UNIVERSITAS INDONESIA

ANALISIS DAN MITIGASI *RELIGION BIAS* PADA *DATASET* DAN *EMBEDDING NLP* BERBAHASA INDONESIA

TESIS

MUHAMMAD ARIEF FAUZAN
2106774881

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
DEPOK
2023

UNIVERSITAS INDONESIA

ANALISIS DAN MITIGASI *RELIGION BIAS* PADA *DATASET* DAN *EMBEDDING NLP* BERBAHASA INDONESIA

TESIS
Diajukan sebagai salah satu syarat untuk memperoleh gelar
Magister Ilmu Komputer

MUHAMMAD ARIEF FAUZAN
2106774881

FAKULTAS ILMU KOMPUTER
PROGRAM STUDI ILMU KOMPUTER
DEPOK
JULI 2023

HALAMAN PERNYATAAN ORISINALITAS

Tesis ini adalah hasil karya saya sendiri,
dan semua sumber baik yang dikutip maupun dirujuk
telah saya nyatakan dengan benar.

Nama	:	**Muhammad Arief Fauzan**
NPM	:	**2106774881**
Tanda Tangan	:	
Tanggal	:	**24 Juli 2023**

HALAMAN PENGESAHAN

Tugas Akhir ini diajukan oleh	:	
Nama	:	Muhammad Arief Fauzan
NPM	:	2106774881
Program Studi	:	Magister Ilmu Komputer
Judul	:	ANALYSIS AND MITIGATION OF RELIGION BIAS IN INDONESIAN NLP DATASETS AND EMBEDDINGS

Telah berhasil dipertahankan di hadapan Dewan Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Magister Ilmu Komputer pada Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Indonesia

DEWAN PENGUJI

Pembimbing	:	Dr. Indra Budi, S.Kom., M.Kom.	(Nilai telah diberikan melalui SISIDANG pada 17-07-2023, 17:40:27) (Revisi telah disetujui melalui SISIDANG pada 24-07-2023, 08:03:55)
Pembimbing	:	Ari Saptawijaya, S.Kom., M.Sc., Ph.D.	(Nilai telah diberikan melalui SISIDANG pada 21-07-2023, 10:06:58) (Revisi telah disetujui melalui SISIDANG pada 24-07-2023, 05:22:20)
Penguji 1	:	Ir. Suryana Setiawan M.Sc., Ph.D.	(Nilai telah diberikan melalui SISIDANG pada 17-07-2023, 17:58:46) (Revisi telah disetujui melalui SISIDANG pada 23-07-2023, 21:38:05)
Penguji 2	:	Prof. Dr. Achmad Nizar Hidayanto, S.Kom., M.Kom.	(Nilai telah diberikan melalui SISIDANG pada 17-07-2023, 17:54:40) (Revisi telah disetujui melalui SISIDANG pada 20-07-2023, 09:43:13)
Penguji 3	:	Dr. Ika Alfina, S.Kom., M.Kom.	(Nilai telah diberikan melalui SISIDANG pada 17-07-2023, 17:57:37) (Revisi telah disetujui melalui SISIDANG pada 20-07-2023, 11:50:04)

Ditetapkan di : Depok, Jawa Barat

Tanggal : 24 Juli 2023

KATA PENGANTAR

Syukur Alhamdulillah penulis panjatkan kepada Allah SWT yang karena-Nya penulis dapat menyelesaikan tesis berjudul 'Analisis dan Mitigasi *Religion Bias* pada *Dataset* dan *Embedding* NLP berbahasa Indonesia' sebagai salah satu syarat untuk memperoleh magister ilmu komputer di FASILKOM UI dengan tepat waktu. Pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Bapak Ari Saptawijaya, S.Kom., M.Sc., Ph.D. dan Dr. Indra Budi, S.Kom., M.Kom. sebagai pembimbing tesis, yang telah menuntun penulis pada proses pembentukan tesis ini serta menyediakan banyak waktu dan tenaga sehingga tesis ini dapat dapat selesai tepat waktu.

2. Bapak Ir. Suryana Setiawan M.Sc., Ph.D., Bapak Prof. Dr. Achmad Nizar Hidayanto, S.Kom., M.Kom., dan Ibu Dr. Ika Alfina, S.Kom., M.Kom. sebagai penguji tesis, yang telah memberikan banyak masukan untuk tesis ini sehingga meningkatkan kualitas penulisan tesis.

3. PT Maybank Indonesia, sebagai sponsor penulis dalam menjalankan studi magister ini. Terlebih untuk anggota divisi Data Analytics - Rahmat, Intan, Ryan, Yessi, Briston, Faqih, Hendri, Retno, terimakasih sudah menjadi rekan kerja terbaik, tempat pengeluaran keluh kesah penulis, serta bantuan cover untuk *daily work* pada proses penulisan tesis ini.

4. Orang tua dan keluarga besar penulis, yang senantiasa mendukung dan menyemangati penulis untuk menyelesaikan skripsi tepat waktu.

5. Adib, Dion, Naufal, Akmal, seluruh anggota Zoo dan Sewer Dwellers, serta berbagai penghuni Discord penulis lainnya yang sangat dapat diandalkan oleh penulis untuk melepas penat selama masa perkuliahan.

Akhir kata, terima kasih. Semoga tesis ini dapat bermanfaat baik bagi pembaca maupun penulis kedepannya.

Depok, 24 Juli 2023

Muhammad Arief Fauzan

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Indonesia, saya yang bertanda tangan di bawah ini:

Nama	:	Muhammad Arief Fauzan
NPM	:	2106774881
Program Studi	:	Ilmu Komputer
Fakultas	:	Ilmu Komputer
Jenis Karya	:	Tesis

demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Indonesia **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty Free Right*)** atas karya ilmiah saya yang berjudul:

Analisis dan Mitigasi *Religion Bias* pada *Dataset* dan *Embedding NLP* berbahasa Indonesia

beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini Universitas Indonesia berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan memublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyatan ini saya buat dengan sebenarnya.

Dibuat di : Depok
Pada tanggal : 24 Juli 2023
Yang menyatakan

(Muhammad Arief Fauzan)

ABSTRAK

Nama	:	Muhammad Arief Fauzan
Program Studi	:	Ilmu Komputer
Judul	:	Analisis dan Mitigasi *Religion Bias* pada *Dataset* dan *Embedding NLP* berbahasa Indonesia
Pembimbing	:	Ari Saptawijaya, S.Kom., M.Sc., Ph.D
	:	Dr. Indra Budi, S.Kom., M.Kom

Riset terdahulu menunjukkan adanya misrepresentasi identitas agama pada media Indonesia. Menurut studi sebelumnya, misrepresentasi identitas marjinal pada *dataset* dan *word embedding* untuk *natural language processing* dapat merugikan identitas marjinal tersebut, dan karenanya harus dimitigasi. Tesis ini menganalisis keberadaan bias agama pada beberapa *dataset* dan *word embedding* NLP berbahasa Indonesia, dampak bias yang ditemukan pada *downstream performance*, serta proses dan dampaj *debiasing* untuk *dataset* dan *word embedding*. Dengan menggunakan metode uji PMI untuk deteksi bias pada *dataset* dan *word similarity* untuk deteksi bias pada *word embedding*, ditemukan bahwa dua dari tiga *dataset*, serta satu dari empat *word embedding* yang digunakan pada studi ini mengandung bias agama. Model *machine learning* yang dibentuk dari *dataset* dan *word embedding* yang mengandung bias agama memiliki dampak negatif untuk *downstream performance* model tersebut, yang direpresentasikan dengan *allocation harm* dan *representation harm*. *Allocation harm* direpresentasikan oleh performa *false negative rate* (FNR) dan *true positive rate* (TPR) model *machine learning* yang lebih buruk untuk identitas agama tertentu, sedangkan *representation harm* direpresentasi oleh kesalahan model dalam mengasosiasikan kalimat non-negatif yang mengandung identitas agama sebagai kalimat negatif. Metode *debiasing* pada *dataset* dan *word embedding* mampu memitigasi bias agama yang muncul pada *dataset* dan *word embedding*, tetapi memiliki performa yang beragam dalam mitigasi *allocation* dan *representation harm*. Dalam tesis ini, akan digunakan 5 metode *debiasing*: *dataset debiasing* dengan menggunakan *sentence templates*, *dataset debiasing* dengan menggunakan kalimat dari Wikipedia, *word embedding debiasing*, *joint debiasing* dengan *sentence templates*, serta *joint debiasing* menggunakan kalimat dari Wikipedia. Dari 5 metode *debiasing*, *joint debiasing* dengan *sentence templates* memiliki performa yang paling baik dalam mitigasi *allocation harm* dan *representation harm*, sehingga menjadi metode *debiasing* yang terbaik.

Kata kunci:

Natural language processing, bias sosial, *debiasing*

ABSTRACT

Name	:	Muhammad Arief Fauzan
Study Program	:	Ilmu Komputer
Title	:	Analysis and Mitigation of Religion Bias in Indonesian NLP Datasets and Embeddings
Counsellor	:	Ari Saptawijaya, S.Kom., M.Sc., Ph.D
	:	Dr. Indra Budi, S.Kom., M.Kom

Previous researches have shown the existence of misrepresentation regarding various religious identities in Indonesian media. Misrepresentations of other marginalized identities in natural language processing (NLP) resources have been recorded to inflict harm against such marginalized identities, and as such must be mitigated. This thesis analyzes several Indonesian language NLP datasets and word embeddings to see whether they contain unwanted bias, the impact of bias on downstream performance, the process of debiasing datasets or word embeddings, and the effect of debiasing on them. By using the PMI test to detect dataset bias and word similarity to detect word embedding bias, it is found that two out of three datasets and one out of four word embeddings contain religion bias. The downstream performances of machine learning models which learn from biased datasets and word embeddings are found to be negatively impacted by the biases, represented in the form of allocation and representation harms. Allocation harm is represented by worse false negative rate (FNR) and true positive rate (TPR) of models with respect to certain religious identities, whereas representation harm is represented by the misprediction of non-negative sentences containing religious identity terms as negative sentences. Debiasing at dataset and word embedding level was found to correctly mitigate the respective biases at dataset and word embedding level. Nevertheless, depending on the dataset and word embedding used to train the model, the performance of each debiasing method can vary highly at dowsntream performance. This thesis utilizes 5 debiasing methods: dataset debiasing using sentence templates, dataset debiasing using sentences obtained from Wikipedia, word embedding debiasing, joint debiasing using sentence templates, as well as joint debiasing using sentences obtained from Wikipedia. Out of all 5 debiasing techniques, joint debiasing using sentence templates perform the best on mitigating both allocation and representation harm. As such, it is concluded that combining dataset debiasing using sentence templates and word embedding debiasing is the best performing debiasing method.

Key words:

Natural language processing, social bias, debiasing

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF CODES

CHAPTER 1
INTRODUCTION

This chapter describes the motivation for this thesis, in particular showing the lack of research regarding bias in Indonesian-based natural language processing (NLP) models as well as existing researches that shows how bias may possibly exist in Indonesian-based models, then describes the research questions and objectives.

1.1 Research Background

As natural language processing models become more ubiquitous in human life, there has been growing interest in ensuring that they perform fairly across all walks of life. However, researches show 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 on certain groups of people based on harmful stereotypes (Ball-Burack et al., 2021; Hendricks et al., 2018).

The form of 'bias' considered in this thesis is external bias (Olteanu et al., 2019), defined as biases stemming from factors outside the media platform that contains textual data. Some examples of factors that contribute to external biases are interaction between individuals with different ideological leanings and cultural backgrounds, as well as external social pressures that may drive each individual to interact with each other. These biases are represented in the corpus, whether caused by the platform where the datapoints are stored, the sampling process to obtain the dataset from the platform, or the content of the datapoints itself.

Since these corpuses are used to create datasets and word embeddings, various researches have found the existence of external bias in datasets and word embeddings. Wiegand et al. (2019) analyzed various datasets of abusive language detection and show that datasets collected by querying specific topics coinciding with abusive content (thereby referred by the paper as biased sampling) tend to contain more implicit abuse (abusive language conveyed through subtle means, e.g, harmful stereotypes and jokes). Additionally, they found that high classification performances on the models that learn from said datasets are caused by the model learning the implicit biases.

Since word embeddings are generated from corpuses, NLP models that utilize

word embeddings, e.g.: neural network-based models, may also encounter another source of bias from the word embeddings itself. Multiple researches have shown that word embeddings from different corpuses learn unwanted social bias through word associations, as shown by Bolukbasi et al. (2016) on learning unwanted gender stereotypes through news articles taken from Google, and Manzini et al. (2019) on learning unwanted religious and racial stereotypes through social media sources.

The biases contained in the datasets and word embeddings is found to negatively impact the downstream performance of NLP models that learn from them. As an example, machine learning models that learn from biased datasets and word embeddings in which certain marginalized identities only exists in negativity-related datapoints was found to constantly perform worse for said marginalized identities by mispredicting sentences containing them as negativity-related classes (Dixon et al., 2018; Wiegand et al., 2019; Ball-Burack et al., 2021). Additionally, machine learning models that learn from biased datasets and word embeddings in which certain marginalized identities are negatively stereotyped was found to propagate said stereotypes on their prediction results (Kiritchenko and Mohammad, 2018). As such, this highlights the importance of debiasing social biases that exist in datasets and word embeddings, in order to mitigate the biases of models that learn from them.

To this end, there exist researches dedicating to developing debiasing methods used to mitigate biases found in datasets and word embeddings. For datasets, Dixon et al. (2018) augments biased datasets by adding non-negative sentences obtained from an external source assumed to not contain the same bias contained in the original datasets. In the word embedding case, Bolukbasi et al. (2016) developed a procedure used to remove biases contained in the word embeddings by defining a subspace which represent certain social biases, e.g.: gender bias, and projecting words to the orthogonal subspace of said subspace in order to remove the bias component from said word embeddings.

Most of the existing researches on NLP bias, whether on detecting and mitigating biases, focuses around English, and to the best of the author's knowledge, there has not been any NLP bias researches for Indonesian, whether in dataset bias or in word embedding bias. Prior researches on non-English word embedding bias exists, whether in bilingual setups concerning non-English word embeddings aligned to existing English word embeddings (Zhou et al., 2019; Zhao et al., 2020), or in monolingual non-English word embeddings (Takeshita et al., 2020; Pujari et al., 2019; Sahlgren and Olsson, 2019). Additionally, there have been a lack on analyzing social biases other than gender bias, which existing researches heavily focuses

on (Sambasivan et al., 2021).

1.2 Research Problem

Since text medias (whether through social media or other news sources) are common corpuses to be used as training data for Indonesian NLP implementations, whether in the form of datasets to train models or word embeddings to represent Indonesian in vector form, religion biases encoded in said sources can potentially manifest in Indonesian NLP implementations that learn from them.

NLP researches and implementations in Indonesia gained popularity around the 2017 governor election (Pilkada) and 2019 presidential election (Pilpres), and various sentiment analysis and abusive language detection researches use datasets that contain topics about said elections, typically gathered by sampling from social media sources.

However, the effects of algorithmic enclaves, where multiple groups of social media users created as a byproduct of amplifying online interactions, each with their own shared online identities, beliefs, and outsider threats interact to silence other groups, dominate the online discussion space regarding the topics of 2017 governor election (Lim, 2017). This can potentially bring unwanted bias to the dataset in the form of abusive messages regarding various religious and racial identities. The limited representations of marginalized identities on media aside from conflicts and celebrations, as reported by Remotivi (2021) can possibly exacerbate the bias issue for marginalized religious or racial identities, creating a biased representation on which certain racial and religious identities are only used as insults or other negativity-related contents. A variation of this effect is shown in Wiegand et al. (2019), where specific neutral terms, such as *announcer* (in *I can't handle the female play by play announcer on espn*) and *commentator* (in *this female commentator is killing me*), became indicators of abusive language in a dataset gathered by biased sampling on social media sources, querying for specific topics likely to contain abusive language content.

