A Multi-Labeled Dataset for Indonesian Discourse: Examining Toxicity, Polarization, and Demographics Information

Lucky Susanto[*,1] Musa Wijanarko[*,1] Prasetia Pratama[4] Zilu Tang[2]
Fariz Akyas[1] Traci Hong[2] Ika Idris[1] Alham Aji[3] Derry Wijaya[1,2]
[1]Monash University Indonesia [2]Boston University [3]MBZUAI
[4]Independent Researcher [*]Equal Contributor

Abstract

Online discourse is increasingly trapped in a vicious cycle where polarizing language fuels toxicity and vice versa. Identity, one of the most divisive issues in modern politics, often increases polarization. Yet, prior NLP research has mostly treated toxicity and polarization as separate problems. In Indonesia, the world's third-largest democracy, this dynamic threatens democratic discourse, particularly in online spaces. We argue that polarization and toxicity must be studied in relation to each other. To this end, we present a novel multi-label Indonesian dataset annotated for toxicity, polarization, and annotator demographic information. Benchmarking with BERT-base models and large language models (LLMs) reveals that polarization cues improve toxicity classification and vice versa. Including demographic context further enhances polarization classification performance.

1 Introduction

While ideological differences are natural in a healthy democracy, extreme polarization deepens divisions, often escalating into hostility and societal fragmentation (McCoy and Somer, 2018). In such cases, opposing groups begin to see each other as existential threats, making reconciliation increasingly difficult (Kolod et al., 2024; Milačić, 2021). At the same time, online toxicity disproportionately affects minority groups (Alexandra and Satria, 2023), leading to self-censorship (Midtbøen, 2018) and undermining public discourse, particularly in journalism (Löfgren Nilsson and Örnebring, 2020; Williams et al., 2019).

Indonesia, the third-largest democracy in the world, is home to 277 million people from diverse backgrounds (Data Commons, 2024), making it an appropriate case study. The 2024 presidential election saw intense political competition alongside a sharp rise in divisive and toxic online discussions. For example, CSIS (2022) found that in 2019, 1.35% of 800,000 online texts contained toxic language, while in 2024, AJI (2024) reported that 13.8% of 1.45 million texts were toxic, which is a tenfold increase. This surge highlights the growing toxicity in Indonesian discourse.

Previous research has explored toxicity and polarization as separate problems extensively, yet their complex relationship remains largely unstudied. This gap limits our understanding of how hostile online environments evolve. While political polarization can intensify toxicity, not all polarized discourse is toxic, and not all toxic speech is politically polarized. A dataset that captures both distinctions allows for a clearer differentiation between divisive yet civil discussions and interactions that escalate into outright hostility. To address this, **we introduce the first Indonesian dataset with multiple labels that includes toxicity, polarization, and demographic information of the annotator**[1]. This dataset serves as a foundation for investigating how these factors interact in online discourse, offering insights into the broader implications of digital polarization and toxicity.

2 Interplay Of Toxicity, Political Polarization, and Identity

Online discourse is increasingly characterized by a vicious cycle in which polarization fuels toxic language and vice versa. Social media platforms exacerbate these dynamics by allowing unopposed expression of opinions, thereby deepening societal divisions (Romero-Rodríguez et al., 2023; Vasist et al., 2024; Schweighofer, 2018).

2.1 Toxicity and Polarization

Toxicity is defined as rude, disrespectful, or unreasonable language that manifests as insults, harassment, hate speech, or other abusive commu-

[1]Dataset and Toxicity Code Experiment available at https://huggingface.co/datasets/Exqrch/IndoDiscourse

Findings of the Association for Computational Linguistics: ACL 2025, pages 18863–18890

Dataset	Entry	Language	Toxic	Polar	Identity
Ours	**28K**	**Indonesian**	✓	✓	✓
Davidson et al. (2017)	25K	English	✓	✗	✗
Moon et al. (2020)	9K	Korean	✓	✗	✓
Vorakitphan et al. (2020)	67K[a]	English	✗	✓	✗
Kumar et al. (2021)	107K	English	✓	✗	✓
Sinno et al. (2022)	1K[p]	English	✗	✓	✗
Szwoch et al. (2022)	16k[a]	Polish	✗	✓	✗
Hoang et al. (2023)	11K	Vietnamese	✓	✗	✓
Lima et al. (2024)	6M*	Brazilian Portuguese	✓	✗	✗

Table 1: Comparison of Datasets. Unless specified, entry counts are at the sentence/comment level. The superscripts [a] and [p] denote the **"Article"** and **"Paragraphs"** level data, respectively. Lima et al. (2024) utilizes Perspective API (cjadams et al., 2017) for automatic labeling.

nication intended to harm or disrupt communities (Jigsaw and Google, 2017). In contrast, **polarization** refers to the degree of divergence in opinions between groups on substantive issues (DiMaggio et al., 1996).

Specifically for polarization, recent work has shifted focus from ideological to identity-based polarization (Schweighofer, 2018). Political polarization creates a divide in the population between political groups on either side of the political orientation spectrum (Weber et al., 2021). Polarizing messages, driven to reinforce inter-group biases and invoke a strong in-group identity, occasionally take the form of toxicity, as defined by Donohue and Hamilton (2022). While the converse is also true (see Appendix C), the two phenomena remain distinct.

2.2 Non-Toxic Polarization

Diverse opinions are essential to democracy (Powell, 2022). Yet, without a willingness to compromise (Axelrod et al., 2021), even civil exchanges can generate polarization. This nontoxic polarization can erode the common ground (DiMaggio et al., 1996), foster echo chambers (HOBOLT et al., 2024), and normalize extreme positions (Turner and Smaldino, 2018).

2.3 How Identities Shape Discourse Dynamics

Identity plays a pivotal role in shaping online discourse by influencing opinion formation and interaction patterns. Research shows that identity issues are among the strongest drivers of polarization (Milačić, 2021). In diverse societies, variations in cultural, social, and political identities can intensify divisions. Initially, exposure to diversity can reduce both in-group and out-group trust (Putnam, 2007), affecting constructive dialogue. Moreover, heightened polarization is often linked with increased online toxicity, frequently directed at vulnerable and minority groups (Alexandra and Satria, 2023). However, Putnam (2007) also states that sustained outer-group interaction beyond a critical threshold can foster inclusive encompassing identities and potentially mitigate polarization.

In summary, the interplay between toxicity, polarization, and demographic identities remains a critical yet understudied aspect of online discourse. By integrating demographic factors into our analysis, we aim to provide a nuanced understanding of how identities shape discourse dynamics and develop targeted strategies for mitigating both polarization and toxicity in digital environments.

3 Available Datasets

Prior polarization datasets are largely US-centric (KhudaBukhsh et al., 2021; Sinno et al., 2022), although some have addressed other contexts, such as Brexit (Vorakitphan et al., 2020) and Poland (Szwoch et al., 2022) (see Table 1). Meanwhile, toxicity detection is a more popular and mature field, where datasets vary in labeling schema—ranging from continuous scales (Kumar et al., 2021) to discrete classes like *Hate*, *Offensive*, and *Neither* (Davidson et al., 2017). More recent efforts include low-resource languages such as Brazilian Portuguese (Lima et al., 2024), Vietnamese (Hoang et al., 2023), and Korean (Moon et al., 2020). However, existing datasets rarely annotate both toxicity and polarization. Our dataset is the first to offer multi-label annotations for both phenomena in a non-Western language.

4 Dataset Creation

4.1 Annotation Instrument

To help annotators identify texts containing toxicity and/or polarization, whether explicit (e.g., direct

Demographic	Group	Count
Disability	With Disability	3
	No Disability	26
Ethnicity	Chinese-descent	3
	Indigeneous	25
	Other	1
Religion	Islam	18
	Christian or Catholics	4
	Hinduism or Buddhism	4
	Ahmadiyya or Shia	2
	Traditional Beliefs	1
Gender	Male	13
	Female	16
Age	18 - 24	9
	25 - 34	8
	35 - 44	9
	45 - 54	2
	55+	1
Education	PhD Degree	1
	Master's Degree	6
	Bachelor's Degree	12
	Associate's Degree	2
	High School Degree	8
Job Status	Employed	18
	College Student	8
	Unemployed	3
Domicile	Greater Jakarta	10
	Sumatera	7
	Bandung Area	4
	Javanese-Region	2
	Other	6
Presidential Vote	Candidate no. 1	9
	Candidate no. 2	9
	Candidate no. 3	8
	Unknown or Abstain	3

Table 2: The demographic background of the 29 annotators in coarser granularity. The ethnicity demographic information that we have are more fine-grained where *Indigenous* group here refers to several ethnic Indonesian groups: Java, Minang, Sunda, Bali, Dayak, Bugis, etc. with 1-2 annotators per ethnicity.

insults) or implicit (e.g., sarcasm) (Krippendorff, 2018), we developed an annotation instrument. Based on literature review and consultations with representatives from vulnerable communities, we designed a comprehensive codebook (see Appendix B) that explains definitions and guide the coders in detecting both toxic (Sellars, 2016, p.25–30) and polarizing content (Donohue and Hamilton, 2022; Weber et al., 2021). This instrument addresses the nuanced, context-dependent expressions of toxicity, an aspect that remains underexplored in prior NLP research (ElSherief et al., 2021).

4.2 Data Collection and Preprocessing

We compile our dataset by gathering Indonesian texts from multiple social media platforms. Texts from X (formerly Twitter) were collected using Brandwatch (Brandwatch, 2021), while Facebook and Instagram were scraped via CrowdTangle (Team, 2024). In addition, we retrieved online news articles from CekFakta[2], a collaborative fact-checking initiative in Indonesia. The data, spanning from September 2023 to January 2024, was scraped using a curated list of keywords indicative of hate speech targeting vulnerable groups. These keywords were derived from literature reviews, expert consultations, and focus group discussions with community representatives (see Appendix A.1). Preprocessing involved quality filtering (removing duplicates, spam, and advertisements using keyword and regex-based filters as detailed in Appendix A.2) and excluding texts with fewer than four words. This processing resulted in an initial corpus of 42,846 texts comprising 36,550 tweets, 1,548 Facebook posts, 3,881 Instagram posts, and 867 news articles.

4.3 Recruitment and Validation Metrics

To ensure diverse perspectives, we recruited 28 annotators from varied demographic backgrounds, and one from our research team (totaling 29; see Table 2). Annotators were compensated at a rate of 1.14 million IDR per 1,000 texts. As a comparison, the average monthly wage in Indonesia is approximately 3.5 million IDR (BPS-Statistics, 2024). For quality control, we employed inter-coder reliability (ICR) metrics. Although Cohen's Kappa is frequently used for toxicity annotations (Aldreabi and Blackburn, 2024; Ayele et al., 2023; Vo et al., 2024), we opted for Gwet's AC1 due to its robustness in the presence of class imbalance (Ohyama, 2021; Wongpakaran et al., 2013), which is suitable for our tasks.

