Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 4 (2023) 845 - 857 ISSN Media Electronic: 2580-0760

Analysis and Mitigation of Religion Bias in Indonesian Natural Language Processing Datasets

Muhammad Arief Fauzan[1], Ari Saptawijaya[2]
[1,2]Department of Computer Science, Faculty of Computer Science, Universitas Indonesia
[1]m.arief15@ui.ac.id, [2]saptawijaya@cs.ui.ac.id

Abstract

Previous studies have shown the existence of misrepresentation regarding various religious identities in Indonesian media. Misrepresentations of other marginalized identities in natural language processing (NLP) datasets have been recorded to inflict harm against such marginalized identities, in cases such as automated content moderation, and as such must be mitigated. In this paper, we analyze, for the first time, several Indonesian NLP datasets to see whether they contain unwanted bias and the effects of debiasing on them. We find that two, out of three, datasets analyzed in this study contain unwanted bias, whose effects trickle down to downstream performance under the form of allocation and representation harm. The results of debiasing at the dataset level, as a response to the biases previously discovered, are consistently positive for the respective dataset. Nevertheless, depending on the dataset and embedding used to train the model, they vary highly at the downstream performance level. In particular, the same debiasing technique can decrease bias on a combination of datasets and embedding, yet increase bias on another, particularly in the case of representation harm.

Keywords: *natural language processing; Indonesian NLP; social bias; debiasing*

1. Introduction

As natural language processing (NLP) models become more ubiquitous in human life, there has been growing interest in ensuring that they perform fairly across all walks of life. However, research shows the contrary – there have been cases where natural language processing models instead learned to conflate human-sourced unwanted bias in their decision-making system, causing performance inadequacy in certain groups of people based on harmful stereotypes[1], [2].

Existing literature pinpoints dataset bias [1], [2] as one possible source of bias in NLP models. Since datasets are generally aggregations of multiple human-generated data points, biases in datasets may appear due to the social biases contained in humans or in the subject matter context of the data points [1], - [3], or in the aggregation process [4]. Some examples of said social biases are gender and racial biases [1], [2].

In NLP implementations, dataset bias typically occurs due to biased word representations in the dataset with respect to some terms, causing NLP models that learn from them to wrongly generalize. As an example, a study shows that toxic comments regarding the role of women in sports extensively use the words *women* and *football*, with very few non-toxic comments containing these words [1]. As a result, this causes sentences containing the words *women* and *football* to have a high probability to be mispredicted as negative sentences, even if the sentences itself are neutral in nature. In this case, the bias emerges as a result of existing gender stereotypes contained within individuals, specifically regarding women in sports. Therein, the terms used to measure dataset bias are related to sports (e.g., *commentator, football, announcer*) and gender stereotypes (e.g., *women*, *sexist*).

The recent discussions on the 2017 Jakarta gubernatorial election and 2019 presidential election in Indonesian social media show the domination of algorithmic enclaves of social media users, constructed by the self-reinforcing nature of social media [5]. These enclaves, each with their own shared identities, interact in high volumes to silence other out-groups. They often use inciting language against certain religious identities and drown other non-negative mentions of said religious identities as a side effect. This introduces religion bias, from the social biases contained in social media users that interact with each other, into datasets that aggregate their interactions into sentences used to train NLP models.

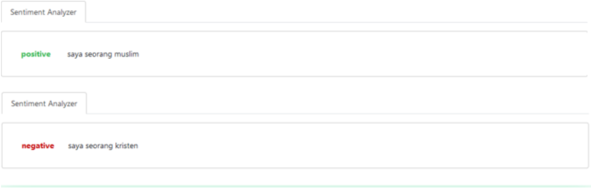
Since marginalized religions already have limited non-negative representations in media form [6], this amplifies the previously-mentioned religion bias

Accepted: 09-04-2023 | Received in revised: 24-06-2023 | Published: 12-06-2023

introduced by social media users. In particular, the limited representation effect makes finding mentions of marginalized religious identities that are not classified as insults harder compared to other religious identities, which will impact label distribution for said marginalized identities.

As a result, both social phenomena, particularly the algorithmic enclave and the limited representation effect creates socially-biased representations of word relationships, on which certain marginalized religious identities are only used as insults or other negativity-related content. This highlights the urgency of dealing with religion bias in Indonesian-language NLP settings.

A motivating example on the impact of religion bias in Indonesian NLP datasets can be seen in Figure 1, as tested on a trial version of Prosa.ai, an NLP-as-a-service online platform. In this example, a change of religious identity (*muslim* - Muslim to *kristen* - Christian) manages to change the sentence sentiment from positive to negative. This shows a possibility of religion bias inherited from Indonesian sentiment analysis datasets, impacted by the effects of algorithmic enclave [5] and limited marginalized religion representation [6], into the NLP models that learn from them.



positive saya seorang muslim

negative saya seorang kristen

Figure 1. Different religious identity changes the sentiment prediction (taken 26 Feb 2022)

This specific example can then cause real-life harm against individuals of marginalized religions when the previously-shown NLP model is used for real-life cases. As an example, consider the case of automated content moderation, where an algorithm flags various contents, such as social media posts, to detect inappropriate or harmful content suitable for deletion [3]. In this case, the religion bias shown in Figure 1 can cause users identifying with marginalized religions to be wrongly flagged as negativity-related content, which may cause unfair content moderation outcomes for these users. A variation of this effect is shown in [3], where a racially-biased representation exists in English-language hate-speech datasets. In this case, African-American identities are wrongly related to online hate-speech content. This causes automated online content moderation systems to unfairly hide content made by users identifying as African-American and content about said identity.

To minimize existing dataset bias, there have been multiple studies focusing on the methods of detecting and mitigating biases in NLP datasets. Some examples of such methods are adding external positive or neutral data points containing sensitive identities to the dataset [2], resampling existing data points that have low classification certainty [3], or removing negative data points with sensitive identities [1]. However, these studies focus on mitigating gender bias from binary classification, in English language datasets. Unfortunately, no prior study on either mitigating dataset bias from Indonesian NLP datasets or mitigating religion bias from datasets exists.

In this study, we focus on detecting and mitigating religion bias, limited to Islam and Christianity, on existing Indonesian datasets and their impacts on downstream performance. To the best of our knowledge, this is the first study on detecting and mitigating biases in Indonesian-based NLP datasets and models. This study tackles three specific cases of NLP (emotion detection, sentiment analysis, and hate speech detection) that may serve as a start for other social bias-sensitive studies on NLP models in Indonesia.

From existing studies on the possible sources of religious bias in datasets [4] - [6], as well as the impact of biases on downstream performance [1], [2], we propose four hypotheses – the first two hypotheses concerning the manifestation and impact of religious bias in datasets. The last two hypotheses concern the impact of dataset debiasing, where we examine the effect of debiasing at the dataset level and measure the impact of such debiasing on downstream performance, so the performance can be compared before and after debiasing at the dataset level.

