IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language

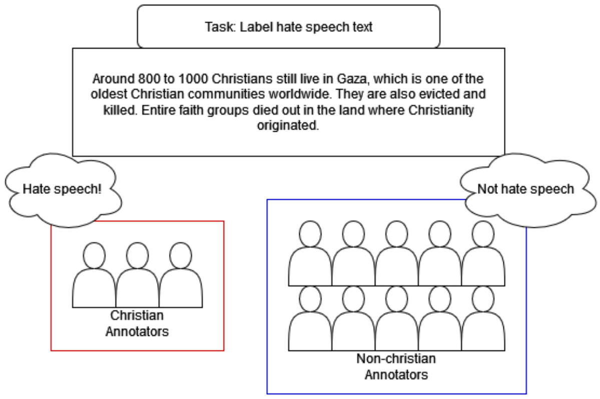
Anonymous EMNLP submission

Abstract

Hate speech poses a significant threat to social harmony. Over the past two years, Indonesia has seen a ten-fold increase in the online hate speech ratio, underscoring the urgent need for effective detection mechanisms. However, progress is hindered by the limited availability of labeled data for Indonesian texts. The condition is even worse for marginalized minorities, such as Shia, LGBTQ, and other ethnic minorities because hate speech is underreported and less understood by detection tools. Furthermore, the lack of accommodation for subjectivity in current datasets compounds this issue. To address this, we introduce IndoToxic2024, a comprehensive Indonesian hate speech and toxicity classification dataset. Comprising 43,692 entries annotated by 19 diverse individuals, the dataset focuses on texts targeting vulnerable groups in Indonesia, specifically during the hottest political event in the country: the presidential election. We establish baselines for seven binary classification tasks, achieving a macro-F1 score of 0.78 with a BERT model (IndoBERTweet) fine-tuned for hate speech classification. Furthermore, we demonstrate how incorporating demographic information can enhance the zero-shot performance of the large language model, gpt-3.5-turbo. However, we also caution that an overemphasis on demographic information can negatively impact the fine-tuned model performance due to data fragmentation.

1 Introduction

In the rapidly evolving digital landscape of Indonesia, a disturbing **ten-fold increase in hate speech** ratio has been observed in just two years (CSIS, 2022; AJI, 2024). Left alone, this surge threatens social harmony (Williams et al., 2019), and is especially harmful to minority groups (Alexandra and Satria, 2023), because it could lead to societal polarization (Unlu and Kotonen, 2024). One potential solution comes in the form of an automated hate speech detection system. However, many significant challenges to the development of this system exist. One of these challenges is the lack of comprehensive and up-to-date data. The existing Indonesian datasets (Alfina et al., 2017; Ibrohim and Budi, 2018) are considerably dated and consist only of around 3,000 labeled Indonesian texts. Furthermore, these datasets lack crucial information, such as the demographic information of annotators, which is key in training systems since subjectivity is inherent in hate speech annotations (Fleisig et al., 2024) (Figure 1). Subjectivity in the annotation of text often occurs with latent content (e.g., sarcasm) which contains "under the surface" information that requires human annotators' mental scheme to decipher their meaning. In contrast, manifest content (e.g., how many words are in the sentence) is "surface information" that requires minimal interpretation (Lombard et al., 2002). Arguably, the more pertinent information is found in latent content. However, human annotator's judgments are affected by not only their background, geography, and personal experiences (Armstrong et al., 1997) but also by their perceived status in relation to other human coders working on the research (Campbell



Around 800 to 1000 Christians still live in Gaza, which is one of the oldest Christian communities worldwide. They are also evicted and killed. Entire faith groups died out in the land where Christianity originated.

Figure 1: The perception of hate speech is influenced by the identity of a person or a group. A text may be considered hate speech by one group of people, while another group may not view it as such. More divisive example available in appendix A.

et al., 2013).

As a first step to tackle this issue, we release **IndoToxic2024**, a hate speech dataset annotated across various demographics, where each entry is accompanied by ten-dimensional demographic information. Unlike most hate speech datasets that are single-labeled, IndoToxic2024's labels are preserved per annotator. This opens up the potential for studying subjectivity in annotations. This dataset consists of 43,692 entries annotated by 19 diverse annotators, aimed at classifying hate speech and the types of toxic behaviors. It is a human-annotated collection, assembled by gathering posts from various social media platforms using keywords related to Indonesian vulnerable groups. Our contributions are three-fold:

- **Creation of IndoToxic2024 Dataset,** enabling the creation of better hate speech detection systems (IndoBERTweet fine-tuned for hate speech classification), specifically in the Indonesian language.
- **Exploring the Role of Demographic Information in Hate Speech Classification,** demonstrating that gpt-3.5-turbo's hate speech classification performance can be significantly improved when provided with the demographic information of the annotator. This insight suggests that demographic information can be a valuable addition to improve model performance.
- **Analyzing the Impact of Excessive Demographic Information on Fine-Tuned Model Performance,** providing an extensive analysis that shows an addition of demographic information can lead to fragmentation of the dataset (fewer training data in each demographic, thus worse performance).

2 Background & Related Work

2.1 What is Hate Speech?

Hate speech definitions evolved over time. Initially defined as language intended to demean others (Delgado, 1982), it has expanded to include public speech or writing inciting hatred towards demographic groups (Greenawalt, 1989; Nations, 2023). In this work, we adopt the definition by Indonesia's National Human Rights Commission, which includes any communication motivated by hatred against people based on their identities, intending to incite violence, death, and social unrest (Paramadina and Mafindo, 2023).

2.2 Detecting Hate Speech

Deep learning techniques have been promising in automatic hate speech detection (cjadams et al., 2017; Das et al., 2021). However, these techniques still fall short in real-world scenarios due to the evolving nature of hate speech and the complexity of the task. Recently, Fleisig et al. (2024) demonstrated that incorporating demographic information of annotators can improve model performance. Fleisig et al. (2024) approaches the problem as a regression instead of a classification, where texts are labeled based on the severity of their toxicity.

2.3 Available Hate Speech Datasets

Initially, Indonesian hate speech datasets (Alfina et al., 2017), consisted of texts with a binary label indicating is a hate speech or not. As the field evolved, datasets began to incorporate different levels of toxicity. For instance, the datasets by Ibrohim and Budi (2018) and Mathew et al. (2022) introduced varying degrees of toxicity, with the latter focusing on explainable hate speech classification. Following this trend, datasets started to include types of hate speech, as seen in the challenge run by cjadams et al. (2017). More recently, demographic information has been incorporated into hate speech datasets. To our knowledge, the earliest dataset to contain more than just text and labels was created by Kumar et al. (2021) and was recently utilized by Fleisig et al. (2024). Altough, it's important to note that most modern datasets focus on the English language. Aside from the private CSIS (2022) dataset, we have not found new Indonesian hate speech datasets in the past five years.

2.4 Datasets With Demographic Information

Datasets incorporating demographic information are typically utilized for tasks associated with an individual's behavior or circumstances. More often than not, this demographic data is employed in predicting aspects such as insurance premiums (Patil et al., 2024) or an individual's purchasing power (Olodo et al., 2022). However, considering the subjectivity inherent in hate speech, we only find a single hate speech dataset that includes demographic information (Kumar et al., 2021). Most hate speech datasets only go as far as providing annotator IDs, as in the case of Mathew et al. (2022), without any demographic information of the text annotator.

3 Dataset Creation

3.1 Data Collection

We obtain our text data from some of the popular social media platforms in Indonesia (Kemp, 2023) including Facebook, Instagram, and Twitter (or X). Additionally, we also retrieve articles from CekFakta[1], a movement that focuses on clarifying misinformation that spreads on the internet.

Different tools are utilized to gather data from platforms. We use Brandwatch (2021) to obtain tweets, replies, and quotes from X, Crowdtangle (Team, 2024) to obtain posts from Instagram and Facebook. Additionally, we retrieved articles from Cek Fakta[2], a fact-checking collaboration across online media and fact-checking organizations in the country. We collect data from September 2023 to January 2024, following the 2024 Indonesian presidential election timeline (detikcom, 2022), as hate speech was found to intensify in Indonesia during a similar election in 2019 (Iswatiningsih et al., 2019).

We obtain data using keywords that were previously used to express hate toward vulnerable groups in Indonesian texts, compiled based on various sources such as literature research, discussions with experts, and a focus group discussion (FGD) with representatives from vulnerable communities[3]. We provide the keywords in Appendix B. The data was then sampled for annotation by *coders* (i.e. annotators from diverse demographic backgrounds).

3.2 Recruitment and Validation Metrics

18 people from various demographic backgrounds and 1 from our research team were recruited to annotate the data. From the FGD, each of the vulnerable groups proposed their representatives to annotate the data, ensuring representations from each group. Table 1 gives us a coarse-grained overview of this diversity. Subsequently, contracts were drafted, and each annotator was compensated with 1.5 million IDR for every 1,000 texts they annotated. For comparison, an average monthly wage in Indonesia across sectors is 3.5 million IDR in 2024 (BPS-Statistics, 2024).

Inter-coder Reliability Metrics Coders were trained on a codebook and their agreement on applying the codebook on the data was determined by calculating inter-coder reliability (ICR). High ICR score means that the annotators consistently categorized the text similarly. Although Cohen's Kappa is a popular inter-coder reliability test used in several hate speech works (Aldreabi and Blackburn, 2024; Ayele et al., 2023; Vo et al., 2024), Gwet's AC1 has been suggested as a more stable metric that is robust against class imbalance (Ohyama, 2021; Wongpakaran et al., 2013). This is especially relevant for social media platforms with a high volume of data where the majority is non-hate speech.

Demographic	Group	Count
Disability	With Disability	3
	No Disability	16
Ethnicity	Chinese	3
	Indigeneous	15
	Other	1
Religion	Islam	9
	Christian or Catholics	4
	Hinduism or Buddhism	3
	Ahmadiyya or Shia	2
	Traditional Beliefs	1
Gender	Male	6
	Female	13
LGBTQ+	Yes	1
	No	18
Age	18 - 24	9
	25 - 34	5
	35 - 44	3
	45 - 54	2
Education	Master's Degree	3
	Bachelor's Degree	7
	Associate's Degree	2
	High School Degree	7
Job Status	Employed	9
	College Student	8
	Housewife or Unemployed	2
Presidential Vote	Candidate no. 1	4
	Candidate no. 2	7
	Candidate no. 3	5
	Unknown or Abstain	3

Table 1: The demographic background of the 19 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.

3.3 Annotation Instrument

Our goal is to capture text that contains toxicity, be it explicit (manifest content such as inclusion of offensive words) or implicit (latent content such as sarcasm) (Krippendorff, 2018). This nuanced and contextual hate speech has not yet been confronted in a consistent and unified manner in the NLP community (ElSherief et al., 2021).

Based on the literature review of hate speech and

[1] https://cekfakta.com/

[2] https://mafindo.or.id

[3] identified minority groups in Indonesia that have been the target of hate speech in previous elections including disability, LGBTQ+, Chinese, Ahmadiyya, Shia, Catholics, and Christian groups.

discussions with vulnerable groups in Indonesia, we create a codebook (Appendix C) as a guide for annotators to identify text as hate speech when certain criteria are met (Sellars, 2016, p.25-30). The codebook helps annotators recognize text typically seen as hate speech and text that seems normal but is indeed harmful to a specific vulnerable group.

3.4 Annotation Process

The annotation process is divided into two stages: the training phase and the main annotation phase. The annotators were trained on the codebook during the training phase. They were instructed to identify whether the text contains hate speech (or toxicity). If yes, they were asked to identify the types of the hate speech (i.e., whether it is an insult, threat, profanity, identity attack, or sexually explicit text).

During the first training session, the annotators were given 100 randomly-sampled text to code (i.e., annotate), but ICR was not met. Hence, a second (with another 100 randomly-sampled text) and a third (with another 249 randomly-sampled text) training session were held to further clarify the codebook and resolve any confusion. A satisfactory ICR score was met after the third training session: a Gwet's AC1 score of 0.61 for the toxicity label. Following Quantitative Content Analysis (QCA) in communication research, which is a commonly used method to derive replicable and meaningful inferences from texts (Krippendorff, 2018), once the ICR was met, the annotators continued to code more texts independently in the main annotation phase. Our final data is comprised of 43,692 texts that were annotated from the three sessions of training and the main annotation phase[4].