A further motivating example for considering sentiment analysis and hatespeech as two main cases can be seen in Figure 1.1, as tested on Prosa.ai, an NLP-as-a-service platform online. In this example, a change of religious identity in an otherwise neutral template sentence (*saya seorang muslim/kristen*, tl: *i am a muslim/christian*), managed to change the sentiment from positive (with *muslim*) to negative (with *kristen*). This shows a possibility of bias existing in the model, manifesting as unwanted relationship between religious identities and sentiment learned

by the model. The existence of bias in the output can cause harmful outcomes when an NLP model generated from this platform is used for real-life cases. As an example, the biased representations between marginalized religious identities and negativity can cause unfair automated content moderation outcomes. This is depicted in Ball-Burack et al. (2021), where biased representation between African-American identities and hate-speech content causes automated content moderation to unfairly hide contents sourced from, or related to, said identities.

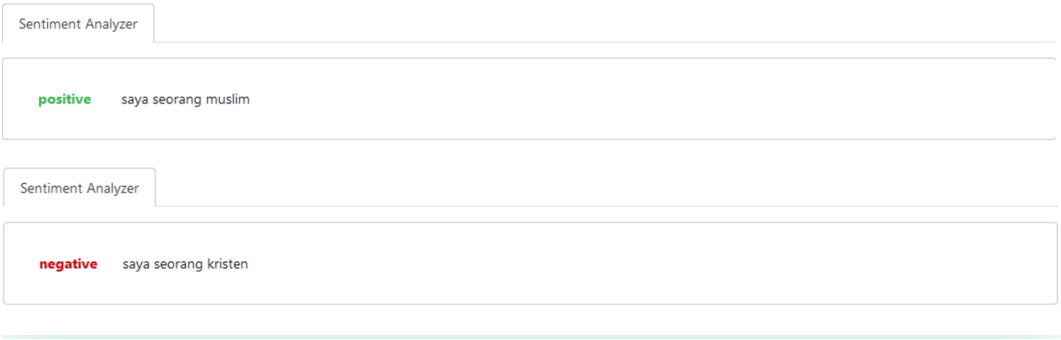

Figure 1.1: Different religious identity changes the sentiment prediction (taken 11 May 2023)

Preliminary research on the existence and impact of religion bias in Indonesian NLP datasets and word embeddings shows that bias exists in them, and results in neutral and positive sentiment sentences being mispredicted as negative, much like the example shown at Figure 1.1 (Fauzan, 2022). As an example, in the preliminary research done prior to this thesis, the sentence *saya menganut agama kristen*, originally a neutral-sentiment sentence, was mispredicted as negative-sentiment.

Debiasing datasets and word embeddings was done using techniques shown in Dixon et al. (2018) and Manzini et al. (2019) to debias datasets and word embeddings respectively, mitigating religion bias as shown by downstream performance results to some degree. In particular, debiasing managed to reduce misprediction rate of sentences containing religious terms, both for sentences originally contained in the dataset and sentences not contained in the dataset. However, the debiasing procedure was only done on datasets and word embeddings separately, without combining both procedures together. Additionally, the low amount of sentences used to debias datasets causes dataset debiasing to reduce overall accuracy score, unlike word embedding debiasing which was shown to both improve model accuracy and mitigate religion bias.

1.2.1 Research Questions

This research aims to answer these questions:

1. How does religion bias manifest on various Indonesian NLP datasets and word embeddings?

This research question concerns the manifestation of religion bias on Indonesian-language NLP datasets as well as word embeddings, owing to the possibility of biased representations caused by social media contents (Lim, 2017; Remotivi, 2021). The hypothesis is as follows: the effect of algorithmic enclaves (Lim, 2017) and the limited representation, both in content and in quantity, of marginalized religions in Indonesian media (Remotivi, 2021) introduce unwanted religion bias to Indonesian NLP datasets and word embeddings, where identities related to marginalized religions are wrongfully related to negativity. In this thesis, this research question is tackled at Sections 4.1 and 4.2.

2. How does religion bias in Indonesian NLP datasets and word embeddings impact downstream performance?

This research question concerns the impact of religion bias that exists (if any) in Indonesian-language NLP datasets and word embeddings on downstream performance. The hypothesis is as follows: NLP models that learn from Indonesian datasets and word embeddings that contain unwanted religion biases inherits said biases, which introduces allocation and representational harms (Blodgett et al., 2020) against marginalized religious identities. In this thesis, this research question is tackled at Section 4.3.

3. How does bias mitigation on dataset level impact the manifestations of religion bias on various Indonesian NLP datasets?

This research question concerns the impact of bias mitigation on Indonesian-language NLP datasets. The hypothesis is as follows: bias mitigation at dataset level reduces the impact of religion bias in Indonesian-language NLP datasets. In this thesis, this research question is tackled at Sections 5.1 and 5.2.

4. How does bias mitigation on word embedding level impact the manifestations of religion bias on various Indonesian NLP word embeddings?

This research question concerns the impact of bias mitigation on Indonesian-language word embeddings. The hypothesis is as follows: bias mitigation at word embedding level reduces the impact of religion bias in Indonesian-language word embeddings. In this thesis, this research question is tackled at Section 5.3.

5. How do bias mitigation on dataset and word embedding level, either independently or jointly, impact downstream performance?

This research question concerns the impact of bias mitigation on dataset and word embedding level, either independently or jointly, on downstream performance. The hypothesis is as follows: bias mitigation on dataset and word embedding level, either independently or jointly, mitigate allocation and representational harms that exists as a result of biases in said datasets and word embeddings. Additionally, due to the effect of cascaded debiasing (Ghai et al., 2022), bias mitigation on joint dataset and word embedding level works better compared to independent bias mitigation.

1.2.2 Research Objective

The objective of this research are as follows:

1. Analyze the manifestations of religion bias on various Indonesian NLP datasets and word embeddings
2. Analyze the impact of religion bias in Indonesian NLP datasets and word embeddings on downstream performance
3. Analyze the impact of bias mitigation on dataset level on various Indonesian NLP datasets
4. Analyze the impact of bias mitigation on word embedding level on various Indonesian NLP word embeddings
5. Analyze the impact of bias mitigation on dataset and word embedding level, either independently or jointly, on downstream performance

1.3 Research Benefits

To the best of the author's knowledge, this thesis is the first to research about possible biases in Indonesian sentiment analysis and hate speech detection datasets and word embeddings, as well as on debiasing endeavors of Indonesian-language NLP datasets and word embeddings. Since Indonesian is still classified as a low-resource language (Le et al., 2016), this thesis also explores one example of bias detection and mitigation efforts on such language.

this thesis is created with hope that it can motivate other Indonesian-based NLP researches to consider the effect of social biases in NLP models. On top of analyzing the impact of prior debiasing methods in Indonesian-language NLP resources,

which was untested prior to this thesis, this thesis also analyzes how unwanted social biases can manifest in Indonesian NLP resources. As an end result, this thesis can facilitate the start of developing fairer Indonesian NLP models, particularly on the case of sentiment analysis and hate speech detection tasks.

1.4 Research Position

The thesis adapts the work of Bolukbasi et al. (2016) and Manzini et al. (2019) on debiasing word embeddings for religion bias on Indonesian context, as well as Kiritchenko and Mohammad (2018) and Dixon et al. (2018) on debiasing datasets concerning religious identities. There exists a research gap on debiasing datasets for religion bias, as well as debiasing Indonesian-language word embeddings, which this thesis covers. In order to do so, this thesis formulates a new framework in which religion bias may manifest in Indonesian NLP datasets and word embeddings (Lim, 2017; Remotivi, 2021).

Specifically for debiasing datasets using dataset augmentation, this thesis explores two cases: using sentence templates as introduced by Kiritchenko and Mohammad (2018), adapted into Indonesian-based religion bias context, and using sentences from Wikipedia articles (Dixon et al., 2018), adapted for the case of sentiment analysis and hate-speech detection task. Additionally, this thesis expands the usage of pointwise mutual information (PMI) used to detect dataset bias, as introduced by Dixon et al. (2018) and Wiegand et al. (2019) for single-label, binary-classification cases into a multi-class and multi-label case fit for this thesis

Another research gap found in literatures regarding debiasing in NLP is on combining both methods of debiasing, which this thesis aims to cover. Existing literatures only considers debiasing either dataset only (Dixon et al., 2018; Ball-Burack et al., 2021) or word embeddings only (Bolukbasi et al., 2016; Manzini et al., 2019). Similar gaps was found to exist for debiasing machine learning models in general, as reported by Ghai et al. (2022) which analyzes the impact of implementing multiple debiasing methods. The results of Ghai et al. (2022) is implemented as a basis for analyzing debiasing dataset and word embeddings both separately and jointly.

Tables 1.1 and 1.2 explains where this thesis stands in more detail.

Table 1.1: Current standings for this research

Bias detection and mitigation aspects	Source (existing literature)	Existing method	Adaptation
Detecting dataset bias	Dixon et al. (2018); Wiegand et al. (2019)	Uses PMI to detect social bias in a binary-classification case, e.g. toxic language detection and abusive language detection	Extends PMI use into multi-class and multi-label case, with contextual knowledge of religious representation in Indonesia (Lim, 2017; Remotivi, 2021)
Parity metrics used to detect downstream performance results	Dixon et al. (2018); Kiritchenko and Mohammad (2018)	Uses various parity metrics to measure allocational harms in a binary-classification case, e.g. toxic language detection and abusive language detection	Adapted for religion bias and for multi-class/multi-label case, used to test representation harms in downstream performance. Additionally, considers the domain specificity and contexts when choosing which metrics to choose, as suggested by Leben (2020) and Selbst et al. (2019)
Sentence templates used to detect downstream performance results	Kiritchenko and Mohammad (2018)	Used to test gender and racial bias in sentiment analysis models	Adapted for religion bias, used to test representation harms in downstream performance

Table 1.2: Current standings for this research (continued)

Augmenting dataset for dataset debiasing	Dixon et al. (2018)	Uses sentences in Wikipedia articles containing specific racial and gender terms	Uses, and compares separately, sentences in Wikipedia articles containing religious terms and sentence templates (Kiritchenko and Mohammad, 2018), adapted for religion bias
Detecting word embedding bias	Bolukbasi et al. (2016); Manzini et al. (2019)	Uses word analogies of certain gender, racial, and religious term in English	Uses word analogies and cosine similarity of religious terms, translated into Indonesian
Debiasing word embeddings	Bolukbasi et al. (2016); Manzini et al. (2019)	Uses Hard Debiasing method, for gender and racial bias	Uses Hard Debiasing method, adapted for religion bias
Debiasing procedure	Bolukbasi et al. (2016); Manzini et al. (2019) for word embedding, Dixon et al. (2018); Wiegand et al. (2019) for dataset	Separately debiases datasets or word embeddings	Debiases datasets and word embeddings, either independently or jointly, and compares the result using Ghai et al. (2022) as basis

1.5 Research Limitations

This thesis considers the existence of religion bias in 3 datasets and 4 word embeddings, all of them in Indonesian. The religion identities are limited to Islam (representing religious majority) and Christianity (representing marginalized religions). This is done due to the discussions regarding algorithmic enclaves (Lim, 2017) mainly focusing on algorithmic enclaves representing Islamic and Christianity-related identities.

For the dataset case, this thesis focuses on the case of emotion detection using EmoT dataset (Saputri et al., 2018), sentiment analysis using SmSA dataset (Purwarianti and Crisdayanti, 2019) and hate-speech detection using Hate Speech dataset(Ibrohim and Budi, 2019). These datasets of interest due to three main reasons. First, all three datasets utilize social media sources as main sources, where hateful comments on various religious, race, and other identities often take place (Lim, 2017; Kiritchenko and Mohammad, 2018; Remotivi, 2021) and may affect the performance of the NLP models that learn from them (Dixon et al., 2018; Kiritchenko and Mohammad, 2018; Wiegand et al., 2019; Ball-Burack et al., 2021). Second, both EmoT and SmSA datasets are currently used as benchmarks in which various Indonesian-language NLP models are tested on (Wilie et al., 2020), and as such analyzing and mitigating possible religion biases in these datasets is of importance in order to ensure minimize religion bias in future benchmarking endeavors. Additionally, these datasets are single-label, multi-class datasets, which allows adaptation of existing dataset detection procedures for multi-class cases. Third, the Hate Speech datasets are used to detect abusive and hate-speech sentences, which existing researches focusing on social bias in NLP focuses on (Dixon et al., 2018; Wiegand et al., 2019; Ball-Burack et al., 2021). This dataset is multi-label, binary-class datasets, which allows adaptation of existing dataset detection procedures for the multi-label case.

For the word embedding case, this thesis focuses on four existing Word2Vec word embeddings, each trained on a different corpus: Twitter data (Saputri et al., 2018), Wikipedia [1], Tempo online news media (Kurniawan, 2019), and Common Crawl data used for CoNLL (Zeman et al., 2018). These word embeddings are chosen to analyze whether word embeddings created with different sources manifests religious bias differently, due to the minimal number of abusive sentences using religious identities in more curated informational sources (e.g., news articles for Tempo word embedding, Wikipedia entries for Wikipedia word embedding).

[1] https://medium.com/@diekanugraha/membuat-model-word2vec-bahasa-indonesia-dari-wikipedia-menggunakan-gensim-e5745b98714d

As shown on the fifth research question, one aim of this thesis is to explore the impact of debiasing methods on downstream performance, as opposed to creating a procedure to maximize mitigation capability. The debiasing procedure is done whether the datasets or word embeddings are actually biased, in order to analyze the possible effects of debiasing on unbiased datasets and word embeddings.

Additionally, since this thesis is the first research to detect and mitigate religion bias in Indonesian NLP datasets and word embeddings, there are currently no existing point of reference in which the results of new proposed methods to detect and mitigate religion bias in Indonesian NLP datasets and word embeddings can be compared to. As such, the detection and mitigation results done in Indonesian NLP datasets and word embeddings are measured only using the proposed metrics done in this research.

1.6 Thesis Structure

This thesis is organized as follows. Chapter 2 covers existing works on both dataset and word embedding bias, the effects of each bias on downstream performance, as well as existing debiasing methods. We explain our research methods, particularly the datasets and word embeddings chosen, as well as the debiasing methods and metrics used to measure downstream performance pre- and post-debiasing in Chapter 3. The analysis of bias detection results is shown on Chapter 4, whereas the analysis of debiasing results is shown on Chapter 5. We then conclude the findings of this thesis as well as highlight possible future work directions on Chapter 6.

CHAPTER 2
LITERATURE REVIEW

This chapter describes the relevant literatures for this thesis, starting from the NLP systems used in this thesis, how unwanted bias may show up in NLP systems, and how to debias them.

2.1 Word Embedding

In NLP models, various representations of textual data into numerical vector forms (vector space models) are often used to train language models. One such representation is word embedding, where textual data are trained on a neural network for certain tasks, then its hidden layer, which contains values corresponding to each word in the original textual data, is taken as textual representations to be used on other machine learning models.

Mikolov et al. (2013a) introduced Word2Vec, a word embedding that is commonly used as vector representations of textual data. To train a Word2Vec word embedding, a training set containing a list of target words and words that appear around them (context words) obtained from an unlabeled textual corpus is used to train a neural network architecture to predict the probability that a word is likely to be the context words of a given target words. The neural network architecture used in Word2Vec consist of an input layer, one fully-connected hidden layer whose dimension is tunable as a hyperparameter, and a softmax probability layer as an output.