4.4 Annotation Process

The annotation proceeded in two phases. During the **Training Phase**, annotators attended a comprehensive workshop on the codebook and annotated a pilot set of texts to identify toxicity (and its subtypes, such as insults, threats, profanity, identity attacks, and sexually explicit content) as well as polarized texts. Figure 1 shows that after three training sessions, annotators achieved a satisfactory Gwet's AC1 based on 250 sample texts, which is comparable to prior studies (Waseem and Hovy, 2016; Davidson et al., 2017). In the **Main Annotation Phase**, annotators were assigned texts using

[2] https://cekfakta.com

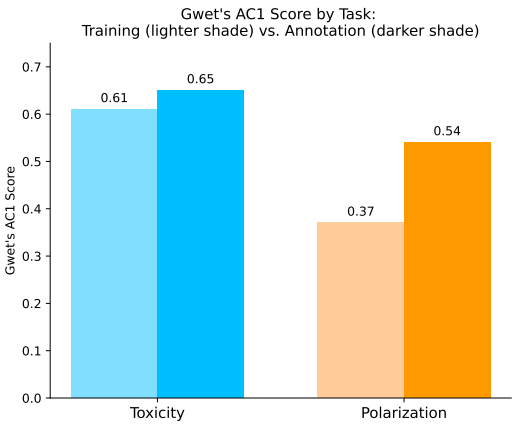

Figure 1: Gwet's AC1 Score post-training vs post-annotation.

stratified random sampling with respect to social media platform, resulting in a final annotated set of 28,477 unique texts, where a higher Gwet's AC1 score is observed. On average, each annotator contributed approximately 1,850 labels, with the note that some annotators completed only portions of their assignments due to the inherent mental burden of the task.

An AC1 value of 0.61 or higher is often considered a substantial agreement in practical contexts. Meanwhile, an AC1 value of 0.21 or higher is only considered fair. However, this threshold is arbitrary and should not replace contextual judgment (Gwet, 2014), which we provide in Appendix D.

4.5 Dataset Properties

Among the 28,477 unique texts, 55.4% were annotated by a single coder, with the specific annotator varying across entries. Figure 2 summarizes the distribution of toxicity and polarization labels aggregated via majority vote, where texts with perfect disagreement were excluded. Full breakdown is available at Appendix E.

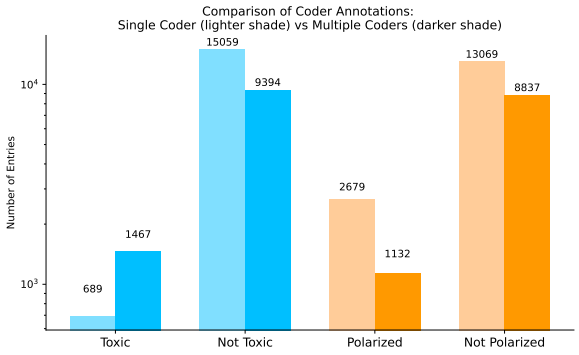

Figure 2: Dataset annotation statistics based on majority aggregation.

5 Experiment Setup and Results

Metric	Full Data	Toxic Exp	Polar Exp
Kendall-Tau	0.28	0.30	0.40
$P(t = 1 \mid p = 1)$	0.25	0.57	0.25
$P(p = 1 \mid t = 1)$	0.48	0.48	0.64
AUC: $t \rightarrow p$	0.68	0.69	0.71
AUC: $p \rightarrow t$	0.60	0.71	0.59

Table 3: Directional comparison of metrics across differing dataset splits.

Our dataset exhibits a strong imbalance toward non-toxic and non-polarized texts. To mitigate this, we balance each classification task separately by maintaining a **1:3 ratio between positive and negative instances**. Specifically, for toxicity detection, we sample[3] three non-toxic texts for every toxic text, resulting in 2,156 toxic texts after balancing in the **Toxic Exp** dataset. We sample our polarization detection data the same way, yielding 3,811 polarized texts in the **Polar Exp** dataset.

For annotation consistency, we employ a majority voting strategy, denoted by **(AGG)**: a text is labeled as toxic or polarized if more than half of the annotators agree on the label. In most cases, this rule is strictly followed, but exceptions exist, which are discussed in relevant sections. To reduce ambiguity, both **Toxic Exp** and **Polar Exp** datasets exclude texts where annotators exhibit perfect disagreement (i.e., cases where exactly half of the annotators assigned one label while the other half assigned the opposite label). Table 3 shows statistical information of the original **Full Data** and the sampled data.

5.1 Baseline

We compare transformer BERT-based models (Koto et al., 2021; Wang et al., 2024; Wongso et al., 2025) and Large Language Models (LLMs) (OpenAI et al., 2024; Aryabumi et al., 2024; Grattafiori et al., 2024; Nguyen et al., 2024), both opaque and open-sourced, for toxicity and polarization detection. BERT-based models were evaluated using stratified 5-fold cross-validation[4] where we report the averaged results, whereas LLMs were evaluated in a zero-shot setup (see Appendix J for two-shot results) without any fine-tuning. All prompts are provided in Appendix K.

[3]Utilized pandas. sample (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sample.html) with a seed of 42.

[4]Utilizing scikit-learn's package (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html), with set seed of 42.

Metric	IndoBERTweet	NusaBERT	Multi-e5	Llama3.1-8B	Aya23-8B	SeaLLMs-7B	GPT-4o	GPT-4o-mini
Toxicity Detection								
Accuracy	**.844 ± .008**	.841 ± .005	.834 ± .007	.646	.750	.512	.829	.819
Macro F1	**.791 ± .011**	.779 ± .006	.776 ± .011	.631	.429	.505	.776	.775
Precision@1	.692 ± .022	**.704 ± .018**	.675 ± .015	.405	.000	.311	.649	.613
Recall@1	.681 ± .037	.627 ± .013	.650 ± .028	**.892**	.000	.781	.688	.750
ROC AUC	**.790 ± .015**	.769 ± .006	.773 ± .013	–	–	–	–	–
Polarization Detection								
Accuracy	.801 ± .009	**.804 ± .010**	.800 ± .009	.440	.750	.750	.555	.542
Macro F1	.731 ± .013	.732 ± .016	**.735 ± .011**	.440	.429	.411	.553	.540
Precision@1	.608 ± .019	**.615 ± .019**	.597 ± .018	.302	.000	.268	.356	.347
Recall@1	.579 ± .027	.574 ± .038	**.612 ± .025**	.942	.000	.781	.968	.946
ROC AUC	.727 ± .014	.727 ± .018	**.737 ± .012**	–	–	–	–	–

Table 4: Baseline model performance on toxicity and polarization detection across various models. **ROC AUC** scores are not available for LLMs.

For open-sourced models (non-GPT-4o family), we follow their respective open source licenses as available from their respective hugging-face webpage. GPT-4o (OpenAI et al., 2024) usage is subject to OpenAI's API terms. Table 4 shows that IndoBERTweet (Koto et al., 2021) performs the best when averaged among other BERT-based models, although Multi-e5 (Wang et al., 2024) slightly outperforms it in polarization detection. Meanwhile, GPT-4o and GPT-4o-mini have the highest performance among LLMs for both tasks.

Although GPT-4o and GPT-4o-mini (OpenAI et al., 2024) perform well in toxic text detection, their performance drops significantly in polarization detection, indicating that polarization detection is a harder task than toxicity detection. Notably, Aya23-8B (Aryabumi et al., 2024) classifies all texts as non-toxic and non-polarized.

This discrepancy suggests that polarization detection is more challenging than toxicity detection. A possible explanation is that many models are explicitly trained to avoid generating toxic outputs, passively learning about toxicity detection, while polarization detection is largely neglected during training. Furthermore, toxicity detection benefits from extensive research and datasets, unlike polarization detection, leading to models struggling with the nuances of polarizing linguistic features.

Based on model performance, we conducted subsequent experiments only with IndoBERTweet and GPT-4o-mini. IndoBERTweet was selected for its strong reputation and the comparable performance of BERT-based models. GPT-4o-mini was preferred over GPT-4o due to negligible performance differences and significantly lower cost.

5.2 Wisdom of the Crowd

Each entry of our dataset is annotated by a varied number of coders due to our annotation process (see Figure 2). This allows us to explore the impact of coder counts when it comes to dataset creation and how it affects model performance.

Multiple-Coder Data Enhances Recall in Toxicity Detection For toxicity detection, training exclusively on single-coder data yields a conservative model characterized by high precision but low recall (see Table 5). In contrast, models trained on data annotated by multiple coders resulted in a broad-net model, achieving higher recall, albeit with a reduction in precision. **Notably**, even though the multiple-coder subset comprises less than half of the original training data, its performance is comparable to the baseline, achieving significantly higher recall (more than one standard deviation compared to baseline) despite lower precision.

Maintaining Performance with Only Single-Coder Data in Polarization Detection For polarization detection, the effects are reversed. Training on single-coder data results in a broad-net model and a marginally higher macro F1 score relative to the baseline. Conversely, training solely on multiple-coder data produces a model with substantially lower recall and diminished performance overall. **Interestingly**, when we modify the labeling rule from a majority vote **(AGG)** to an **(ANY)** rule (an entry is labeled as polarizing if at least one annotator flags it), we obtain a model that performs only slightly below the baseline, even though it only utilizes roughly one-third of the original training data.

Metric	IndoBERTweet (Baseline)	Single Coders	+Norm	Multiple Coders	+Norm	Multiple Coders (ANY)	+Norm
Toxicity Detection							
Accuracy	**.844 ± .008**	.831 ± .006	.824 ± .008	.827 ± .014	.835 ± .006	.828 ± .010	.780 ± .014
Macro F1	**.792 ± .011**	.746 ± .016	.728 ± .017	.785 ± .014	.782 ± .009	.786 ± .009	.709 ± .013
Precision@1	.692 ± .022	**.736 ± .011**	.736 ± .022	.628 ± .033	.666 ± .016	.627 ± .024	.560 ± .029
Recall@1	.681 ± .037	.507 ± .041	.463 ± .039	.767 ± .034	.686 ± .033	**.773 ± .036**	.573 ± .021
ROC AUC	.790 ± .015	.723 ± .018	.704 ± .017	.807 ± .013	.785 ± .013	**.810 ± .011**	.711 ± .010
Polarization Detection							
Accuracy	**.801 ± .009**	.796 ± .006	.793 ± .003	.787 ± .005	.781 ± .005	.767 ± .004	.778 ± .009
Macro F1	.731 ± .013	**.736 ± .008**	.723 ± .005	.674 ± .011	.636 ± .023	.706 ± .007	.702 ± .011
Precision@1	.608 ± .019	.585 ± .012	.589 ± .008	.617 ± .019	**.627 ± .010**	.528 ± .008	.559 ± .022
Recall@1	.579 ± .027	**.637 ± .019**	.577 ± .017	.395 ± .030	.304 ± .051	.625 ± .043	.547 ± .048
ROC AUC	.727 ± .014	**.743 ± .009**	.721 ± .006	.657 ± .012	.622 ± .020	.719 ± .014	.701 ± .015

Table 5: Performance of each setup for the "Wisdom of the Crowd" experiment on Toxicity and Polarization tasks, with and without distribution normalization **+Norm** on the training data discussed in Section 6.2.