3.5 Dataset Properties

Out of 43,692 texts in IndoToxic2024, 6,894 are labeled as toxic. From Table 2, we can observe that almost half of the toxic texts are insults.

Toxicity Types	# Texts
Insults	3140
Threat / Incitement to Violence	2837
Profanity or Obscenity	1271
Identity Attack	1061
Sexually Explicit	224

Table 2: Number of toxic texts labeled with types. These categories are not mutually exclusive, since a toxic comment could exhibit more than one type of hate speech.

We can further explore how different demographic groups annotate the texts. When we aggregate the dataset, we see that the distribution of toxicity labels differs between genders. Males labeled 19.3% of their data as toxic, whereas females labeled 13%. We explore this topic in more depth in the next section.

4 Dataset Analysis: Existence of Subjectivity

4.1 Subjectivity in Hate Speech Annotations

The research by Fleisig et al. (2024) indicates that combining demographic data with potential hate speech improves toxicity prediction models. Kumar et al. (2021) further emphasizes the subjective nature of toxic text, as shown by annotator disagreements. Our study expands on these insights by exploring the topic from three new perspectives.

Through Distribution Assumptions In the annotation process, texts are randomly assigned without considering the annotator's gender, education, disability, or domicile. We use chi-square testing to verify the null hypothesis: "If texts are randomly assigned and there's no subjectivity, each annotator should receive a similar proportion of hate speech texts". However, Figure 2 shows that the null hypothesis is only valid for the "domicile" group. Therefore, we reject the null hypothesis, confirming subjectivity in most demographic groups. It's important to note that some texts were assigned based on the religion or ethnicity of the annotator, so these groups are excluded from this analysis.

Through ICR Score We calculate the ICR score within each demographic group (within-group ICR score) and between two demographic groups (between-group ICR score) (refer to Appendix G for more details). It's generally assumed that the between-group ICR scores would be lower than the within-group ICR scores due to subjectivity within the group. However, our results mostly contradict this assumption. For example, while the within-group ICR score for females is 0.61 and for males is 0.54, the between-group ICR score is 0.58, which is not lower than both within-group scores. Further analysis reveals the role of intersectional identity. For instance, a high ICR score contributor among the female-male pairs belongs to other similar demographic groups (both are disabled annotators and identify as Christians). This suggests that demographics are intersecting factors, not mutually

[4] Due to the long annotation process and the complexity and human toll of the task, some annotators complete only parts of their assignments during the main annotation phase.

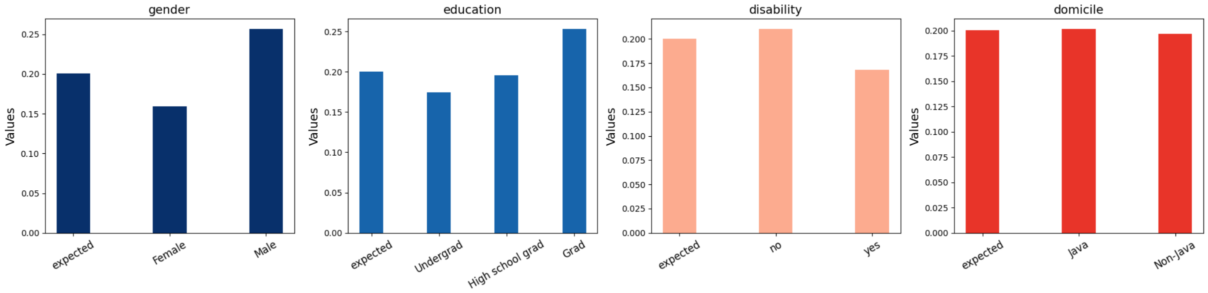


Figure 2: **Subjectivity affects the annotation process.** Expected ratio of hate speech vs. non-hate speech labeled data against the ratio of each demographic group's annotations using chi-square testing. Blue colors indicate a lower p-value (more significant difference to the expected ratio), and red colors indicate a higher p-value (less significant difference to the expected ratio).

Train on	**Test on**	
	Non-Islam	Islam
Non-Islam	0.605	0.597
Islam	0.628	0.66

Table 3: **Subjectivity affects performance of models.** F1-scores of IndoBERTweet, trained on *Islam* and *Non-Islam* annotators' labels, using 5-fold cross-validation when the training and testing data are from the same group's annotations (i.e., the diagonals). It can be seen that the performance of training on one group's annotations and testing on another is worse than training and testing on the same group's annotations.

exclusive aspects.

Through Modeling Results Fine-tuning a model using a subjective or biased dataset can result in a model that inherits that subjectivity (Sengupta and Srivastava, 2022). To ensure the availability and quantity of the training dataset, we use a coarser granularity for some of our demographic categories. For instance, we use age groups instead of specific age numbers for the age demography, and a coarser grouping of Islam and Non-Islam for the religious demography. We then fine-tune IndoBERTweet model (Koto et al., 2021) by limiting the training data to texts that are annotated by the same group (e.g., Islam), and test on texts that are annotated by the other group in the demographic category (e.g., Non-Islam). As a result, a model fine-tuned on *Non-Islam* annotators' labels tends to perform worse when tested on *Islam* annotators' labels compared to when tested on their own labels (using 5-fold cross-validation) and vice versa (see Table 3). This trend holds true for other demographic categories such as disability, ethnicity, and religion. The result indicates that for some demographic groups, there may exist other demographic groups that annotate the texts differently due to differences in their identities.

4.2 Texts with Differing Annotations

Accompanying previous results, we rank texts based on their divisiveness. A text is highly divisive when groups of annotators in the same demographic category largely disagree on their annotations. For example, the text in Figure 1 where Christian annotators unanimously agree that the text *is* hate speech while the non-Christians unanimously disagree. More examples can be found in appendix A. The text illustrates how the interpretation of hate speech can vary depending on the annotator's demography, in this case their religion. However, it is crucial to consider the whole picture. There are instances where a text appears to be on the topic of a specific demographic category, such as female, yet annotators are split not along gender but along their (last) education level.

4.3 How Topic Affects Subjectivity

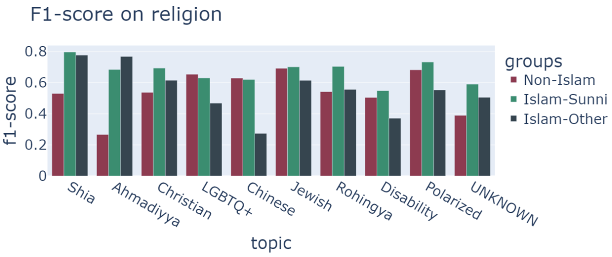


Figure 3: **Topics can act as a stabilizer or a catalyst.** Some topics can be *easier* for a certain demography to annotate. For example, texts relating to Shia are *easier* for Islamic group of people to annotate while *harder* for non-Islam people to annotate.

We trained a hate speech classifier on texts annotated by each demographic group and then checked how well each model did on different topics. Topics of a text are target demographics of the text, identified based on keywords mentioned in the text. The keywords-to-topics mapping is in appendix D.

An example of how models performed across topics is shown in Figure 3. One model was trained on a dataset annotated by people *with* non-Muslim religion ("Non-Islam"), one was trained on a dataset annotated by people *with* Islam Sunni religion ("Islam-Sunni"), and the third one trained on a dataset annotated by people *with* "Shia" or "Ahmadiyya" as their religion ("Islam-Other"). For topics relating to Shia or Ahmadiyya, both of the "Islam" models perform similarly while the model trained on "Non-Islam" dataset perform relative poorly. However, for topics relating to LGBTQ+, or Chinese, there's a jump in performance for the "Non-Islam" model. This suggests that there's less disagreement among people of the same group when they understand or relate to the topics, and higher disagreement when they are unfamiliar with the topics.

5 Benchmark Results, Experiments, and Analysis

5.1 Baseline Model Performance per Task

We benchmark 6 classification tasks to identify: whether a text contains hate speech, and the five types of hate speech. We fine-tune IndoBERTweet (Koto et al., 2021) on our dataset. To enhance the performance of our baselines, we merged our dataset with that of CSIS (2022) and generated synthetic hate speech data using gpt-3.5-turbo (Brown et al., 2020) through 10-shot generation. This synthetic data was only used for training. The stratified 10-fold performance and data breakdown for each task are reported in Table 4. The prompt given to gpt-3.5-turbo to generate synthetic texts is available in appendix E.

For most tasks, our baseline achieved a macro F1-score above 0.7. However, for the "incitement to violence" classification task, we only achieved a macro F1-score of 0.53. It's important to note that the data reported in Table 4 were used to achieve the best baseline performance reported here.

5.2 Incorporating Demographic and Topic Information

Fleisig et al. (2024) shows that incorporating metadata such as demographic information and survey results can increase a hate speech classifier's performance. For the following set of experiments, we report on the performance of three models: IndoBERTweet (Koto et al., 2021), gpt-3.5-turbo (Brown et al., 2020), and SeaLLM-7B-v2.5 (Nguyen et al., 2023). IndoBERTweet and SeaLLM are pre-trained with a focus on Indonesia and SouthEast Asian (SEA) languages respectively.

Without any demographic or topic information, IndoBERTweet, fine-tuned only on IndoToxic2024, has the best performance for hate speech classification with a macro-F1 of 0.718 (5-fold cross-validation). We also report the zero-shot performance of the other models in Table 5a. The performance of IndoBERTweet, after incorporating topic and/or demographic information, is detailed in Table 5b.

To incorporate the topic (t) and demographic information (d) along with the texts (w), the input is formatted as follows:

$$d_1 \ldots d_n \; t_1 \ldots t_n \, [SEP] \, w_1 \ldots w_n$$

For instance, a complete input given to IndoBERTweet might be: *"Reader information: Chinese ethnicity, Christian, Male, Millennial. Topic: Christian. [SEP] [TEXT]"*.

To use gpt-3.5-turbo and SeaLLM-7B-v2.5, a custom prompt is used, with its system prompt set to: *"You may only respond with either a 0 or a 1"*. We also enforce the statement in the user input, which follows this pattern: *"Is the Indonesian text below a hate speech [Demographic Information]? If yes, respond with 1, otherwise respond with 0. [TEXT]"*.

Each text in our dataset is accompanied by demographic information about the annotator. This information includes disability status, domicile, ethnicity, gender, generation (age group), part of the LGBTQ+ communities, last education, religion, work status, and political leaning. Our experiments are divided into three parts: one with no demographic information provided, one with all demographic information provided, and one with only a single piece of demographic information provided, totaling 12 experiments. These experiments are conducted on three models: IndoBERTweet, IndoBERTweet with topic information given, and gpt-3.5-turbo. The performance of these models can be found in Figure 4.

Our findings indicate that for IndoBERTweet models, providing demographic information does not yield any significant improvement. In fact, it may even negatively impact the model's performance. However, in all instances, the performance of gpt-3.5-turbo improved when demographic information was provided. We hypothesize that the additional information adds a level of complexity

Classification Task	Metrics				Data Statistic				
	Accuracy	F1	Precision	Recall	Our-0	Our-1	CSIS-0	CSIS-1	Synthetic-1
Hate Speech	0.89	0.78	0.80	0.76	4014	1338	21125	7112	1325
Identity Attack	0.75	0.80	0.74	0.88	461	648	-	-	228
Incitement to Violence	0.77	0.53	0.55	0.52	345	115	945	315	890
Insult	0.79	0.85	0.81	0.88	705	423	149	1142	785
Profanity or Obscenity	0.81	0.70	0.70	0.72	824	373	-	-	370
Sexual Explicit	0.91	0.80	0.88	0.77	123	41	-	-	82

Table 4: Performance of IndoBERTweet across various binary classification tasks, utilizing a combination of our annotated data, data from CSIS, and synthetically generated data via GPT-3.5-turbo. The term **-x** represents the quantity of data associated with that label for a specific task. Our sampling strategy ensures that the quantity of 0-labeled (e.g., non-hate speech) data is at most three times that of 1-labeled (e.g., hate speech) data.