As shown in the left-hand side of Figure 2.1, target-context word pairs are first shown, generated from the sentence *i like apple juice*. In this example, the window size is set to be 2, which means context words can be as far as two words before or after the target word. An illustration of this can be shown in the highlighted cells of the target-context table, where the context words of *apple* are *like*, *i*, and *juice*, which are 1 word before, 2 word before, and 1 word after the target word respectively. The target-context word pairs are then transformed into the training data and label forms required for generating the Word2Vec word embedding. Using the previous example, the training input for the word *apple* is a vector where the *i*-th element corresponds to the alphanumerically-ordered *i*-th word in the corpuses used to generate the word embedding, and that the value of the element where *apple*

is located is 1, and 0 elsewhere. The label corresponding to this input uses the same vector format, but elements with non-zero values instead corresponds to context words of *apple*, as shown in the right-hand side of Figure 2.1

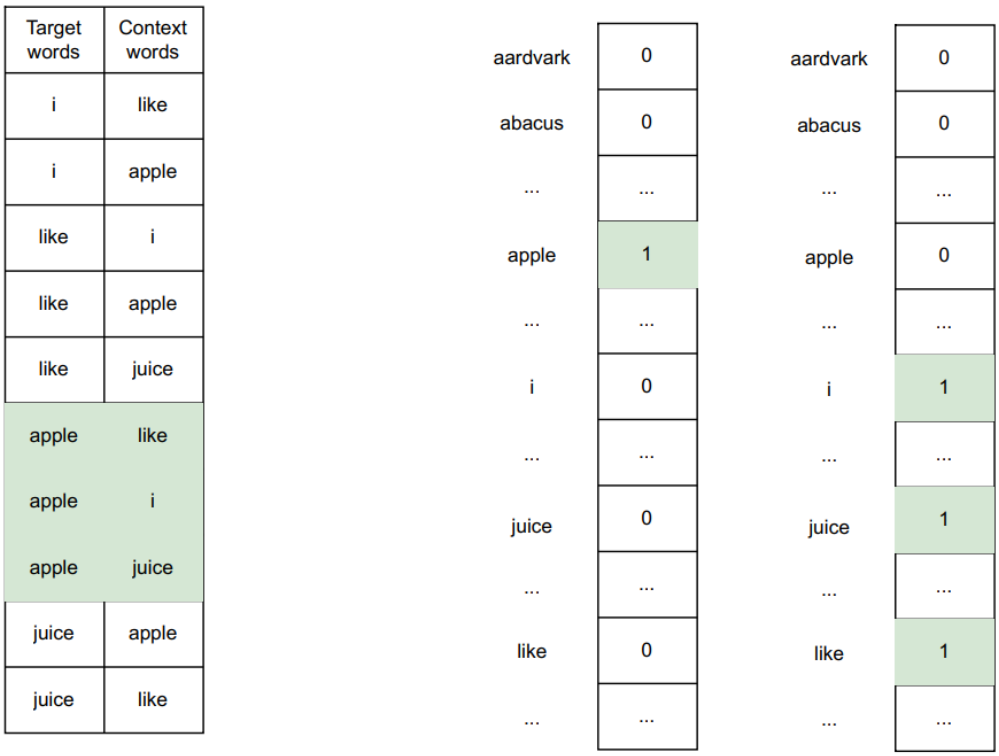

Figure 2.1: (L) Example of target-context word pairs (R) Example of training data and corresponding target, matching to a target word

Figure 2.2 shows the architecture used to obtain Word2Vec word embeddings, using a classification task to predict context words given a target word. The associated Word2Vec word embedding is located at the hidden layer of this architecture, where each word in the corpus is assigned a word vector, whose dimension is tuneable as a parameter at the training process. The word embedding can then be used in other NLP machine learning models, in order to translate words in an input sentence to a representative numeric form.

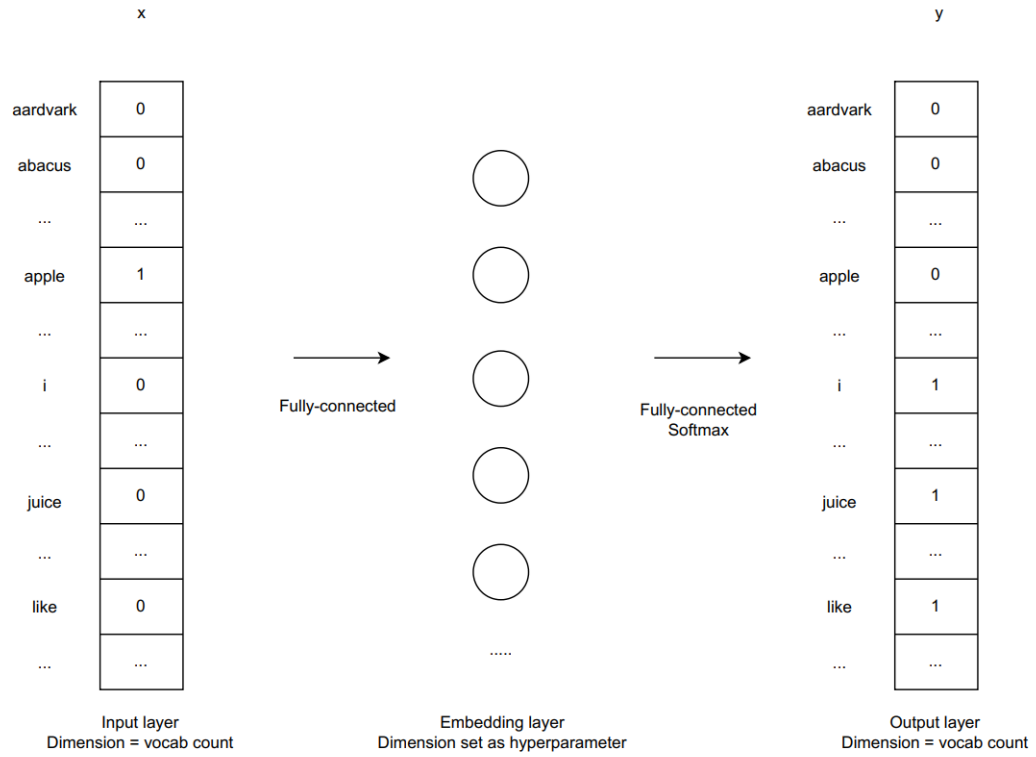

Figure 2.2: Architecture used to obtain Word2Vec word embeddings

In their follow-up research, Mikolov et al. (2013b) introduced two additional methods to improve their original Word2Vec implementations in both training speed and quality: taking out words with low occurrence in the dataset, and the usage of negative sampling on updating weights. In negative sampling, instead of a target word updating the weights of all words at the backpropagation stage, a target word only updates the weights of all context words as well as a random subset of non-context words.

An interesting characteristic of Word2Vec word embeddings is the ability to capture learned relationships between words in the form of word analogies: given word vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{x}$, and the word analogy "*a* is to *x* as *b* is to *y*", finding the word vectors closest to $\boldsymbol{b} - \boldsymbol{a} + \boldsymbol{x}$ results in a vector $\boldsymbol{y}$ that answer the analogy – in essence, $\boldsymbol{a} - \boldsymbol{x} \approx \boldsymbol{b} - \boldsymbol{y}$ in the word embedding space. To use an analogy example "*man* is to *king* as *woman* is to $\boldsymbol{y}$", the operation $\boldsymbol{woman} - \boldsymbol{man} + \boldsymbol{king}$ is closest to the word vector representing the word *queen*, which shows the ability of Word2Vec word embeddings to learn the relationship between gender and the associated title (*king* or *queen*). The relationship is often described using cosine similarity – given word vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{x}$ and the analogy "*a* is to *x* as *b* is to *y*", we can find *y* as the word vector that maximizes the cosine similarity function described in Equation 2.1:

$$cos(\boldsymbol{a}-\boldsymbol{x},\boldsymbol{b}-\boldsymbol{y}) = \frac{(\boldsymbol{a}-\boldsymbol{x}) \bullet (\boldsymbol{b}-\boldsymbol{y})}{\|\boldsymbol{a}-\boldsymbol{x}\|\|\boldsymbol{b}-\boldsymbol{y}\|} \tag{2.1}$$

This form of relationship between words would then be used by Bolukbasi et al. (2016) as well as other word embedding bias literatures to discover bias in word embeddings.

2.2 BiLSTM for Sentiment Analysis

Recurrent neural networks (RNNs) are a form of neural network where the state and outputs of a hidden layer at a given timestep depends on the values on the previous timestep. It is often used in NLP models, where each input layer corresponds to each word in a sentence, and that the contextual information of a preceding word is passed over to process the next word in a sentence. However, standard RNNs is especially susceptible of vanishing/exploding gradient problem, where the partial derivative of weights eventually reaches either very small or very high, divergent values, limiting their usage for longer training epochs (Bengio et al., 1994).

To this end, Hochreiter and Schmidhuber (1997) developed long short-term memory (LSTM), a neural network cell consisting of four interacting layers grouped into three gates: *forget*, *input*, and *output* gate, which has the capability to forget and remember certain information as time passes, allowing the overall model to retain attention to certain parts of the sentence, while clipping gradients to avoid both vanishing and exploding gradient. Figure 2.3 shows an illustration of a singular LSTM block, and Figure 2.4 shows an example of an LSTM layer that receives sentences containing up to 4 words as input, with the maximum sentence length adjustable as a hyperparameter while constructing the neural network architecture. In practice, these inputs are generally way longer than 4 words, and sentences shorter than the maximum limit are padded to reach the limit.

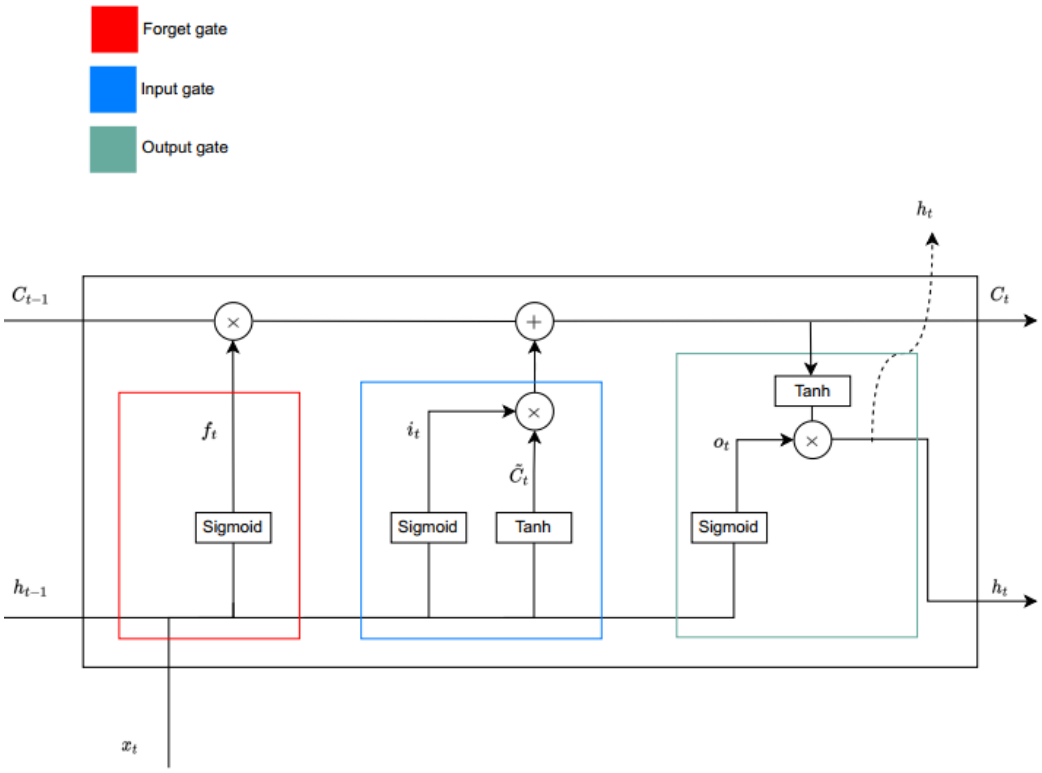

Figure 2.3: Illustration of a single LSTM block

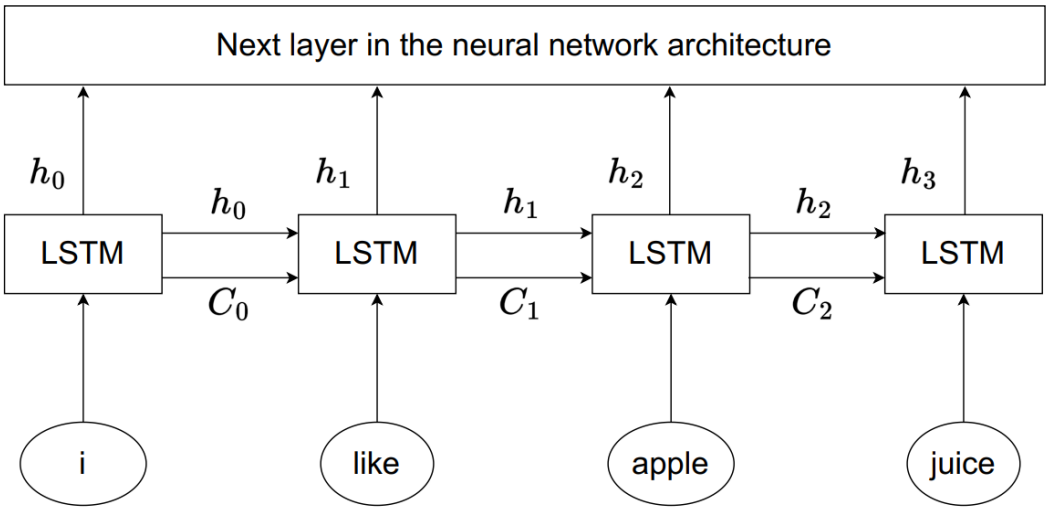

Figure 2.4: Illustration of an LSTM layer

At a certain timestep t, the t-th LSTM block receive both the LSTM results of the previous word in the input sentence $\boldsymbol{h}_{t-1}$ and $\boldsymbol{C}_{t-1}$ as well as the current input $\boldsymbol{x}_t$. As an example, when processing the sentence *i like apple juice*, the fourth ($\boldsymbol{t} = 4$) LSTM block receive the LSTM results ($\boldsymbol{h}_3$ and $\boldsymbol{C}_3$) of the previous word *apple*, processed by the previous iteration, as well as the current word $\boldsymbol{x}_4$ *juice* as

an input. Both of these enter the forget gate first, which consists of a sigmoid neural network layer whose output is between 0 and 1, essentially determining how much information from the previous timestep should be kept and used for the current timestep, based on the contextual knowledge given by both previous and current words. Equation 2.2 shows the calculation of both inputs inside a forget gate, where $\boldsymbol{W}_{fh}, \boldsymbol{W}_{fx}, \boldsymbol{b}_f$ are parameters to be optimized in the neural network learning process.

$$\begin{aligned} \sigma(\ \boldsymbol{x}) \ &= \frac{1}{1+e^{-\boldsymbol{x}}} \\ \boldsymbol{f}_t &= \sigma(\ \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{b}_f) \end{aligned} \tag{2.2}$$

The same input enters the input gate next, which consists of two separate layers: a sigmoid layer that acts as a multiplier on how much information from the current timestep should be used in the final cell state, and a tanh layer that calculates the candidate for current timestep, whose value ranges from -1 to 1. Each of these layers have their own parameters to be optimized in the learning process ($\boldsymbol{W}_{ih}, \boldsymbol{W}_{ix}, \boldsymbol{b}_i$ in sigmoid layer, W_{ch}, W_{cx}, b_c in the tanh layer). Both of these are used to calculate the final cell state C_t using a linear combination between the candidate cell state (the contextual knowledge of the current word in the input sentence – *juice*) and the previous cell state (the contextual knowledge of the previous word in the input sentence – *apple*), each scaled by the sigmoid outputs of the respective gates. as shown in 2.3.

$$\begin{aligned} \tanh(\ \boldsymbol{x}) \ &= \frac{e^{\boldsymbol{x}} - e^{-\boldsymbol{x}}}{e^{\boldsymbol{x}} + e^{-\boldsymbol{x}}} \\ \boldsymbol{i}_t &= \sigma(\ \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{b}_i) \\ \tilde{\boldsymbol{C}}_t &= tanh(\ \boldsymbol{W}_{ch}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{cx}\boldsymbol{x}_t + \boldsymbol{b}_c) \\ \boldsymbol{C}_t &= \boldsymbol{f}_t \times \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \times \tilde{\boldsymbol{C}}_t \end{aligned} \tag{2.3}$$

The current cell state is processed at the output gate, which is pushed through a tanh layer at first, then scaled by a sigmoid layer to express the importance of the current timestep. After scaling, the cell state is passed as a hidden state for the current timestep, to be used in the next time step. The calculations happening in the output gate is detailed in 2.4 below, where $\boldsymbol{W}_{oh}, \boldsymbol{W}_{ox}, \boldsymbol{b}_o$ are parameters to be learnt.

$$\begin{aligned} \boldsymbol{o}_t &= \sigma(\, \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{b}_o) \\ \boldsymbol{h}_t &= \boldsymbol{o}_t * tanh(\, \boldsymbol{C}_t) \end{aligned} \tag{2.4}$$

Graves and Schmidhuber (2005) proposes the usage of bi-directional LSTM, where a sentence is fed into two concurrent LSTM chains that get aggregated together for output, each observing the same sentence from opposite directions, on processing textual data. This approach comes from the intuition that a word may make sense after further contexts in time, e.g., how the contextual meaning of *apple* in *i use apple products* and *i like apple juice* is only revealed after the word *apple* is mentioned (e.g: *apple products* and *apple juice*). Bi-LSTM was shown by Graves and Schmidhuber (2005) to give coherence improvements compared to singular direction Bi-LSTMs, which increases their popularity to be used for various NLP modelling approaches.