Our findings suggest that polarization detection is inherently more subjective than toxicity detection. In a large enough annotator pool, at least one person will likely perceive a text as polarizing. This observation aligns with our dataset creation: despite efforts to standardize coder interpretations of toxicity and polarization, inter-annotator agreement for polarization is significantly lower. Consequently, models trained on polarization data with multiple annotations may struggle to generalize, as the increased annotation variability introduces more noise than informative patterns. These findings suggest that an **(AGG)** rule may not be ideal for polarization detection. **A more permissive strategy**, while not as naive as an **(ANY)** rule, **could yield better results and is worth exploring.**

5.3 Cross-task Label As A Feature

Each entry in our dataset contains coder annotations for both toxicity and polarization. This allows us to examine the relationship between the two by using one as a feature when predicting the other. To use the cross-task label as a feature, we average the annotations, following $\frac{\sum_{i=1}^{n} A_i}{n}$, where for an entry with n coders, we convert the i^{th} coder's annotation A_i to a binary value where "1" represents the toxic/polar text.

To incorporate these values into GPT-4o-mini, we prepend the input with the text: "Average [toxicity/polarization] value (range 0 to 1): [value]." For IndoBERTweet, we use the Indonesian translation and separate the added segment from the main input using a "[SEP]" token. As shown in Figure 3, IndoBERTweet benefits substantially from this additional information, exhibiting notable gains in both accuracy and macro F1. In contrast, GPT-4o-mini shows minimal to no change, indicating that it does not effectively utilize the provided scalar values.

These findings highlight a deeper correlation between toxicity and polarization, potentially driven by the rise of toxic and polarizing texts in online discussions. The strong performance boost in IndoBERTweet suggests that **jointly modeling these phenomena** could be a promising direction for future research.

5.4 Incorporating Demographic Information

To incorporate demographic information into our models, we first **explode** the dataset by splitting each text annotated by n coders into n separate entries, each linked to a single annotator's demographic profile. Although this creates duplicate texts, each instance is uniquely associated with its coder's attributes. Similar to the previous subsection, we prepend the annotator's information alongside the input text (see Appendix K for full information).

IndoBERTweet shows a strong reliance on demographic information. Shown in Table 6, when trained on the exploded dataset *without* demographic inputs (baseline), the model fails to distinguish between toxic or polarizing content. However, when demographic details are provided, performance improves significantly.

The best-performing setup includes *ethnicity, domicile, and religion*, achieving the highest scores across evaluation metrics. In contrast, the worst-performing setup, where the model only receives information about whether the coder is disabled, leads to the weakest results. For polarization detection, the best-performing setup also outperforms IndoBERTweet trained on the *non-exploded* dataset, suggesting that demographic information contributes meaningfully to polarization detection.

For GPT-4o-mini, however, incorporating demographic information does not significantly impact performance. We attribute this to the rar-

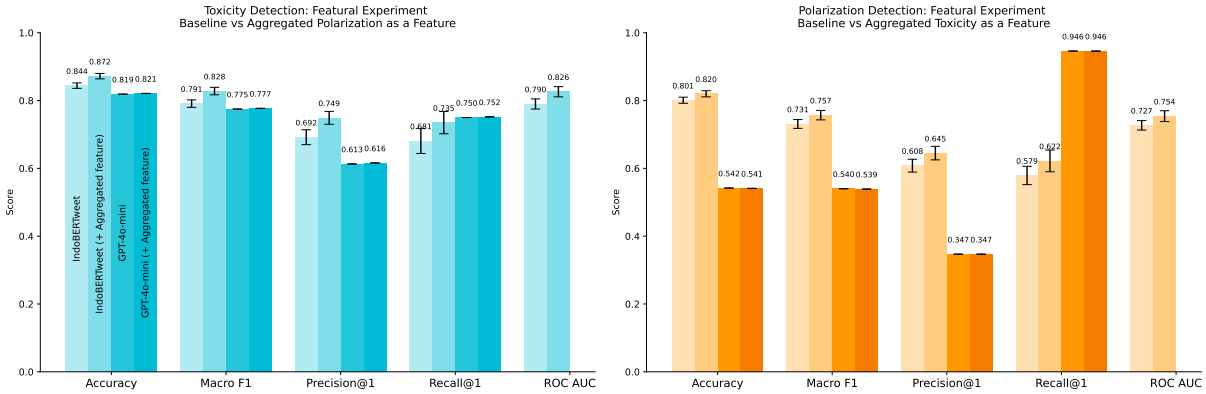

Figure 3: **Cross-task Label As A Feature (Featural):** Performance of each model under different experiment setups. From lightest to darkest shade: IndoBERTweet (Baseline), IndoBERTweet with aggregated feature, GPT-4o-mini (Baseline), and GPT-4o-mini with aggregated feature, respectively. Full table performance available at Appendix F.3

Metric	IndoBERTweet			GPT-4o-mini		
	No Demographic	Best	Worst	No Demographic	Best	Worst
Toxicity Detection						
Accuracy	.680 ± .007	**.832 ± .006**	.788 ± .011	.805	.806	.803
Macro F1	.405 ± .002	**.806 ± .004**	.757 ± .008	.789	.797	.788
Precision@1	.000 ± .000	**.744 ± .023**	.671 ± .025	.712	.686	.710
Recall@1	.000 ± .000	.728 ± .022	.671 ± .027	.753	**.833**	.751
ROC AUC	.500 ± .000	**.805 ± .003**	.757 ± .008	–	–	–
Polarization Detection						
Accuracy	.820 ± .010	**.864 ± .004**	.836 ± .005	.530	.542	.527
Macro F1	.450 ± .003	**.750 ± .008**	.687 ± .009	.529	.540	.526
Precision@1	.000 ± .000	**.655 ± .040**	.562 ± .027	.349	.352	.345
Recall@1	.000 ± .000	.525 ± .019	.407 ± .022	**.967**	.962	.966
ROC AUC	.500 ± .000	**.732 ± .007**	.669 ± .009	–	–	–

Table 6: Performance of IndoBERTweet and GPT-4o-mini with different demographic setups. **No Demographic** uses an exploded dataset with no demographic information. **Best** includes the coder's ethnicity, domicile, and religion. **Worst** (IndoBERTweet) includes whether the coder is disabled, while **Worst** (GPT-4o-mini) includes only the coder's age group.

ity of these information in its training data. Though GPT-4o has been used to simulate human users, its performance has been left wanting (Salewski et al., 2023; Choi and Li, 2024; Jiang et al., 2023). Compounded with the fact that this data is in Indonesian, it potentially ignores the provided demographic information. A notable **exception occurs in toxicity detection under the best setup**, where recall improves substantially at the cost of lower precision, even though each of these information alone does not contribute any significant changes (see Appendix F.4). However, this does not explain why GPT-4o-mini's performance remains unchanged when provided with polarization annotations for toxicity classification and vice versa. This suggests that the model may selectively prioritize certain features over others, a behavior that warrants further investigation. Additional information on GPT-4o-mini's "persona" with respect to Indonesian identities can be found in Appendix M.

5.5 Combining Featural and Demographic Information

Both featural information and demographic information improve model performance compared to their respective baseline. **By using the exploded dataset**, we investigate and find that combining both types of information leads to further improvements in IndoBERTweet, where the full results are available in Appendix F.5. We excluded GPT-4o-mini in this experiment due to its consistently unchanging performance across previous experiments.

For toxicity classification, combining featural and demographic information yields the best results, achieving an F1@1 score of 0.765, significantly higher than using only demographic (0.735) information alone. Similarly, polarization classifi-

cation benefits from this combination significantly, with F1@1 score increasing to 0.748, compared to only having demographic information (0.582). Notably, IndoBERTweet's performance on polarization classification is nearly on par with toxicity classification when both information types are provided, suggesting that the model learns a shared representation for both tasks.

Overall, these results indicate that featural and demographic information complement each other, enhancing the model's ability to detect toxic and polarizing texts more effectively than when using either information type alone.

6 Ablation and Discussion

6.1 How Related Are Polarization and Toxicity

The strongest theoretical link between toxicity and polarization manifests as toxic polarization (Milačić, 2021; Powell, 2022). Kolod et al. (2024) define toxic polarization as "a state of intense, chronic polarization marked by high levels of loyalty to a person's ingroup and contempt or even hate for outgroups." This state deepens societal divisions, making it evident that some polarizing texts in our dataset are also toxic.

From this work, Table 3 and Experiment 5.2 also demonstrate that toxicity can aid in predicting polarization and vice versa, thereby confirming the existence of a relationship. Table 3 further shows that using logistic regression to predict toxicity solely from the polarization label yields an AUC-ROC score exceeding 0.68 in all splits, although the results for polarization are more variable. This finding indicates that incorporating polarization as a feature for toxicity detection is more advantageous than the converse.

Notably, approximately 48% of toxic texts during Indonesia's 2024 Presidential Election were used for polarizing purposes. Given that only 25% of polarizing texts are toxic, our dataset suggests that **Indonesia is becoming polarized at a faster rate than it is becoming toxic**. This trend is particularly alarming, as Indonesia, the world's third-largest democracy, has not only seen a tenfold increase in toxicity since 2019, but also a significant portion of this toxicity may be linked to toxic polarization

6.2 Wisdom of the Crowd on Normalized Distribution

We confirmed that the pattern observed in Result 5.2 is not due to distribution shifts between entries annotated by one coder and those annotated by multiple coders. This was verified by normalizing the distribution (via up-sampling or down-sampling as appropriate) to maintain a consistent class ratio of one "toxic/polarizing" entry to three "not toxic/not polarizing" entries.

Table 5 shows that, despite normalization, the original pattern persists in many cases. However, new patterns emerged in both toxicity and polarization tasks. Following normalization, both toxicity's "Multiple Coders" condition and polarization's "Multiple Coders (ANY)" condition achieved balanced precision@1 and recall@1, albeit with a lower macro F1 in each instance.

This validates the results in Table 5, indicating that polarization detection may be inherently more subjective than toxicity detection. Moreover, further analysis on whether polarization detection should adhere to the same strict dataset creation protocols as toxicity detection should be done, especially given our finding that a more permissive strategy than an **(AGG)** rule may yield better result for polarization detection.

6.3 Indonesian's Polarizing Identities

Our dataset reveals identity groups characterized by high in-group agreement and significant out-group disagreement. We define these as polarizing identities, as they contribute to pronounced social divisions, measured by the gap between in-group agreement and out-group disagreement.