Model	F1
IndoBERTweet	**0.718**
GPT-3.5-turbo	0.627
SeaLLM-7B-v2.5	0.517

(a) Baseline performance

Information	F1
Baseline	0.718
+ Demographic	0.672
+ Topic	**0.755**
+ Topic & Demo	0.709

(b) IndoBERTweet performance with augmented input.

Table 5: IndoBERTweet, when given only topic information, performs best based on the macro-F1 metric. Models are trained using only **IndoToxic2024** dataset.

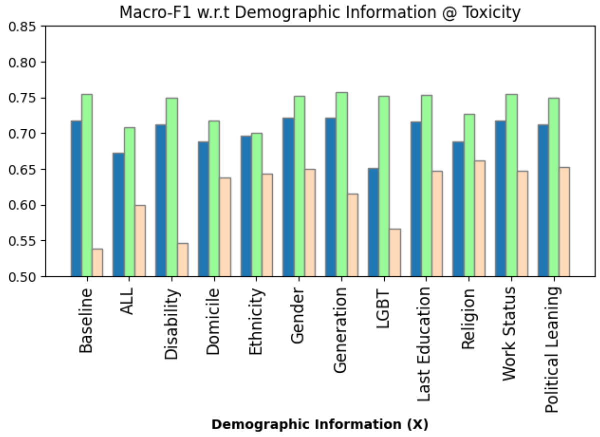


Figure 4: Comparison of macro-F1 Score from IndoBERTweet (dark green/left bars), IndoBERTweet with topic (light green/middle bars), and gpt-3.5-turbo (orange/right bars) given varying degrees of demographic information. Baseline means no demographic information was given.

during the fine-tuning of the IndoBERTweet model that exceeds what the dataset can support.

5.3 Ablation of Impact per Topic

Leveraging its Pre-training, gpt-3.5-turbo Understands Cultural and Value Differences. As indicated by its performance in Figure 5, gpt-3.5-turbo, thanks to its pre-training data, inherently understands the cultural and value differences among people based on their identities. The figure shows that gpt-3.5-turbo significantly improves when provided with the annotator's domicile, without any fine-tuning. For instance, it achieved a 0.47 macro-F1 improvement when the annotator resided in **Medan** and the text targeted **Rohingya** refugees. We attribute this improvement to the training data, which might include articles about Rohingya refugees often first arriving in Medan (Suryono, 2024). Another example is **Makassar** and text targeting the **Disabled**, where the model achieves a 0.2 macro-F1 improvement when given demographic information. This could be due to the model's understanding of the unique relationship of **Disabled** people in **Makassar** compared to other locations (Post, 2023). Without demographic information, we assume that the model defaults to an inherent bias. For instance, when domicile information is not provided, gpt-3.5-turbo seems to assume the reader's location is either the Indone-

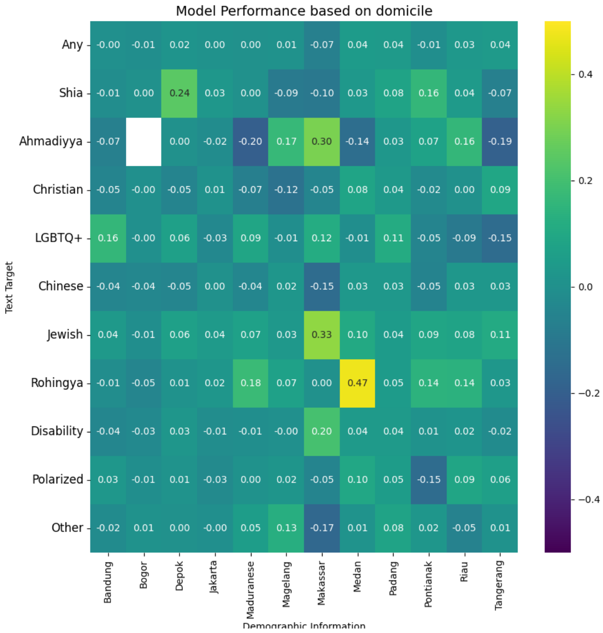


Figure 5: The effect (Δ F1) of giving domicile information to gpt-3.5-turbo on its hate speech text classification compared to not giving it any information.

sian capital **Jakarta** or another big city, **Bogor**, as suggested by the non-significant performance gap visible in Figure 5. Visualizations for other cases are available in the appendix F.

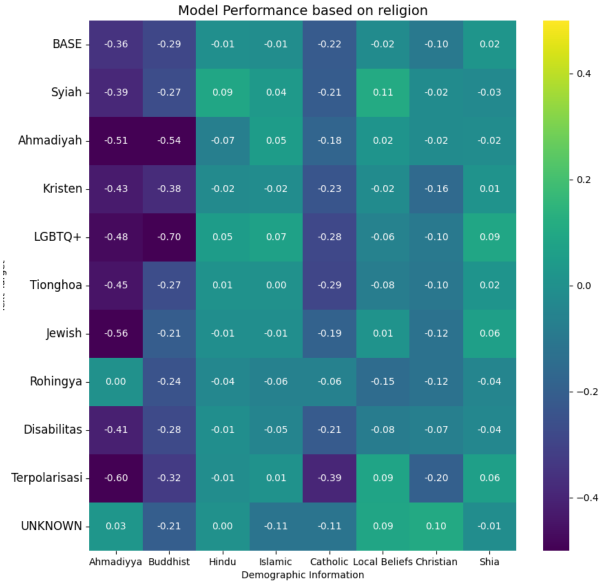


Figure 6: The effect (Δ F1) of giving religion information to IndoBERTweet on its hate speech text classification compared to not giving it any information.

However, demographic information hurts IndoBERTweet's classification performance. We attribute this to the diversity and subjectivity of hate speech texts. Though we hoped to increase the model's performance by providing demographic information, this instead fragmented the training data, with fewer training data for each demographic. In other words, we add another dimension(s) to the input without increasing the sample size. When the data is too few, this fragmentation hurts the model instead (Figure 6). Though the model learns the difference between each religion, there is very few data for some of them, such as **Ahmadiyya** and **Buddhist** which are the two least represented religion demographics in our dataset consisting of only 346 and 1,368 annotations respectively. Without this demographic information, the model only has the text as its input, which may explain the baseline's generally good performance.

Topic information increases IndoBERTweet's performance. Focusing on the row entries from Figure 7, we observe that the model's performance increases in most cases. Though we hypothesized that the model would innately learn the topic of the text by itself, it failed to do so in practice, potentially due to a lack of data. Therefore, adding topic information improves the model's performance.

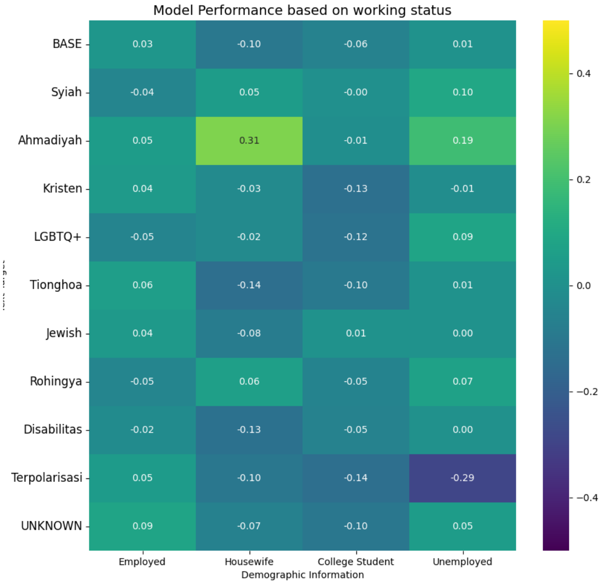


Figure 7: The effect (Δ F1) of giving topic information to IndoBERTweet on its hate speech text classification compared to not giving it any information.

6 Discussion

The surge in online hate speech (CSIS, 2022; AJI, 2024) poses societal risks (Williams et al., 2019). The challenge lies in defining hate speech and the debate over censoring such content. Some fear censorship could have a domino effect (Franco and Warburton, 2013), while others argue for immediate action due to existing societal damage (Laaksonen et al., 2020). Despite the complexity of defining hate speech, we can't remain idle. We propose that researchers focus on creating hate speech detection systems focusing on protecting vulnerable minority groups, often targeted by hate speech and hate crimes, a crucial step for their protection.

To develop a hate speech detection system that filters texts targeting vulnerable minority groups, datasets with demographic information, like those in Kumar et al. (2021) and ours, are needed. However, such datasets are surprisingly scarce. Given the potential benefits to these groups, we call on researchers to promptly create similar datasets.

Furthermore, demographic datasets may have a larger role than expected. Our analysis indicates that large language models like gpt-3.5-turbo may contain biases. These biases can be reduced with metadata, which can form a "Persona" for effective model interaction. Our tests show that demographic information makes gpt-3.5-turbo a better hate speech detector. There could be other unexplored scenarios where demographic information is beneficial for different tasks.

Limitations

Our work has the potential to pave the way for future research in creating better hate speech and toxic text detection. However, our work is not without flaws.

Baseline Performance Relies on a Private Dataset. The baseline performance for the seven binary classification tasks was established by integrating our IndoToxic2024 dataset with a private dataset from (CSIS, 2022), accessed through collaboration.

Comparison of Fine-tuned and Zero-shot Models. We compared the performance of a fine-tuned IndoBERTweet model against zero-shot gpt-3.5-turbo and SeaLLM-7B-v2.5 models. Due to the extensive number of experiments conducted with gpt-3.5-turbo, we were unable to fine-tune it for a direct comparison. Additionally, we discontinued the use of SeaLLM-7B-v2.5 due to its subpar performance relative to the other models.

Inconsistent Annotator Count Across Annotation Phases. The varying number of annotators across different annotation phases may lead to inconsistencies in the data distribution. This could potentially result in an incomplete representation of the overall demographic traits, as not all demographic dimensions may be uniformly captured.

Controlled Topic Distribution in Main Annotation Phase 1. Annotators did not received a completely random texts when they annotate the 1000-text batch in main annotation phase 1. Instead, they received 50% text *that* mentions their group, and 50% random text that *did not* mentions their groups. This applies to all dataset except the Shia & Ahmadiyya dataset,who received only 39.3% text of their group due to the scarcity of existing data.

Use of a Naive Keyword-based Approach for Topic Extraction. Our approach to extracting featured topics relies on predefined keywords, which may overlook nuanced or emergent themes. This limitation could restrict the scope of our analysis and prevent the identification of more subtle or complex topics present in the text. A more dynamic and flexible topic extraction approach could potentially enhance the richness and accuracy of topic identification within IndoToxic2024.

Ethics Statement

The creation of this dataset exposes annotators to potentially harmful hate speech 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 of texts if they feel unable to proceed. We recognize the potential misuse of such datasets, which could include training models to generate more hate speech. Yet, it's worth noting that without these datasets, it is alarmingly straightforward to train a model to produce toxic content, as the source of their training data, the internet, still consists of hate speech and toxic 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 hate speech targeted at specific demographic groups is still green, 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.

References

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

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.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The place of inter-rater reliability in qualitative research: An empirical study. *Sociology*, 31(3):597–606.

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/`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

John L Campbell, Charles Quincy, Jordan Osserman, and Ove K Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological methods & research*, 42(3):294–320.

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

CSIS. 2022. Hate speech dashboard.

Arijit Das, Somashree Nandy, Rupam Saha, Srijan Das, and Diganta Saha. 2021. Analysis and detection of multilingual hate speech using transformer based deep learning. *Jadavpur University*.

R. Delgado. 1982. Words that wound: A tort action for racial insults, epithets, and name-calling. *Harvard Civil Rights-Civil Liberties Law Review*, 17:133–181.

detikcom. 2022. Jadwal pemilu 2024 lengkap, termasuk jadwal pilpres jika 2 putaran. Accessed November 8th, 2023.

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.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2024. When the majority is wrong: Modeling annotator disagreement for subjective tasks.

