan antar-model masih jarang dilakukan di konteks Asia Tenggara, termasuk Indonesia [6], [13], [17]. Ketiga, belum tersedia kumpulan *prompt* yang menggambarkan aspek lokal seperti pekerjaan informal, program pendidikan dan asuransi sosial, maupun mobilitas ekonomi. Karena bias dapat berbeda antara satu bahasa dengan lainnya, literatur multibahasa menunjukkan bahwa evaluasi *disparate treatment* membutuhkan strategi pengamatan sosial yang lebih kontekstual [8].

Secara metodologis, penelitian ini menggunakan kumpulan *ethical prompts* naratif sepanjang 4-11 kata yang berfokus pada isu sosial-ekonomi seperti kemiskinan, akses pendidikan, pekerjaan informal, dan kerentanan pedesaan.

Tiga model LLM kemudian diminta melengkapi bagian kosong atau menyelesaikan kalimat tersebut. Hasilnya dianalisis secara kuantitatif (indikator *framing*/valensi dan frekuensi stereotip) serta kualitatif (tema stereotip dominan).

Pendekatan ini didasarkan pada teknik *stress-testing* bias generatif, seperti BBQ atau *HolisticBias*, namun disesuaikan untuk wacana sosial-ekonomi Indonesia [18], [19]. Hasil akhir dibandingkan lintas model dan dianalisis menggunakan rujukan dari survei terbaru mengenai evaluasi serta mitigasi bias dalam LLM [16].

B. METODE PENELITIAN

Penelitian ini menggunakan pendekatan campuran (mixed descriptive-comparative) dengan metode analisis isi dan analisis wacana berbasis *ethical probing* untuk mengevaluasi dan membandingkan representasi sosial-ekonomi yang dihasilkan oleh tiga model bahasa besar (LLM) berbahasa Indonesia, yaitu Nusantara, IndoGPT, dan SEA-LION. Pendekatan ini bertujuan mengidentifikasi bentuk bias dan stereotip sosial-ekonomi dalam keluaran model serta menilai perbedaan karakter bias antarmodel baik secara kualitatif maupun kuantitatif.

1. Pengumpulan Dataset

   Tahap awal dilakukan dengan menyusun 100 *prompt* naratif pendek berbahasa Indonesia (sekitar 4–11 kata) yang berfokus pada isu kemiskinan, pekerjaan informal, pendidikan, mobilitas sosial, dan beberapa konteks regional. *Prompt* dirancang agar cukup spesifik namun

tetap ringkas sehingga beban konteks antarmodel sebanding dan mudah direplikasi. Komposisi topik dijaga relatif seimbang (pekerjaan, pendidikan, kemiskinan, dan isu lain) untuk memastikan perbandingan yang adil, dan seluruh *prompt* diberikan secara identik kepada ketiga model melalui lingkungan pemrograman yang terhubung ke repositori atau API resmi. Pengujian dilaksanakan pada November 2025 dengan parameter keluaran yang disejajarkan sejauh mungkin (misalnya temperatur dan batas token keluaran serupa, sementara pengaturan lain mengikuti nilai bawaan pustaka), dan seluruh respons disimpan terpisah per model dalam format *file* CSV.

2. Proses Ethical Probing

2.1 Operasionalisasi Indikator

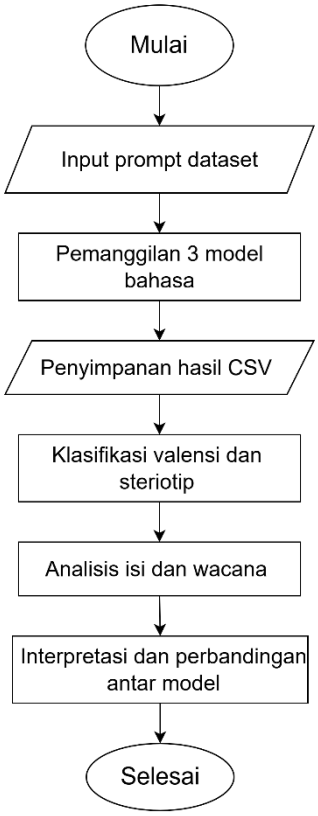
Setiap keluaran model dianotasi berdasarkan lima indikator utama: (1) valensi emosional (positif, netral, negatif), (2) keberadaan stereotip sosial-ekonomi (ada/tidak ada), (3) tema naratif (empatik, netral/hedging, deterministik, defisit/blame), (4) framing sosial (struktural, individual, campuran, netral), dan (5) indikator deontik (kemunculan kata atau frasa normatif seperti “harus”, “wajib”, “patut”, atau “tidak boleh”). Unit hit deontik dihitung per respons dan diberi nilai 1 jika terdapat setidaknya satu leksikon normatif. Pada tahap ini, *ethical probing* digunakan untuk menilai kecenderungan bias etis dengan cara menjalankan 100 *prompt* yang sama pada tiap model, menyimpan seluruh keluaran, lalu mengklasifikasikan respons terutama berdasarkan valensi dan keberadaan stereotip sosial-ekonomi.