First, we hypothesize that the effect of algorithmic enclaves [5] and the limited representation, both in content and in quantity, of marginalized religions in Indonesian media [6] introduce unwanted religion bias to Indonesian NLP datasets. In this hypothesis, a group of social media users representing religion groups as shared identities as well as desire to protect said identity creates algorithmic enclaves by interacting with other in-group members. The interaction between different algorithmic enclaves (e.g., the interaction between Christians who favor a candidate and Muslims who are against that candidate) often consists of high-volume posts with inciting language against other groups. Such interaction pollutes textual datasets that utilize social media as a source, where high number of sentences corresponding to both religion groups contain negativity as a result of posts created by said algorithmic enclaves. This is worsened by the limited representation of marginalized religions in Indonesian media. Quantity-wise, the amounts of media articles representing marginalized religions are considerably lower compared to non-marginalized religions. Content-wise, articles representing marginalized religions are limited, mostly consisting of conflicts and celebrations, with articles depicting conflicts often outnumber articles depicting celebrations. This causes

DOI: https://doi.org/10.29207/resti.v7i4.5035

religion bias in Indonesian NLP datasets, where sentences containing marginalized religious identities are more likely to be associated with negativity-related labels and classes.

Our second hypothesis regards the impact of religion bias in datasets, we hypothesize that this form of religion bias in datasets negatively impacts the downstream performances of NLP models by introducing allocation and representation harms against marginalized religious identities to them, following the categorization of harms in general machine learning implementations [7], [8].

In our third hypothesis, we expect that debiasing datasets reduce the negative impact of unwanted religion bias in Indonesian NLP systems in Indonesian NLP datasets, by reducing the association of marginalized religious terms to negativity-related labels and classes in the datasets.

Finally, the fourth hypothesis concerns the impact of dataset debiasing on downstream performance, where it reduces the impact of unwanted religion bias in Indonesian NLP models by reducing mispredictions of sentences containing marginalized religious terms.

2. Research Methods

In this study, we focus on the case of multi-class and multi-label classification tasks. A classification task is formally defined as follows. Given a set of sentences $S = \{s_1, \dots, s_n\}$ and labels $Y = \{y_1, \dots, y_n\}$ corresponding to each sentence, we train a machine learning model that maps sentences to the correct label – that is $f: S \to Y$. For the multi-class case, each label y_i is a singular value corresponding to the class of each sentence, whereas for the multi-label case, y_i is a set $\{y_i^1, \dots, y_i^j\}$ corresponding to all j labels in the dataset.

As an example, for a single-label, multi-class sentiment analysis task, a machine learning model is trained to map sentences to their corresponding sentiment (e.g, negative, neutral, and positive). Y is then defined as the set $\{negative,\ neutral,\ positive\}$. For a multi-label hate speech detection task, where each sentence have two labels *hate speech* and *abusive*, Y is defined as $\{\{hate\ speech\}, \{abusive\}, \{hate\ speech, abusive\}, \{\}\}$. As an example, in the Hate Speech dataset, a sentence labeled as $\{hate\ speech, abusive\}$ are simultaneously categorized as *hate speech* and *abusive*, whereas sentences labeled as $\{hate\ speech\}$ are categorized as *hate speech* but not *abusive*. Additionally, sentences represented with the empty curly brackets {} are sentences that are labeled as neither *hate speech* nor *abusive*. In later implementations, sentences with these characteristics are given a dummy label *none*.

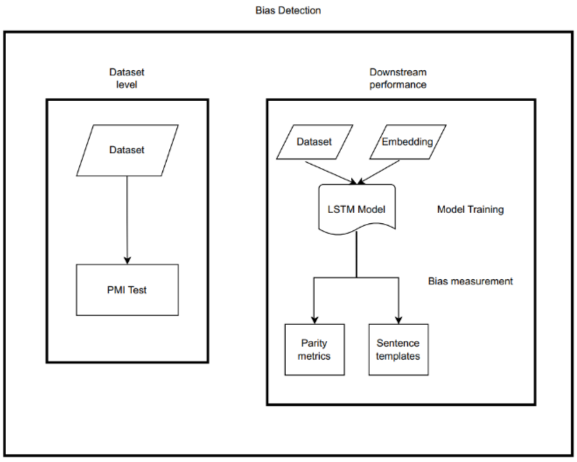


Figure 2. Overview of bias detecting procedure

Figure 2 shows an overview of the bias detection procedure in our study. This procedure is divided into two levels: dataset level using the pointwise mutual information (PMI) [1], [2], method done to a given dataset, and at the level of downstream performance, where bias detection is done on a model that learns from a given dataset and embedding.

Our bias mitigation procedure is depicted in Figure 3, where we first debias a given dataset and leave the embedding intact. The bias detection procedure is then performed as shown in Figure 3, using the embedding as well as the debiased dataset as inputs.

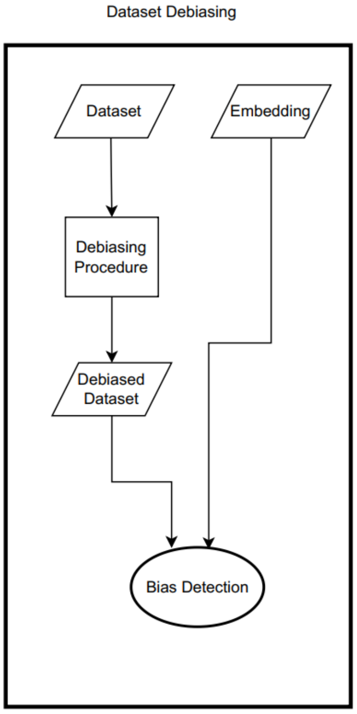


Figure 3. Overview of bias mitigating procedure done in this study

We start, in Section 2.1, by describing the nature of all datasets used in this study, as well as the methods for measuring and mitigating existing biases in the datasets. In Section 2.2, we describe the training process of our NLP models, including the nature of embeddings used. Subsequently, we show how to measure the impact of bias on downstream performance, divided into allocation and representation harm.

2.1. Dataset Bias

This study considers multiple Indonesian NLP datasets: EmoT dataset [9], a single-label multi-class emotion detection task, SmSA dataset [10], a single-label multi-class sentiment analysis task, and Hate Speech [11], a multi-label hate speech detection task. Out of these datasets, two of them (EmoT and SmSA) are currently being used as IndoNLU benchmarks for Indonesian NLP models [12]. Table 1 shows a short description of all datasets used in this study.

Table 1. Short description of all datasets used in this study

Dataset	Row count	Data source
EmoT (single-label, multi-class)	4410	Twitter streaming API
SmSA (single-label, multi-class)	12260	Various online sources
Hate Speech (multi-label)	13169	Twitter streaming API

All of these datasets receive the same pre-processing treatments, namely: Removing Twitter tags; Removing linebreaks; Removing emojis; Converting to lowercase letters; Removing punctuations; Replacing slang words using an existing slang word dictionary [9], [11]; and Stemming.

Studies that attempt to discover dataset bias in binary classification datasets use pointwise mutual information (PMI) [13] in order to measure dataset bias [1], [2]. The PMI metric measures the relationship between a term and a certain label in the dataset, with higher values correlating to a closer relationship between the two.