Joshua Franco and Nigel Warburton. 2013. Should there be limits on hate speech? *Index on censorship*, 42(2):150–152.

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

K. Greenawalt. 1989. *Conflicts of Law and Morality*. Oxford University Press, New York.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Daroe Iswatiningsih, Eggy Fajar Andalas, and Nina Inayati. 2019. Hate speech by supporters of indonesian presidential candidates on social media. In *6th International Conference on Community Development (ICCD 2019)*, pages 130–133. Atlantis Press.

Simon Kemp. 2023. Digital 2023: Indonesia — datareportal – global digital insights. Accessed June 14th, 2024.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization.

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. *arXiv (Cornell University)*, pages 299–318.

Salla-Maaria Laaksonen, Jesse Haapoja, Teemu Kinnunen, Matti Nelimarkka, and Reeta Pöyhtäri. 2020. The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data*, 3.

Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, 28(4):587–604.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. Hatexplain: A benchmark dataset for explainable hate speech detection.

United Nations. 2023. Hate speech and real harm | United Nations.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms – large language models for southeast asia.

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.

Hameedat Olodo, Bukola, Moriam Aremu, and Mustafa Raji. 2022. Demographic variables: A predictor of consumer buying behaviour in formal and informal retail outlets. 8:63–76.

T. P. Paramadina and Mafindo. 2023. *Buku Panduan Melawan Hasutan Kebencian dan Hoax Edisi Perluasan*. PUSAD Paramadina, Jakarta.

Prof Patil, Kulkarni Sanika, and Khurpe Sanjana. 2024. Medical insurance premium prediction with machine learning. *International Journal of Innovations in Engineering Research and Technology*, 11:5–11.

Jakarta Post. 2023. ASEAN produces 'Makassar Recommendations' for persons with disabilities - Front Row - The Jakarta Post.

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

Kinshuk Sengupta and Praveen Ranjan Srivastava. 2022. Causal effect of racial bias in data and machine learning algorithms on user persuasiveness & discriminatory decision making: An empirical study.

Mitra Salima Suryono. 2024. Rohingya refugees risk dangerous sea route to Indonesia in search of safety and freedom.

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

Ali Unlu and Tommi Kotonen. 2024. Online polarization and identity politics: An analysis of Facebook discourse on Muslim and LGBTQ+ communities in Finland. *Scandinavian political studies*.

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.

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).

Appendix

A Example of Divisive Texts

A list of divisive text is available in Table 6.

B 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.

C Annotation Guidelines

C.1 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

Divisive Text	Comment
Around 800 to 1000 Christians still live in Gaza, which is one of the oldest Christian communities worldwide. They are also evicted and killed. Entire faith groups died out in the land where Christianity originated.	**3 of 3 Christians or Catholics**, and **0 of 10 non-Christians or Catholics** annotators annotate this text as hate speech.
Be a woman who is crazy about work. Because there are enough women crazy about boys. Even though people keep seeing us hanging out.	**1 of 4 undergraduates** and **3 of 3 high school graduates** annotators annotate this text as hate speech. However, **2 of 5 females** and **2 of 3 males** annotators annotate this text as hate speech.
"GANJA(Weed) FOR ALL CEBONG" with all their 45 spirit putting up billboards, they forgot one character and turned it into something completely different... hahahaah, this is what happens when you want to scam people.	**2 of 3 Ganjar voters (Group 3)**, **5 of 5 Prabowo voters (Group 2)**, and **0 of 3 Anies voters (Group 1)** annotate this text as hate speech.
Everybody knows that most PDIP supporters are Christians and that anyone gaining the support of PDIP will also get the support of Christians. However, this chances because of Ganjar Pranowo's stance on banning Israel's soccer team from coming to Indonesia to compete, which makes Christians unhappy.	**3 of 4 males** and **0 of 7 females** annotators annotate the text as hate speech.
I used to praise Jokowi but not anymore, why? Because he used his child for evil people. If Gibran was made to become a candidate for president, it would have been okay. But, Gibran was made as a candidate for vice president, used only to gain supporters for Golkar political group.	**7 of 8 non-Chinese** and **0 of 3 Chinese** annotators annotate this text as hate speech.
Goooo GaMa (Word play on the president and vice president candidate of Group 3)! Legally defective products will continue to create defective products.	**6 of 7 Gen Z** and **0 of 4 Gen X and Millenials** annotate this text as hate speech.

Table 6: Examples of divisive texts and the demographic group in which they are divisive.

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.

C.2 Manual Annotation

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 contain toxicity (hate speech)?

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

Q4: What is the type of toxicity?

Note: Code up to two or more types. Consider the following sentences as an example: *"PDIP Provokasi Massa pendukungnya geruduk kediaman Anies"*. This headline should be coded as both threat and incitement to violence.

Q4-1: Does the message contains profanity/obscenity?

- Yes
- No

Q4-2: Does the message contain threat / incitement to violence?

- Yes
- No

Q4-3: Does the message contain insults?

- Yes
- No

Q4-4: Does the message contain an identity attack?

- Yes
- No

Q4-5: Does the message contain sexually explicit?

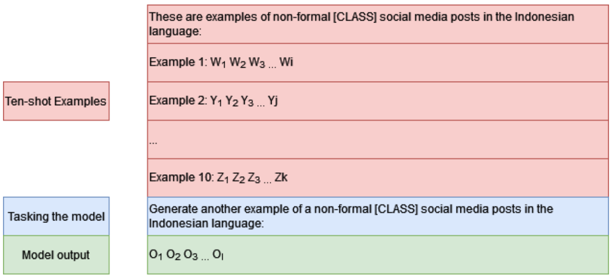

Ten-shot Examples
These are examples of non-formal [CLASS] social media posts in the Indonesian language:
Example 1: W_1 W_2 W_3 ... Wi
Example 2: Y_1 Y_2 Y_3 ... Yj
...
Example 10: Z_1 Z_2 Z_3 ... Zk
Tasking the model
Generate another example of a non-formal [CLASS] social media posts in the Indonesian language:
Model output
O_1 O_2 O_3 ... O_l

Figure 8: The template we use to prompt gpt-3.5-turbo through ten-shot prompting.

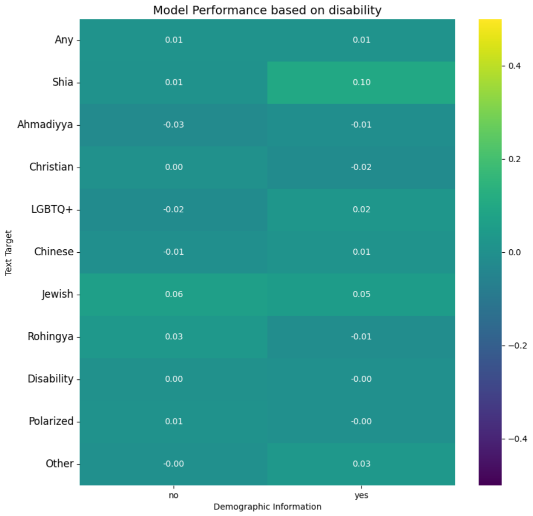
- Yes
- No

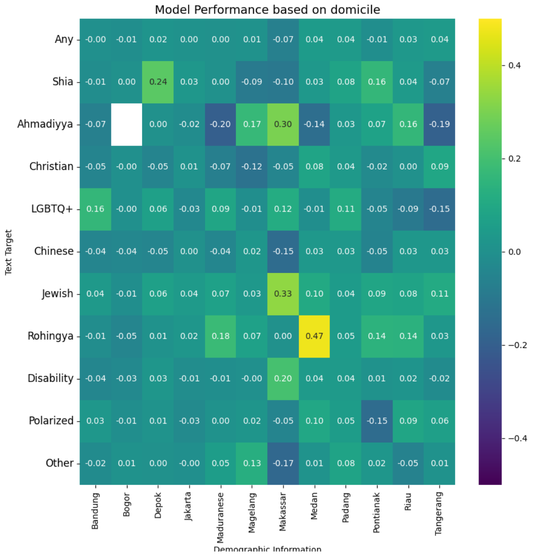
D Mapping of Keywords-to-Topics

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

E Prompt for Synthetic Text Generation using gpt-3.5-turbo

To generate synthetic data from gpt-3.5-turbo to help increase model performance for various binary classification task, we utilize the prompt visualized in Figure 8





F Effect of Demographic and Topic Information to Model Performance

F.1 gpt-3.5-turbo

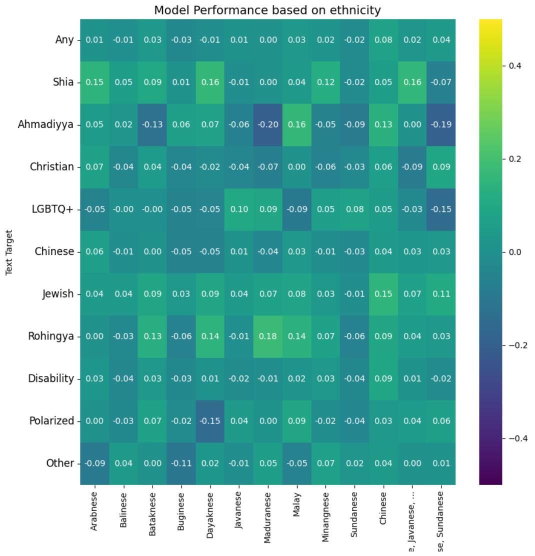
F.2 IndoBERTweet with Demographic Information

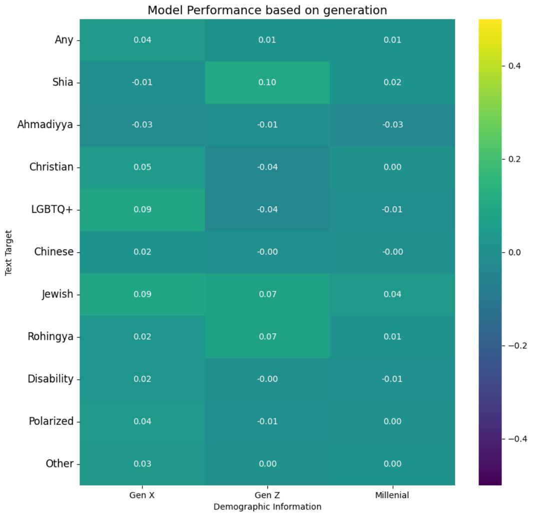
F.3 IndoBERTweet with Topic

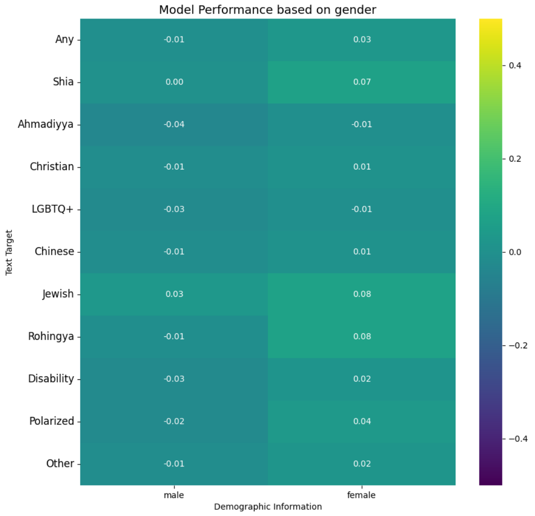
F.4 IndoBERTweet with Demographic and Topic

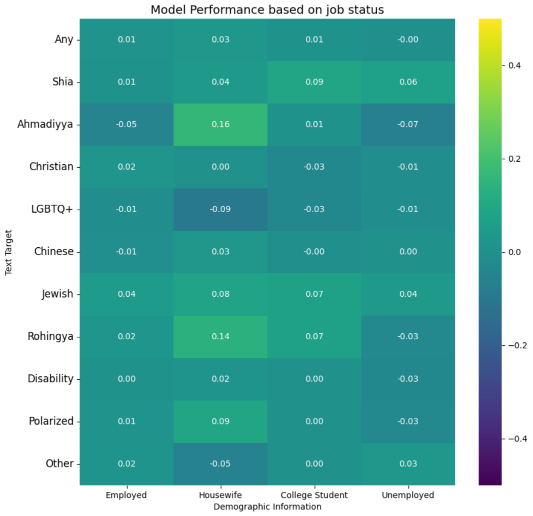
G Within & Between Group ICR Score Calculation

To compute the toxicity ICR score for a demographic group, we calculate the weighted average of Gwet's AC1 score for every pairwise combina-









tion of annotators within the same group, using the volume of text in each pair as the weight. This approach maximizes the utilization of available data. A similar calculation method is implemented to find the ICR score between two groups, with the modification that each pair consists of members from different groups. We refer to these metrics as "within-group ICR score" and "between-group ICR score" respectively.