After passing through the LSTM layer and all other layers in the architecture, the input enters a final output layer in the architecture. Two common activation functions for sentiment analysis tasks for this layer, which are classification problems in nature, are sigmoid and softmax functions, both giving probabilities of class membership. Table 2.1 shows the mathematical form of sigmoid and softmax functions, as well as their usages and characteristics.

Table 2.1: Characteristics of sigmoid and softmax functions

Characteristics	Sigmoid	Softmax
Mathematical function ($f(x_i)$)	$\frac{1}{1+e^{-x_i}}$	$\frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}}$
Output range	$0 \leq f(x_i) \leq 1$	$0 \leq f(x_i) \leq 1$
Probability characteristics	Probability of all class does not sum to 1	Probability of all class sums to 1
Usage	Multi-label classification tasks	Multi-class classification tasks

2.3 Analyzing and Measuring Unwanted Bias in NLP

Previous literatures have documented the existence of unwanted bias in NLP, generally split into bias in datasets and bias in word embeddings, detailed in further subchapters. While word embeddings learn from various corpuses, which implies bias in word embeddings happen as direct consequences to bias in datasets, their sec-

tions on this thesis are split into separate sections. The decision owes to the practice of using pre-trained word embeddings as inputs for model implementations, which are often created using large, sometimes inaccessible amounts of corpuses in such a way that analyzing each corpus for dataset bias is intractable.

Measuring unwanted bias in NLP happens in multiple stages. Depending on the measurement method, bias measuring happens in datasets and word embeddings where the dataset or word embedding itself is checked for unwanted biases. After training the NLP model, bias measurement happens in the downstream performance where the performance of a model is checked for bias using performance metrics separated by certain identity groups (e.g., false positives of sentences with terms related to Islam and Christianity religions), commonly referred as parity metrics. Debiasing NLP model happens in the dataset or word embedding stage using various methods, where after debiasing the downstream performance is checked again to ensure whether debiasing successfully mitigate prior biases or not.

One common theme of these measurement and debiasing methods is that the majority of them require manual additions of social contexts relevant to the social bias at hand. This makes domain knowledge about the possible sources of bias that may emerge from the dataset, word embedding, and the machine learning methods to be used rank as important as the actual measurement and debiasing methods itself (Bender and Friedman, 2018; Sambasivan et al., 2021).

2.3.1 Dataset Bias

Various researches have elaborated the existence of unwanted bias in dataset for NLP, which comes under the form of biased representation. Biased representation concerns cases where certain terms with multiple possible usages on a sentence only exist for some usages in the dataset, and may be influenced by pre-existing social biases related to the context of the dataset. As an example in Indonesian datasets related to conversation in social media, the word *anjing* mostly exists as part of an insult, without any usage to describe the species. One example of biased representations in a dataset is shown by Hutchinson et al. (2020), where models incorrectly classify non-toxic sentences containing mention of disabilities as toxic due to the massive number of sentences using disability-related terms solely for insults, and no other neutral representations of disabilities on the dataset exists.

Unwanted bias may also come from different aspects of data collecting (Olteanu et al., 2019). One such aspect is data queries, as shown by Wiegand et al. (2019) where datasets obtained with queries of specific terms that correlate with abusive content tend to contain more unwanted bias. Another possible source of unwanted

bias may come from the dataset annotators as discovered by Geva et al. (2019) where NLP models trained on natural language understanding datasets gain increasing performance when given the information of annotators. Additionally, said performance increase does not generalize to annotators of testing data, suggesting that the training data of those datasets may encode annotator bias.

Bias in datasets is measured by calculating the distribution of sensitive identities related to the to-be analyzed bias on the labels, looking for possible misrepresentations or stereotypes that might occur. A common metric to use for this is pointwise mutual information (PMI). Given a term $\boldsymbol{x}$ and a label y, PMI calculates the probability of their co-occurrence given their individual distributions assuming independence between terms and labels (Church and Hanks, 1990), i.e., how often term $\boldsymbol{x}$ is being labelled y.

$$PMI(\boldsymbol{x}, y) = log\frac{p(\boldsymbol{x}|y)}{p(\boldsymbol{x})} \tag{2.5}$$

Higher PMI score implies that the term x is more likely to be associated with label y in the corpus, and is leveraged by Wiegand et al. (2019) to find that neutral terms (e.g., women and announcer) has strong correlation to abusive labels. It is also used by Dixon et al. (2018) to find disproportionate amounts of gender and racial identity terms used in toxic languages, causing the model to generalize sentences that contain them as abusive languages.

2.3.2 Embedding Bias

Bolukbasi et al. (2016) first showed that a word embedding learned from the Google News corpus managed to learn unwanted gender bias, creating stereotyped analogies such as *man is to a computer programmer as woman is to homemaker*. Utilizing the usage of word analogies to measure bias found in their research, Caliskan et al. (2017) developed a statistical test to measure association bias in word embeddings called word embedding Association Test (WEAT). Both papers proved to be seminal on the field of measuring bias in word embeddings, and multiple researches have since expanded on their methods (Manzini et al., 2019) or applied their methods on less represented languages (Sahlgren and Olsson, 2019; Pujari et al., 2019; Takeshita et al., 2020). Another group of researches involve on analyzing the shortcomings of said seminal methods as well as creating new methods that cover them, whether on the type of word embedding bias not covered in previous researches (Gonen and Goldberg, 2019), or on the metrics used to measure word embedding

bias (Ethayarajh et al., 2019).

Bias in word embeddings exist under the form of bias-by-projection (Bolukbasi et al., 2016; Manzini et al., 2019) and bias-by-neighborhood (Gonen and Goldberg, 2019). Word analogies utilizing cosine similarity is a common tool to measure bias-by-projection in word embeddings – by using word analogies catered to certain sensitive identities, multiple researches have found that these analogies contain unwanted social biases and stereotypes. Some examples of researches that utilize word analogies to discover social biases are Bolukbasi et al. (2016) on gender stereotypes and Manzini et al. (2019) on racial and religion stereotypes.

On the other hand, bias-by-neighborhood utilizes the findings that words that encode similar social stereotypes are often close together in the word embedding space (Gonen and Goldberg, 2019) to define a per-word neighborhood metric $\boldsymbol{h}$ as the ratio of top $\boldsymbol{k}$ other words containing similar social stereotypes as the original word. As an example, consider the word *nurse*, and occupational gender bias (i.e., occupations that are stereotyped to be jobs for certain genders) as the bias to be analyzed. Using $\boldsymbol{k} = 10$ as an example, the neighborhood metric first finds the 10 closest words of *nurse* $\boldsymbol{n}$ on the word embedding. If, among these closest words, 2 of them also contain the same occupational gender bias as *nurse* (e.g., *babysitter* and *secretary*, where both are also stereotyped to be female occupations much like *nurse*), then the neighborhood metric for the word *nurse* is $\boldsymbol{h} = 0.2$.

A pseudocode to calculate word analogies and neighborhood metric is described on Algorithm 2.1 and 2.2 respectively.

```
INPUT:  word embedding W ⊂ ℝ^n
        words a, b, x ∈ W
        analogy candidate amount k

analogy_candidates=∅
#create analogy a : b :: x : y
for y ∈ W :
    if y ∉ {a, b, x} :
        #get cosine similarity of analogy
        analogy_candidates.append(cos(a − x, b − y))

#sort
sort analogy_candidates from highest to lowest

#get top k elements
similar_words ← top k words from analogy_candidates

OUTPUT: similar_words
```

Algorithm 2.1: Pseudocode to generate word analogies

```
INPUT:  word embedding W ⊂ ℝ^n
        words X ⊂ W that contain groups of unwanted social
            stereotypes
        word  x ∈ X to be analyzed
        labelling function y : W → {0, 1, ..., p} corresponding to their
            stereotypes
        number of neighbors k

#fit k-nearest neighbor model using all words X
M = KNearestNeighbor(X, y(X))
n = top k neighbors of x using M


h = |{w∈n : y(w) = y(x)}| / k

OUTPUT: h
```

Algorithm 2.2: Pseudocode to generate neighborhood metric of a word x

2.3.3 Downstream Performance

Following Barocas et al. (2017) on impacts of unwanted bias in machine learning implementations, Blodgett et al. (2020) separated the negative impact of unwanted bias in NLP implementations into two types: allocation harms (where the performance of said implementations are tied to memberships of social groups) and representational harms (where the performance of said implementations are tied to stereotypes and other types of misrepresentations of social groups). Representational harm is further separated into multiple subtypes, two of them being under-representation and mis-representation (Crawford, 2017), which are relevant to this thesis.

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 amount 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.

Table 2.2 shows several examples of researches depicting the impact of unwanted bias in NLP, divided into allocation and representational harms. Note that these harms by definition can intersect, and each can cause the other (De-Arteaga et al., 2019), depending on the point of view. Using the earlier example from De-Arteaga et al. (2019), an NLP model that performs well but propagates unwanted stereotypes for certain religions may indirectly result in a performance difference between religion groups. These characteristics make identifying specific types of harms that exist in NLP implementations difficult. This is especially relevant for representational harms, where detecting it requires contextual knowledge of certain social stereotypes with regards to the implementation and the identities being harmed (Blodgett et al., 2020; Sambasivan et al., 2021).

Table 2.2: Examples of existing works depicting bias in NLP, separated into types of harms

Type of harm	Prior research	Description
Allocation	De-Arteaga et al. (2019)	NLP models trained to predict occupations perform worse when gender indicators are removed from the sentences
Allocation	Ball-Burack et al. (2021)	NLP models incorrectly classifies more African-American sourced tweets as toxic compared to other sources
Representational	Kiritchenko and Mohammad (2018)	The performance of multiple emotion detection models agrees with certain unwanted social bias regarding gender and race
Representational	Mehrabi et al. (2021)	Open-source commonsense knowledge bases, often used for question-answering and other NLP uses, contain various stereotypes regarding religion, gender, race, and occupations

Existence of unwanted bias in downstream performance is typically marked by statistically significant difference in model performance of different identity groups, often referred to as parity conditions, which is a common measurement in other fairness in machine learning researches. Some common metrics to be used as parity conditions are false positive rate (FPR) and false negative rate (FNR) (Dixon et al., 2018; Kiritchenko and Mohammad, 2018), as well as true positive rate (TPR) and demographic parity (DP) (Kleinberg et al., 2017), detailed in Table 2.3 below. A significant difference in these metrics over different identity groups that correlates to pre-existing unwanted social bias is often used to imply that said model inflicts allocational harm to certain identity groups by performing worse, or giving worse outcomes than other advantaged groups.

However, Kleinberg et al. (2017) proved that all of these parity conditions cannot be simultaneously achieved together unless in trivial conditions. According to their research, at least one of these parity conditions will not be satisfied. As an example, a model that satisfies false positive, false negative, and true positive parity will not satisfy demographic parity. To this end, Leben (2020) and Selbst et al. (2019) characterized various form of parity metrics into normative principles of fair distribution, and suggests choosing the proper normative principle (i.e, the parity conditions associated with them) based on domain-specific knowledge, context,

and applications of the model.

Table 2.3: Examples of parity conditions used to evaluate allocation bias in downstream performance

Metric name	Mathematical definition	Textual description
False positive rate parity (FPR)	$\frac{FP_A}{TN_A+FP_A} = \frac{FP_B}{TN_B+FP_B}$	The rate of a false positive class prediction should be the same over groups A and B
False negative parity rate (FNR)	$\frac{FN_A}{TP_A+FN_A} = \frac{FN_B}{TP_B+FN_B}$	The rate of a false negative class prediction should be the same over groups A and B
True positive parity rate (TPR)	$\frac{TP_A}{TP_A+FN_A} = \frac{TP_B}{TP_B+FN_B}$	The rate of a true positive class prediction should be the same over both groups A and B
Demography parity (DP)	$\frac{TN_A+FN_A}{TP_A+FP_A+TN_A+FN_A} = \frac{TN_B+FN_B}{TP_B+FP_B+TN_B+FN_B}$	The probability of a negative class prediction should be the same over both groups A and B

Another method used to measure bias in downstream performance is sentence templates Kiritchenko and Mohammad (2018). In this method, sentences are generated from a template with a pre-defined label (e.g: *saya menganut agama [agama]*, originally labelled as neutral in sentiment) and a list of identities to fill said template (e.g: *saya menganut agama* islam and *saya menganut agama* kristen). Since the identities are supposed to be sentiment-neutral, both sentences should be neutral in sentiment, following the original label. Kiritchenko and Mohammad (2018) leveraged this by using prediction difference between sentences from the same template generated with different identities as indicators of bias, particularly one caused by representational harms in the form of misrepresentation or stereotyping (Crawford, 2017).

Using the example above, *saya menganut agama* islam and *saya menganut agama* kristen should both be labeled neutral, since *islam* and *kristen* are by definition sentiment-neutral. Prediction difference between both sentences can then be used as one possible indicator of bias. As an example, if *saya menganut agama* kristen was predicted to be more negative than *saya menganut agama* islam, there is an

indication that the model used to output prediction inflicts representational harm by misrepresenting terms related to Christianity as more negative than terms related to Islam. This method is used by Kiritchenko and Mohammad (2018) to discover that, among other findings, sentence templates filled with female-presenting identities are consistently predicted by sentiment analysis models submitted as a competition entry as angrier than sentence templates filled with male-presenting identities. This finding aligns with pre-existing gender stereotypes, in which women are presupposed to be more emotional than men (Kiritchenko and Mohammad, 2018). Since the performance of these sentiment analysis models mirror existing gender biases, these models are said to inflict representational harms.

2.4 Debiasing Unwanted Bias in NLP

Existing researches pinpoint datasets and word embeddings as two main components in NLP implementations that can contain unwanted social biases, and that these biases can impact downstream performances of models that learn from them. As such, many bodies of work have been dedicated on researching how to debias unwanted social biases from them, which will be detailed in further subchapters.

2.4.1 Dataset Bias

Dataset debiasing methods aim to influence the distribution of datapoints containing sensitive identities on certain labels, in order to neutralize the association of sensitive identities towards labels. This can be done by either adding external positive/neutral datapoints containing sensitive identities to the dataset (Dixon et al., 2018), resample existing datapoints that has low classification certainty, defined by the highest difference of prediction probability $|p_i(\boldsymbol{x}) - p_j(\boldsymbol{x})|$ for a datapoint $\boldsymbol{x}$ and all pair of labels (i, j) in the dataset (Ball-Burack et al., 2021), or removing negative datapoints with sensitive identities (Wiegand et al., 2019).

2.4.2 Embedding Bias

Debiasing word embeddings are split into three steps: identifying the bias subspace, neutralizing bias-neutral words, and equalizing words in equality sets. In order to find the bias subspace, we first define multiple equal-sized defining sets of bias-defining words $D = \{D_i\}, |D_i| = |D_j| \forall i, j$, then calculate defining direction as the difference between each element of the set. The contents of each defining set D_i depends on what types of bias are being measured, e.g:

$\{man, woman\}, \{father, mother\}$ for gender bias between man and woman (Bolukbasi et al., 2016) and $\{church, mosque\}, \{priest, imam\}$ for religious bias between Christianity and Muslim (Manzini et al., 2019).

The k-dimension bias subspace $\boldsymbol{B}$ is then defined as the first k principal component analysis (PCA) components of all previous defining directions as calculated in Algorithm 2.3. Then, all bias-neutral words in the word embedding are debiased by projecting them to the orthogonal subspace of $\boldsymbol{B}$ in the neutralizing step in Algorithm 2.4. To obtain bias-neutral words, a set of bias-specific words is first defined. Then, a set of bias-specific words is obtained by taking the complement between said bias-specific word set and all words in the word embedding.