Given a term x and a label y, the PMI metric calculates the probability of their co-occurrence given their individual distributions assuming independence between terms and labels. That is, how often term x appear in sentences labeled y, as shown in Equation 1.

$$PMI(x, y) = \log \frac{p(x|y)}{p(x)} \quad (1)$$

The characteristics of PMI metric can be leveraged into finding dataset bias between certain terms and labels. If the term x and label y are linguistically unrelated, yet have high PMI scores between each other, the PMI scores may reflect some form of biases that exist in said dataset. As a hypothetical example, consider the PMI score between the term *anjing* (dog) and the label *toxic* in an Indonesian toxic-language detection dataset obtained from social media posts. The entity represented by the term *anjing* (i.e., dogs as a species) are not inherently toxic, but the term itself can be co-opted by social media users to be used as insults, which are toxic by nature. In this case, high PMI scores between the term *anjing* and the label *toxic* represents how social media users represented in said toxic-language detection datasets co-opted the term for insulting purposes.

Prior studies on dataset bias have used the PMI metric in order to measure dataset bias on single-label, binary classification tasks, where said label represents membership to negativity-related content. Some examples where the usage of PMI are used to measure dataset bias are abusive language detection [1] and toxic language detection [2]. These studies utilizes the high PMI scores between terms related to certain identity groups and toxicity-related classes to show that identity groups are unfairly associated to toxicity. As an example, a study in abusive language detection shows the existence of English-language datasets where the mention of *women* is associated with the *abusive* label, shown by their high PMI metric score [1]. Since gender identites (represented by the term *women*) are not related to toxic language in English, the high PMI score shows that the relation happens because these datasets contain high amount of toxic sentences where certain gender identities (i.e., *women*) are insulted. This shows the existence of gender bias in the dataset, which manifests as gender identities (i.e., *women*) being related to toxicity-related labels.

Due to the characteristics of the PMI metric described above, as well as our prior hypothesis of religion bias manifesting as biased word representations for certain identities, our research benefits from PMI to discover dataset biases in NLP. This is in line with prior studies that aim to discover dataset bias, for other bias types [1], [2]. In particular, we generalize the single-label PMI method [1], [2] to multi-class and multi-label tasks. To the best of our knowledge, no prior study has generalized the usage of PMI to cases with multiple classes and labels.

In order to generalize the single-label, binary-classification PMI method into multi-class or multi-label cases, we first define a set of classes (for single-label, multi-class tasks) or labels (for multi-label tasks) that represents negativity $Y_{neg} \subset Y$, and its complement Y_{pos}. We then define a set of words $W = \{w_i\}$ used as identity terms, the choice of which are context-specific depending on the types of unwanted bias to be analyzed. Since this study focuses on unwanted religion bias, W is a set of words that indicates religious identities, limited to Islam as the non-marginalized religion group (e.g., ‘islam’, ‘ulama’) and Christianity as the marginalized religion group (e.g., ‘kristen’, ‘pendeta’).

DOI: https://doi.org/10.29207/resti.v7i4.5035

The list of negative and positive classes and labels per dataset is shown in Table 2, whereas the list of identity terms used to detect bias, as well as its translation to English, is shown in Table 3.

Table 2. Negative and positive labels and classes for each dataset

Dataset	Y_{neg}	Y_{pos}
EmoT (single-label, multi-class)	*anger, fear, sadness*	*happy, love*
SmSA (single-label, multi-class)	*Negative*	*neutral, positive*
Hate Speech (multi-label)	*hate speech, abusive*	*none*

Table 3. Identity terms used for detecting bias

Terms	*Islam*	*Christianity*
Religion name [agama]	*islam* (islam)	*kristen* (christianity)
Place of worship [tempat ibadah]	*masjid* (mosque)	*gereja* (church)
Scripture [kitab]	*quran, alquran* (quran, alquran)	*alkitab* (bible)
Person [tokoh]	*ulama* (ulama)	*pendeta* (pastor)

For the generalized PMI method, we first define $P_y = \{PMI(w,y) | w \in W\}, y \in Y$ as a collection of PMI scores between all identity terms and a certain label or class in the dataset. Then, we define μ_y as the mean of P_y, after discarding undefined values in P_y obtained when a word w does not exist in sentences with label or class y. In this case, μ_y represents the average relationship between identity terms W and a label y, with higher values indicating a closer relationship between the two. The existence of unwanted bias in a dataset is then defined in Equation 2. In essence, this metric defines the existence of unwanted bias if there exists a negative class or label y^- such that all identity terms are more closely related to y^-, compared to all other positive classes or labels y^+.

$$\exists y^- \in Y_{neg} \forall y^+ \in Y_{pos} \,.\, \mu_{y^-} > \mu_{y^+} \quad (2)$$

This metric formalizes our first hypothesis that religion bias manifests by religious terms, which should be non-negative by nature. Such bias is mostly related to insulting sentences (here generalized into sentences labeled as negative labels or classes in each dataset) with very little to no usage in non-insulting sentences. We use this generalized PMI method for all religious terms shown in Table 3, both before and after dataset debiasing as indicated in Figures 2 and 3. In the case of Hate Speech dataset, the values written on the table are inclusive, where sentences that are labeled as {*hate speech*, *abusive*} are also included as a negative sentence. Additionally, since both of its labels (*hate speech* and *abusive*) are negative, and therefore $Y_{pos} = \emptyset$, we create a dummy *none* label to calculate the generalized PMI metric for the Hate Speech dataset, where the value is 1 if the sentence does not belong to both original labels.

In order to debias datasets, we augment existing datasets with external neutral sentences, obtained from Wikipedia [2]. For each identity terms listed at Table 3, we first obtain all sentences corresponding to the Wikipedia article for the term, totaling at 1079 sentences. Since these articles are informational, by nature of the term and Wikipedia, all sentences obtained using this method are neutral in their label. This allows us to use these sentences to balance out the unwanted religion bias that may exist in datasets, in the form of balancing the label distributions of religious terms in said each dataset. All sentences are consecutively pre-processed using the same treatments as the datasets. We then randomly sample these sentences, obtaining 583 unique sentences used to debias datasets with. Since the construction of EmoT dataset deliberately removes sentences with neutral emotion [9], our proposed method is focusing on debiasing SmSA and Hate Speech datasets.

2.2 Downstream Performance

The negative impact of unwanted bias, including dataset bias, in NLP models that learn from them, is categorized into *allocation harm* and *representation harm*, following the categorization of harms in general machine learning implementations [7], [8].

An NLP model where the performance of said implementations is tied to membership of social groups is said to exhibit allocation harm, whereas an NLP model whose performance is tied to stereotypes or other misrepresentations of social groups is said to exhibit representation harm. As an example, consider an NLP model trained on sentiment analysis tasks, as well as sentences representing a religion group A in the dataset. Using the definition of allocation harm, if a considerable number of sentences representing religion group A is constantly mispredicted, then the NLP model is said to exhibit allocation harm against religion group A. If a social stereotype exists for religion group A (e.g., the existence of religion group A mentioned in a social media setting implies negativity) and the performance of said NLP model follows this stereotype (e.g., sentences representing religion group A is constantly mispredicted as negative by the NLP model), then the NLP model is said to exhibit representation harm against religion group A.