Based on this definition, disability emerges as the most polarizing identity in Indonesia, with a **Gwet's AC1 agreement gap** of 0.37 for toxicity and 0.46 for polarization. The second most polarizing identity is residence in Jakarta, as annotators from Jakarta exhibit a high Gwet's AC1 agreement gap, even compared to those from other regions within Java. The third is membership in the Gen X age group, which shows a substantial agreement gap for toxicity but a polarization agreement gap of 0 relative to other age groups. Beyond these three, most identities do not exhibit strong polarization, with education level showing the lowest agreement gap for toxicity (0.01). Full results are provided in Appendix N.

Metric	IndoBERTweet (Baseline)	(AGG)	+Pred	(ANY)	+Pred
Toxicity					
Accuracy	.844 ± .008	**.872 ± .008**	.869 ± .007	.867 ± .009	.834 ± .016
Macro F1	.791 ± .011	**.828 ± .011**	.824 ± .009	.823 ± .012	.722 ± .045
Precision@1	.692 ± .022	.749 ± .019	.743 ± .023	.734 ± .024	**.856 ± .020**
Recall@1	.681 ± .037	.735 ± .033	.727 ± .034	**.735 ± .029**	.406 ± .090
ROC AUC	.790 ± .015	**.826 ± .015**	.821 ± .013	.823 ± .014	.691 ± .041
Polarization					
Accuracy	.801 ± .009	**.820 ± .009**	.811 ± .005	.808 ± .009	.808 ± .005
Macro F1	.731 ± .013	**.757 ± .014**	.716 ± .018	.742 ± .014	.713 ± .020
Precision@1	.608 ± .019	.645 ± .020	**.679 ± .017**	.620 ± .019	.666 ± .014
Recall@1	.579 ± .027	**.622 ± .032**	.468 ± .052	.602 ± .031	.470 ± .064
ROC AUC	.727 ± .014	**.754 ± .016**	.697 ± .020	.739 ± .015	.695 ± .024

Table 7: Ablation study of Featural models on Toxicity and Polarization tasks. Performance of Predictor models are available in Appendix L.

6.4 Non-ideal cases for Featural Experiments

Experiment 5.2 is done under an ideal situation **(AGG)**. A more realistic setup would include a simpler feature, such as utilizing a predictor or under a less-ideal format, such as **(ANY)** where the independent variable is featured as a binary value following $\max(A_1, A_2, ..., A_n)$. Table 7 showcases these results, showing that under the **(ANY)** rule, the model still performs better than the baseline. However, utilizing a predictor (see Appendix L) degrades the performance massively below the baseline when it comes to both precision@1 and recall@1, with **Toxic (AGG) + Pred** being the only exception.

Through ablation, we show that even under non-ideal conditions, including polarization as a feature for toxicity detection and vice versa, can be helpful. Moreover, it is plausible to create a predictor for the independent variable, removing the need for human labels. However, creating a predictor through simple methods, as done in this work, may not be adequate and is a potential area for future work.

7 Conclusion and Future Work

We present a novel multi-labeled Indonesian discourse dataset of 28,477 texts annotated for toxicity, polarization, and annotator demographics. Our analysis yields the following findings:

Polarization detection is more subjective than toxicity detection. Despite extensive training, coder agreement is significantly lower for polarization, reflecting its inherent subjectivity. However, polarization labels still enhance toxicity detection, even in non-ideal conditions (Sections 5.2, 6.4). We hypothesize that jointly modeling polarization and toxicity detection through other means, such as using soft labels instead of binary labels, may lead to a better model on both tasks.

Demographic information aids classification but is less effective than cross-task features. While demographics improve both toxicity and polarization detection, using polarization as a feature for toxicity (and vice versa) has a greater impact.

Combining demographic and cross-task features further boosts performance. This hybrid approach (Appendix F.5) improves precision@1 and recall@1, allowing polarization detection to reach performance levels comparable to toxicity detection (F1@1 = 0.748).

GPT-4o-mini does not effectively utilize demographic information. Likely due to its training data limitations, GPT-4o-mini ignores demographic attributes except in one setup, where recall improves at the cost of precision (Appendix F.4). Its inability to leverage polarization for toxicity detection (and vice versa) suggests selective feature prioritization, warranting further investigation (Appendix M).

Limitations

Our work faces several limitations, some of which reflect broader challenges in the field while others are specific to our dataset.

Unused Ambiguous Cases We did not use entries where a clear consensus was not reached. This was done to simplify the analysis in this work. However, ambiguous cases are particularly interesting, as shown by work such as Akhtar et al. (2021), because they may provide insights towards in-group vs out-group dynamic in Indonesia.

Low Inter-Coder Reliability for Polarization Detection Our dataset exhibits a relatively low ICR for polarization tasks; even after maintaining a 1:3 ratio of polar to non-polar texts, the ICR only increases to 0.39. Although this low score may partly be attributed to the inherent subjectivity of polarization judgments, as suggested by our "Wisdom of the Crowd" experiment, it also implies that the polarization labels may be noisy. Despite this, Table 3 showcase a moderate correlation between polarization and toxicity features exists, which proves beneficial in our cross-task experiments (Section 5.2).

Annotation Bias While our pool of 29 annotators is larger than that used in many non-crowdsourced toxicity datasets (Davidson et al., 2017; Moon et al., 2020; Hoang et al., 2023), Indonesia's cultural and linguistic diversity means that this number may still be insufficient to capture all perspectives, potentially introducing bias into the annotations. Although the toxicity labels reached Gwet's AC1 scores comparable to other studies, the lower reliability for polarization suggests that additional or more diverse annotators could improve consistency. **Additionally**, the same set of annotators is tasked to annotate both toxicity and polarization labels at the same time, which may lead to additional biases.

Lack of Comparable Datasets As the first dataset to label both toxicity and polarization in this context, our work lacks a comparative baseline. This novelty makes it impossible to benchmark our models against existing resources, as they simply do not exist. The development of similar datasets in the future will be essential for contextualizing and validating our results.

Ethics Statement

Balancing Risk and Benefit The creation of this dataset exposes annotators to potentially harmful texts. To avoid excessive mental strain, we intentionally extended the annotation duration to two and a half months. Individuals are preemptively warned and asked for consent during the initial recruitment process. Furthermore, annotators are permitted to quit the annotation process if they feel unable to proceed. We recognize the potential misuse of such datasets, which could include training models to generate more toxic and polarizing text. Yet, it's worth noting that even without these datasets, it is alarmingly straightforward to train a model to produce toxic content, as the source of their training data, the internet, contain many of such texts. This has been demonstrated by numerous researchers who have attempted to reduce toxic output or identify vulnerabilities in large language models (refer to Gehman et al. (2020); Wen et al. (2023)). On the other hand, the area of developing models to detect and moderate toxicity and polarizing texts, targeted at specific demographic groups is still growing, with a notable lack of available data, especially in Indonesia. Weighing these considerations, we firmly believe that the potential benefits of this type of dataset significantly outweigh the possible misuse.

Coders' Data Privacy In regard to coders' data privacy, we have ensured that all publicly available demographic information of each coder are not personally identifiable. Even with all the information combined, identifying any one of our 29 coders among the diverse 277 million population is improbable.

Responsible Use of the Dataset This dataset is made available solely for advancing research in detecting and moderating toxic and polarizing content, with a particular focus on Indonesian context. Users are expected to handle the data with sensitivity and ensure that any models or applications built upon it do not inadvertently promote harmful content or reinforce societal biases. The dataset should not be employed for surveillance, profiling, or any purpose that infringes on individual or community rights. Researchers and developers must implement robust privacy safeguards and conduct thorough impact assessments before deploying any systems based on this data. Any redistribution or modification of the dataset must preserve these ethical guidelines, and users are encouraged to document and share any additional measures taken to ensure its responsible use.

Acknowledgements

This research was supported by the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia (Indonesia–US Research Collaboration in Open Digital Technology), Monash Data Futures Institute Seed Funding, Aliansi Jurnalis Independen (AJI), and Monash University's Action Lab. The authors are solely responsible for the findings and conclusions, which do not necessarily reflect the views of the sponsors.

References

AJI. 2024. 2024 indonesian general election hate speech monitoring dashboard. https://aji.or.id/. Accessed June 14th, 2024.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *Preprint*, arXiv:2106.15896.

Esraa Aldreabi and Jeremy Blackburn. 2024. Enhancing automated hate speech detection: Addressing islamophobia and freedom of speech in online discussions. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '23, page 644–651, New York, NY, USA. Association for Computing Machinery.

Lina A. Alexandra and Alif Satria. 2023. Identifying Hate Speech Trends and Prevention in Indonesia: a Cross-Case Comparison. *Global responsibility to protect*, 15(2-3):135–176.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Robert Axelrod, Joshua J. Daymude, and Stephanie Forrest. 2021. Preventing extreme polarization of political attitudes. *Proceedings of the National Academy of Sciences*, 118(50):e2102139118.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Indonesia BPS-Statistics. 2024. Average of Net Wage/Salary - Statistical Data — bps.go.id.

Brandwatch. 2021. Brandwatch consumer intelligence. https://www.brandwatch.com/suite/consumer-intelligence/.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. *Preprint*, arXiv:2405.02501.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

CSIS. 2022. Hate speech dashboard.

Data Commons. 2024. Indonesia population data. Accessed: 2024-12-19.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Paul DiMaggio, John Evans, and Bethany Bryson. 1996. Have american's social attitudes become more polarized? *American Journal of Sociology*, 102(3):690–755.

William Donohue and Mark Hamilton. 2022. *A Framework for Understanding Polarizing Language*, 1 edition. Routledge.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *Preprint*, arXiv:2009.11462.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, and Diego Garcia-Olano et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kilem L. Gwet. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*, 4 edition. Advanced Analytics, LLC, Gaithersburg, MD. Softcover edition.

Janosch Haber, Bertie Vidgen, Matthew Chapman, Vibhor Agarwal, Roy Ka-Wei Lee, Yong Keong Yap, and Paul Röttger. 2023. Improving the detection of multilingual online attacks with rich social media data from Singapore. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12705–12721, Toronto, Canada. Association for Computational Linguistics.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023.

ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

SARA B. HOBOLT, KATHARINA LAWALL, and JAMES TILLEY. 2024. The polarizing effect of partisan echo chambers. *American Political Science Review*, 118(3):1464–1479.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Preprint*, arXiv:2206.07550.

Jigsaw and Google. 2017. Toxic comment classification challenge. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge. Accessed: 2025-04-19.

Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom Mitchell. 2021. We don't speak the same language: Interpreting polarization through machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14893–14901.

Sue Kolod, Nancy Freeman-Carroll, William Glover, Cemile Serin Gurdal, Michelle Kwintner, Tamara Lysa, Lizbeth Moses, Jhelum Podder, Hossein Raisi, Silvia Resnizky, Gordon Yanchyshyn, Alena Zhilinskaya, and Heloisa Zimmermann. 2024. Thinking labs: Political polarization and social identity. Accessed: 2024-12-19.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318. USENIX Association.