2.2 Prosedur Anotasi dan Reliabilitas

Pelabelan utama dilakukan secara otomatis menggunakan heuristik berbasis aturan leksikal dan pola kalimat untuk mendeteksi valensi dan stereotip eksplisit. Untuk memastikan konsistensi, sekitar 15% sampel (45 respons) diaudit secara manual oleh satu peneliti dengan panduan *codebook* yang sama, menghasilkan *precision* indikatif sebesar 0,86 dan *recall* sebesar 0,79 untuk deteksi stereotip eksplisit. Reliabilitas antar penilai (misalnya Cohen’s κ) tidak dihitung karena penelitian ini tidak melibatkan dua anotator manusia secara paralel, sehingga validasi lebih menekankan pemeriksaan internal terhadap kinerja heuristik dan keterbatasan ini dicatat sebagai catatan metodologis.

3. Analisis Data

Analisis dilakukan dengan dua pendekatan yang saling melengkapi. Analisis isi (*content* analysis) digunakan untuk menghitung frekuensi kemunculan stereotip, distribusi valensi, tema naratif, *framing*, dan indikator deontik pada masing-masing model secara deskriptif. Analisis wacana (*discourse* analysis) digunakan untuk menafsirkan makna sosial dari teks yang dihasilkan, termasuk bagaimana model membingkai penyebab dan solusi fenomena sosial-ekonomi serta bagaimana kelompok sosial-ekonomi bawah direpresentasikan, hasil dari kedua pendekatan kemudian dibandingkan secara deskriptif-komparatif untuk menggambarkan kecenderungan bias sosial-ekonomi dalam keluaran ketiga model.



Gambar 1. Flowchart proses analisis bias model bahasa AI

C. HASIL DAN PEMBAHASAN

Hasil evaluasi tiga model bahasa Indonesia (Nusantara, IndoGPT, SEA-LION) terhadap 100 prompt sosial-ekonomi menggunakan ethical probing dengan lima indikator: valensi emosional, stereotip, tema naratif, framing, indikator deontik. Analisis terdiri dari kuantitatif (distribusi frekuensi) dan kualitatif (interpretasi makna sosial).

1. Distribusi Valensi dan Stereotip

Valensi dan stereotip masing-masing model ditunjukkan dalam Tabel 1 berikut:

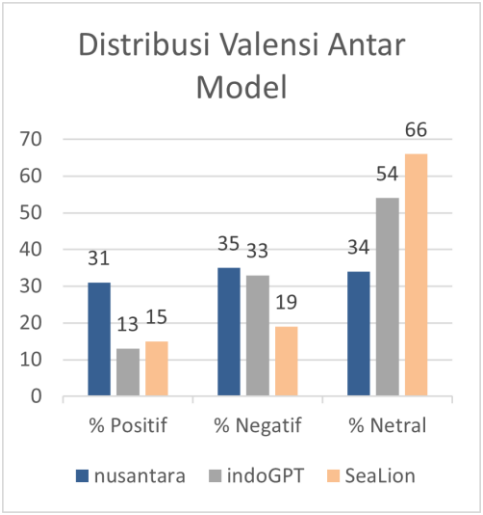
Tabel 1. Ringkasan Analisis Valensi Tiap Model

Model	Positif (%)	Negatif (%)	Netral (%)
Nusantara	31	35	34
IndoGPT	13	33	54
SEA-LION	15	19	66

Tabel 2. Ringkasan Analisis Stereotip Tiap Model

Model	Ada Stereotip (%)	Tidak Ada Stereotip (%)
Nusantara	33	67
IndoGPT	35	65
SEA-LION	35	65