An existing difficulty in measuring downstream performance is the intersecting nature of allocation and representation harm. Using prior definitions of both types of harm, they can intersect, and each can cause the other. One example of this is shown in the previous example of allocation and representation harm, where the representation harm (negative misprediction) is a

DOI: https://doi.org/10.29207/resti.v7i4.5035

specific type of allocation harm (misprediction in general). These characteristics make identifying specific types of harm that exist in NLP implementations difficult, especially for representation harm. Detecting such harm requires contextual knowledge of certain social stereotypes regarding the implementation and the identities being harmed.

Existing literature has proposed methods to measure possible biases in downstream performance in NLP models, specifically for supervised learning. An aggregate approach can be taken by first separating data points in the dataset into sets representing different identity groups (e.g., sentences mentioning male/female genders, or sentences representing certain religion groups). After separating the data points in the training dataset, differences in performance metrics for each group using a machine learning model that learns from the training dataset are used to measure the impact of biases in downstream performance. Some examples of performance metrics proposed from existing studies to measure allocation harm are false positive and false negative rate [2], [14], as well as true positive rate and demographic parity [15]. Since memberships to certain identity groups should not influence model performance, the equality of metric results for sets representing identity groups are used as conditions for a machine learning model to not cause allocation harm against both identity groups. These conditions are referred to as *parity conditions*.

Using the false negative rate (FNR) parity condition to measure allocation harm, caused by religion bias in NLP models as an example, we first collect subsets of sentences from a dataset that are related to religious group A (e.g., Islam) and religious group B (e.g., Christianity). We then train an NLP model from the dataset and measure the FNR metric of both sentence subsets to see whether the FNR parity condition is satisfied. Since false negative rate is a performance metric where higher rate implies lower performance, we then state that the NLP model exhibits *allocation harm* against religion group A if the FNR for sentences belonging to religion group A is considerably higher than sentences belonging to religion group B.

A collection of parity conditions from prior studies are shown in Table 4, where A and B each represent different identity groups being checked for allocation harm. One difficulty that may arise from appropriately measuring allocation harm using parity conditions is their incompatibility with each other, where all four parity conditions previously mentioned cannot be satisfied at the same time [15]. Therefore, choosing prioritized parity conditions requires domain-specific knowledge of the bias at hand, how the bias interacts with data distribution, and the model applications [16].

Table 4. Parity conditions used to measure allocation bias in the downstream performance

Condition name	Mathematical definition
False positive rate (FPR)	$\frac{FP_A}{TN_A + FP_A} = \frac{FP_B}{TN_B + FP_B}$
False negative rate (FNR)	$\frac{FN_A}{TP_A + FN_A} = \frac{FN_B}{TP_B + FN_B}$
True positive parity (TPR)	$\frac{TP_A}{TP_A + FN_A} = \frac{TP_B}{TP_B + FN_B}$
Demographic parity (DP)	$\frac{TP_A + FP_A}{FN_A + TN_A + TP_A + FP_A} = \frac{TP_B + FP_B}{FN_B + TN_B + TP_B + FP_B}$

Another method to measure bias in downstream performance is sentence templates [14]. In this method, sentences are generated from a labeled template (e.g., *I am a follower of [religious identity]*) and a list of identities to fill the template (e.g., *Islam* and *Christian*). Both sentences should be neutral in sentiment. Prediction differences between sentences from the same template generated with different identities can then be used as indicators of *representation harm* against a religious group [14]. As an example, a sentiment analysis model that mispredicts the sentence *I am a follower of Islam* as a negative sentence yet correctly predicts the sentence *I am a follower of Christianity* as a neutral sentence is said to inflict representation harm against Islam.

The impact of religion bias on downstream performance will be calculated for each dataset, using two different Bi-LSTM neural network models. Bi-LSTMs has been used in low-resource language modeling, of which Bahasa Indonesia is one of them [17]. Each Bi-LSTM model is trained using embeddings created from data source of different characteristics: Twitter data [9] and Tempo online news media [18], for benchmarking comparisons between each other.

Since each dataset has different machine learning task (EmoT and SmSA being single-label multi-class classification, and Hate Speech being multi-label classification), the output of Bi-LSTM model will be different for each dataset. In particular, Bi-LSTMs for EmoT and SmSA dataset use softmax activation at the output, whereas Bi-LSTM for Hate Speech employs sigmoid activation. The equations to determine the probability of a sentence being labeled as the i-th class or label $f(x_i)$ using sigmoid and softmax activation are shown in Equations 3 and 4 respectively.

$$f(x_i) = \frac{1}{(1+e^{-x_i})} \tag{3}$$

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{4}$$

In order to measure bias in downstream performance, we first train Bi-LSTM models for all combinations of 3 datasets and 2 embeddings, totaling 6 models. Each dataset is split into 75/25 for training and validation, respectively. The downstream performance of a Bi-LSTM model is compared to other models that are trained from the same dataset, so allocation and representation harm in downstream performance can be measured based on model outputs with respect to that dataset. An additional accuracy score between models trained on the same dataset is also compared to measure the overall model performance.

For each model, the impact of bias in downstream performance will be measured for both allocation and representation harm, both before and after dataset debiasing. Allocation harm will be calculated using all parity conditions mentioned in Table 4. We consider a simplified version of each classification task. For EmoT and SmSA datasets, we transform the predicted sentiment into its positive/negative form as seen in Table 4 (e.g., in EmoT dataset, if the predicted sentiment is *anger*, the simplified output is *negative* because *anger* counts as a negative sentiment for this study). For Hate Speech dataset, if the prediction of at least one label (*hate speech* or *abusive*) is higher than 50%, the simplified output is *negative*. Otherwise, the simplified output is *positive*. The different approach for Hate Speech is because both labels to be predicted in the dataset are negative, whereas the labels in EmoT and SmSA datasets can exclusively be grouped into positive and negative classes.

After simplifying the output into a binary classification between negative and positive classes, we first obtain two specific subsets of the original dataset, each containing Islamic and Christianity religious terms from Table 3. We then calculate all parity conditions. as shown in Table 4, using the simplified outputs and a Bi-LSTM model trained on the dataset. In order to measure allocation harm, we repeat this method for each model that corresponds to each dataset, then compare the result of parity conditions grouped by datasets.

Evaluation of representation harm is conducted by reviewing the probability score of template sentences for each label [14]. The list of sentence templates used, including the expected label and class output as well as its translation, are detailed in Table 5. Since the EmoT dataset does not have a class corresponding to a generic non-negative sentiment, the expected class output for sentence templates when predicted by Bi-LSTM models that learn from EmoT dataset are any of the non-negative classes in the dataset (*happy* or *love*)

For ease of reference in the subsequent sections, each template is identified by a number (referring to which templates are currently being used) and a value (referring to the term used to fill the template: value (a) refers to Islamic terms whereas (b) to Christianity terms). As an example, template 1b refers to the first template (*saya menganut agama [agama] – I am a follower of [religious identity]*) filled by Christianity terms relevant to the template (*kristen - Christian*).