For specific bias-defining word sets (referred to as family sets by Bolukbasi et al. (2016)) required by the user, an additional equalizing step as shown in Algorithm 2.5 is taken for those words, ensuring that neutral words are also equidistant to said set. Taking from Bolukbasi et al. (2016) and Manzini et al. (2019), after debiasing unwanted religion bias from a word embedding, we may want to maintain the contextual relationship of the recently-debiased word *baca* to certain religion-defining words, e.g., being closer to $\{quran, alkitab\}$ than to $\{masjid, gereja\}$. By using family sets $E_1 = \{quran, alkitab\}, E_2 = \{masjid, gereja\}$ as an input for Algorithm 2.5, we maintain the relationship between *baca* and the mean of E_1 (which contain the contextual knowledge of religious scriptures given by *quran* and *alkitab*), as well as the mean of E_2 (which contain the contextual knowledge of places of worship given by *masjid* and *gereja*) while being equidistant from elements of E_1 and E_2 to maintain the religion-neutrality of *baca*.

A short pseudocode to explain the overall process of debiasing word embeddings, taken from Bolukbasi et al. (2016), is detailed from Algorithm 2.3 to 2.5 below.

INPUT: word embedding $\boldsymbol{W} \subset \mathbb{R}^n$
set of defining sets $D = \{D_i\}, D_i = \{\boldsymbol{w}_j\}, 1 \leq i \leq p, 1 \leq j \leq k$
bias subspaced dimension $k \geq 1$

$\boldsymbol{C} = \emptyset$
for $1 \leq i \leq p$:
    $\boldsymbol{\mu}_i = \sum_{\boldsymbol{w}_j \in D_i} \frac{\boldsymbol{w}_j}{|D_i|}$
    for $\boldsymbol{w}_j \in D_i$:
        $\boldsymbol{C} \rightarrow \boldsymbol{w}_j - \boldsymbol{\mu}_i$

OUTPUT: $\boldsymbol{B} \rightarrow$ first k PCA components of $\boldsymbol{C}$

Algorithm 2.3: Pseudocode to generate bias subspace

INPUT: word embedding $\boldsymbol{W} \subset \mathbb{R}^n$
word to debias $\boldsymbol{w} \in \boldsymbol{W}$
bias subspace $\boldsymbol{B}$ with orthogonal basis $b_i, 1 \leq i \leq k$

$\boldsymbol{w}_B = \boldsymbol{0}$
for $1 \leq i \leq k$:
    $\boldsymbol{w}_B = \boldsymbol{w} + (\boldsymbol{w}_B \bullet \boldsymbol{b}_i)\boldsymbol{b}_i$
$\boldsymbol{w}' = \frac{\boldsymbol{w} - \boldsymbol{w}_B}{\|\boldsymbol{w} - \boldsymbol{w}_B\|}$
OUTPUT: Neutralized word $\boldsymbol{w}'$

Algorithm 2.4: Pseudocode to neutralize words

INPUT: word embedding $\boldsymbol{W} \subset \mathbb{R}^n$
bias subspace $\boldsymbol{B}$ with orthogonal basis $b_i, 1 \leq i \leq k$
set of defining sets $D = \{D_i\}, D_i = \{\boldsymbol{w}_j\}, 1 \leq i \leq p, 1 \leq j \leq k$
set of family sets to equalize $E = \{E_1, E_2, \ldots, E_m\}, E_j \subset W$

for $1 \leq j \leq m$:
$\boldsymbol{\mu}_i = \sum_{\boldsymbol{w}_j \in D_i} \frac{\boldsymbol{w}_j}{|D_i|}$
$\boldsymbol{\mu}_B = \boldsymbol{0}$

for $1 \leq i \leq k$:
$\boldsymbol{\mu}_B = \boldsymbol{\mu} + (\boldsymbol{\mu}_B \bullet \boldsymbol{b}_i)\boldsymbol{b}_i$
$\boldsymbol{v} = \boldsymbol{\mu} - \boldsymbol{\mu}_B$

for $\boldsymbol{w} \in E_j$:
$\boldsymbol{w}_B = \boldsymbol{0}$
for $1 \leq i \leq k$:
$\boldsymbol{w}_B = \boldsymbol{w}_B + (\boldsymbol{w}_B \bullet \boldsymbol{b}_i)\boldsymbol{b}_i$
$\boldsymbol{w}' = \boldsymbol{v} + \sqrt{1 - \|\boldsymbol{v}\|^2} \frac{\boldsymbol{w}_B - \boldsymbol{\mu}_B}{\|\boldsymbol{w}_B - \boldsymbol{\mu}_B\|}$

OUTPUT: All equalized words $\boldsymbol{w}' \in E_j \forall j$

Algorithm 2.5: Pseudocode to equalize words

Figure 2.5 to 2.7 shows an illustration of debiasing religion bias from English word embeddings utilizing VERB, a tool to visualize the impact of word embedding debiasing on English word embeddings (Rathore et al., 2021). This example uses defining set $D = \{\{islam, christian\}, \{quran, bible\}\}$, family set $E = D$, and sample religion-neutral words $w = \{violent, terrorist, conservative, judgmental\}$, following Manzini et al. (2019).

Figure 2.5 shows the result of bias subspace projection, as a result of Algorithm 2.3. Here, the defining set D is used to create a 1-dimension bias subspace B, and words W are projected to the subspace. As seen on the image, *islam* and *quran* are located on the right side of the subspace origin, whereas *christian* and *bible* are located on the left side. As found by Manzini et al. (2019), the words *judgmental*, *terrorist*, and *violent* are closer to Islamic-representing words, whereas *conservative* is closer to the Christianity-representing word. Since all words in W are religion-neutral in meaning, yet are closer to certain religious identities in the bias subspace, it follows from Bolukbasi et al. (2016) that this word embedding contains unwanted religion bias. In this case, the religion bias comes from the stereotypes of certain religious identities commonly seen in various media.

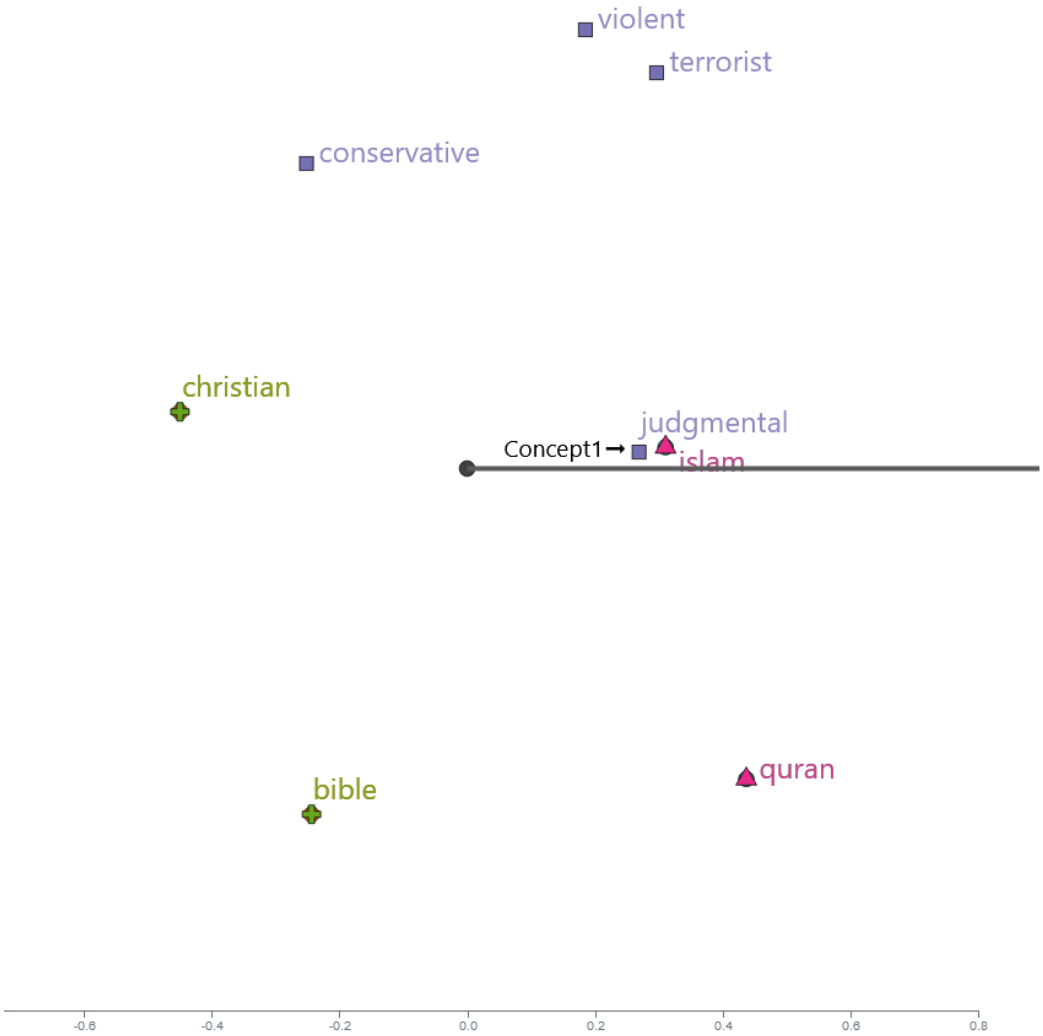

Figure 2.5: Example of projecting words to a bias subspace

Figure 2.6 shows the result of bias neutralizing, following Algorithm 2.4. Since the bias after projecting to the subspace is represented by words being on either side of the origin, debiasing is therefore done by bringing all religion-neutral words to the origin. This debiases the religion neutral words, with respect to the bias subspace B at hand.

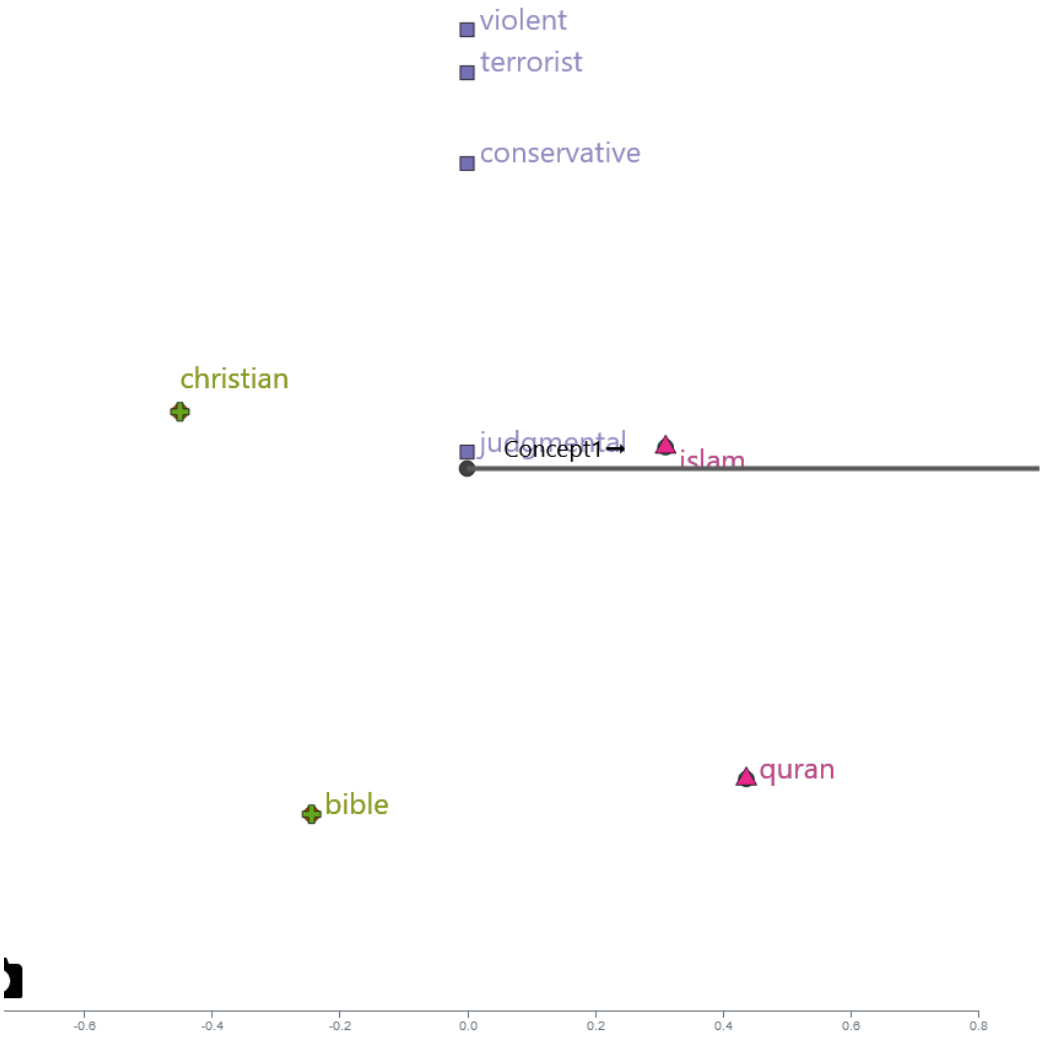

Figure 2.6: Example of neutralizing religion-neutral words

Figure 2.7 shows the result of bias equalization, following Algorithm 2.5. This algorithm focuses on the religion-specific words in the family set $E = D = \{\{islam, christian\}, \{quran, bible\}\}$. The result maintains the distance between all four words in the family set, and as such maintains semantic meaning between all of them, while equalizing all words in family sets with respect to the bias space.

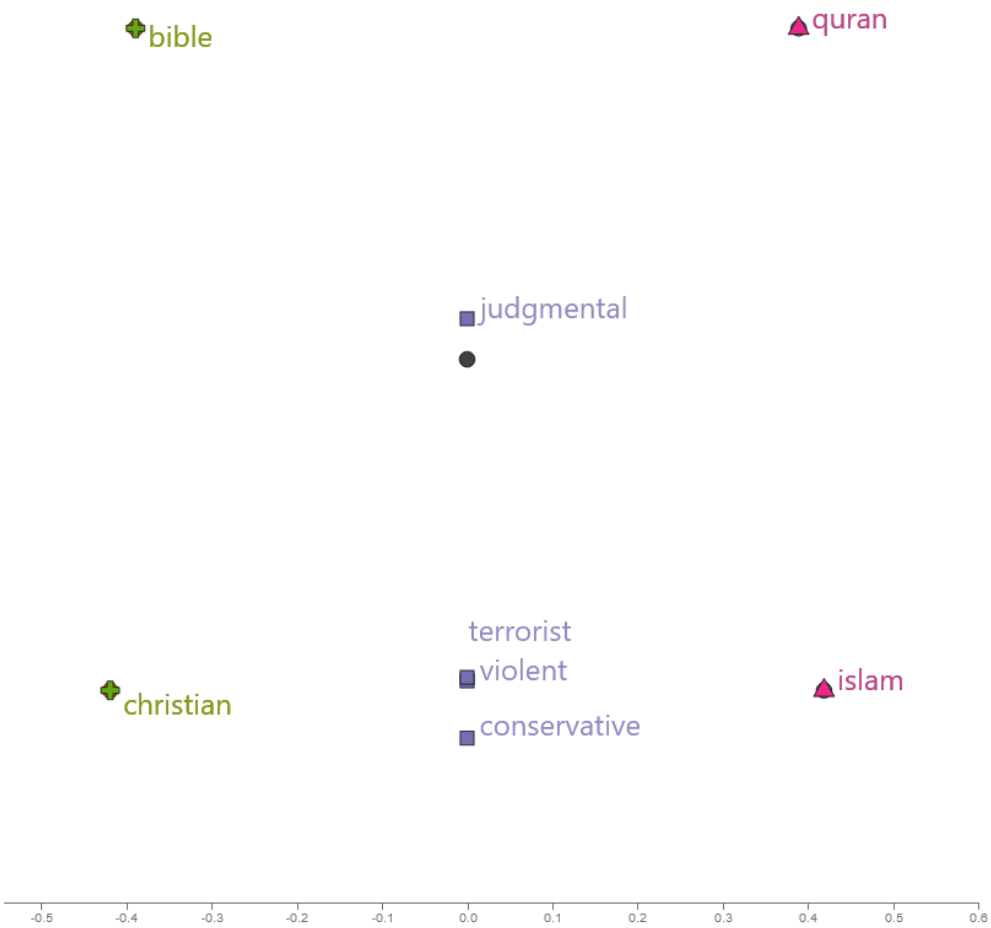

Figure 2.7: Example of equalizing religion-specific words

CHAPTER 3
RESEARCH METHODOLOGY

This section describes the experimental setups used in this thesis, ranging from the datasets and embeddings used, bias measurements, and debiasing methods.