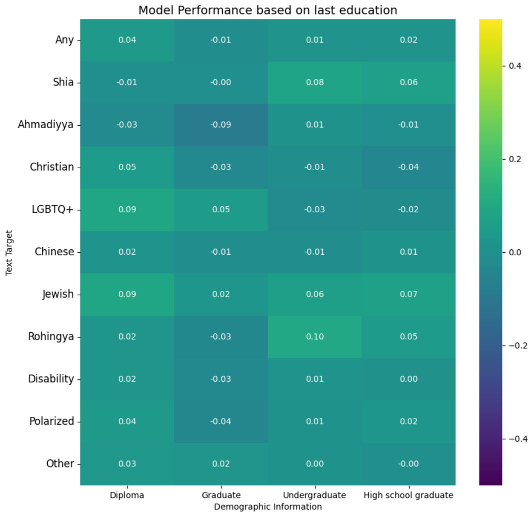
We can rigorously define an equation to compute ICR score within a group as

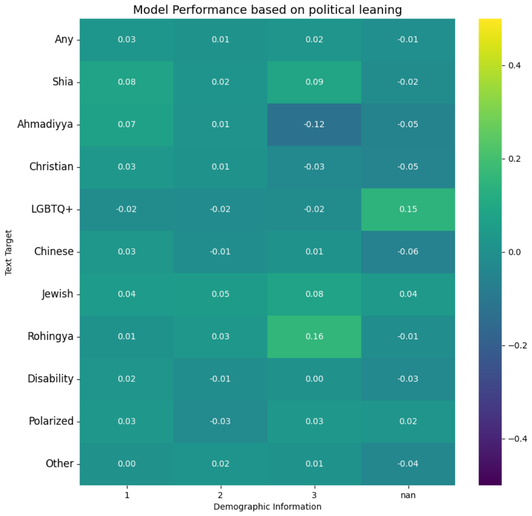
$$\gamma(g) = \frac{\sum\limits_{i,j \in g} dim(v_{ij}) \cdot \text{Gwet}(\phi_i(v_{ij}), \phi_j(v_{ij}))}{\sum\limits_{i,j \in g} dim(v_{ij})}$$

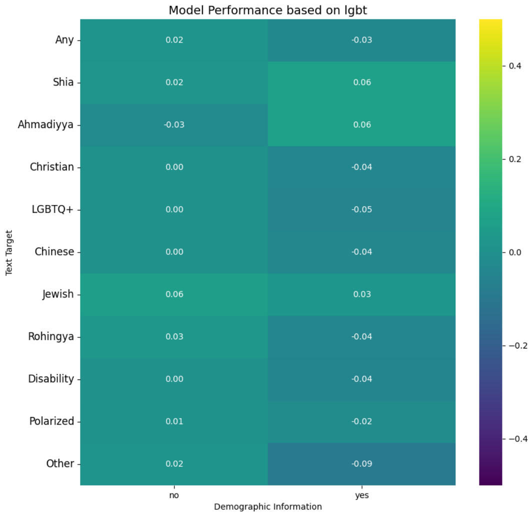
Where g are an arbitrary groups in a demographic; v_{ij} are set of text that both mutually by annotators i,j; and ϕ_i, ϕ_j are annotation result from i, j. To calculate ICR score between two groups, we slightly modified the equation above into

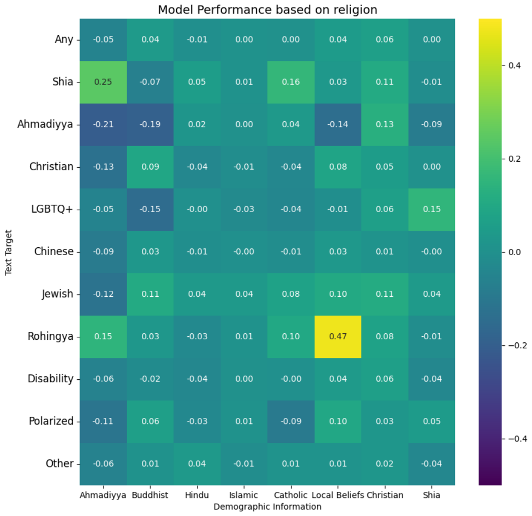
$$\Gamma(g_1, g_2) = \frac{\sum\limits_{i \in g_1, j \in g_2} dim(v_{ij}) \cdot \text{Gwet}(\phi_i(v_{ij}), \phi_j(v_{ij}))}{\sum\limits_{i \in g_1, j \in g_2} dim(v_{ij})}$$

Where g_1,g_2 are arbitrary two groups in a demographic; v_{ij} are set of text that both mutually by annotators i,j; and ϕ_i, ϕ_j are annotation result from i, j.



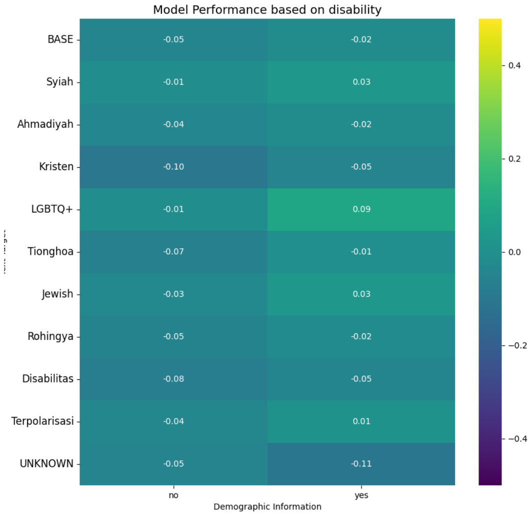


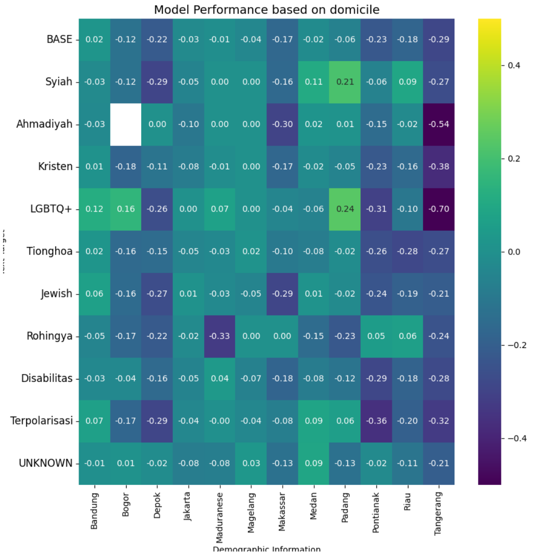


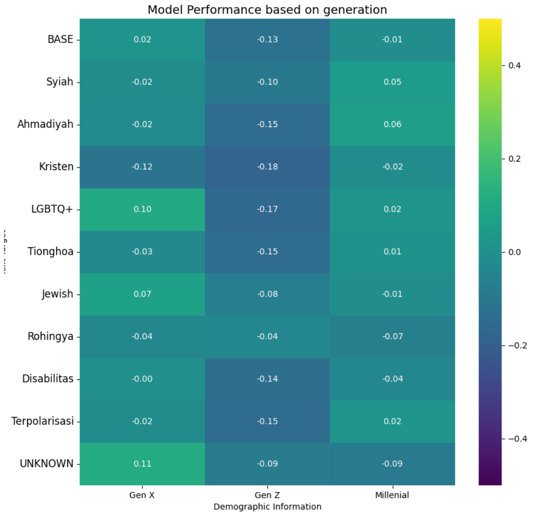


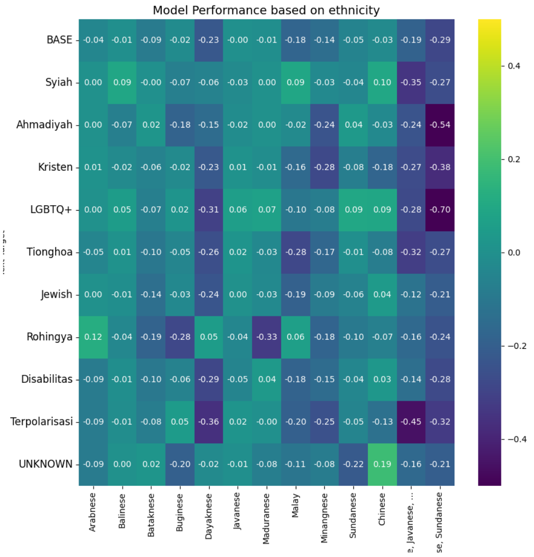
H Performance of Fine-tuned Models on Multiple Topics

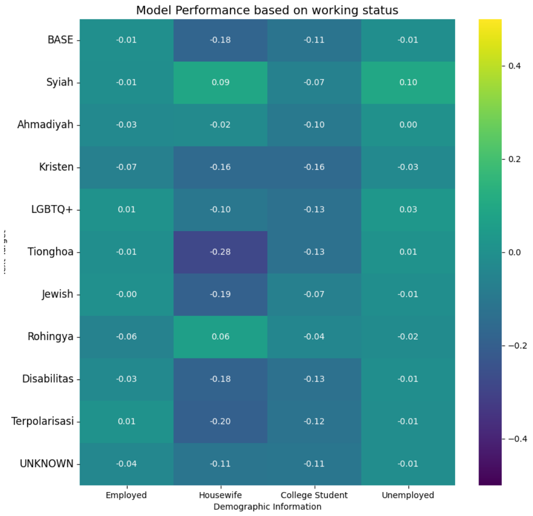
Figure ??,??,??,,?? shows how the models that fine-tuned per group