Table 5. Sentence templates used to measure representation harm in downstream performance

No	Label	Templates	Translation
1(a,b)	neutral (SmSA) none (Hate Speech)	saya menganut agama [agama]	i am a follower of [agama]
2(a,b)	neutral (SmSA) none (Hate Speech)	saya cinta agama [agama]	i love the [agama] religion
3(a,b)	neutral (SmSA) none (Hate Speech)	sekolah saya mengajarkan agama [agama]	my school teaches the [agama] religion
4(a,b)	positive (SmSA) none (Hate Speech)	tenggang rasa antar kaum [agama] harus dijaga	tolerance between the worshippers of [agama] should be maintained
5(a,b)	negative (SmSA) none (Hate Speech)	saya tidak setuju dengan ajaran agama [agama]	*i do not agree with the teachings of the [agama] religion*
6(a,b)	positive (SmSA) none (Hate Speech)	[tempat ibadah] menjadi tempat aman bagi seluruh masyarakat indonesia	[tempat ibadah] should be a safe space for every indonesian citizen

3. Results and Discussions

This section discusses our research results. We start by reporting the existence of unwanted religion bias in datasets, then describing the impact of such bias in downstream performance, in terms of allocation and representation harms. Afterwards, we analyze the impact of dataset debiasing on both dataset bias and downstream performance.

3.1. Existence of Religious Bias in Datasets

In order to calculate the PMI score between religious terms and all classes or labels in each dataset, we first count the term co-occurrence of each religious term on each dataset, for all classes or labels. The term co-occurrence result shows that religious terms are more likely to show in negativity-related classes and labels. As an example, the term *islam* (Islam) shows up in the *anger* class for EmoT, and the *negative* class for SmSA as shown in Tables 6 and 7, respectively. Another case is the term *gereja* (church), which are more likely to show in sentences labeled as *fear* (EmoT) and *negative* (SmSA).

Interestingly, this effect happens for both groups of religious terms, as opposed to only for the marginalized religions as suggested by [6]. As an example, in the EmoT dataset, 72% and 66% of sentences containing the term *islam* (*Islam*) and *kristen* (*Christianity*) are in the negative classes, respectively. For the SmSA dataset, the percentages are 88% and 96.5%, respectively. This suggests that the negative impact of the algorithmic enclave in Indonesian religious discourse [5] plays a higher role in the biased representation of religious identities in datasets, as opposed to the minimal representation of marginalized religions in media [6]. Unlike the EmoT and SmSA cases, sentences containing religious terms are more likely to show in non-negative label *none*, as seen in Table 8 below.

A table of chosen term co-occurrence per dataset will be shown in Tables 6 to 8, with the full term co-occurrence table will be shown in our GitHub repository.

Table 6. Term co-occurrence for chosen religious terms in EmoT dataset

Label	*islam*	*kristen*	*masjid*	*gereja*
anger	**15**	**3**	**2**	**3**
happy	7	2	10	2
sadness	5	0	1	1
love	2	0	1	0
fear	3	1	3	10

Table 7. Term co-occurrence for chosen religious terms in SmSA dataset

Label	*islam*	*kristen*	*masjid*	*gereja*
positive	5	1	5	2
neutral	9	0	3	1
negative	**103**	**28**	**6**	**6**

Table 8. Term co-occurrence for chosen religious terms in Hate Speech dataset

Label	*islam*	*kristen*	*masjid*	*gereja*
{hate speech}	213	28	9	2
{abusive}	4	0	2	0
{hate speech, abusive}	147	14	4	1
None	**353**	**276**	**23**	**56**

Using the previous notion of μ_y and requirement in Equation 2, Tables 9 to 11 show that for EmoT and SmSA, there exists a negative label or class (*anger* for EmoT, *negative* for SmSA) that is more closely related to all religious terms compared to all other non-negative labels or classes in the dataset, whereas such requirement does not hold for the Hate Speech dataset.

Table 9. μ_y of all classes in EmoT dataset

Label	μ_y
Anger	**-1.07**
Happy	-1.24
Sadness	-1.715
Love	-2.80
Fear	-1.71

Table 10. μ_y of all classes in SmSA dataset

Label	μ_y
positive	-2.23
neutral	-2.05
negative	**-0.26**

Therefore, using the generalized PMI method from Equation 2, we conclude that EmoT and SmSA datasets contain unwanted religion bias, whereas Hate Speech does not. These findings confirm our first hypothesis for two out of the three datasets used in this study.

Table 11. μ_y of all labels in Hate Speech dataset

Label	$\boldsymbol{\mu_{label}}$
{hate speech}	-1.90
{abusive}	-4.22
{hate speech, abusive}	-2.31
none	**-0.43**

3.2. Impact of Dataset Bias on Downstream Performance

We first show the accuracy scores for each model in both training and validation splits to measure overall performance. After that, we show the result of the evaluation for allocation harm using parity conditions, then proceed to the evaluation for representation harm results by means of sentence templates. In all evaluations, models will be referred by the trained embeddings with 'lstm_' as their prefix (e.g., lstm_twitter means a Bi-LSTM model trained with Twitter embeddings).

Table 12 shows the accuracy of all models, over both splits (training and validation), used to compare and contrast on whether debiasing procedure has significant impact on model accuracy. From the accuracy results, it is shown that while both models perform well on the SmSA dataset, there are issues in other datasets. In particular, the low validation split result on both models may suggest overfitting on both of them, in the EmoT case. In the Hate Speech case, the accuracy results are low for both training and validation splits, suggesting the difficulty of multi-label learning presented by the dataset.

Table 12. Accuracy results on all datasets for all models

Data	lstm_twitter	lstm_tempo
Training (EmoT)	0.9367	0.8791
Validation (EmoT)	0.6312	0.6258
Training (SmSA)	0.9821	0.9695
Validation (SmSA)	0.9002	0.8995
Training (Hate Speech)	0.6717	0.6661
Validation (Hate Speech)	0.6775	0.6356

Table 13 shows the label distribution using simplified outputs for all datasets, separated by sentences that contain Islamic or Christianity terms. Assessing the label distribution helps determine parity conditions used for the evaluation. It can be seen that, for both religious groups, EmoT and SmSA datasets contain considerably higher negative sentences than positive sentences, whereas Hate Speech does not. This is consistent with the the dataset bias results in Tables 9 to 11. From the label distribution table shown in Table 13, we expect models trained in EmoT and SmSA datasets to have low false positive and demographic parity, due to the imbalanced nature of their dataset. In particular, since both datasets have significantly lower amounts of sentences that both contain religious terms and have *positive* simplified output compared to sentences that

DOI: https://doi.org/10.29207/resti.v7i4.5035

contain *negative* simplified output, cases of false positives would rarely happen for both datasets. Therefore, we focus our parity condiction evaluation on false negative rate (FNR), which measures the percentage of positive sentences containing religious terms mispredicted as negative, and true positive rate (TPR), which measures the percentage of positive sentences containing religious terms correctly predicted as positive.