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic content detection in online social networks: a new dataset from Brazilian Reddit communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 472–482, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Monica Löfgren Nilsson and Henrik Örnebring. 2020. Journalism under threat. *Taylor and Francis*, pages 217–227.

Jennifer McCoy and Murat Somer. 2018. Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The ANNALS of the American Academy of Political and Social Science*, 681(1):234–271.

Arnfinn H Midtbøen. 2018. The making and unmaking of ethnic boundaries in the public sphere: The case of norway. *Ethnicities*, 18(3):344–362.

Filip Milačić. 2021. The negative impact of polarization on democracy. *Friedrich–Ebert-Stiftung. https://library. fes. de/pdf-files/bueros/wien/18175. pdf.*

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Tetsuji Ohyama. 2021. Statistical inference of gwet's ac1 coefficient for multiple raters and binary outcomes. *Communications in Statistics - Theory and Methods*, 50(15):3564–3572.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, and Amin Tootoonchian et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

John A. Powell. 2022. Overcoming toxic polarization: Lessons in effective bridging. *Law & Inequality*, 40(2):247.

Robert Putnam. 2007. E pluribus unum: Diversity and community in the twenty-first century – the 2006 johan skytte prize lecture. *Scandinavian Political Studies*, 30:137 – 174.

Luis Romero-Rodríguez, Bárbara Castillo-Abdul, and Pedro Cuesta-Valiño. 2023. The process of the transfer of hate speech to demonization and social polarization. *Politics and Governance*, 11(2):109–113.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Preprint*, arXiv:2305.14930.

Simon Schweighofer. 2018. *Affective, Cognitive and Social Identity Related Factors of Political Polarization*. ETH Zurich, Salzburg.

Andrew Sellars. 2016. Defining hate speech. *Social Science Research Network*.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political ideology and polarization: A multi-dimensional approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 231–243.

Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. 2022. Creation of Polish online news corpus for political polarization studies. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 86–90, Marseille, France. European Language Resources Association.

CrowdTangle Team. 2024. Crowdtangle. Facebook, Menlo Park, Californmia, United States. 1816403,1824912.

Matthew A. Turner and Paul E. Smaldino. 2018. Paths to polarization: How extreme views, miscommunication, and random chance drive opinion dynamics. *Complexity*, 2018(1):2740959.

Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2024. The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, 26(2):663–688.

Cuong Nhat Vo, Khanh Bao Huynh, Son T. Luu, and Trong-Hop Do. 2024. Exploiting hatred by targets for hate speech detection on vietnamese social media texts. *Preprint*, arXiv:2404.19252.

Vorakit Vorakitphan, Marco Guerini, Elena Cabrio, and Serena Villata. 2020. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 219–224, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

T.J. Weber, Chris Hydock, William Ding, Meryl Gardner, Pradeep Jacob, Naomi Mandel, David E. Sprott, and Eric Van Steenburg. 2021. Political polarization: Challenges, opportunities, and hope for consumer welfare, marketers, and public policy. *Journal of Public Policy & Marketing*, 40(2):184–205.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2019. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. 2013. A comparison of cohen's kappa and gwet's ac1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(1).

Wilson Wongso, David Samuel Setiawan, Steven Limcorn, and Ananto Joyoadikusumo. 2025. NusaBERT: Teaching IndoBERT to be multilingual and multicultural. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 10–26, Online. Association for Computational Linguistics.

A Data Scraping and Preprocessing

A.1 Keywords Used for Scraping

cina, china, tionghoa, chinese, cokin, cindo, chindo, shia, syiah, syia, ahmadiyya, ahmadiyah, ahmadiya, ahmadiyyah, transgender, queer, bisexual, bisex, gay, lesbian, lesbong, gangguan jiwa, gangguan mental, lgbt, eljibiti, lgbtq+, lghdtv+, katolik, khatolik, kristen, kris10, kr1st3n, buta, tuli, bisu, budek, conge, idiot, autis, orang gila, orgil, gila, gendut, cacat, odgj, zionis, israel, jewish, jew, yahudi, joo, anti-christ, anti kristus, anti christ, netanyahu, setanyahu, bangsa pengecut, is ra hell, rohingya, pengungsi, imigran, sakit jiwa, tuna netra, tuna rungu, sinting.

A.2 Keywords Used for Removing Spam Texts

#openBO, #partnerpasutri, #JudiOnline, Slot Gacor, #pijat[a-z]+, #gigolo[a-z]+, #pasutri[a-z]+, pijit sensual, #sangekberat, #viralmesum, "privasi terjamin 100%", privasi 100%, ready open, ready partner, ready pijat, ready sayang, #sangeberat, obat herbal, no minus, new produk

B Annotation Guidelines

B.1 Toxic Messages Definition

Toxic comments is a post, text, or comment that is harsh, impolite, or nonsensical, causing you to become silent and unresponsive, or that is filled with hatred and aggression, provoking feelings of disgust, anger, sadness, or humiliation, making you want to leave the discussion or give up sharing your opinion.

Profanity or Obscenity The message / sentence on social media posts contains offensive, indecent, or inappropriate in a way that goes against accepted social norms. It often involves explicit or vulgar language, graphic content, or inappropriate references. Essentially, it's a message that is likely to be considered offensive or objectionable by most people.

Threat / Incitement to Violence The message / sentence on social media posts conveys an intent to cause harm, danger, or significant distress to an individual or a group. It often includes explicit or implicit threats of violence, physical harm, intimidation, or any action that creates a sense of fear or apprehension.

Insults The message / sentence on social media posts contains offensive, disrespectful, or scornful language with the intention of belittling, offending, or hurting the feelings.

Identity Attack The message / sentence on social media posts deliberately targets and undermines a person's sense of self, identity, or personal characteristics. This can include derogatory comments, or harmful statements aimed at aspects such as one's race, gender, sexual orientation, religion, appearance, or other defining attributes.

Sexually Explicit The message / sentence on social media posts contains explicit and detailed descriptions or discussions of sexual activities, body parts, or other related content.

B.2 Polarizing Messages Definition

Polarizing Messages is a post, text, or comment with purpose to promote conflict between two or more groups of people, often by presenting a highly biased or extreme perspective on a particular topic. A polarizing messages are designed to provoke strong reactions and attract individuals with similar beliefs, while potentially alienating or opposing those with differing perspectives.

B.3 Manual Annotation

Table 8 shows the list of questions that was asked to annotators for the annotation tasks.

Annotation Form		
Q1	**Does this text appear to be random spam or lack context?**	• Yes • No
Q2	**Does this text related to Indonesian 2024 General Election?**	• Yes • No
Q3	**Does this text polarized?**	• Yes • No
Q4	**Does this text contain toxicity?** *Note*: Irrelevant toxicity or hate speech includes hate speech that is meant as a joke among friends or is not considered hate speech by the recipient. Thus, it will be coded as "No".	• Yes • No
Q5	**What is the type of toxicity?** *Note:* Checkmark one or more types. Consider the following sentences as an example: *"PDIP Provokasi Massa pendukungnya geruduk kediaman Anies"* (*"Political party PDIP incites their supporters to storm Anies' residence"*). This headline should be coded as both threat and incitement to violence.	☐ Insults ☐ Threat ☐ Profanity ☐ Identity Attack ☐ Sexually Explicit

Table 8: List of questions given to annotators for every text.

C Example of Toxic, Politically Polarizing, and Both

Toxic	Polarizing	Toxic and Polarizing
Ngibuuuulll ngiibuuulll Syiah di percaya mah bisa kelar dah... Foolssss foolssss trusting Syiah is just...	Le klian setuju ga sama ada nya Rohingya di Indonesia, apa mreka msih ada di Aceh sampe skrang Yo you guys agree with Rohingya in Indonesia, are they still in Aceh till now	Alkitab orng kristen Hanya sebuah karangn pendeta Nyata nya udah brtahun\" enggk hapal\" isi nya The Christian bible is just a fake story, in reality its been years since pastors "can't remember" its content
lgbt adalah manusia paling pengecut yg pernah ada, bahkan dirinya sendiri tidak bisa menerima, aplg org lain melawan Tuhan lgbt are the most coward human in existence, they themself can't accept, especially others that oppose God	Gara2 shopee china gak bisa jualan lg. Mau belin case hp bagus, murah dan unik susah Because of shopee, china can't sell anything. Wanted to buy a good handphone case that's cheap and unique, and it is hard.	artis2 ga terkenal mah bodoamat, klo artis2 sekaligus aktifis yg citrnya pinter tp dukung zionis ya mungkin aja lg pd lolong, but wait, im not racis If its just non-popular influences then who cares, if they are also activists who seems smart but support zionist well they are currently being stupid, but wait, I'm not racist
Tapi Israel emang anjeeeengggg sih But Israel is really such a dogggg	AHY DAN DEMOKRAT GERUDUK RUMAH ANIES BASWEDAN AHY [leader of Indonesia's democratic party] AND DEMOCRATS RAIDED ANIES BASWEDAN'S HOME	Rakyat Jawa Barat merasa nyaman dengan sikap tegas Anies - Cak Imin [presidential candidate number 1] dalam menolak pengaruh LGBT yang dianggap bertentangan dengan norma masyarakat West Java population feels comfortable with Anies - Cak Imin's harsh stance on rejecting LGBT influence who are thought to be against societal norms.
Temen gw ngaku b0tita biar dapat modusin cewek-cewek. Ternyata dia womanizer njir My friend confess he claimed he's queer to scam girls. In reality, he's a womanizer mannn	Muslim Indonesia dukung Ganjar yang tolak timnas Israel Indonesian muslims supports Ganjar [presidential candidate number 3] who rejected Israel's national [soccer] team.	Yang pasti sih cawapresnya hasil pelanggaran berat sidang etik. Alias produk cacat It is obvious that the vice presidential candidate is the result of a huge law ethic violation. Essentially defective product
Yang jual ODGJ (Orang Dengan Gen Jawa) The seller is ODGJ [should be short for: "Person with mental instability"] (Person with Javanese Genetics)	Kristen, Hindu, Islam dapat perlakuan istimewa dari pak Anies Ncep ketar-ketir Christian, Hindu, Islam all get special treatment from mr Anies, Ncep [Indonesian influencer] is panicking.	Rohingya imigran gelap, bukan pengungsi. Rohingya imigran gelap, bukan pengungsi. Rohingya are illegal immigrants, not refugees. Rohingya are illegal immigrants, not refugees.

Figure 4: Samples of Toxic, Polarizing, alongside both Toxic and Polarizing texts.