3.1 Datasets and Embeddings

This thesis uses three Indonesian NLP datasets for measuring dataset bias and downstream performance. The NLP tasks represented by these datasets are emotion detection (Saputri et al., 2018) that classifies a sentence into emotions (anger, fear, happy, love, sadness), sentiment detection (Purwarianti and Crisdayanti, 2019) that classifies a sentence into sentiments (positive, negative, neutral), and multi-label hate-speech and abusive language detection (Ibrohim and Budi, 2019) that classifies sentences into hate-speech, abusive, both, or neither. A short description of all datasets used in this thesis is shown in Table 3.1. These datasets are chosen particularly because of their similarity of source – all three uses Twitter as their source for sentences, as well as additional social media sources (Instagram, Facebook) and other websites (Zomato, TripAdvisor, Qraved) for SmSA. Two of these datasets (EmoT and SmSA) are also used in IndoNLU (Wilie et al., 2020) as a benchmark for Indonesian NLU systems, particularly their sentiment analysis implementations.

Table 3.1: Short description of all datasets used in this thesis

Dataset Name	Literature	Positive labels	Negative labels	Source
EmoT	Saputri et al. (2018)	happy, love	anger, fear, sadness	Twitter Streaming API from June 1-14, 2018, filtering Indonesian geolocation coordinates
SmSA	Purwarianti and Crisdayanti (2019)	neutral, positive	negative	Crawling and cleaned texts from Twitter, Zomato, TripAdvisor, Facebook, Instagram, and Qraved
Hate Speech	Ibrohim and Budi (2019)	none	hate speech, abusive	Twitter Search API obtained from March to September 2018

Before the training process, all of these datasets receive the same pre-processing treatments, namely:

1. Removing Twitter tags (username, 'RT')
2. Removing linebreaks
3. Removing emojis
4. Converting to lowercase letters
5. Removing punctuations
6. Replacing slang words with an existing slang word dictionary, obtained from Saputri et al (2018) as well as Ibrohim and Budi (2019)
7. Stemming

In order to analyze embedding bias, this thesis uses multiple open-source Word2Vec word embeddings, each trained on different Indonesian sources: Twitter data (Saputri et al., 2018), Wikipedia, Tempo online news media (Kurniawan, 2019), and Common Crawl data used for CoNLL (Zeman et al., 2018). For ease of writing and reference, starting from this point, the mention of embeddings refer to Word2Vec word embeddings.

3.2 Bias Framework and Measurements

In this thesis, it is hypothesized that the impact of algorithmic enclave on religious discourse (Lim, 2017) as well as the limited representation of marginalized religions in media (Remotivi, 2021) causes representational harm in the form of underrepresentation of positive content for marginalized religious identities in various sources (e.g: Twitter, news sources). These sources would then be used to create NLP datasets and train word embeddings, which introduces biases in said datasets and embeddings. NLP models that train from these biased resources inherits said unwanted religious bias from the datasets and embeddings, becoming biased NLP models themselves. The bias contained in this NLP model would cause allocation harm, where models performing worse for sentences containing marginalized religious identities compared to sentences without. In parallel, the bias contained in the NLP model would also reinforce the representational harms caused by the datasets and embeddings in the form of misrepresentation, where terms corresponding to

marginalized religious identities are wrongly related to negativity, as shown in Figure 1.1.

Figure 3.1 shows the framework concerning unwanted religious bias sources and harms caused by said bias used in this thesis.

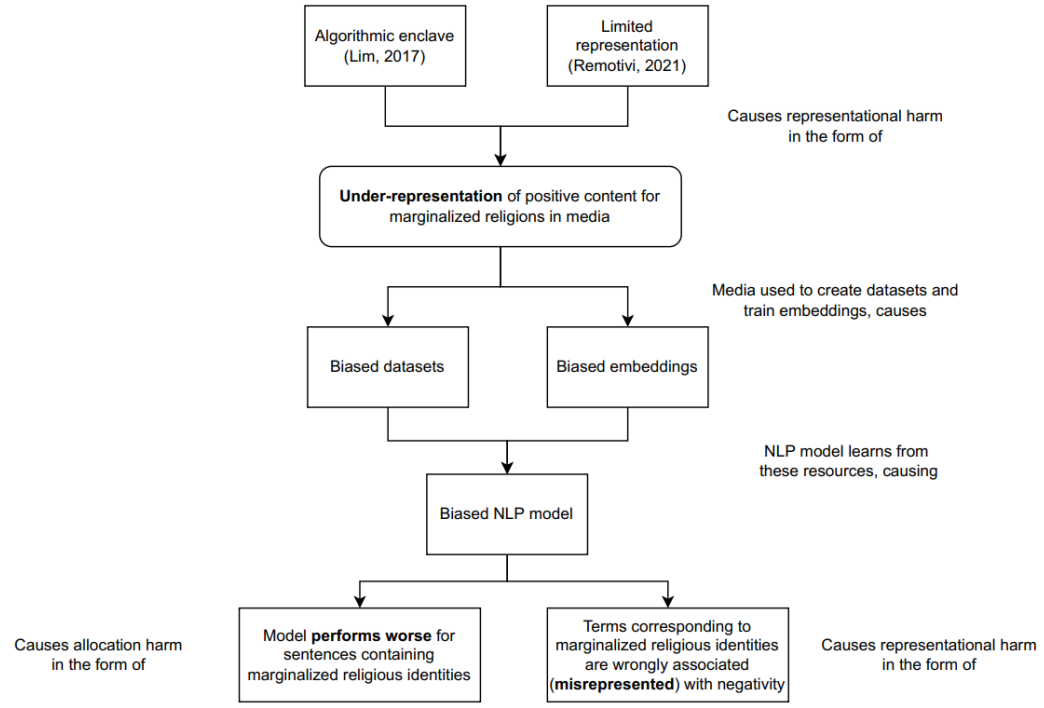

Figure 3.1: Framework of sources of bias and harms caused by said bias

Figure 3.2 shows an overview of the bias detection method done in this study. We conduct PMI test to detect religion bias at dataset level, whose processes is detailed on Figure 3.3. Word analogy and word similarity is done to measure religion bias at embedding level, which is detailed on Figure 3.4. For downstream performance level, allocation harm is analyzed using parity metrics, detailed on Figure 3.5, whereas representation harm is analyzed using sentence templates as shown on Figure 3.6.

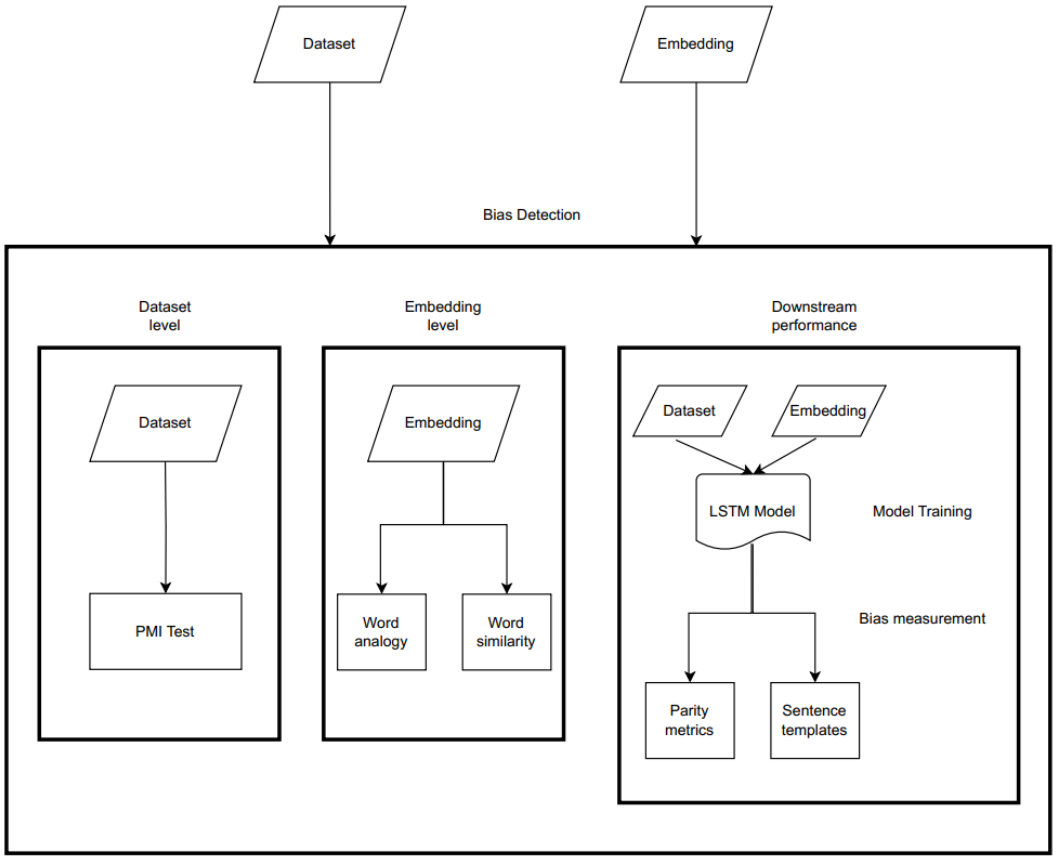

Figure 3.2: Flowchart on detecting religious bias

The following subsections detail the measurement of bias in datasets, embeddings, as well as downstream performance.

3.2.1 Dataset Bias

This thesis uses PMI (Church and Hanks, 1990) to measure unwanted dataset bias between a set of religious terms W (detailed in Table 3.2) and labels that exist in each dataset. The usage of PMI as a method to detect dataset bias, as shown by Dixon et al. (2018) and Wiegand et al. (2019), aligns to the case of religion bias in Indonesia. This comes from the bias source framework shown in Figure 3.1, where due to the effects of algorithmic enclaves and minimal positive media representation (Lim, 2017; Remotivi, 2021), religious identities are more related to negativity-related sentences due to the amount of sentences which uses religious identities as insults. As such, the association between terms representing religious identities to insults can be used to detect dataset bias, much like the case of toxic language detection (Dixon et al., 2018) and abusive language detection (Wiegand et al., 2019).

Additionally, this thesis utilizes PMI instead of PPMI (positive pointwise mutual

information, where negative values of PMI scores are brought to zero) as a result of the preliminary research preceding this thesis (Fauzan, 2022), where it is found that although terms representing religious identities have higher PMI scores to negative labels compared to positive labels, the PMI scores between religious terms are still in the negatives. As an example, for the term *islam* in the SmSA dataset case, Fauzan (2022) found that the PMI score between said term to the *positive* label is -3.15, whereas the PMI score for the *negative* label is -0.13. Here, the negative scores are caused by the low mention of the term *islam* over the whole document, but is majorly used in *negative*-labeled sentences, which explains the higher PMI score to the *negative* label.

Table 3.2: Religious terms W to be used to detect dataset bias using PMI

	Religious Terms			
Religion	Religion Name	Place of Worship	Scripture	Person
Islam	islam	masjid	quran	ulama
Christian	kristen	gereja	alkitab	pendeta

Existing works revolving on dataset bias have used PMI to determine the existence of unwanted bias against certain identities (Dixon et al., 2018; Wiegand et al., 2019). However, their usage are limited to single-label, binary-classification task where the unwanted social bias is represented by membership to a certain class. As an example, for an abusive language detection case (Wiegand et al., 2019), each sentence being analyzed is represented by a single-label, binary-class between 'is abusive' or 'is not abusive', where unfair membership between certain identities to the 'is abusive' class determines the existence of bias in said dataset. However, for multi-class and multi-label cases, additional criteria needs to be added in order to consider which classes and which labels contribute to the existence of dataset bias using PMI. To this end, we turn to the under-representations of certain religious identities in media besides conflicts and celebrations (Lim, 2017; Remotivi, 2021) as the rationale for definitions of unwanted dataset bias in Table 3.3.

As shown on Figure 3.1, it is hypothesized that misrepresentation of religious identities, as well as the under-representation of positive content for religious minorities as described in Lim (2017) and Remotivi (2021) is reflected by the dataset having disproportionately more 'negative'-labeled sentences that contain certain religious terms than 'positive'-labeled sentences. The under-representation of positive content on religious terms, especially for marginalized religions, impacts the PMI score between religious terms and the negative labels in the dataset. These

negativities may manifest in multiple sentiment categories; some examples are fear-mongering campaigns unfairly targeting certain religious identities, or inciting hate-speech comments aiming to anger oppositions.

Manually defining the types of negativities to be used on a given dataset to define dataset bias requires the contextual knowledge of what specific types of negativities exist at the data collection method, in the sources containing said data (Olteanu et al., 2019). This is not feasible in bigger datasets, which feature sentences obtained from multiple textual sources and long time periods, and as such may contain multiple types of negativity against a certain religion identity. Therefore, we use the existence of at least one type of negativity-related classes and labels outrepresenting all existing positivity-related classes and labels in the dataset to define the existence of dataset bias.

In this definition, a dataset is said to contain unwanted religious bias for a certain word if the PMI score between said word and at least one negative label (defined as a label that represents 'negativity', e.g, *negative* in SmSA, *anger*, *fear* and *sadness* in EmoT, *hate speech* and *abusive* in Hate Speech dataset) is higher than the PMI score between said word and all other non-negative labels (e.g, *positive* in SmSA, *happy* and *love* in EmoT, *none* in Hate Speech dataset). We then choose terms that represents certain religious identities depicted in Table 3.2, then aggregate the PMI results between labels by averaging over all words, ending up with the dataset bias definition on Table 3.3. In the case where no sentences with a certain label and word exists, causing an infinite value on its PMI calculation, the word is taken out at the aggregation step for that label. This expands on the usage of PMI to detect dataset bias (Dixon et al., 2018; Wiegand et al., 2019) for tasks with multi-class label.

Since each dataset has different labels, we then define unwanted bias at a dataset level. Given $P_{label} = \{PMI(w, label)|w \in W\}$ as a set containing all PMI values between all terms in W and the label *label*, and μ_{label} as the average of all elements in P_{label}, we proceed with the unwanted bias definition as shown in Table 3.3 below:

Table 3.3: Definitions of unwanted dataset bias

Dataset	Definition
EmoT	$(\mu_{anger} > \mu_{happy}\ AND\ \mu_{anger} > \mu_{love})\ OR\ (\mu_{fear} > \mu_{happy}\ AND\ \mu_{fear} > \mu_{love})\ OR\ (\mu_{sadness} > \mu_{happy}\ AND\ \mu_{sadness} > \mu_{love})$
SmSA	$\mu_{negative} > \mu_{positive}\ \ OR\ \ \mu_{negative} > \mu_{neutral}$
Hate Speech	$\mu_{hatespeech} > \mu_{none}\ \ OR\ \ \mu_{abusive} > \mu_{none}$

Figure 3.3 shows the flowchart to detect dataset bias using PMI test, as previously mentioned. Given a dataset, P_{label} for all labels in the input dataset is obtained using the religious terms W defined at Table 3.2. Then, μ_{label} is obtained as the average of all elements in P_{label}. Dataset bias is defined by checking whether the bias condition seen in Table 3.3 for the dataset currently being analyzed is fulfilled.