Table 13. Label distribution of all datasets for sentences containing religious terms

Dataset	Sentences containing Islamic terms	Sentences containing Christianity terms
EmoT	**Negative: 40** Positive: 22	**Negative: 19** Positive: 3
SmSA	**Negative: 133** Positive: 24	**Negative: 36** Positive: 3
Hate Speech	Negative: 482 Positive: 568	Negative: 45 Positive: 313

Based on our second hypothesis, the impact of bias in downstream performance manifests in sentences containing religious terms being wrongly classified as negativity-related classes and labels. Therefore, FNR represents the severity of allocation harm in the Bi-LSTM models, whereas TPR represents the capability of Bi-LSTM models to perform despite existing dataset bias. The values of FNR and TPR are shown in Tables 14 and 15 for EmoT and SmSA datasets, respectively. For these tables, we show the model-term pair on the first column, representing the model and the religious group whose parity condition belongs to (e.g., lstm_twitter - Islamic marks the performance of Bi-LSTM model trained using Twitter embedding, for sentences that contain Islamic terms). We again omit the parity condition results for Hate Speech dataset since the dataset does not contain bias.

We note that the parity condition results for the EmoT dataset differ between models. For lstm_twitter, higher FNR and lower TPR for Christianity terms show the existence of allocation harms against Christianity, whereas the results of lstm_tempo show the existence of allocation harms against Islamic terms. However, the 100% TPR as well as 0% FNR score for lstm_twitter on Christianity terms may be indicative of overfitting. This is shown in the stark accuracy difference between training and validation for EmoT as seen in Table 12, and may affect the prediction results. The results of Table 14 show that the existence of allocation harm in downstream performance is directly influenced by imbalanced label distribution in the dataset used to train the model. In this case, the imbalance happens due to the nature of religious discourse in social media [5], [6].

The parity condition results for the SmSA dataset is consistent across Bi-LSTM models. Since models tend to have significantly higher FNR and lower TPR on Islamic terms than Christianity, this implies an initial allocation harm against Islamic terms. However, this finding is influenced by the very imbalanced label distribution of sentences containing Islamic and Christianity terms in this dataset as shown in Table 7. Since the vast majority of religious terms are contained in sentences with negative labels, as seen in Table 7, both lstm_twitter and lstm_tempo may assume that the existence of religious terms in sentences are indicators of negative sentences.

Table 14 and 15 confirms our second hypothesis in the case of allocation harm, where sentences containing religious groups are more likely to be mispredicted as negativity-related labels. Additionally, both tables show that, unlike dataset bias, only certain religious groups were harmed, instead of both at the same time. However, the results of Table 14 show that even in the same dataset, different models can cause allocation harm against different religious groups.

Table 14. Parity conditions result of EmoT dataset, in percentage

Model - Term	False negative rate in percentage (FNR)	True positive rate in percentage (TPR)
lstm_twitter – Islamic	4.3478	95.6522
lstm_twitter – Christianity	66.6667	33.3333
lstm_tempo - Islamic	4.3478	95.6522
lstm_tempo - Christianity	0	100

Table 15. Parity conditions result of SmSA dataset, in percentage

Model - Term	False negative rate in percentage (FNR)	True positive rate in percentage (TPR)
lstm_twitter – Islamic	8.3333	91.66667
lstm_twitter – Christianity	0	100
lstm_tempo - Islamic	8.3333	91.66667
lstm_tempo - Christianity	0	100

The assessment of representation harm will be done using the prediction result of chosen sentence templates, as shown in Table 5. The prediction results for sentence templates are shown in Table 16 and 17 for EmoT and SmSA, respectively. The Hate Speech dataset is once again omitted since the dataset does not contain bias..

The chosen templates to be shown in this paper are 1(a,b), 5(a,b), and 6(a,b) from Table 5. The 1(a,b) template is used to check whether neutral sentences are impacted by dataset bias, representing all neutral templates 1(a,b), 2(a,b), and 3(a,b). The 6(a,b) template is used to check whether positive sentences are impacted by dataset bias, representing other positive templates 4(a,b). As the only template with negative label, template 5(a,b) is used to detect the impact of bias on negative sentences.

Using Table 16 and 17 as references, we note that models trained using datasets with unwanted religion bias (EmoT and SmSA) assign negative labels to neutral sentiment template 1(a,b), which shows that representation harm occurs in downstream performance. As an example, Table 17 shows that

DOI: https://doi.org/10.29207/resti.v7i4.5035

template 1a (*I am a follower of Islam*) and template 1b (*I am a follower of Christianity*) are both mispredicted as negative sentences with high probability, both for lstm_twitter and lstm_tempo. Since this template is neutral in sentiment, this misprediction shows that the addition of religious identity to an otherwise neutral template manages to change the sentiment to negative. This confirms our second hypothesis in the case of representation harm, where sentences containing religious terms are more likely to be mispredicted as negativity-related labels and classes.

Table 16. Prediction results of chosen templates for models trained with the EmoT dataset

Template	lstm_twitter	lstm_tempo
1a	anger, 0.9198	anger, 0.9628
1b	anger, 0.9104	anger, 0.8689
5a	**anger, 0.9821**	anger, 0.9865
5b	**anger, 0.9852**	anger, 0.96663
6a	happy, 0.9993	happy, 0.9988
6b	happy, 0.9989	happy, 0.9924

Table 17. Prediction results of chosen templates for models trained with the SmSA dataset

Template	lstm_twitter	lstm_tempo
1a	negative, 0.9956	negative, 0.9828
1b	negative, 0.9998	negative, 0.9989
5a	negative, 0.9902	negative, 0.8862
5b	negative, 0.9883	negative, 0.9795
6a	positive, 0.8288	positive, 0.9502
6b	positive, 0.8283	positive, 0.8411

Unlike the case of allocation harm, and in line with dataset bias, the impact of algorithmic enclaves [5] plays a higher role in the impact of dataset bias on downstream performance as opposed to the limited representation of marginalized identities in media [6]. This causes both religious groups to be equally impacted by representation harm, shown by the little to no variation between mispredictions over different religious groups on the same template.

However, the impact of representation harm may also be influenced by term occurences, as seen in template 6(a,b) being correctly assigned positive labels in most cases. As shown in Tables 6 and 7, this may be influenced by the fact that the terms used to fill the template (*masjid – mosque* or *gereja - church*) do not exhibit label imbalance as much as the case of *islam* (*Islam*) or *kristen* (*Christianity*), as the terms used to fill template 1(a,b).