D Notes on Agreement Score

To establish a clearer understanding of what considered as a *good ICR score*, we conducted literature review on several sources. However, due to variations in measurement methods and to ensure a more robust comparison, we recalculated the ICR metric internally. However, some of the datasets only present the aggregated annotation, and as a result, we are unable to compute some of the ICR scores for these datasets. Table 9 shows us the comparison between our datasets and some other previous works, with additional information on the number of annotated texts and the number of toxicity label categories.

$$n = \frac{\frac{z^2 p(1-p)}{e^2}}{1 + \left(\frac{z^2 p(1-p)}{e^2 N}\right)}$$

Figure 5: This equation is used to calculate sample size n, where z represents the Z-score associated with the confidence level, p is the probability of a positive label, e is the margin of error, and N is the population size.

While the number of texts in our datasets may seem relatively low compared to others, Equation in the Figure 5 shows that with a population of 42,846 texts, under the assumption that 20% of the scraped texts were toxic, and setting the 95% confidence level ($\alpha = 0.05$) with a 5% margin error, we find that the minimum number of required samples to represent the population is 245 texts. This showcase that while relatively small, our sample size is statistically representative.

Dataset	details	Gwet's AC1	Fleiss Kappa
Waseem and Hovy (2016)	• #texts: 6,654 • categories: 2	0.78	0.57
Ours	• #texts: 250 • categories: 2	0.61	-
Davidson et al. (2017)	• #texts: 22,807 • categories: 3	-	0.55
Haber et al. (2023)	• #texts: 15,000 • categories: 2	-	0.31
Kumar et al. (2021)	• #texts: 107,620 • categories: 2	0.27	0.26

Table 9: The distribution of text that annotated by one or more annotators.

E Dataset Properties

E.1 Annotation Statistics

Table 10 shows more fine-grained distribution on number of texts annotated by number of annotators.

#annotators	#texts	% of total
1	15,748	55.36
2	7,907	27.79
3	2,352	8.27
4	1,755	6.17
5	21	0.07
6	215	0.76
7	1	0.0
11	26	0.09
12	2	0.01
13	150	0.53
14	1	0.0
15	146	0.51
16	2	0.01
17	97	0.34
19	25	0.09

Table 10: The distribution of text that annotated by one or more annotators.

E.2 Label Statistics

Table 11 shows more detailed toxicity and polarization label distribution under different aggregation setup, while Table 12 and Table 13 respectively shows the statistics of labeled data for toxicity types and related to election. **Any** aggregation is where an entry is labeled as positive if at least one annotator flags it, and **Consensus** aggregation is where we only consider texts with 100% agreement of annotation.

#coder(s)	aggregation setup	Toxicity			Polarization		
		#toxic	#non-toxic	Total	#polarizing	#non-polarizing	Total
1	-	689	15,059	15,748	2,679	13,069	15,748
2+	Majority	1,467	9,394	10,861	1,132	8,847	9,969
	Any	4,684	8,116	12,700	5,286	7,414	12,700
	Consensus	726	8,116	8,842	664	7,414	8,078

Table 11: Number of toxic and polarizing texts based on several aggregation setup.

#coder(s)	aggregation setup	Toxicity Types				
		#insults	#threat	#profanity	#identity-attack	#sexually-explicit
1	-	326	63	105	318	6
2+	Majority	422	25	155	455	44
	Any	2,593	1,029	1,158	2,201	241
	Consensus	188	9	57	183	8

Table 12: Number of texts per toxic types based on several aggregation setup. Keep in mind that one texts can contain multiple toxicity types.

#coder(s)	aggregation setup	Related to Election		
		#related	#not-related	Total
1	-	922	14,826	15,748
2+	Majority	1,010	10,761	11,771
	Any	2,403	10,297	12,700
	Consensus	719	10,297	11,016

Table 13: Number of texts with "Related to Election" label based on several aggregation setups.

F Full Model Performance

F.1 Baseline Experiment

Metric	IndoBERTweet	NusaBERT	Multi-e5	Llama3.1-8B	Aya23-8B	SeaLLMs-7B	GPT-4o	GPT-4o-mini
Toxicity Detection								
Accuracy	.844 ± .008	.841 ± .005	.834 ± .007	.646	.750	.512	.829	.819
Macro F1	.791 ± .011	.779 ± .006	.776 ± .011	.631	.429	.505	.776	.775
F1 (Class 0)	.896 ± .006	.896 ± .004	.890 ± .005	.705	.857	.565	.885	.875
F1 (Class 1)	.686 ± .019	.663 ± .009	.662 ± .018	.557	.000	.445	.668	.675
Precision (Class 1)	.692 ± .022	.704 ± .018	.675 ± .015	.405	.000	.311	.649	.613
Recall (Class 1)	.681 ± .037	.627 ± .013	.650 ± .028	.892	.000	.781	.688	.750
ROC AUC	.790 ± .015	.769 ± .006	.773 ± .013	–	–	–	–	–
Precision-Recall AUC	.551 ± .019	.534 ± .011	.527 ± .017	–	–	–	–	–
Polarization Detection								
Accuracy	.801 ± .009	.804 ± .010	.800 ± .009	.440	.750	.750	.555	.542
Macro F1	.731 ± .013	.732 ± .016	.735 ± .011	.440	.429	.411	.553	.540
F1 (Class 0)	.869 ± .006	.870 ± .006	.866 ± .006	.422	.857	.423	.585	.571
F1 (Class 1)	.593 ± .020	.593 ± .026	.604 ± .018	.457	.000	.399	.521	.508
Precision (Class 1)	.608 ± .019	.615 ± .019	.597 ± .018	.302	.000	.268	.356	.347
Recall (Class 1)	.579 ± .027	.574 ± .038	.612 ± .025	.942	.000	.781	.968	.946
ROC AUC	.727 ± .014	.727 ± .018	.737 ± .012	–	–	–	–	–
Precision-Recall AUC	.457 ± .017	.460 ± .022	.462 ± .016	–	–	–	–	–

Table 14: Combined model performance on toxicity and polarization detection. ROC AUC and Precision-Recall AUC scores are not available for the LLMs.

F.2 Wisdom of the Crowd Experiment

Metric	Baseline	Baseline (ANY)	Single Coder	Multiple Coders	Multiple Coders (ANY)
Toxicity Detection					
Accuracy	.844 ± .008	.769 ± .012	.831 ± .006	.827 ± .014	.828 ± .010
Macro F1	.791 ± .011	.715 ± .011	.746 ± .016	.785 ± .014	.786 ± .009
F1 (Class 0)	.896 ± .006	.839 ± .011	.893 ± .003	.880 ± .012	.880 ± .008
F1 (Class 1)	.686 ± .019	.591 ± .017	.599 ± .028	.690 ± .019	.692 ± .012
Precision (Class 1)	.692 ± .022	.532 ± .023	.736 ± .011	.628 ± .033	.627 ± .024
Recall (Class 1)	.681 ± .037	.668 ± .042	.507 ± .041	.767 ± .034	.773 ± .036
ROC AUC	.790 ± .015	.735 ± .014	.723 ± .018	.807 ± .013	.810 ± .011
Precision-Recall AUC	.551 ± .019	.438 ± .015	.496 ± .019	.539 ± .023	.541 ± .014
Polarization Detection					
Accuracy	.801 ± .009	.792 ± .006	.796 ± .006	.787 ± .005	.767 ± .004
Macro F1	.731 ± .013	.736 ± .006	.736 ± .008	.674 ± .011	.706 ± .007
F1 (Class 0)	.869 ± .006	.857 ± .006	.862 ± .004	.866 ± .003	.840 ± .004
F1 (Class 1)	.593 ± .020	.614 ± .012	.610 ± .012	.481 ± .021	.572 ± .016
Precision (Class 1)	.608 ± .019	.572 ± .013	.585 ± .012	.617 ± .019	.528 ± .008
Recall (Class 1)	.579 ± .027	.664 ± .037	.637 ± .019	.395 ± .030	.625 ± .043
ROC AUC	.727 ± .014	.749 ± .011	.743 ± .009	.657 ± .012	.719 ± .014
Precision-Recall AUC	.457 ± .017	.464 ± .009	.463 ± .011	.395 ± .012	.424 ± .011

Table 15: Performance of **IndoBERTweet** Wisdom-of-the-Crowd Setup on toxicity and polarization detection.

F.3 Cross-task Label As A Feature

Metric	IndoBERTweet	IndoBERTweet-featural	GPT-4o	GPT-4o-featural	GPT-4o-mini-featural
Toxicity Detection					
Accuracy	.844 ± .008	.872 ± .008	.829	.829	.821
Macro F1	.791 ± .011	.828 ± .011	.776	.776	.777
F1 (Class 0)	.896 ± .006	.915 ± .005	.885	.885	.876
F1 (Class 1)	.686 ± .019	.741 ± .018	.668	.667	.678
Precision (Class 1)	.692 ± .022	.749 ± .019	.649	.648	.616
Recall (Class 1)	.681 ± .037	.735 ± .033	.688	.687	.752
ROC AUC	.790 ± .015	.826 ± .015	-	-	-
Precision-Recall AUC	.551 ± .019	.617 ± .020	-	-	-
Polarization Detection					
Accuracy	.801 ± .009	.820 ± .009	.555	.553	.541
Macro F1	.731 ± .013	.757 ± .014	.553	.551	.539
F1 (Class 0)	.869 ± .006	.881 ± .006	.585	.582	.571
F1 (Class 1)	.593 ± .020	.633 ± .022	.521	.520	.508
Precision (Class 1)	.608 ± .019	.645 ± .020	.356	.355	.347
Recall (Class 1)	.579 ± .027	.622 ± .032	.968	.967	.946
ROC AUC	.727 ± .014	.754 ± .016	-	-	-

Table 16: Performance of IndoBERTweet and GPT-4o in the Featural setup for toxicity and polarization detection.