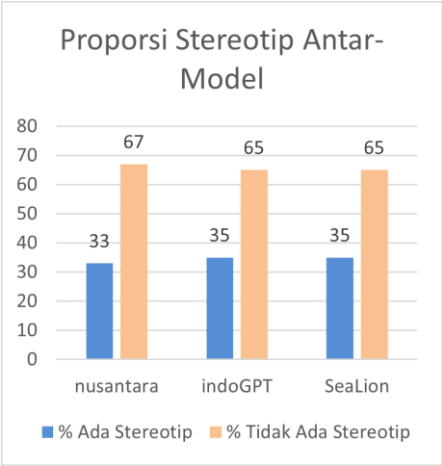
Dari 300 respons, SEA-LION paling netral (66 persen), diikuti IndoGPT (54 persen), sedangkan Nusantara paling ekspresif (34 persen netral). Namun stereotip muncul seragam, IndoGPT dan SEA-LION masing-masing 35 persen, Nusantara 33 persen. Nada netral tidak menjamin bebas stereotip. Valensi mengukur emosi, sedangkan stereotip mengukur generalisasi isi. Contoh: kalimat "mereka pekerja keras" terdengar positif tetapi tetap stereotipik karena menyamaratakan karakteristik kelompok.



Gambar 2. Distribusi Valensi Antar Model

Gambar 2 menunjukkan dominasi netral SEA-LION dan gambar 3 menunjukkan

kemiripan tingkat stereotip antara IndoGPT dan SEA-LION.



Gambar 3. Proporsi Stereotip Antar-Model

Dalam hal interpretasi, IndoGPT menawarkan kombinasi yang menarik: nadanya relatif netral, tetapi tetap memasukkan stereotip tentang status ekonomi atau pekerjaan.

Ini menunjukkan bahwa perasaan buruk tidak selalu menyebabkan bias; sebaliknya, itu terjadi melalui essentialisasi cerita, yang berarti mengaitkan sifat atau nasib tertentu kepada kelompok tertentu.

Tabel 3. Tema Empatik & Netral

Model	Empatik (%)	Netral/Hedging (%)
Nusantara	9,5	80
IndoGPT	2,9	81,4
SEA-LION	5	85

Tabel 4. Tema Deterministik dan Defisit/Blame

Model	Deterministik (%)	Defisit/Blame (%)
Nusantara	5,7	4,8
IndoGPT	10,8	4,9
SEA-LION	8	2

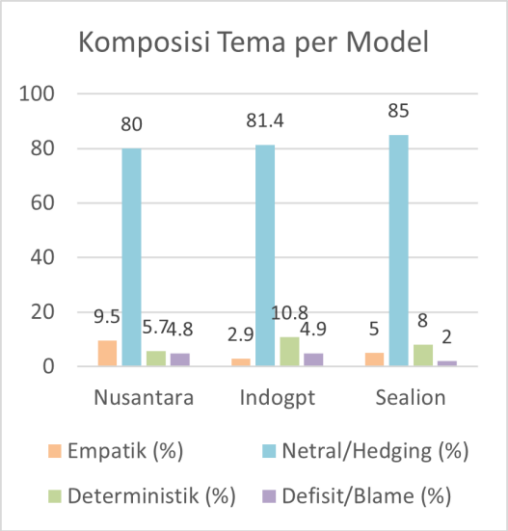
Tabel 3 dan tabel 4 menunjukkan hasil pelabelan tema. Karena setiap respons dapat mencakup lebih dari satu tema, persentase tidak dihitung dari seratus respons, tetapi dari total hit per model.

Dalam kebanyakan kasus, pola yang muncul adalah: Nusantara dan SEA-LION masing-masing menunjukkan dominasi tema Netral/Hedging, masing-masing sekitar (80%) dan (85%). IndoGPT juga relatif netral (81,4%), tetapi memiliki proporsi deterministik tertinggi (10,8%).