The result of template 5(a,b) is *anger* for EmoT dataset, and *negative* for SmSA dataset, both with high probability scores. The high probability of *anger* prediction on EmoT dataset here shows that sentences with negative sentiment but are otherwise unbiased are also influenced with the dataset bias. In this case, since sentences containing *islam* in the EmoT dataset are mostly labeled *anger*, sentence templates that contain *islam* are likely to be labeled as *anger*, regardless of emotion showed in the sentence.

The results for both allocation and representation harms show that while allocation and representation harms do exist in downstream performance, the exact manifestation varies between models. The difference between manifestations of allocation and representation harm on models trained on the same dataset shows the need to separate both types of harm to get a clearer picture of the bias at hand [6]. This also implies that the embeddings themselves also play a role in the manifestations of allocation and representation harms in downstream performance [19], [20], since embeddings are required to create Bi-LSTM models along with datasets.

3.3. Impact of Dataset Debiasing

Table 18 shows the impact of dataset debiasing on SmSA, using the generalized PMI method shown in Equation 2. As shown in the table, dataset debiasing managed to reduce the association in terms of μ_l between all religious terms and the *negative* label originally shown in Table 10. This confirms our third hypothesis, where dataset debiasing is able to reduce associations between religious identities and negativity-related class or labels. The difference between $\mu_{positive}$ and $\mu_{neutral}$ after debiasing is likely caused by the lack of sentences with *positive* label augmented at the debiasing procedure, due to the nature of informational sentences obtained from Wikipedia having neutral sentiment.

Table 18. μ_y of all classes in SmSA dataset after debiasing

Label	$\boldsymbol{\mu_y}$
positive	-3.62
neutral	-0.42
negative	**-1.92**

In addition, Table 19 shows the impact of dataset debiasing on Hate Speech, the only dataset used in this study that does not contain dataset bias. Here, it shows that the non-negative label *none* is closer to all religious terms using Equation 2, which is the same label before debiasing as shown in Table 11.

Table 19. μ_y of all classes in Hate Speech dataset after debiasing

Label	$\boldsymbol{\mu_y}$
{hate speech}	-2.68
{abusive}	-4.95
{hate speech, abusive}	-2.87
{none}	-0.24

This shows that on top of being able to reduce dataset bias, our proposed dataset augmentation method does not introduce additional dataset bias.

Table 20 shows the accuracy of all models, over both splits (training and validation), after dataset debiasing. Compared to the accuracy results before debiasing in Table 12, for the SmSA case, dataset debiasing improves training and validation set accuracy for lstm_twitter and lstm_tempo respectively, while maintaining accuracy for the other splits. This shows that the variety of sentences added by the debiasing procedure in the SmSA case allows models to generalize better, causing an overall accuracy increase. This is especially true for the lstm_tempo case, where the increase in validation split implies that the model is able to perform better on sentences unseen in the training process beforehand, when compared to pre-debiasing accuracy scores.

Table 20. Accuracy results on all datasets for all models after dataset debiasing

Data	lstm_twitter	lstm_tempo
Training (SmSA)	0.985	0.9674
Validation (SmSA)	0.8979	0.9128
Training (Hate Speech)	0.6959	0.6576
Validation (Hate Speech)	0.7185	0.628

The results of parity conditions after dataset debiasing is shown in Table 21 for SmSA dataset, using the same parity conditions (FNR and TPR) as before debiasing. It shows that for lstm_twitter, there is an improvement in FNR and TPR for Islamic sentences, whereas the FNR and TPR for lstm_tempo is the same for both Islamic and Christianity sentences.

Table 21. Parity conditions result of SmSA dataset, in percentage, after dataset debiasing

Model - Term	False negative rate in percentage (FNR)	True positive rate in percentage (TPR)
lstm_twitter – Islamic	4.16667	95.83333
lstm_twitter – Christianity	0	100
lstm_tempo - Islamic	**8.33333**	**91.6667**
lstm_tempo - Christianity	0	100

The parity conditions imply that dataset debiasing managed to reduce allocation harm in lstm_twitter, but maintains the same level of allocation harm in lstm_tempo, keeping the exact same FNR and TPR score before debiasing. This shows an unclear result for our fourth hypothesis in the case of allocation harm. For allocation harm, our fourth hypothesis holds true for lstm_twitter, but does not hold for lstm_tempo. This result corroborates the results from Table 14, where the religious groups impacted by allocational harm differ per model, albeit in a different dataset. In particular, this effect shows that models react differently to the same dataset debiasing procedure done on the same dataset.

Table 22 shows the impact of dataset debiasing for representation harm in the SmSA dataset, which varies across models. Dataset debiasing managed to decrease the negative label probability of template 1(a,b), thus reducing the representation harms that exist in downstream performance. Unlike the case of allocation harm, our fourth hypothesis is confirmed in the case of representation harm.

Much like the variety of reactions shown in Table 21 for allocational harm, the reduction of representation harm varies per model, where the decrease is stronger for lstm_tempo compared to lstm_twitter. As an example, template 1a (*I am a follower of Islam*) shows a 16% decrease from 99% negative to 83% for lstm_twitter, whereas the decrease for lstm_tempo is 12%. In some cases, the debiasing effect is strong enough to correct the label prediction, as shown by 1b (*I am a follower of Christianity*) changing labels from *negative* (pre-debiasing) to *neutral* (post-debiasing) in the lstm_twitter case.

Table 22. Prediction results of chosen templates for models trained with the SmSA dataset after dataset debiasing

Template	lstm_twitter	lstm_tempo
1a	negative, 0.8320	negative, 0.8664
1b	**neutral, 0.5748**	negative, 0.6537
5a	**positive, 0.7133**	negative, 0.9949
5b	**positive, 0.8668**	negative, 0.9861
6a	neutral, 0.5691	neutral, 0.9913
6b	positive, 0.9636	neutral, 0.9961

For template 6(a,b), both models are able to maintain the non-negative label prediction after debiasing. In most cases, the label predictions change from *positive* to *neutral*, with roughly equal percentages. As an example, the template 6a was originally predicted as 95% positive in the lstm_tempo case, and 99% neutral after debiasing. This shows that on top of being able to reduce mispredictions, dataset debiasing procedure does not introduce additional representation harms to correctly-predicted instances. The label changes from *positive* to *neutral* may be impacted by the influx of *neutral* sentences added as part of the debiasing procedure, due to the nature of Wikipedia sentences being neutral sentiment.

An interesting outcome is that the dataset debiasing procedure managed to change the labels of 5(a,b) from *negative* to *positive* on the lstm_twitter case. This does not happen for the lstm_tempo case, which again shows how different models may react differently to the same debiasing procedure.

Much like the previous results of bias detection, the results of Table 21 and 22 show that the overall impact of dataset debiasing, both for allocation and representation harm, varies per Bi-LSTM model. Since models differ by word embeddings, this implies that the different natures of the embeddings, which are created using different sources of corpora, impact dataset debiasing results. Understanding the exact reason for this effect would likely require extensive domain knowledge on how religious representation manifests on different types of media, on top of the data collection methods used to create said embeddings, and how they may interact with each other.