F.4 Demographical

F.4.1 IndoBERTweet

Model	Accuracy	Macro F1	F1 (Class 0)	F1 (Class 1)	Precision (Class 1)	Recall (Class 1)	ROC AUC	PR AUC
Toxicity Detection								
Age Group	.803 ± .008	.774 ± .006	.855 ± .008	.692 ± .008	.692 ± .018	.693 ± .023	.774 ± .007	.578 ± .009
Baseline	.680 ± .007	.405 ± .002	.809 ± .005	.000 ± .000	.000 ± .000	.000 ± .000	.500 ± .000	.320 ± .007
Disability	.788 ± .011	.757 ± .008	.844 ± .011	.670 ± .008	.671 ± .025	.671 ± .027	.757 ± .008	.555 ± .010
Domicile	.808 ± .007	.773 ± .008	.862 ± .006	.684 ± .015	.724 ± .020	.650 ± .040	.766 ± .013	.582 ± .005
Ethnicity	.825 ± .008	.797 ± .011	.873 ± .006	.721 ± .018	.737 ± .020	.707 ± .036	.794 ± .013	.615 ± .017
Ethnicity-Domicile-Religion	.832 ± .006	.806 ± .004	.877 ± .007	.735 ± .004	.744 ± .023	.728 ± .022	.805 ± .003	.628 ± .009
Gender	.792 ± .008	.762 ± .005	.847 ± .009	.676 ± .009	.675 ± .021	.679 ± .029	.762 ± .006	.561 ± .010
LGBT	.788 ± .010	.756 ± .008	.844 ± .010	.667 ± .011	.672 ± .021	.664 ± .032	.755 ± .009	.553 ± .009
Education	.798 ± .008	.768 ± .006	.851 ± .009	.684 ± .011	.687 ± .021	.683 ± .034	.768 ± .008	.570 ± .010
President Vote Leaning	.799 ± .008	.765 ± .005	.854 ± .008	.677 ± .008	.698 ± .019	.657 ± .026	.761 ± .006	.568 ± .007
Religion	.796 ± .010	.766 ± .008	.850 ± .009	.682 ± .009	.682 ± .023	.683 ± .023	.766 ± .008	.567 ± .011
Employment Status	.793 ± .010	.764 ± .006	.847 ± .011	.681 ± .005	.674 ± .026	.689 ± .025	.765 ± .004	.563 ± .011
Polarization Detection								
Age Group	.846 ± .005	.709 ± .004	.908 ± .004	.509 ± .008	.596 ± .025	.445 ± .008	.689 ± .003	.365 ± .015
Baseline	.820 ± .010	.450 ± .003	.901 ± .006	.000 ± .000	.000 ± .000	.000 ± .000	.500 ± .000	.180 ± .010
Disability	.836 ± .005	.687 ± .009	.903 ± .004	.472 ± .019	.562 ± .027	.407 ± .022	.669 ± .009	.336 ± .020
Domicile	.850 ± .005	.716 ± .003	.911 ± .004	.522 ± .008	.612 ± .035	.457 ± .019	.696 ± .005	.377 ± .016
Ethnicity	.857 ± .005	.738 ± .005	.915 ± .003	.561 ± .009	.632 ± .039	.506 ± .018	.721 ± .005	.408 ± .018
Ethnicity-Domicile-Religion	.864 ± .004	.750 ± .008	.919 ± .003	.582 ± .016	.655 ± .040	.525 ± .019	.732 ± .007	.429 ± .024
Gender	.838 ± .007	.695 ± .011	.904 ± .005	.487 ± .022	.566 ± .029	.429 ± .032	.678 ± .012	.346 ± .021
LGBT	.837 ± .006	.684 ± .007	.904 ± .004	.465 ± .015	.569 ± .028	.393 ± .011	.664 ± .006	.333 ± .019
Education	.844 ± .007	.707 ± .006	.907 ± .005	.507 ± .013	.588 ± .024	.448 ± .032	.689 ± .011	.362 ± .010
President Vote Leaning	.847 ± .004	.708 ± .010	.909 ± .003	.506 ± .019	.602 ± .032	.437 ± .015	.687 ± .008	.365 ± .023
Religion	.844 ± .006	.710 ± .006	.907 ± .004	.512 ± .009	.588 ± .027	.455 ± .022	.692 ± .008	.366 ± .012
Employment Status	.836 ± .009	.689 ± .012	.902 ± .006	.476 ± .022	.559 ± .009	.416 ± .036	.672 ± .015	.338 ± .013

Table 17: Performance of IndoBERTweet demographic-aware models on toxicity and polarization detection.

F.4.2 GPT-4o-mini

Model	Accuracy	Macro F1	F1 (Class 0)	F1 (Class 1)	Precision (Class 1)	Recall (Class 1)
Toxicity Detection						
Age Group	.804	.788	.846	.730	.710	.752
Baseline	.806	.790	.847	.732	.712	.753
Disability	.804	.789	.846	.731	.710	.754
Domicile	.806	.791	.848	.734	.713	.756
Ethnicity	.805	.789	.847	.731	.711	.753
Ethnicity-Domicile-Religion	.807	.797	.841	.753	.687	.834
Gender	.804	.789	.846	.731	.710	.754
LGBT	.805	.790	.847	.732	.712	.754
Education	.805	.790	.847	.732	.712	.753
President Vote Leaning	.805	.790	.847	.732	.712	.754
Religion	.804	.789	.846	.731	.711	.752
Employment Status	.806	.790	.847	.733	.712	.755
Polarization Detection						
Age Group	.527	.527	.545	.509	.346	.967
Baseline	.530	.530	.547	.513	.349	.968
Disability	.529	.528	.546	.510	.346	.967
Domicile	.534	.534	.551	.516	.352	.967
Ethnicity	.535	.534	.552	.517	.352	.968
Ethnicity-Domicile-Religion	.542	.540	.565	.516	.352	.962
Gender	.529	.528	.546	.510	.346	.967
LGBT	.535	.534	.551	.517	.353	.968
Education	.531	.531	.548	.514	.350	.968
President Vote Leaning	.528	.527	.545	.509	.346	.966
Religion	.534	.534	.551	.517	.353	.968
Employment Status	.529	.528	.546	.510	.346	.967

Table 18: Performance of GPT-4o-mini demographic-aware models on toxicity and polarization detection.

F.5 Demographic + Featural

Model	Accuracy	Macro F1	F1 (Class 0)	F1 (Class 1)	Precision (Class 1)	Recall (Class 1)	ROC AUC	PR AUC
Toxicity Detection								
IndoBERTweet	.844 ± .008	.791 ± .011	.896 ± .006	.686 ± .019	.692 ± .022	.681 ± .037	.790 ± .015	.551 ± .019
Best-featural	.872 ± .008	.828 ± .011	.915 ± .005	.741 ± .018	.749 ± .019	.735 ± .033	.826 ± .015	.617 ± .020
Best-demo only	.832 ± .006	.806 ± .004	.877 ± .007	.735 ± .004	.744 ± .023	.728 ± .022	.805 ± .003	.628 ± .009
Age Group	.818 ± .005	.790 ± .003	.867 ± .006	.714 ± .006	.720 ± .023	.710 ± .024	.790 ± .004	.604 ± .010
Baseline	.680 ± .007	.405 ± .002	.809 ± .005	.000 ± .000	.000 ± .000	.000 ± .000	.500 ± .000	.320 ± .007
Disability	.808 ± .007	.782 ± .002	.857 ± .009	.707 ± .008	.693 ± .030	.724 ± .041	.786 ± .006	.589 ± .008
Domicile	.836 ± .006	.809 ± .006	.881 ± .007	.737 ± .012	.761 ± .034	.718 ± .048	.805 ± .012	.635 ± .005
Ethnicity	.837 ± .007	.812 ± .007	.881 ± .006	.744 ± .010	.750 ± .020	.739 ± .018	.811 ± .006	.637 ± .015
Ethnicity-Domicile-Religion	.850 ± .005	.827 ± .004	.890 ± .005	.765 ± .004	.768 ± .016	.762 ± .015	.827 ± .004	.661 ± .007
Gender	.813 ± .006	.788 ± .005	.861 ± .007	.714 ± .009	.701 ± .026	.730 ± .033	.791 ± .006	.597 ± .012
LGBT	.811 ± .010	.784 ± .008	.861 ± .009	.708 ± .008	.703 ± .022	.713 ± .019	.785 ± .008	.593 ± .011
Education	.814 ± .008	.788 ± .006	.861 ± .009	.716 ± .004	.701 ± .027	.733 ± .024	.792 ± .003	.599 ± .012
President Vote Leaning	.824 ± .006	.797 ± .006	.872 ± .006	.722 ± .009	.733 ± .021	.713 ± .022	.795 ± .006	.614 ± .012
Religion	.815 ± .008	.790 ± .006	.862 ± .009	.717 ± .007	.704 ± .028	.733 ± .026	.793 ± .005	.601 ± .013
Employment Status	.811 ± .008	.786 ± .007	.859 ± .009	.713 ± .012	.694 ± .024	.735 ± .042	.791 ± .011	.594 ± .010
Polarization Detection								
IndoBERTweet	.801 ± .009	.731 ± .013	.869 ± .006	.593 ± .020	.608 ± .019	.579 ± .027	.727 ± .014	.457 ± .017
Best-featural	.820 ± .009	.757 ± .014	.881 ± .006	.633 ± .022	.645 ± .020	.622 ± .032	.754 ± .016	.496 ± .021
Best-demo only	.864 ± .004	.750 ± .008	.919 ± .003	.582 ± .016	.655 ± .040	.525 ± .019	.732 ± .007	.429 ± .024
Age Group	.818 ± .009	.760 ± .012	.877 ± .006	.643 ± .019	.656 ± .020	.632 ± .025	.757 ± .013	.510 ± .020
Baseline	.739 ± .007	.425 ± .002	.850 ± .004	.000 ± .000	.000 ± .000	.000 ± .000	.500 ± .000	.261 ± .007
Disability	.804 ± .009	.744 ± .016	.868 ± .006	.619 ± .027	.627 ± .019	.612 ± .038	.742 ± .019	.485 ± .025
Domicile	.849 ± .008	.801 ± .011	.898 ± .006	.704 ± .017	.719 ± .014	.690 ± .026	.797 ± .012	.577 ± .018
Ethnicity	.849 ± .009	.804 ± .010	.898 ± .007	.710 ± .013	.711 ± .018	.710 ± .020	.804 ± .010	.580 ± .015
Ethnicity-Domicile-Religion	.871 ± .006	.830 ± .008	.913 ± .004	.748 ± .013	.759 ± .012	.738 ± .021	.827 ± .010	.628 ± .016
Gender	.804 ± .010	.741 ± .014	.869 ± .007	.614 ± .024	.632 ± .017	.599 ± .044	.738 ± .018	.483 ± .020
LGBT	.798 ± .006	.738 ± .013	.863 ± .004	.612 ± .024	.612 ± .009	.613 ± .043	.738 ± .018	.476 ± .021
Education	.816 ± .008	.757 ± .015	.876 ± .005	.637 ± .027	.654 ± .011	.622 ± .048	.753 ± .020	.505 ± .023
President Vote Leaning	.829 ± .006	.773 ± .009	.886 ± .004	.659 ± .015	.687 ± .002	.635 ± .028	.766 ± .012	.531 ± .013
Religion	.829 ± .009	.771 ± .013	.886 ± .006	.655 ± .021	.692 ± .018	.623 ± .035	.762 ± .015	.529 ± .019
Employment Status	.806 ± .008	.746 ± .014	.869 ± .005	.624 ± .024	.630 ± .020	.618 ± .040	.745 ± .017	.489 ± .022

Table 19: Performance of IndoBERTweet-based models on toxicity and polarization detection.