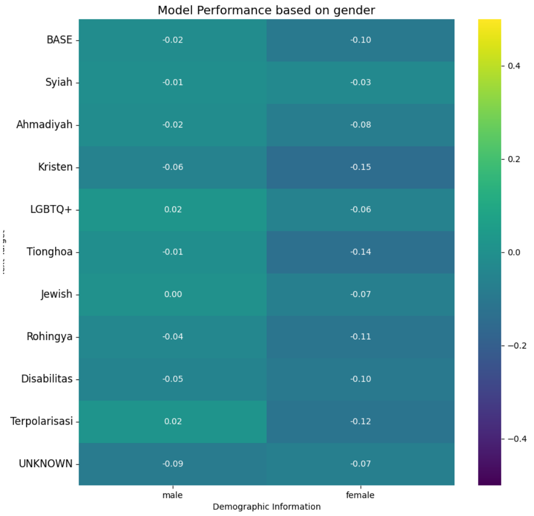


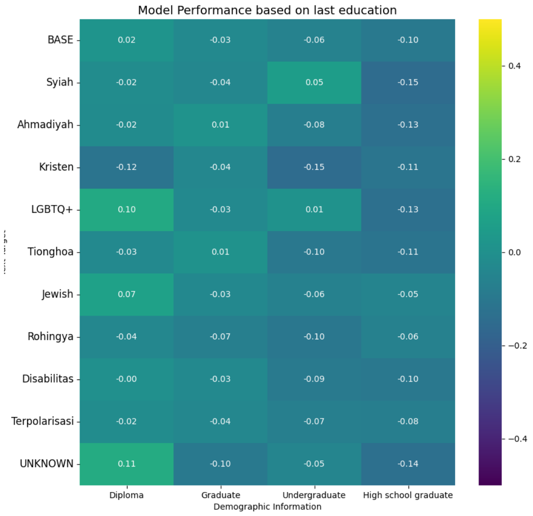

Model Performance based on domicile
BASE 0.02 -0.12 -0.22 -0.03 -0.01 -0.04 -0.17 -0.02 -0.06 -0.23 -0.18 -0.29
Syiah -0.03 -0.12 -0.29 -0.05 0.00 0.00 -0.16 0.11 0.21 -0.06 0.09 -0.27
Ahmadiyah -0.03 0.00 -0.10 0.00 0.00 -0.30 0.02 0.01 -0.15 -0.02 -0.54
Kristen 0.01 -0.18 -0.11 -0.08 -0.01 0.00 -0.17 -0.02 -0.05 -0.23 -0.16 -0.38
LGBTQ+ 0.12 0.16 -0.26 0.00 0.07 0.00 -0.04 -0.06 0.24 -0.31 -0.10 -0.70
Tionghoa 0.02 -0.16 -0.15 -0.05 -0.03 0.02 -0.10 -0.08 -0.02 -0.26 -0.28 -0.27
Jewish 0.06 -0.16 -0.27 0.01 -0.03 -0.05 -0.29 0.01 -0.02 -0.24 -0.19 -0.21
Rohingya -0.05 -0.17 -0.22 -0.02 -0.33 0.00 0.00 -0.15 -0.23 0.05 0.06 -0.24
Disabilitas -0.03 -0.04 -0.16 -0.05 0.04 -0.07 -0.18 -0.08 -0.12 -0.29 -0.18 -0.28
Terpolarisasi 0.07 -0.17 -0.29 -0.04 -0.00 -0.04 -0.08 0.09 0.06 -0.36 -0.20 -0.32
UNKNOWN -0.01 0.01 -0.02 -0.08 -0.08 0.03 -0.13 0.09 -0.13 -0.02 -0.11 -0.21
Bandung
Bogor
Depok
Jakarta
Maduranese
Magelang
Makassar
Medan
Padang
Pontianak
Riau
Tangerang
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

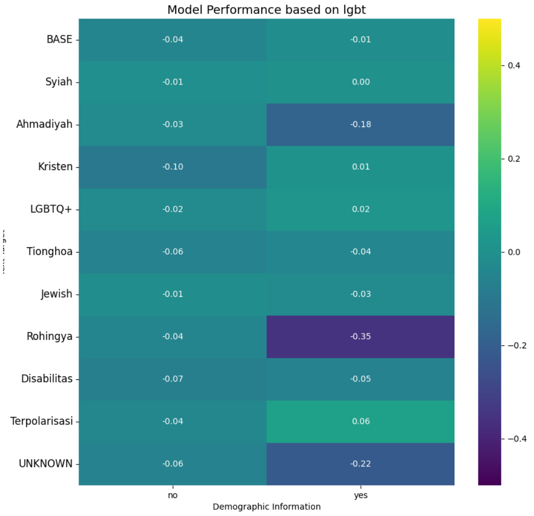

Model Performance based on generation
BASE 0.02 -0.13 -0.01
Syiah -0.02 -0.10 0.05
Ahmadiyah -0.02 -0.15 0.06
Kristen -0.12 -0.18 -0.02
LGBTQ+ 0.10 -0.17 0.02
Tionghoa -0.03 -0.15 0.01
Jewish 0.07 -0.08 -0.01
Rohingya -0.04 -0.04 -0.07
Disabilitas -0.00 -0.14 -0.04
Terpolarisasi -0.02 -0.15 0.02
UNKNOWN 0.11 -0.09 -0.09
Gen X
Gen Z
Millenial
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

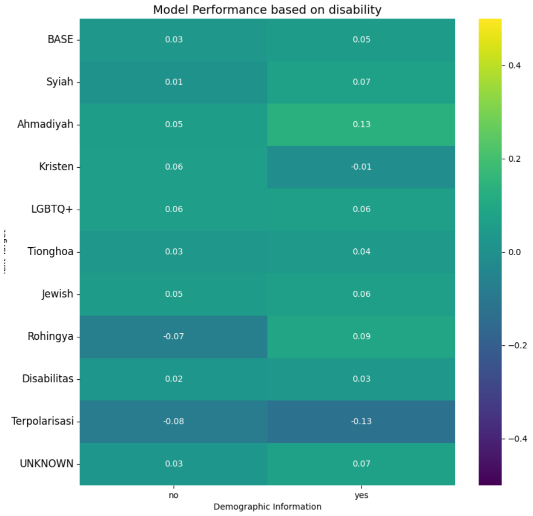

Model Performance based on ethnicity
BASE -0.04 -0.01 -0.09 -0.02 -0.23 -0.00 -0.01 -0.18 -0.14 -0.05 -0.03 -0.19 -0.29
Syiah 0.00 0.09 -0.00 -0.07 -0.06 -0.03 0.00 0.09 -0.03 -0.04 0.10 -0.35 -0.27
Ahmadiyah 0.00 -0.07 0.02 -0.18 -0.15 -0.02 0.00 -0.02 -0.24 0.04 -0.03 -0.24 -0.54
Kristen 0.01 -0.02 -0.06 -0.02 -0.23 0.01 -0.01 -0.16 -0.28 -0.08 -0.18 -0.27 -0.38
LGBTQ+ 0.00 0.05 -0.07 0.02 -0.31 0.06 0.07 -0.10 -0.08 0.09 0.09 -0.28 -0.70
Tionghoa -0.05 0.01 -0.10 -0.05 -0.26 0.02 -0.03 -0.28 -0.17 -0.01 -0.08 -0.32 -0.27
Jewish 0.00 -0.01 -0.14 -0.03 -0.24 0.00 -0.03 -0.19 -0.09 -0.06 0.04 -0.12 -0.21
Rohingya 0.12 -0.04 -0.19 -0.28 0.05 -0.04 -0.33 0.06 -0.18 -0.10 -0.07 -0.16 -0.24
Disabilitas -0.09 -0.01 -0.10 -0.06 -0.29 -0.05 0.04 -0.18 -0.15 -0.04 0.03 -0.14 -0.28
Terpolarisasi -0.09 -0.01 -0.08 0.05 -0.36 0.02 -0.00 -0.20 -0.25 -0.05 -0.13 -0.45 -0.32
UNKNOWN -0.09 0.00 0.02 -0.20 -0.02 -0.01 -0.08 -0.11 -0.08 -0.22 0.19 -0.16 -0.21
Arabnese
Balinese
Bataknese
Buginese
Dayaknese
Javanese
Maduranese
Malay
Minangnese
Sundanese
Chinese
e, Javanese, ...
se, Sundanese
0.4
0.2
0.0
−0.2
−0.4

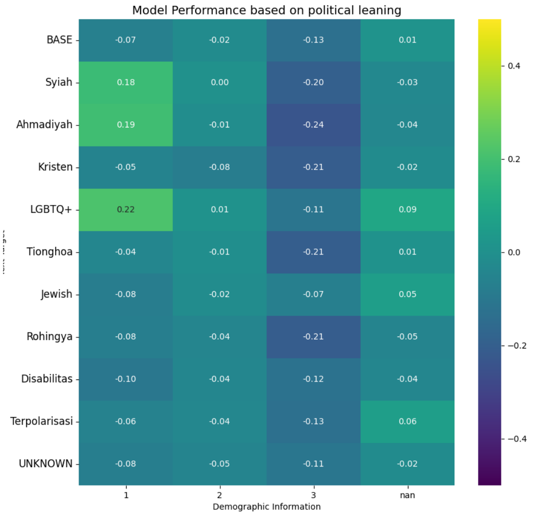

Model Performance based on working status
BASE -0.01 -0.18 -0.11 -0.01
Syiah -0.01 0.09 -0.07 0.10
Ahmadiyah -0.03 -0.02 -0.10 0.00
Kristen -0.07 -0.16 -0.16 -0.03
LGBTQ+ 0.01 -0.10 -0.13 0.03
Tionghoa -0.01 -0.28 -0.13 0.01
Jewish -0.00 -0.19 -0.07 -0.01
Rohingya -0.06 0.06 -0.04 -0.02
Disabilitas -0.03 -0.18 -0.13 -0.01
Terpolarisasi 0.01 -0.20 -0.12 -0.01
UNKNOWN -0.04 -0.11 -0.11 -0.01
Employed
Housewife
College Student
Unemployed
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

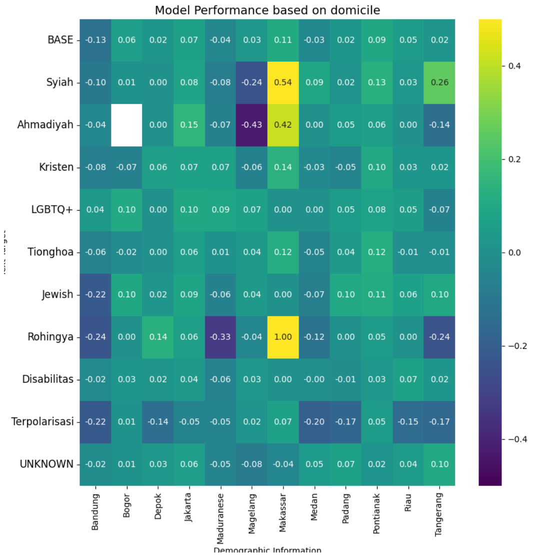

Model Performance based on gender
BASE -0.02 -0.10
Syiah -0.01 -0.03
Ahmadiyah -0.02 -0.08
Kristen -0.06 -0.15
LGBTQ+ 0.02 -0.06
Tionghoa -0.01 -0.14
Jewish 0.00 -0.07
Rohingya -0.04 -0.11
Disabilitas -0.05 -0.10
Terpolarisasi 0.02 -0.12
UNKNOWN -0.09 -0.07
male
female
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

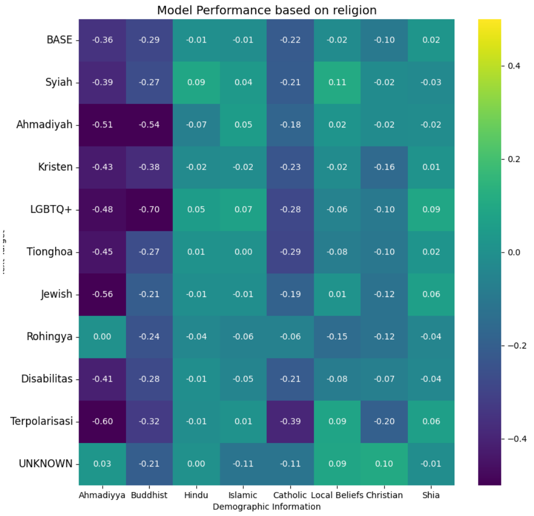

Model Performance based on last education
BASE 0.02 -0.03 -0.06 -0.10
Syiah -0.02 -0.04 0.05 -0.15
Ahmadiyah -0.02 0.01 -0.08 -0.13
Kristen -0.12 -0.04 -0.15 -0.11
LGBTQ+ 0.10 -0.03 0.01 -0.13
Tionghoa -0.03 0.01 -0.10 -0.11
Jewish 0.07 -0.03 -0.06 -0.05
Rohingya -0.04 -0.07 -0.10 -0.06
Disabilitas -0.00 -0.03 -0.09 -0.10
Terpolarisasi -0.02 -0.04 -0.07 -0.08
UNKNOWN 0.11 -0.10 -0.05 -0.14
Diploma
Graduate
Undergraduate
High school graduate
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