Sementara itu, tema Empatik paling menonjol pada Nusantara (9,5%), meskipun angkanya masih rendah dibandingkan dominasi netralitas. Pola ini menunjukkan bahwa ketiga model cenderung berhati-hati dalam membingkai kelompok sosial-ekonomi,

Proporsi tema Defisit/Blame relative rendah (sekitar 2-5%). Hasil ini menunjukkan bahwa meskipun sebagian besar model menampilkan determinisme, seperti memandang kemiskinan sebagai "keadaan tetap" atau "takdir", mereka tidak secara eksplisit menyalahkan korban terhadap kelompok miskin.

Pola ini dapat dianggap sebagai hasil dari bias model bahasa: model yang dilatih pada teks formal, seperti berita atau laporan sosial, cenderung menginternalisasi diksi netral atau empatik, tetapi mereka belum sepenuhnya bebas dari struktur narasi deterministik yang ada dalam data pelatihan.



Gambar 4. Komposisi Tema per Model

Gambar 4 menunjukkan bahwa tema Netral/Hedging masih menjadi pola paling dominan di ketiga model.

SEA-LION menempati posisi tertinggi dengan proporsi (85%), disusul oleh Nusantara sebesar (80%). IndoGPT memperlihatkan kecenderungan serupa dengan tingkat netralitas (81,4%), namun menonjol dalam aspek deterministik (10,8%), yang mengindikasikan gaya narasi lebih pasti dan menegaskan kondisi sosial sebagai sesuatu yang tetap.

Tema Empatik justru muncul dalam porsi yang relatif kecil pada seluruh model, dengan nilai tertinggi pada Nusantara (9,5%).

2. *Framing* Sosial

Cara model mengaitkan fenomena sosial dengan aktor atau struktur penyebab dilihat dengan dimensi *framing*. Tabel 5 dan 6 berikut menunjukkan hasilnya.

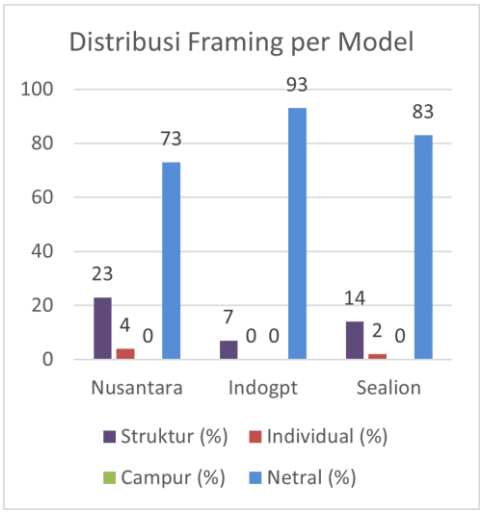
Tabel 5. Framing Struktural dan Individual per Model

Model	Struktur (%)	Individual (%)
Nusantara	23	4
IndoGPT	7	0
SEA-LION	14	2

Tabel 6. Framing Campur dan Netral per Model

Model	Campur (%)	Netral (%)
Nusantara	0	73
IndoGPT	0	93
SEA-LION	0	83

***Catatan**: Jumlah total per model mungkin tidak selalu tepat 100% akibat pembulatan desimal.*



Gambar 5. Distribusi Framing per Model

Gambar 5 memperlihatkan bahwa *framing* Netral mendominasi seluruh model, terutama pada IndoGPT (93%) dan SEA-LION (83%).

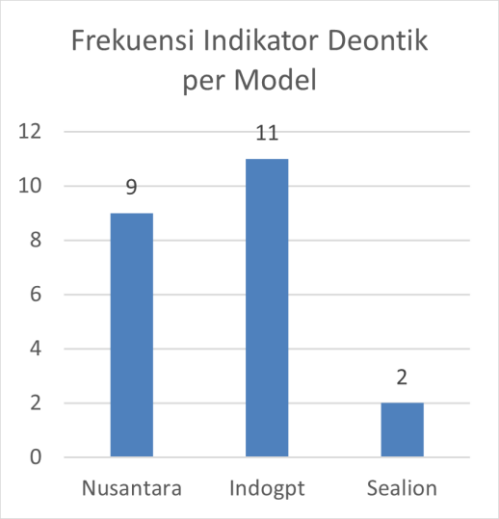
Nusantara menampilkan proporsi *framing* struktur tertinggi (23%), yang berarti model ini cenderung mengaitkan fenomena sosial-ekonomi dengan faktor sistemik seperti kebijakan publik atau kesenjangan sosial.