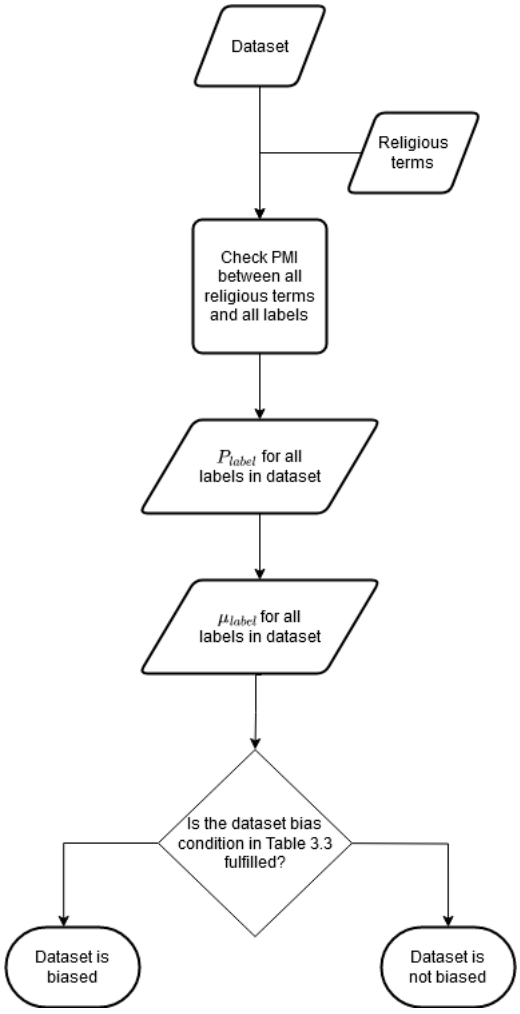

Figure 3.3: Flowchart on detecting religion bias at dataset level

3.2.2 Embedding Bias

In order to measure embedding bias, this thesis utilizes word analogies as with existing embedding bias literatures (Bolukbasi et al., 2016; Manzini et al., 2019). For each word analogy depicted in Table 3.4, we show the top 5 words in each embedding that are the closest to said comparison, and each of said words is manually determined to find any unwanted religious bias.

The analogies used to test religious bias in word embeddings attempt to capture the religion bias that may exist in media form, as impacted by limited representation

of marginalized religions in Indonesian context (Lim, 2017; Remotivi, 2021). The first analogy (*islam* is to *kafir* as *kristen* is to *x*) is used to capture whether the word *kafir*, which has been co-opted by social media interactions to insult other marginalized religions, is wrongly related to other aspects of marginalized religions.

The second analogy (*kristen* is to *kejadian* as *islam* is to *x*) is used to analyze the impact of limited positive representations of marginalized religions in media form (Remotivi, 2021). As is, media coverages that concern marginalized religions mostly cater to either celebrations (e.g., major religion occurrence days) or conflicts (e.g., hate crimes done to them), with minimal variations otherwise. Therefore, there is a possibility that word embeddings may relate marginalized religious identities to negativity, due to the representation issue. Here, the word *kejadian* (*incidents* in English) is used since the word *kejadian* is often used in Indonesian news reports for accidents in Indonesia, whether religion-related or not.

The third analogy (*islam* is to *onta* as *kristen* is to *x*) is used to analyze the impact of algorithmic enclaves (Lim, 2017) in embedding bias. The word *onta*, meaning the animal species *camel*, is co-opted to be used as political insults against individuals identifying as Muslims by algorithmic enclaves that opposes them. Therefore, the equivalence between *islam* and *onta* is used to capture whether word embeddings inherit otherwise religion-neutral terms that are co-opted into political insults, with the expectation that the equivalence between *kristen* and *x* returns other related political insults as *x*.

Table 3.4: Analogies used to test embedding bias

Analogies
islam:kristen::kafir:x
kristen:islam::kejadian:x
islam:kristen::onta:x

Additionally, this thesis uses word-similarity by cosine similarity of certain religious terms, owing to Gonen and Goldberg (2019) on using k-means to detect embedding bias. Their usage of k-means to determine gender bias in word embeddings require all words in an embedding to be manually labeled on which gender they may contain biases for as a ground truth. This is not doable due to lack of resources for this method for religion bias. As an example, it is unclear whether the word *nista* itself is supposed to contain bias against marginalized religions, even if it is often used as part of a sentence that insults them (e.g: *menistakan agama*). However, since k-means and cosine similarity both measure closeness in terms of word vectors, we can implement cosine similarity as an alternative. Owing to Gonen and Goldberg (2019), embedding bias may manifest in religious terms being close to other words that contain religious bias. For each term and each embedding, we take the top 5 words closest to said term using cosine similarity, which measures the 'closeness' of two vectors, then manually determine whether the words close to them contain some sort of religious bias. The existence of religious bias in an embedding is defined by the existence of a term representing common religious insults being close to a term representing religious identity, or vice versa. A table consisting of all religious terms used for word-similarity is shown on Table 3.5.

Table 3.5: Religious terms to be used for word similarity test

Word contexts	Terms
Religious identities	islam, kristen
Common religious insults	radikal, haram, nista, kafir, halal, syiah

Figure 3.4 shows the flowchart to detect religion bias at embedding level, as previously described. Given a word embedding, we first output the top 5 word analogy and word similarity results, using word analogies mentioned at Table 3.4 and terms for word similarity mentioned on Table 3.5. Then, given a test method, if at least one output contain religion bias, the embedding is biased using said method.

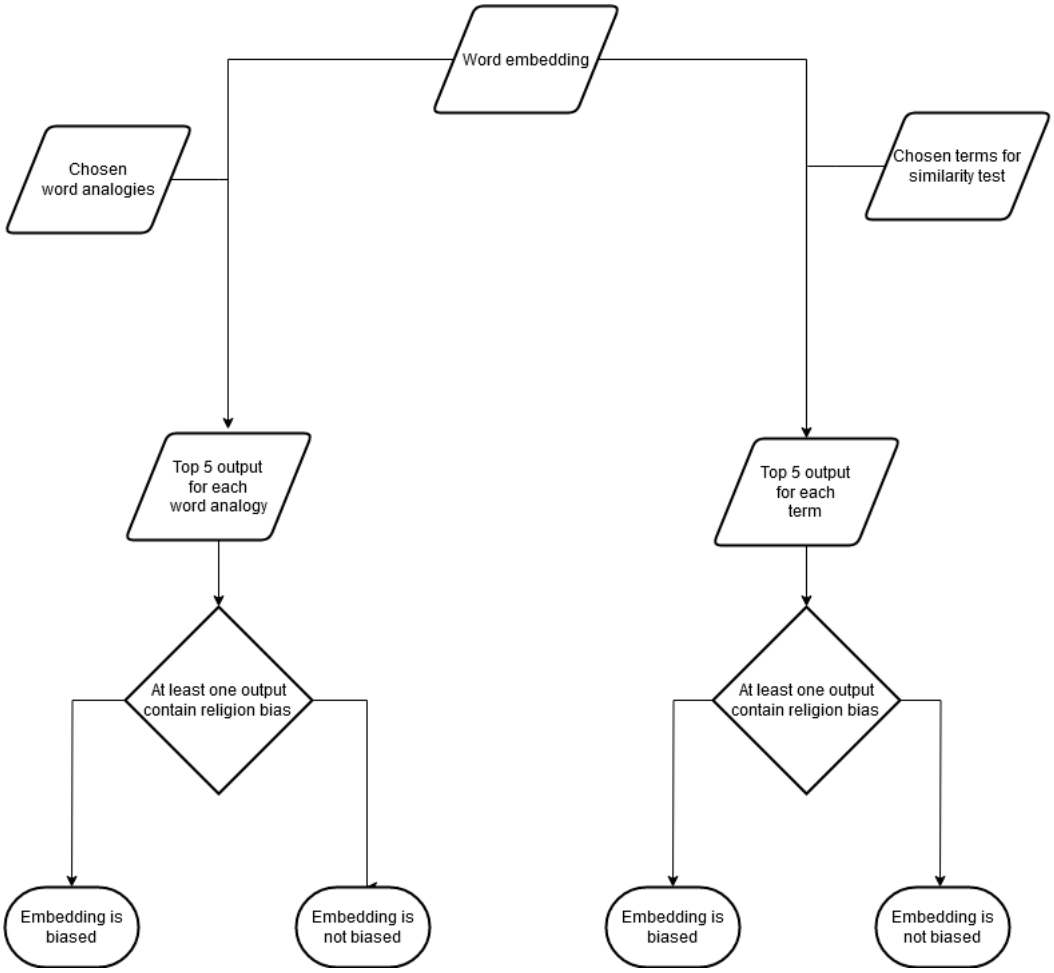

Figure 3.4: Flowchart on detecting religion bias at embedding level

3.2.3 Downstream Performance

The impact of religion bias on downstream performance is calculated for each dataset and embedding, using a Bi-LSTM neural network model with each embedding mentioned in the previous section. The aim of this thesis is to analyze the existence of biases in datasets and embeddings as well as the impact of the models that learn from them, and as such Bi-LSTM is chosen due to their ease of use in model training.

Due to each dataset having different machine learning tasks (EmoT and SMSA being single-label classification, and Hate Speech being multi-label classification), the output of BiLSTM model is different for each dataset – Bi-LSTM for EmoT and SmSA dataset uses softmax activation at the output, whereas BiLSTM for Hate Speech uses sigmoid activation. For each dataset and embedding, the training to create a corresponding Bi-LSTM neural network model uses a five-fold cross-validation approach. For each fold, the training and validation data, as well as the

training and validation accuracy score is stored. The fold that provides the highest validation score after training a Bi-LSTM model is regarded as the best fold for a dataset-embedding combination. For each dataset and embedding, the result of accuracy and downstream performance results correspond to said best model.

For each model and dataset, downstream performance is calculated on both allocation and representational harms, as well as accuracy scores to measure overall model performance. The accuracy scores shown are the training and validation scores of the respective fold the chosen model is trained on. Allocation harm is calculated using all parity conditions in Table 2.3 using a simplified version of each classification task which is described below. The usage of parity metrics to measure allocation harm follows from Dixon et al. (2018) and Ball-Burack et al. (2021), where models that learn from datasets in which certain marginalized identities are used only in negativity-related sentences impact results of parity metrics shown in Table 2.3. This aligns to the bias source framework seen in Figure 3.1, where religion bias manifests in religious identities only used in negativity-related sentences.

For EmoT and SmSA dataset, we simplify both tasks by transforming the predicted sentiment into its positive/negative form as seen in Table 3.1 (e.g: in EmoT dataset, if the predicted sentiment is ‘anger’, the simplified output is ‘negative’ because ‘anger’ counts as ‘negative’ sentiment for this thesis). For Hate Speech dataset, simplification is done by checking both labels: if the prediction of at least one label (i.e: ‘hate speech’ or ‘abusive’) is higher than 50%, the simplified output is ‘negative’, else ‘positive’. The difference in approach for Hate Speech is because both labels to be predicted in the dataset are ‘negative’, whereas EmoT and SmSA dataset has ‘positive’ and ‘negative’ class in their labels. This simplification is done because the exact label prediction does not specifically matter for constituting bias - only the label groups matter. As an example, it does not matter whether a sentence originally labelled as ’happy’ is mislabeled into ’sad’ or ’angry’ to constitute the impact of bias, only that the misprediction from positive label to negative label happens.

After simplifying the output into a binary classification between ‘negative’ and ‘positive’ class, we then obtain two specific subsets of the original dataset, each containing Islamic and Christianity religious terms from Table 3.2, then calculate all parity conditions in Table 2.3 using the simplified outputs. For example, using the false negative parity condition, we measure whether sentences containing Islamic terms are more likely to be mis-predicted as negative than sentences containing Christianity terms. We then repeat this method for all models.

Figure 3.5 shows the flowchart to detect religion bias at downstream perfor-

mance level, focusing on allocation harms. Given a dataset and embedding, a Bi-LSTM model is first trained on those resources. Then, we first obtain the subset of sentences containing Islamic and Christianity terms from the dataset. After that, parity conditions as shown in Table 2.3 is calculated, using the simplified output method. An unequal parity metric result for a certain religious group (Islam or Christianity) is used to detect whether allocation harm is inflicted for that religious group.

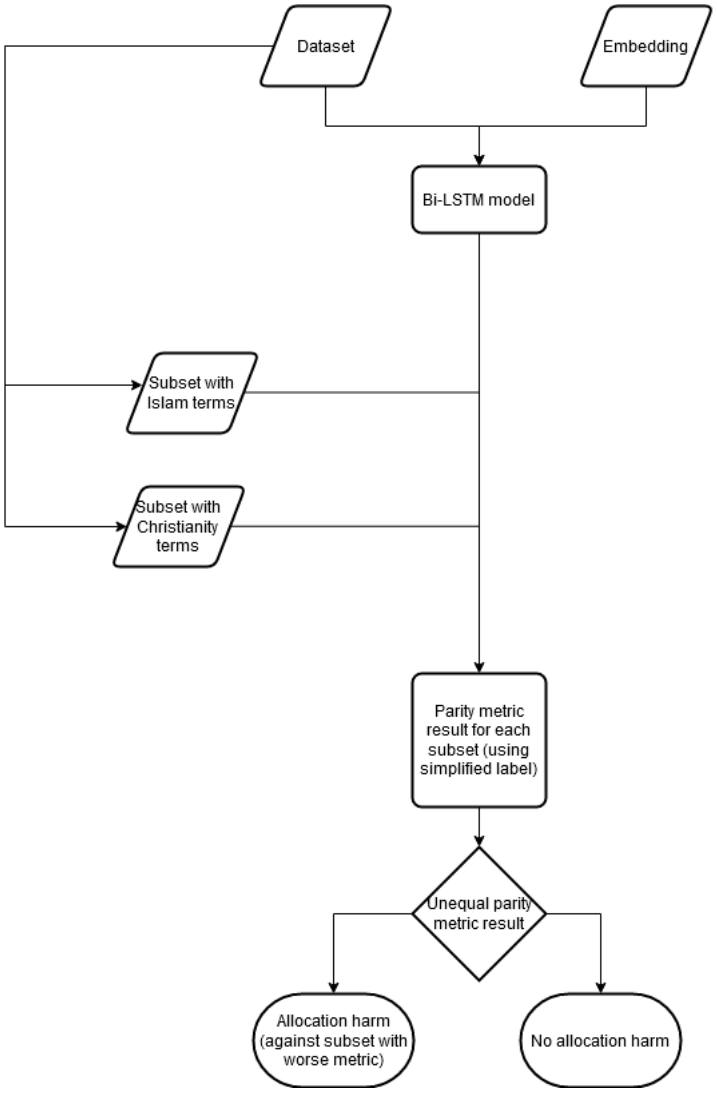

Figure 3.5: Flowchart on detecting religion bias at downstream performance level, focusing on allocation harm

Evaluation of representation harm is done by reviewing the probability score of template sentences for each label – following Kiritchenko and Mohammad (2018), assessment regarding the difference between ground truth labels (neutral/positive, non-hate-speech, non-abusive) and the actual predictions made by the model serves as the basis to determine the existence of said harm. The specific usage of sentence templates method instead of existing sentences from the dataset the model is trained on follows from Kiritchenko and Mohammad (2018), where existing sentences from datasets may contain sentences with multiple social identities (e.g.: religious identities), or may contain other words that can impact prediction results. By carefully constructing neutral sentences with only religious identities being the sole difference, the representation harms caused by the model can be measured.

The list of sentence templates used to evaluate representation harm is detailed in Table 3.6. Note that some sentences are both used to debias datasets and test downstream performance (overlap marked in bold, in Table 3.6). The overlap is added to check whether the results of dataset debiasing are merely the effect of recalling existing datapoints, instead of generalizing from the dataset itself (i.e: debiasing only improves performance for the bolded templates). All of these templates undergo the same cleansing process received by the training data, which may result in the representation harm results showing sentences that don't exactly match the original template, yet are semantically equal (e.g: *saya menganut agama [agama]* to *saya anut agama [agama]* due to stemming).

The exact sentence templates utilized by Kiritchenko and Mohammad (2018), which is used to detect representational harms in the gender/racial bias case, is not applicable to the religion bias case done in this thesis. This is because the construction of templates done in Kiritchenko and Mohammad (2018) stems from the assumption that first names corresponds to racial and gender identities in Western countries, and uses first names as a proxy for racial and gender identity. However, at the time of writing this thesis, no research confirming the correspondence between first names and religion identities in the Indonesian case exists, and as such this thesis opts to create new lists of templates. The author acknowledges the limitations of this approach, particularly in the lack of existing points of reference in which the proposed results for sentence templates can be measured.

The different templates used in Table 3.6 are to test the impact of biases on a range of sentiments. The first three templates (*saya anut agama [agama]*, *saya cinta agama [agama]*, *sekolah saya mengajarkan agama [agama]*) are used to check whether neutral sentences are impacted by biases, whereas the templates *tenggang rasa antar kaum [agama] harus dijaga* and *[tempat ibadah] jadi tempat aman bagi*

seluruh masyarakat indonesia detects the impact of bias on positive sentences. A special interest is given on the template 'saya tidak setuju dengan ajaran agama [agama]', which is used to detect the impact of bias on negative sentences. Since dataset bias is previously defined in the form of over-representing negative sentences that contain religious terms, this template is used to analyze whether sentences that are labelled as negative, but otherwise do not contain religion bias can be impacted by religion bias that exists in datasets and embeddings.