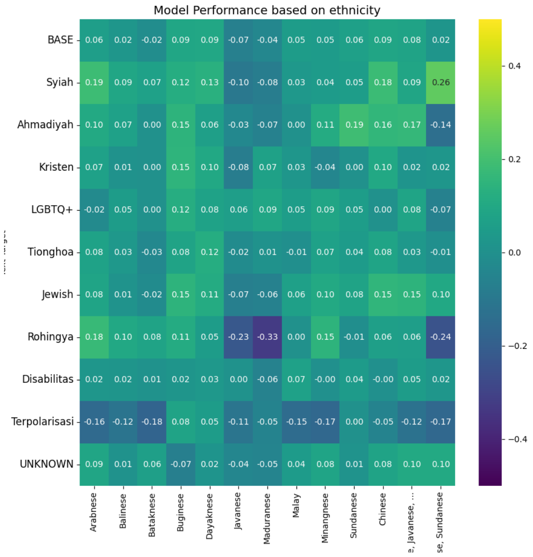

Model Performance based on lgbt

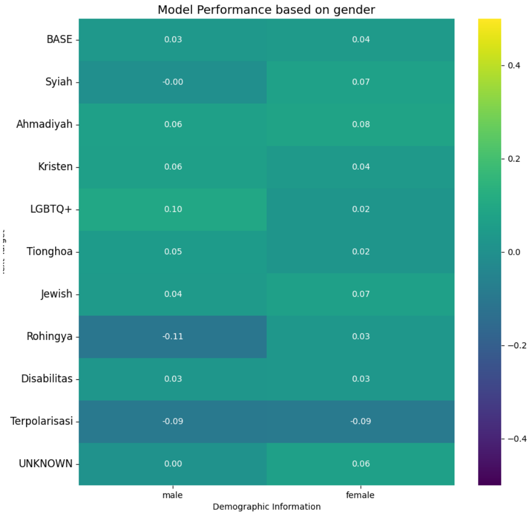

Model Performance based on disability

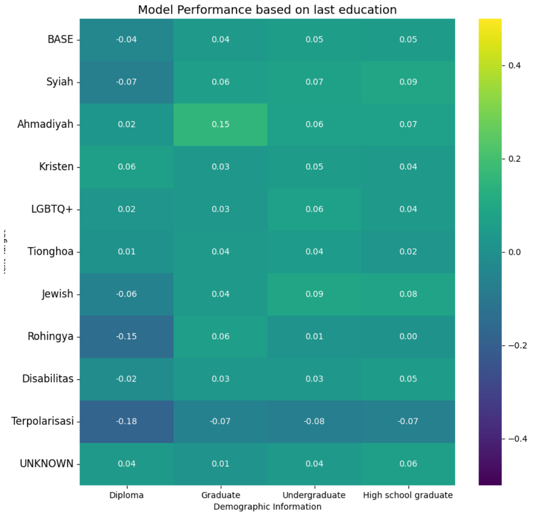

Model Performance based on political leaning

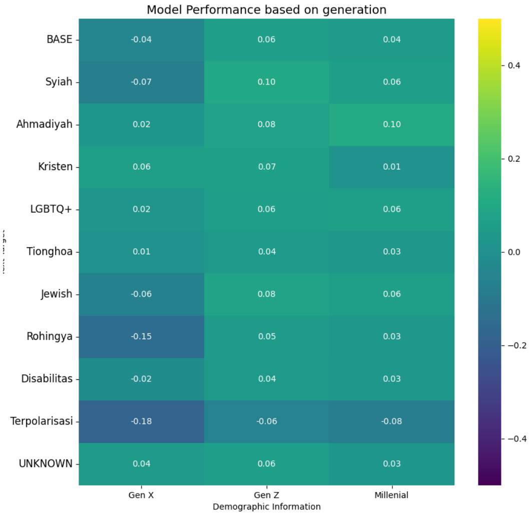

Model Performance based on domicile

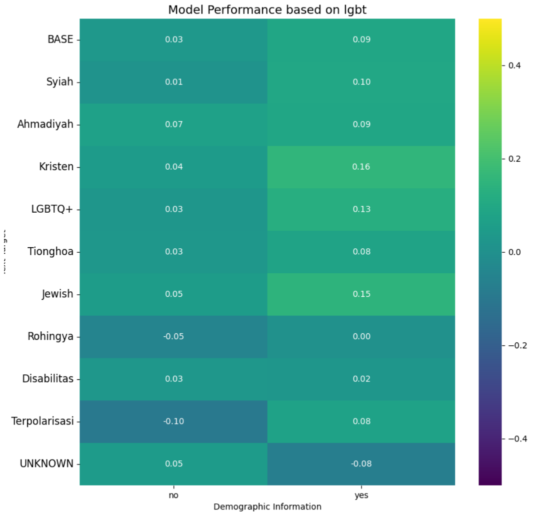

Model Performance based on religion

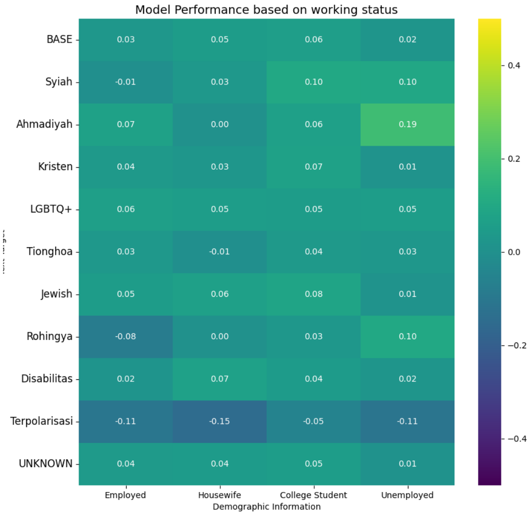

Model Performance based on ethnicity

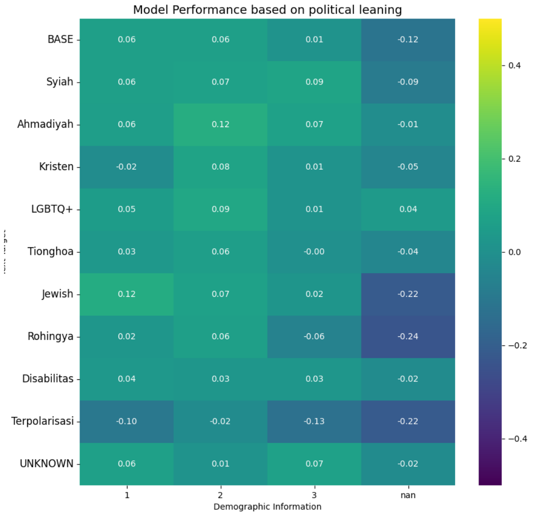

Model Performance based on gender

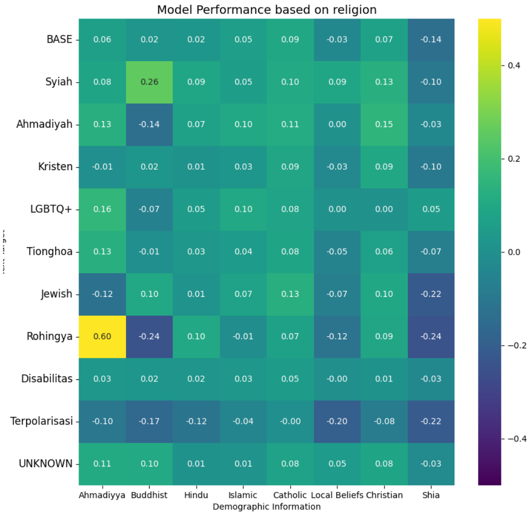

Model Performance based on last education

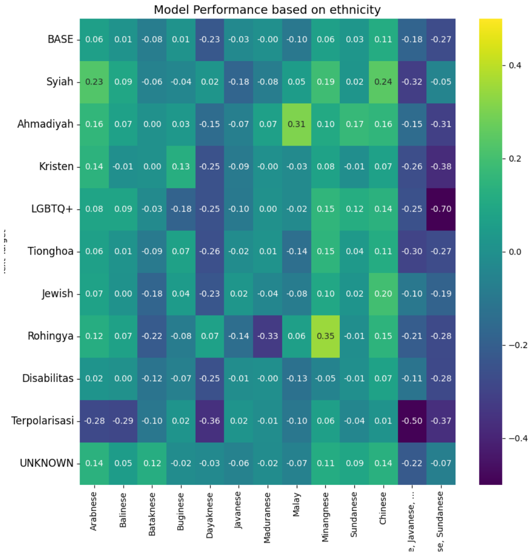

Model Performance based on generation

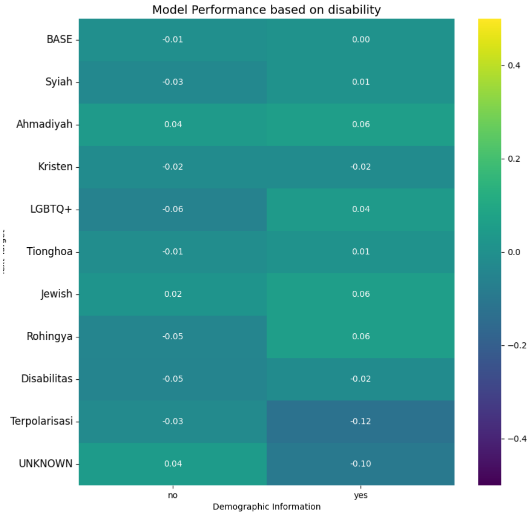

Model Performance based on lgbt

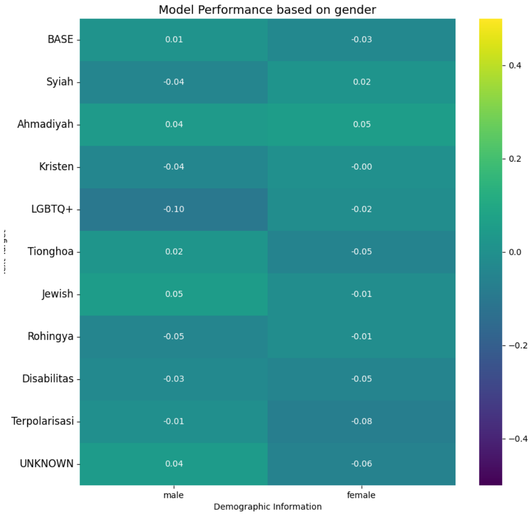

Model Performance based on working status

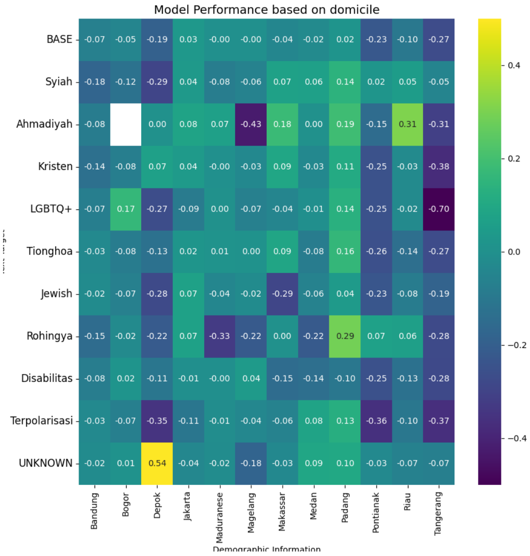

Model Performance based on political leaning

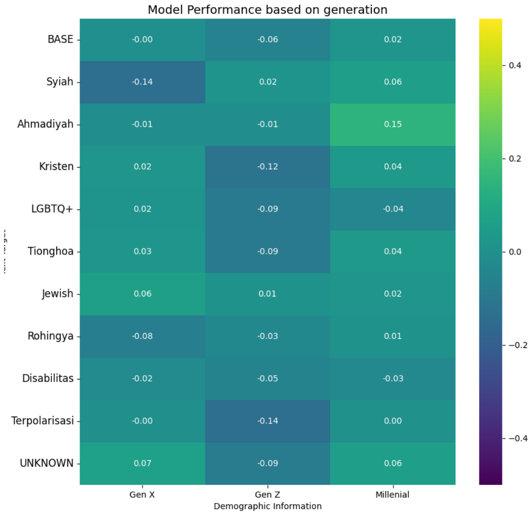

Model Performance based on religion

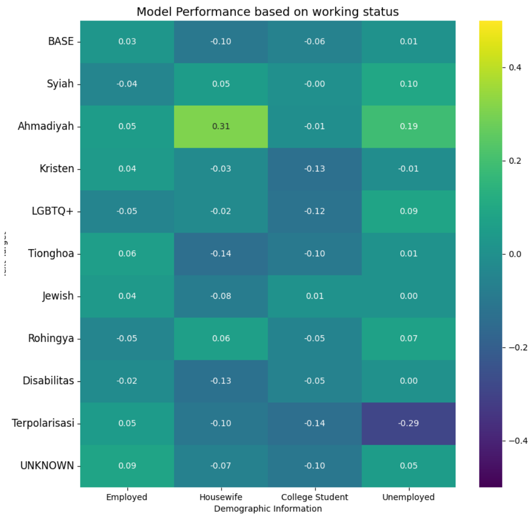

Model Performance based on ethnicity

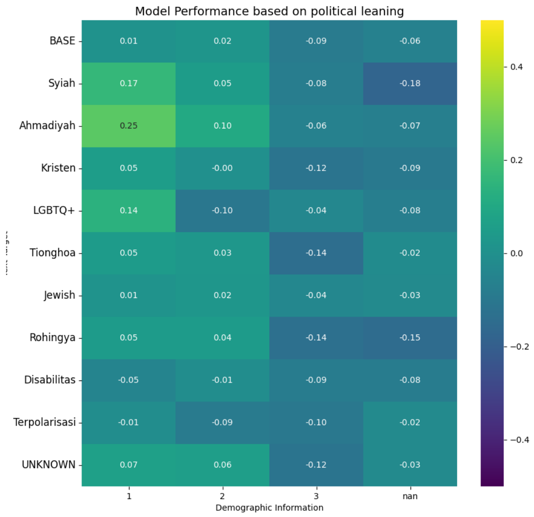

Model Performance based on disability

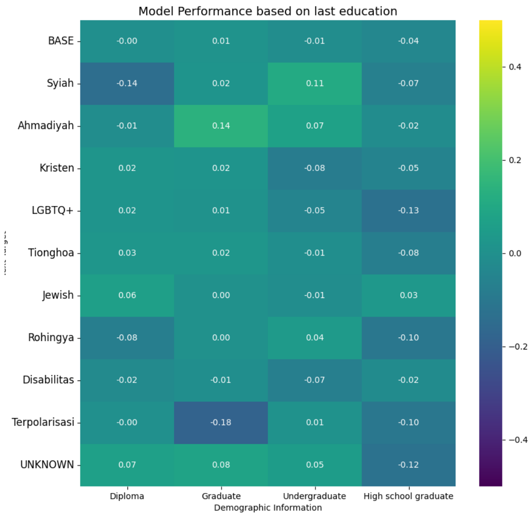

Model Performance based on gender

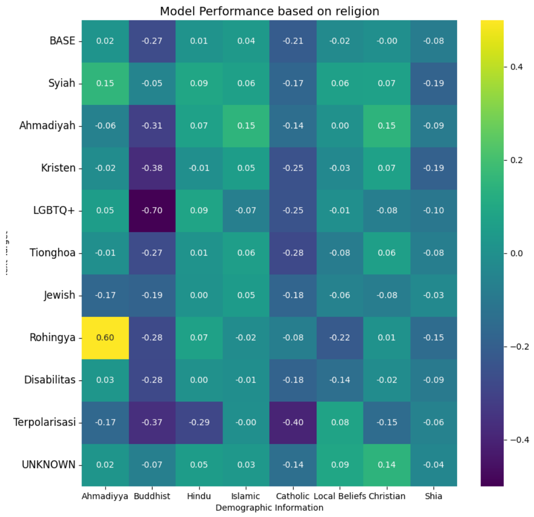

Model Performance based on domicile

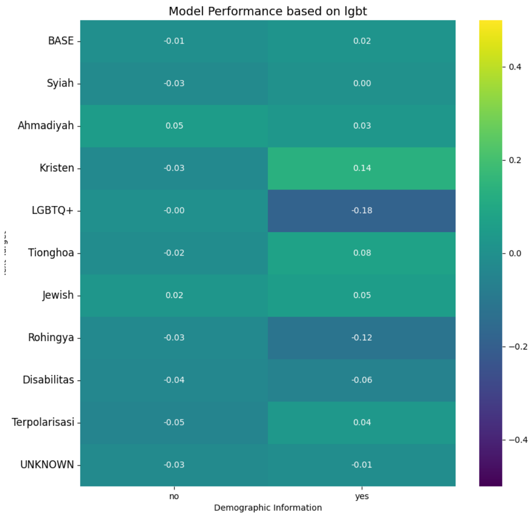

Model Performance based on generation

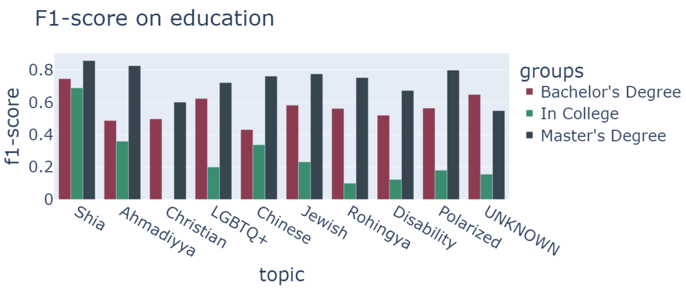

Model Performance based on working status
BASE 0.03 -0.10 -0.06 0.01
Syiah -0.04 0.05 -0.00 0.10
Ahmadiyah 0.05 0.31 -0.01 0.19
Kristen 0.04 -0.03 -0.13 -0.01
LGBTQ+ -0.05 -0.02 -0.12 0.09
Tionghoa 0.06 -0.14 -0.10 0.01
Jewish 0.04 -0.08 0.01 0.00
Rohingya -0.05 0.06 -0.05 0.07
Disabilitas -0.02 -0.13 -0.05 0.00
Terpolarisasi 0.05 -0.10 -0.14 -0.29
UNKNOWN 0.09 -0.07 -0.10 0.05
Employed
Housewife
College Student
Unemployed
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

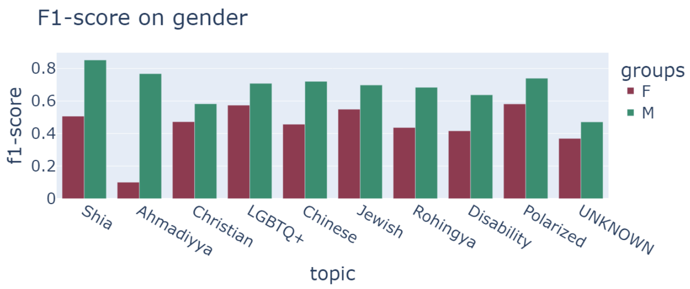

Model Performance based on political leaning
BASE 0.01 0.02 -0.09 -0.06
Syiah 0.17 0.05 -0.08 -0.18
Ahmadiyah 0.25 0.10 -0.06 -0.07
Kristen 0.05 -0.00 -0.12 -0.09
LGBTQ+ 0.14 -0.10 -0.04 -0.08
Tionghoa 0.05 0.03 -0.14 -0.02
Jewish 0.01 0.02 -0.04 -0.03
Rohingya 0.05 0.04 -0.14 -0.15
Disabilitas -0.05 -0.01 -0.09 -0.08
Terpolarisasi -0.01 -0.09 -0.10 -0.02