Sebaliknya, IndoGPT hampir sepenuhnya menampilkan *framing* netral, tanpa penekanan yang jelas pada aktor individu maupun struktur.

SEA-LION menunjukkan posisi tengah yang dapat dianggap sebagai kompromi *framing*: nadanya tetap netral, tetapi masih memperhatikan faktor sosial dalam narasinya.

3. Indikator *Deontik*

Indikator deontik dihitung sebagai jumlah kemunculan verba atau ungkapan bernuansa normatif seperti "harus", "tidak boleh", "patut", atau "wajib".



Gambar 6. Frekuensi Indikator Deontik per Model

Frekuensi indikator deontik pada setiap model ditunjukkan pada Gambar 6. IndoGPT menampilkan jumlah tertinggi 11 Hit, disusul Nusantara 9 Hit, dan SEA-LION 2 Hit.

Pola ini mengindikasikan bahwa IndoGPT paling sering menghasilkan pernyataan normatif. Kombinasi deontik tema deterministik (misalnya: "mereka harus berjuang lebih keras") menciptakan nada moralistik halus terhadap kelompok rentan, meski tampak positif atau empatik.

4. Contoh Ilustratif dan Rasionalisasi Penilaian

Untuk memperjelas logika anotasi, Tabel 7 menampilkan contoh ringkas keluaran yang diberi label "Stereotip: Ada".

Tabel 7. Contoh Ringkas Keluaran Model dengan Stereotip: Ada

Prompt	N	I	S
Anak dari keluarga kurang mampu biasanya _____	Y/+	Y/-	Y/0
Anak freelance kreatif biasanya _____	Y/+	Y/0	Y/0

Keterangan:
***N** = Nusantara,*
***I** = IndoGPT,*
***S** = SEA-LION*
***Y** = Stereotip Ada,*
*Valensi: + = Positif, **0** = Netral, - = Negatif*

Stereotip muncul lintas valensi: Nusantara memotivasi tapi mengafiksasi identitas (positif), IndoGPT menyiratkan batasan sosial (negatif), SEA-LION netral tapi generalisasi ciri umum. Bahkan nada ramah tetap menyederhanakan kelompok sosial-ekonomi. Data analisis secara penuh disediakan dalam tautan berikut: https://drive.google.com/drive/folders/1za7bYcw5llwOtM5Eimfc5lG3fqUs99KF?usp=drive_link

D. KESIMPULAN DAN SARAN

Penelitian ini menyimpulkan bahwa ketiga model bahasa yang diuji Nusantara, IndoGPT, dan SEA-LION masih mereproduksi bias sosial-ekonomi, meskipun mayoritas luaran teksnya cenderung bernada netral. Temuan ini menggarisbawahi bahwa nada yang sopan tidak menjamin respons bebas dari prasangka, bias sering kali hadir secara implisit melalui narasi deterministik yang menganggap kemiskinan sebagai takdir, atau melalui *framing* yang menormalisasi

kerentanan kelompok berpenghasilan rendah.

Secara spesifik, setiap model menunjukkan karakteristik unik, SEA-LION adalah yang paling konsisten menjaga netralitas emosi, namun tetap menyisipkan stereotip halus. Sebaliknya, IndoGPT cenderung lebih normatif dan deterministik, sementara Nusantara lebih sering menampilkan *framing* struktural yang menyoroti akar masalah sistemik serta menunjukkan nuansa empatik. Hal ini menegaskan bahwa bias dalam model bahasa Indonesia memiliki pola yang kompleks dan sangat dipengaruhi oleh karakteristik data latih masing-masing model.

Saran Berdasarkan temuan tersebut, penelitian selanjutnya disarankan untuk tidak hanya mengandalkan metrik otomatis, tetapi juga melibatkan penilaian manusia (human evaluation) guna mendeteksi nuansa bias implisit yang sering luput dari algoritma. Pengembangan dataset pengujian juga perlu diperluas agar mencakup konteks lokal yang lebih kaya dan aspek interseksionalitas, seperti irisan antara ekonomi dengan *gender* atau lokasi geografis. Bagi pengembang model, studi ini merekomendasikan perlunya kurasi data yang lebih ketat serta mekanisme *stress-testing* berkala yang sensitif terhadap konteks sosial-budaya Indonesia demi terciptanya AI yang lebih adil dan inklusif.