A very different effect can be seen on the Hate Speech dataset, which originally does not contain dataset bias as shown in Table 11. For the *hate speech* label, dataset debiasing consistently worsens representation harm by increasing the prediction probability for non-negative sentences containing *islam* (*Islam*) or *kristen* (*Christianity*), up to 10 times its original (mis)prediction probability. Since all of these sentences used to measure representation harms are non-negative, it is expected for the sentence probability in the *hate speech* label to be close to zero. This effect is consistent over both models, which shows that debiasing datasets that do not originally contain biases will instead introduce new biases that can impact downstream performance. An example of this effect will be shown in Table 23, using template 1(a,b).

Table 23. Prediction results of chosen templates for models trained with Hate Speech dataset (hate speech label), before and after dataset debiasing

Template	lstm_twitter	lstm_tempo
1a pre	0.0153	0.369
1a post	**0.1659**	**0.4553**
1b pre	0.0066	0.0017
1b post	**0.0103**	**0.0068**

The full result of our experiment, including all omitted results from the Hate Speech dataset, will be available on our GitHub repository (https://github.com/marieff587/AnalysisMitigationBiasIndonesianNLP/blob/main/Debiasing%20Result.xlsx.

4. Conclusion

Using PMI, adapted from single-label into multi-label and multi-class cases, we show the existence of religion bias on various Indonesian-language NLP datasets, and their effects on downstream performance through allocation and representation harm. In particular, we show that dataset bias exists on two out of the three datasets used for this study and that the dataset bias negatively impacts downstream performances, proving our first two hypotheses. We also show that both religious groups are equally harmed in most cases, which shows that the impact of algorithmic enclaves negatively impacts datasets more than the limited representations of marginalized religious identities.

Throughout this study, we show the variance of debiasing impact, using dataset augmentation to add neutral-labeled sentences from WFikipedia to existing datasets. While dataset debiasing successfully mitigates dataset bias, proving our third hypothesis correct, the result for downstream performance varies per combinations of embedding, dataset, and template, causing mixed results for our fourth hypothesis. In particular, it is shown that for allocation harms, our fourth hypothesis is confirmed for one model (lstm_twitter) but not the other (lstm_tempo). For representation harm, our fourth hypothesis is confirmed, although the exact reduction of representation harm varies per model.

This study only considers the effect of dataset bias. Analyzing the effects of embedding bias, as well as the impact of debiasing them is a line of future work. As shown in our experiment results, models with different embeddings have different manifestations allocation and representation harms. These models also react differently to debiasing, even if they were all trained using the same dataset. This shows the potential of biases that exist in the embeddings themselves impacting the result of dataset debiasing. Since creating word embeddings require large amounts of text corpora, often taken from various sources on the Internet, the social phenomenons that negatively impact religious discourse in Indonesian social media may also cause unwanted religion bias in the embeddings. This is indicated by Table 23, where dataset debiasing for a dataset that originally does not contain dataset bias instead increases the misprediction of sentence templates, and therefore the representation harm caused by the model. For this case, one possible explanation is that dataset debiasing instead amplifies the existing embedding bias contained in each model. This, in turn, may increase the level of representation harm caused by the model.

On top of prior future work recommendations, since two out of three datasets used in this study (EmoT and SmSA) are currently used as benchmarks (IndoNLU) for Indonesian NLP systems, this study doubly works as a partial audit of IndoNLU. A thorough audit of IndoNLU, considering both the datasets used and the resulting pre-trained model, for religious bias and other forms of social bias, can be considered as another line of future work from this study.

References

[1] M. Wiegand, J. Ruppenhofer, and T. Kleinbauer, “{D}etection of {A}busive {L}anguage: the {P}roblem of {B}iased {D}atasets,” in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 602–608, doi: 10.18653/v1/N19-1060.

[2] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, “Measuring and Mitigating Unintended Bias in Text Classification,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 67–73, doi: 10.1145/3278721.3278729.

[3] A. Ball-Burack, M. S. A. Lee, J. Cobbe, and J. Singh, “Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 116–128, doi: 10.1145/3442188.3445875.

[4] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, “Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries ,” *Frontiers in Big Data* , vol. 2. 2019, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2019.00013 .

[5] M. Lim, “Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in Indonesia,” *Crit. Asian Stud.*, vol. 49, no. 3, pp. 411–427, Jul. 2017, doi: 10.1080/14672715.2017.1341188.

[6] M. Heychael, H. Rafika, J. Adiprasetyo, and Y. Arief, “Marginalized Religious Communities in Indonesian Media,” *Remotivi*, 2021. https://www.mediasupport.org/publication/marginalized-religious-communities-in-indonesian-media/.

[7] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (Technology) is Power: A Critical Survey of {``}Bias{''} in {NLP},” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

DOI: https://doi.org/10.29207/resti.v7i4.5035

2020, pp. 5454–5476, doi: 10.18653/v1/2020.acl-main.485.

[8] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019.

[9] M. S. Saputri, R. Mahendra, and M. Adriani, "Emotion Classification on Indonesian Twitter Dataset," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 90–95, doi: 10.1109/IALP.2018.8629262.

[10] A. Purwarianti and I. A. P. A. Crisdayanti, "Improving Bi-LSTM Performance for Indonesian Sentiment Analysis Using Paragraph Vector," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2019, pp. 1–5, doi: 10.1109/ICAICTA.2019.8904199.

[11] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in {I}ndonesian {T}witter," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57, doi: 10.18653/v1/W19-3506.

[12] B. Wilie *et al.*, "{I}ndo{NLU}: Benchmark and Resources for Evaluating {I}ndonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857, [Online]. Available: https://aclanthology.org/2020.aacl-main.85.

[13] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, 1990, [Online]. Available: https://aclanthology.org/J90-1003.

[14] S. Kiritchenko and S. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018, pp. 43–53, doi: 10.18653/v1/S18-2005.

[15] J. M. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *CoRR*, vol. abs/1609.05807, 2016, [Online]. Available: http://arxiv.org/abs/1609.05807.

[16] D. Leben, "Normative Principles for Evaluating Fairness in Machine Learning," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 86–92, doi: 10.1145/3375627.3375808.

[17] T. A. Le, D. Moeljadi, Y. Miura, and T. Ohkuma, "Sentiment Analysis for Low Resource Languages: A Study on Informal {I}ndonesian Tweets," in *Proceedings of the 12th Workshop on {A}sian Language Resources ({ALR}12)*, 2016, pp. 123–131, [Online]. Available: https://aclanthology.org/W16-5415.

[18] K. Kurniawan, "KaWAT: {A} Word Analogy Task Dataset for Indonesian," *CoRR*, vol. abs/1906.09912, 2019, [Online]. Available: http://arxiv.org/abs/1906.09912.

[19] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," *CoRR*, vol. abs/1607.06520, 2016, [Online]. Available: http://arxiv.org/abs/1607.06520.

[20] T. Manzini, L. Yao Chong, A. W. Black, and Y. Tsvetkov, "{B}lack is to Criminal as {C}aucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings," in *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 615–621, doi: 10.18653/v1/N19-1062.