UNKNOWN 0.07 0.06 -0.12 -0.03
1
2
3
nan
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

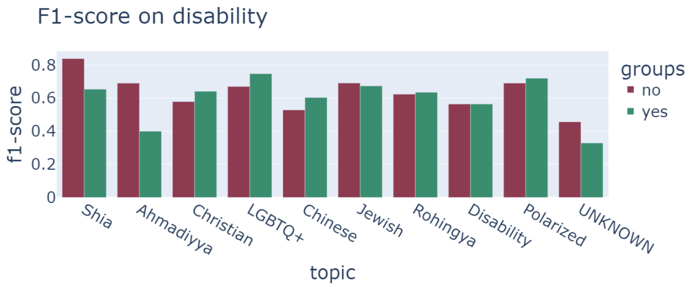

Model Performance based on last education
BASE -0.00 0.01 -0.01 -0.04
Syiah -0.14 0.02 0.11 -0.07
Ahmadiyah -0.01 0.14 0.07 -0.02
Kristen 0.02 0.02 -0.08 -0.05
LGBTQ+ 0.02 0.01 -0.05 -0.13
Tionghoa 0.03 0.02 -0.01 -0.08
Jewish 0.06 0.00 -0.01 0.03
Rohingya -0.08 0.00 0.04 -0.10
Disabilitas -0.02 -0.01 -0.07 -0.02
Terpolarisasi -0.00 -0.18 0.01 -0.10
UNKNOWN 0.07 0.08 0.05 -0.12
Diploma
Graduate
Undergraduate
High school graduate
Demographic Information
0.4
0.2
0.0
−0.2
−0.4

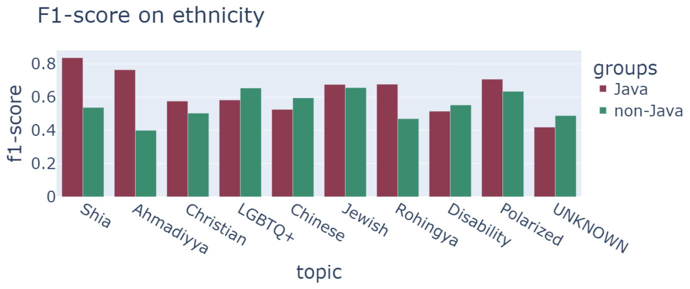

Model Performance based on religion
BASE 0.02 -0.27 0.01 0.04 -0.21 -0.02 -0.00 -0.08
Syiah 0.15 -0.05 0.09 0.06 -0.17 0.06 0.07 -0.19
Ahmadiyah -0.06 -0.31 0.07 0.15 -0.14 0.00 0.15 -0.09
Kristen -0.02 -0.38 -0.01 0.05 -0.25 -0.03 0.07 -0.19
LGBTQ+ 0.05 -0.70 0.09 -0.07 -0.25 -0.01 -0.08 -0.10
Tionghoa -0.01 -0.27 0.01 0.06 -0.28 -0.08 0.06 -0.08
Jewish -0.17 -0.19 0.00 0.05 -0.18 -0.06 -0.08 -0.03
Rohingya 0.60 -0.28 0.07 -0.02 -0.08 -0.22 0.01 -0.15
Disabilitas 0.03 -0.28 0.00 -0.01 -0.18 -0.14 -0.02 -0.09
Terpolarisasi -0.17 -0.37 -0.29 -0.00 -0.40 0.08 -0.15 -0.06
UNKNOWN 0.02 -0.07 0.05 0.03 -0.14 0.09 0.14 -0.04
Ahmadiyya
Buddhist
Hindu
Islamic
Catholic
Local Beliefs
Christian
Shia
Demographic Information
0.4
0.2
0.0
−0.2
−0.4


Model Performance based on lgbt
BASE -0.01 0.02
Syiah -0.03 0.00
Ahmadiyah 0.05 0.03
Kristen -0.03 0.14
LGBTQ+ -0.00 -0.18
Tionghoa -0.02 0.08
Jewish 0.02 0.05
Rohingya -0.03 -0.12
Disabilitas -0.04 -0.06
Terpolarisasi -0.05 0.04
UNKNOWN -0.03 -0.01
no
yes
Demographic Information
0.4
0.2
0.0
−0.2
−0.4


F1-score on education
f1-score
0.8
0.6
0.4
0.2
0
Shia
Ahmadiyya
Christian
LGBTQ+
Chinese
Jewish
Rohingya
Disability
Polarized
UNKNOWN
topic
groups
Bachelor's Degree
In College
Master's Degree


F1-score on gender
f1-score
0
0.2
0.4
0.6
0.8
Shia
Ahmadiyya
Christian
LGBTQ+
Chinese
Jewish
Rohingya
Disability
Polarized
UNKNOWN
topic
groups
F
M


F1-score on disability
f1-score
0
0.2
0.4
0.6
0.8
Shia
Ahmadiyya
Christian
LGBTQ+
Chinese
Jewish
Rohingya
Disability
Polarized
UNKNOWN
topic
groups
no
yes


F1-score on ethnicity
f1-score
0
0.2
0.4
0.6
0.8
Shia
Ahmadiyya
Christian
LGBTQ+
Chinese
Jewish
Rohingya
Disability
Polarized
UNKNOWN
topic
groups
Java
non-Java