G Hyperparameters and Evaluation Setup

G.1 Hyperparameters – LLM

Temperature: 1×10^{-5}

G.2 Hyperparameters – Neural-Based Models

Batch Size: 16
Optimizer: AdamW
Learning Rate: 1×10^{-3}
Weight Decay: 0.01
Betas: (0.9, 0.999)
Epsilon: 1×10^{-8}
Training Epochs: 3
Loss Function: Cross Entropy Loss

G.3 Evaluation Setup

Cross-Validation: `StratifiedKFold(n_splits=5, shuffle=True, random_state=42)` from `sklearn.model_selection`

H Dataset Statistic By Source

Source	Label	Toxic Class	Polarizing Class
Articles	0	1409	1148
	1	44	243
	Ambiguous	107	169
Facebook	0	2392	2062
	1	372	636
	Ambiguous	438	504
Instagram	0	4009	3098
	1	114	546
	Ambiguous	305	784
Twitter	0	16642	15597
	1	1626	2386
	Ambiguous	989	1274

Table 20: Distribution of Toxic and Polarizing Classes Across Platforms and Labels

I Addressing GPT-4o-mini's Zero-Shot Setting

Contextualization is a concept we did not explore extensively in this work. This section showcases that explicitly defining toxicity does not lead to a model improvement. However, defining polarization indeed increase GPT-4o-mini's performance.

I.1 Toxicity Detection Results

Metric	IndoBERTweet	GPT-4o-mini (Base)	GPT-4o-mini (Extended)
Accuracy	0.844 ± 0.008	0.819	0.823
Macro-F1	0.791 ± 0.011	0.775	0.779
F1@0	0.896 ± 0.006	0.875	0.880
F1@1	0.686 ± 0.019	0.675	0.659

Table 21: Toxicity detection performance across models and prompt variants.

Polarization Detection Results

Metric	IndoBERTweet	GPT-4o-mini (Base)	GPT-4o-mini (Extended)
Accuracy	0.801 ± 0.009	0.542	0.668
Macro-F1	0.731 ± 0.013	0.540	0.649
F1@0	0.869 ± 0.006	0.571	0.732
F1@1	0.593 ± 0.020	0.508	0.565

Table 22: Polarization detection performance across models and prompt variants.

I.2 Extended Prompt Definitions

Toxicity Definition Toxicity refers to language that is harmful, offensive, or hostile. Toxic text may include insults, threats, hate speech, or derogatory remarks targeting individuals or groups based on attributes such as race, ethnicity, gender, religion, nationality, or other identity markers. However, critical or controversial opinions that do not contain explicit harm, slurs, or personal attacks should *not* be classified as toxic.

Polarization Definition Polarization refers to text that reinforces division between opposing groups, promotes ideological extremity, or frames issues in a way that discourages compromise. Polarizing text often includes strong "us vs. them" narratives, absolute statements, or language that deepens conflict between different perspectives. However, expressing a strong opinion without dismissing or demonizing the opposing side should *not* be classified as polarizing.

J LLMs' 2-Shot Setup Performance

Toxicity Detection Performance						
Model	**Macro F1**		**Toxic F1**		**Non-Toxic F1**	
	0-shot	2-shot	0-shot	2-shot	0-shot	2-shot
GPT-4o-mini	**0.674**	0.651	**0.456**	0.439	**0.891**	0.863
Llama3.1-8B	**0.511**	0.483	**0.280**	0.262	**0.742**	0.704
SeaLLMs-7B	0.384	**0.454**	0.185	**0.236**	0.583	**0.673**
Aya23-8B	0.536	**0.607**	0.114	**0.336**	0.958	**0.878**

Table 23: Toxicity detection performance of LLMs in 0-shot and 2-shot setups. **Bolded** values highlight the better performing setup (0-shot vs 2-shot) based on the specific metric.

Polarization Detection Performance						
Model	**Macro F1**		**Polar F1**		**Non-Polar F1**	
	0-shot	2-shot	0-shot	2-shot	0-shot	2-shot
GPT-4o-mini	0.536	**0.609**	0.450	**0.512**	0.621	**0.706**
Llama3.1-8B	0.370	**0.485**	0.306	**0.357**	0.434	**0.613**
SeaLLMs-7B	0.354	**0.455**	0.441	**0.343**	0.267	**0.566**
Aya23-8B	0.466	**0.526**	0.013	**0.310**	**0.919**	0.743

Table 24: Polarization detection performance of LLMs in 0-shot and 2-shot setups.

Using a much smaller data subset (see Figure 2's **2+** data count), we conducted preliminary research. We show that for two of the highest performing LLMs (GPT-4o-mini and Llama3.1-8B), their performance degrades for toxicity detection (Table 23). Meanwhile, for polarization detection, their performance improves (Table 24). Due to this difference in behavior, we chose to prioritize the 0-shot setup instead.

K IndoBERTweet Input Setup and GPT-4o-mini Prompts List

Differing experiments require differing setup of the model's input. For IndoBERTweet, we leverage BERT's pre-training schematic and utilize the [SEP] token, following Kumar et al. (2021)'s setup. For GPT-4o-mini, we augment its input by pre-pending specific texts depending on the experiment. These augmentations are available at Table 25.

Experiment	**IndoBERTweet**	**GPT-4o-mini**
Baseline	{TEXT}	"Answer only with ["toxic"/"polarizing"] or ["not toxic"/"not polarizing"]. Is this Indonesian text [toxic/polarizing]? """ {TEXT} """
Featural	Nilai rata-rata [toksisitas/polarisasi]: {VALUE} [SEP] {TEXT}	"Answer only with ["toxic"/"polarizing"] or ["not toxic"/"not polarizing"]. Is this Indonesian text with a [toxicity/polarization] index (range of 0 to 1) of {VALUE} [toxic/polarizing]? """ {TEXT} """
Demographical	"Informasi Demografis: {DEMOGRAPHIC_CLASS_1}: {DEMOGRAPHIC_VALUE_1} ... {DEMOGRAPHIC_CLASS_n}: {DEMOGRAPHIC_VALUE_n} [SEP] {TEXT}	Answer only with ["toxic"/"polarizing"] or ["not toxic"/"not polarizing"]. You are an Indonesian citizen with the following demographic information: {DEMOGRAPHIC_CLASS_1}: {DEMOGRAPHIC_VALUE_1} ... {DEMOGRAPHIC_CLASS_n}: {DEMOGRAPHIC_VALUE_n} Is this Indonesian text [toxic/polarizing]? """ {TEXT} """
Demographical and Featural	Informasi Demografis: {DEMOGRAPHIC_CLASS_1}: {DEMOGRAPHIC_VALUE_1} ... {DEMOGRAPHIC_CLASS_n}: {DEMOGRAPHIC_VALUE_n} Nilai rata-rata [toksisitas/polarisasi]: {VALUE} [SEP] {TEXT}	"Answer only with ["toxic"/"polarizing"] or ["not toxic"/"not polarizing"]. You are an Indonesian citizen with the following demographic information: {DEMOGRAPHIC_CLASS_1}: {DEMOGRAPHIC_VALUE_1} ... {DEMOGRAPHIC_CLASS_n}: {DEMOGRAPHIC_VALUE_n} Is this Indonesian text with a [toxicity/polarization] index (range of 0 to 1) of {VALUE} [toxic/polarizing]? """ {TEXT} """

Table 25: Prompt templates for IndoBERTweet and GPT-4o-mini experiments.

L Predictor Model Performance

Performance of the predictor model on Section 6.4 visible on Table 26. **AGG** represents the independent variable as a value between [0, 1]; while **ANY** represents the independent variable as a binary value of 0 or 1. Because of this, the predictor differs per setup, where on **(AGG)** the predictor is a regressor while on **ANY** it is a classifier.

Toxicity			Polarization		
Metric	**(AGG) Pred**	**(ANY) Pred**	**Metric**	**(AGG) Pred**	**(ANY) Pred**
MSE	0.109	—	MSE	0.072	—
MAE	0.222	—	MAE	0.163	—
$F1_0$	—	0.831	$F1_0$	—	0.907
$F1_1$	—	0.649	$F1_1$	—	0.504
ROC AUC	—	0.736	ROC AUC	—	0.691

Table 26: Comparison of (AGG) and (ANY) Predictor models for Toxicity and Polarization tasks.

M GPT-4o's Persona

Table 27 and 28 present the highest ICR group score from each demographic. To compute the toxicity ICR score for a demographic group, we calculated the weighted average of Gwet's AC1 scores for every pairwise combination between GPT-4o and annotators within the respective group, using the volume of text in each pair as the weight.

demographic	**group**	**Toxicity ICR (avg)**
Ethnicity	Non-indigenous	0.751
Domicile	Greater Jakarta	0.746
Religion	Non-Islam	0.743
Disability	Yes	0.734
Age Group	Gen X	0.731
President Vote Leaning	Candidate No. 2	0.724
Education	Postgraduate Degree	0.715
Job Status	Unemployed	0.707
Gender	Female	0.694

Table 27: GPT-4o's most highest ICR score for toxicity.

demographic	**group**	**Polarized ICR (avg)**
Domicile	Javanese-Region	0.566
President Vote Leaning	Unknown	0.408
Age Group	Gen-X	0.182
Education	Postgraduate Degree	0.108
Disability	No	0.066
Ethnicity	Indigenous	0.065
Job Status	Students	0.061
Gender	Female	0.059
Religion	Islam	0.059

Table 28: GPT-4o's most highest ICR score for toxicity.

N In-group vs Out-group Agreement Gap

index	demographic	group	toxic_gwet	toxic_gwet_diff	polarize_gwet	polarize_gwet_diff	support
0	disability	no	.40	.37	.32	.46	26
1	disability	yes	.77	.37	.78	.46	3
2	general_domicile	Non-Java	.23	.25	.48	.16	6
3	general_domicile	Greater Jakarta	.59	.22	.50	.19	10
4	general_domicile	Java Region	.23	.22	.44	.03	2
5	age group	Gen X	.63	.21	.33	.00	3
6	ethnicity2	Non-Indigenous	.60	.20	.37	.05	4
7	ethnicity2	Indigenous	.40	.20	.32	.05	25
8	job status	Unemployed	.59	.18	.44	.13	3
9	president vote leaning	1	.59	.16	.43	.12	9
10	general_domicile	Sumatera	.56	.13	.43	.08	7
11	general_domicile	Bandung	.56	.13	.62	.28	4
12	religion2	Non-Islam	.52	.11	.41	.12	9
13	religion2	Islam	.41	.11	.29	.12	20
14	education	Postgraduate Degree	.51	.07	.44	.10	7
15	president vote leaning	unknown	.51	.07	.39	.05	3
16	president vote leaning	2	.50	.07	.39	.06	9
17	job status	Students	.41	.06	.29	.13	8
18	president vote leaning	3	.38	.06	.23	.15	8
19	gender	F	.44	.04	.25	.17	16
20	gender	M	.40	.04	.42	.17	13
21	job status	Employed	.44	.03	.39	.09	18
22	age group	Gen Z	.44	.02	.28	.14	12
23	age group	Millennials	.43	.02	.41	.13	14
24	education	Bachelor/Diploma	.43	.01	.41	.11	14
25	education	Highschool Degree	.45	.01	.29	.11	8

Table 29: Demographic Agreement Scores. **ethnicity2** and **religion2** denote higher-level groupings of demographic information (e.g., Christians and Buddhists are grouped as "Non-Islam").