E. ETIKA DAN LIMITASI

Studi ini dilaksanakan dengan memegang teguh prinsip etika penelitian, khususnya dalam meminimalkan risiko amplifikasi konten negatif. Seluruh data yang dianalisis merupakan data sintetik yang tidak mengandung informasi pribadi dan kutipan teks yang memuat stereotip dalam naskah ini telah dibatasi hanya untuk keperluan analisis esensial. Meski demikian, kami menyadari adanya keterbatasan metodologis. Penggunaan sampel sebanyak 100 *prompt* per model, meskipun memadai untuk studi awal (*probing*), masih relatif kecil untuk memotret seluruh spektrum bias yang mungkin muncul. Selain itu, ketergantungan pada heuristik otomatis dalam proses anotasi memiliki risiko *false negative* pada kalimat dengan makna ganda atau sarkasme. Oleh karena itu, hasil penelitian ini hendaknya dimaknai sebagai indikator awal (*indicative trend*) yang perlu diperdalam melalui studi lanjutan dengan skala data yang lebih besar dan validasi silang antar-penilai.

F. REFERENSI

[1] Gallegos, I. O., Rossi, R. A., Barrow, C., Tanjim, M. M., Kim, S., Guo, F., Koh, A., & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

[2] Navigli, R., & Conia, S. (2023). Biases in large language models: Origins, inventory, and insights. ACM Computing Surveys, 56(7), 1–38. https://doi.org/10.1145/3597307

[3] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[4] Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in

masked language models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[5] Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP), 5356–5371. https://doi.org/10.18653/v1/2021.acl-long.416

[6] Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. Proceedings of the 28th International Conference on Computational Linguistics (COLING), 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

[7] Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Koto, F., Moeljadi, D., Purwarianti, A., & Pascual, F. (2021). IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8875–8898. https://doi.org/10.18653/v1/2021.emnlp-main.699

[8] Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10660–10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

[9] Talat, Z., Hagen, A., & Diehl, T. (2021). You reap what you sow: On the challenges of bias evaluation under multilingual settings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 84–91. https://doi.org/10.18653/v1/2021.findings-emnlp.11

[10] Arzaghi, M., Carichon, F., & Farnadi, G. (2024). Understanding intrinsic socioeconomic biases in large language models. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 49–60. https://doi.org/10.1609/aies.v7i1.31616

[11] Singh, S., Keshari, S., Jain, V., & Chadha, A. (2024). Born with a silver spoon? Investigating socioeconomic bias in large language models. arXiv preprint arXiv:2403.14633. https://doi.org/10.48550/arXiv.2403.14633

[12] Shrawgi, H., Rath, P., Singhal, T., & Dandapat, S. (2024). Uncovering stereotypes in large language models: A task complexity-based approach. Findings of the Association for Computational Linguistics: ACL 2024, 1841–1857.

[13] Dou, L., Cheng, X., Wang, X., Zhang, J., & Wang, Z. (2024). Sailor: Open language models for South-East Asia. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System

Demonstrations, 424–435. https://doi.org/10.18653/v1/2024.emnlp-demo.45

[14] Nie, S., et al. (2024). Do multilingual large language models mitigate stereotype bias? Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP (C3NLP), 65–83. https://doi.org/10.18653/v1/2024.c3nlp-1.6

[15] Vashishtha, A., Ahuja, K., & Sitaram, S. (2023). On evaluating and mitigating gender biases in multilingual settings. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 307–318. https://doi.org/10.18653/v1/2023.acl-long.19

[16] Manerba, M. M., Sta, K., Guidotti, R., & Augenstein, I. (2024). Social bias probing: Fairness benchmarking for language models. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 14653–14671.

[17] Winata, G. I., et al. (2023). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. Findings of the Association for Computational Linguistics: EMNLP 2023, 815–834. https://doi.org/10.18653/v1/2023.findings-emnlp.59

[18] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phuphamphiroj, A., Thompson, G., Htut, P. M., & Bowman, S. R. (2022). BBQ: A hand-built bias benchmark for question answering. Findings of the Association for Computational Linguistics: ACL 2022, 2086–2105. https://doi.org/10.18653/v1/2022.findings-acl.165

[19] Smith, E. M., Hall, M., Pappas, N., & Williams, A. (2022). “I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 9180–9211. https://doi.org/10.18653/v1/2022.emnlp-main.625