Table 3.6: Sentence templates used to test dataset bias

Identity	Templates
[agama]	**saya anut agama [agama]** **saya cinta agama [agama]** **sekolah saya mengajarkan agama [agama]** tenggang rasa antar kaum [agama] harus dijaga saya tidak setuju dengan ajaran agama [agama]
[tempat ibadah]	**[tempat ibadah] jadi tempat aman bagi seluruh masyarakat indonesia**

Figure 3.6 shows the flowchart to detect religion bias at downstream performance level, focusing on representation harms. Given a dataset and embedding, a Bi-LSTM model is first trained on those resources. In parallel, we generate sentences obtained for sentence templates by filling the templates with identities, as shown on Table 3.6. Then, we calculate the prediction output for each generated sentence. If there are any mispredicted sentence, then it is said that the Bi-LSTM model inflicts representation harm, against the religious identity represented in the mispredicted sentence.

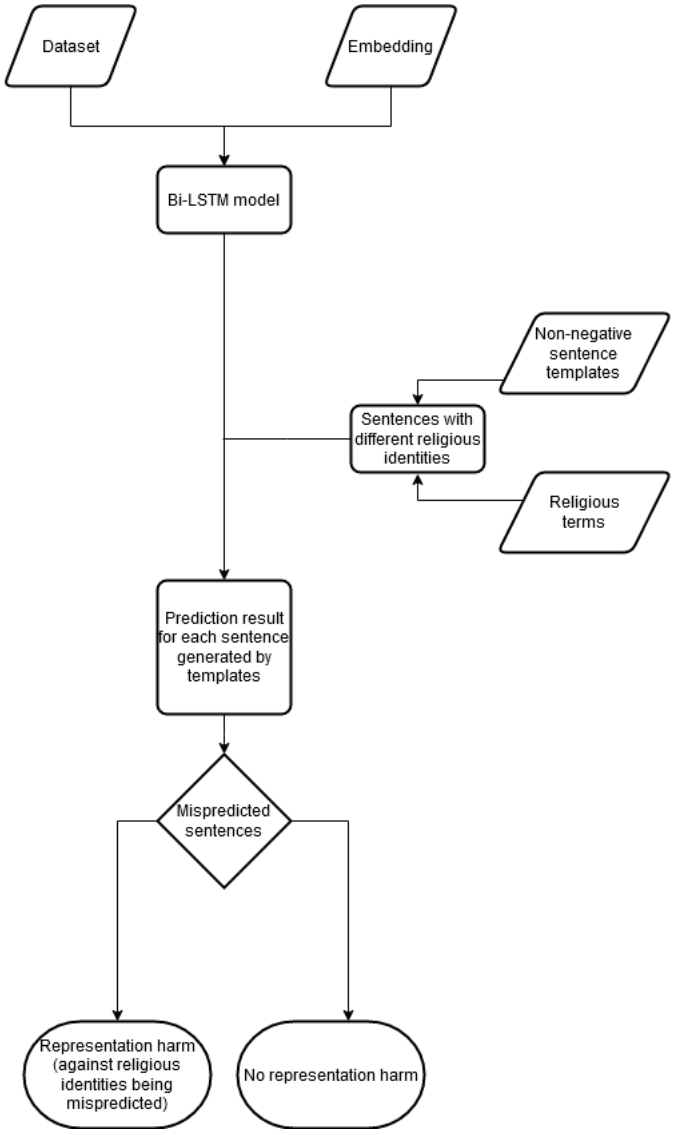

Figure 3.6: Flowchart on detecting religion bias at downstream performance level, focusing on representation harm

3.3 Debiasing Methods

The effect of debiasing is checked by analyzing the existence of religion bias after debiasing dataset only, debiasing embedding only, and debiasing both dataset and embedding. This thesis uses 5 debiasing methods in total, which is shown in Table 3.7

Table 3.7: List of debiasing methods done in this thesis

Debiasing method	Definition	Debiased component
Dataset debiasing using sentence templates	Dataset debiasing using sentence templates	Dataset
Dataset debiasng using Wikipedia	Dataset debiasing using sentences obtained from Wikipedia	Dataset
Embedding debiasing	Embedding debiasing	Embedding
Joint debiasing using sentence templates	Combination of dataset debiasing using sentence templates and embedding debiasing	Dataset and embedding
Joint debiasing using Wikipedia	Combination of dataset debiasing using sentences obtained from Wikipedia and embedding debiasing	Dataset and embedding

For individual debiasing methods, defined as a debiasing method which are not combinations, analysis of bias on the debiased component (dataset or embedding) is done, on top of analyzing bias at downstream level (parity metrics and sentence templates, for allocation and representation harm respectively) after debiasing. In particular, analysis of religion bias at dataset level is done after individual dataset debiasing methods, whereas analysis of religion bias at embedding level is done after embedding debiasing. For joint debiasing methods, defined as a debiasing method which are combinations, analysis of bias is only done at downstream performance, since the analysis of bias for each individual component is already done after each individual debiasing method.

Figure 3.7 shows an overview of what components are being debiased for each method, and what level of biases are checked after debiasing. The methods of dataset debiasing is shown in Figure 3.8 (for dataset debiasing using sentence templates) and Figure 3.9 (for dataset debiasing using sentences obtained from Wikipedia), whereas the methods of embedding debiasing is shown in Figure 3.10

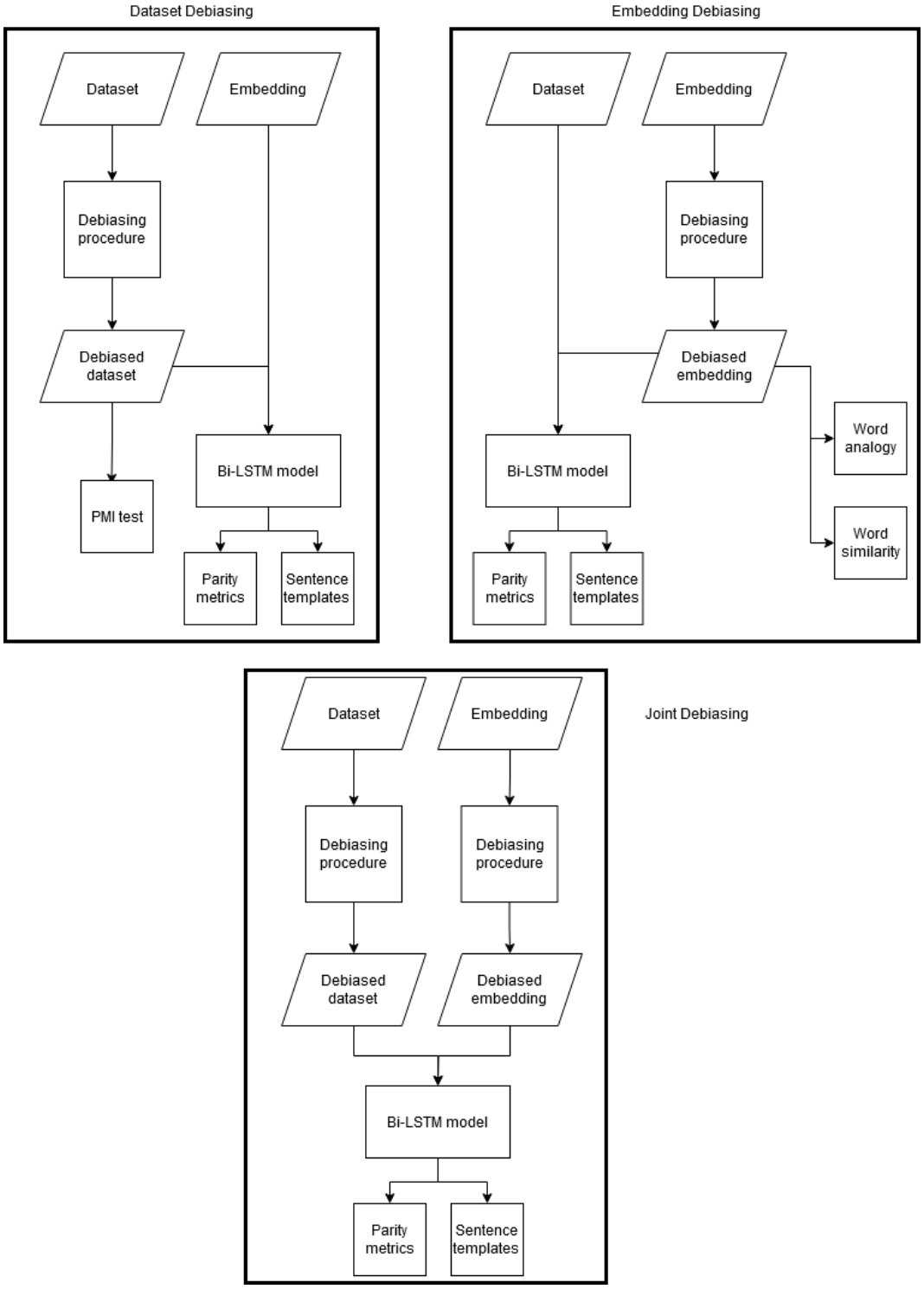

Figure 3.7: Overview on mitigating religion bias

The following subsections describes debiasing methods for datasets and embeddings.

3.3.1 Debiasing Datasets

Dataset debiasing is done by augmenting datasets with positive/neutral datapoints from 2 different sources: using sentence templates as previously done in Fauzan (2022), as well as using sentences in Indonesian Wikipedia articles containing religion-specific terms as mentioned in Table 3.2, which follows from Dixon et al. (2018). Since the bias framework shown in Figure 3.1 posits the manifestation of religion bias in datasets in the form of religious identities mostly existing on negativity-related labels, augmenting datasets in order to rebalance label distributions mitigates the dataset bias. This aligns with the existing work done by Dixon et al. (2018), both in how the dataset bias manifests and in the mitigtion approach, and as such their dataset augmentation method is applicable for this thesis.

In order to augment dataset using sentence templates, we first provide sentence templates to be filled with religious terms, as done by Fauzan (2022). A full list of these templates are shown in Appendix 1, whereas the religious terms referred to are shown in Table 3.2. After filling all sentences with all religious terms, manual inspection is done to remove sentences that do not make sense semantically (e.g., *hormatilah umat islam yang sedang beribadah di gereja*, for a total of 57 sentences. Following our prior work (Fauzan, 2022), we duplicate these sentences 10 times, resulting in 570 sentences total. These sentences are divided into 260 sentences representing Christianity, and 270 sentences representing Islam.

Some of the sentence templates used to debias datasets overlap with the sentence templates used to evaluate representation harm. An example of this is the template *saya anut agama [agama]*, which exists for both debiasing and representation harm evaluation. This addition is done to analyze whether dataset debiasing by sentence templates are affected by overfitting, due to the duplicated sentences done prior to debiasing.

Figure 3.8 shows the flowchart to mitigate dataset bias using sentence templates, as previously described.

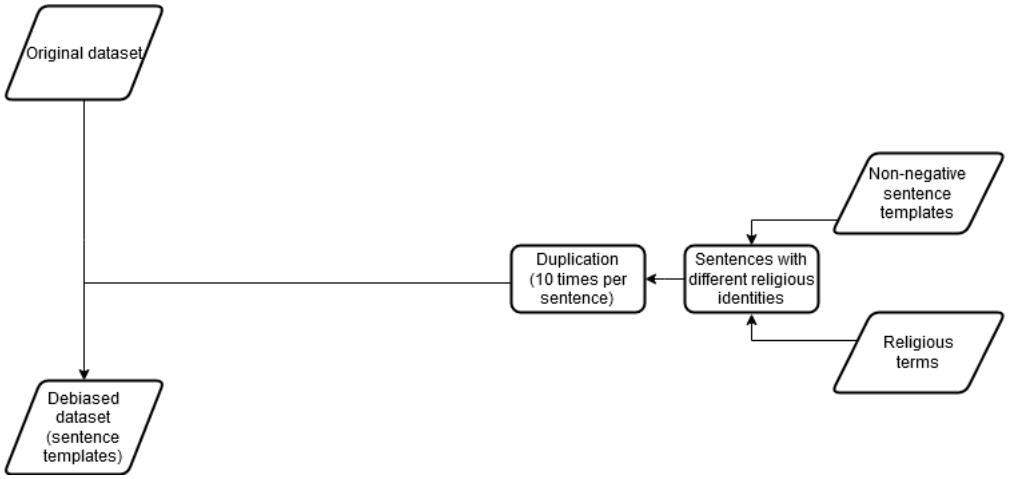

Figure 3.8: Flowchart on mitigating dataset bias, using sentence templates

In order to use Wikipedia sentences to debias datasets, sentences from an existing Indonesian Wikipedia article dump are first cleansed by removing metatags and punctuations, as well as removing accents from texts. For each religious term in Table 3.2, we take sentences containing at least one religious term referred in Table 3.2, as well as longer than 5 words, from a Wikipedia article corresponding to said term. As an example, for the religious term *islam*, we take sentences longer that five words containing the word *islam* from an Wikipedia article exactly titled *Islam*, or the closest if exact title match is not found.

The dataset debiasing method done by Dixon et al. (2018) to debias toxic language detection datasets assumes that Wikipedia moderation ensures that sentences contained in Wikipedia are non-toxic by default. In their example, dataset bias is represented by unfair label representation of sentences containing certain marginalized identities. Therefore, their assumption allows sentences obtained from Wikipedia that contain mentions of certain marginalized identities to re-balance the label distribution of sentences containing certain terms representing marginalized identities, which was the goal of their research.

However, their assumption does not hold for the labels used in this thesis, in which unwanted religious bias is represented by sentences with negative sentiment or emotion (Lim, 2017; Remotivi, 2021). This is because of the existence of sentences that are negative in sentiment or emotion but does not necessarily constitute religious bias. This makes it so that direct adaptation of dataset debiasing using Wikipedia articles, as shown by Dixon et al. (2018), is not feasible due to the aforementioned sentences. As an example, for the term *islam*, directly adapting the Wikipedia debiasing method from Dixon et al. (2018) may result in obtaining, among others, a sentence lamenting the death of an Islamic figure. This sentence

can be considered as a negative-emotion sentence, and as such is unfit to re-balance the label distribution of the term *islam* in the dataset, even if it does not contain bias against Islam.

The limitation of articles required to have the exact title as the religious term is required in order to ensure that the sentences used for debiasing are non-negative, and as such are able to debias datasets. Since the chosen Wikipedia articles corresponding to each religious terms are informational in nature (e.g., the Wikipedia article about Islam contains sentences that explain certain aspects of Islam), the sentences obtained from these articles are non-negative, which allows re-balancing of label distribution using these sentences. Additionally, the limitation of 5 words was done in order to exclude Wikipedia stubs as well as entries of a list, that otherwise contains religious terms.

After obtaining all sentences from all religious terms, which adds up to 1107 sentences, we stratify sample the sentences per article, resulting at 583 unique sentences used for dataset debiasing. All 8 sentences corresponding to the term *pendeta* are omitted prior to stratify sampling since manual inspection of these sentences show that a considerable amount of sentences obtained from Wikipedia also mention *pendeta* in the context of non-Christianity religions, and is such is not relevant to our attempt to debias Islam-Christianity religion bias.

Since different Wikipedia articles contain different amount of sentences, stratified sampling is done to ensure that each religious term are represented in the dataset debiasing method. Using 0.55 as sampling ratio, for each article, we take sentences from said article equal to the sampling ratio, then repeat this for all articles.

In order to test the previous assertion that all of the sentences obtained using this method are non-negative in label, a manual inspection of all 583 sentences are done, and confirms our prior assertion. Table 3.8 shows the religious terms, the corresponding Wikipedia article, as well as the amount of sentences before and after stratified sampling.

Table 3.8: Amount of sentences obtained from Wikipedia per religious term to be used for dataset debiasing

Religious term	Name of Wikipedia article	Total sentences	Sampled sentence count
islam	Islam	114	62
kristen	Kekristenan	386	200
quran	Al-Qur'an	137	72
alkitab	Alkitab	97	49
masjid	Masjid	221	122
gereja	Gereja	129	71
ulama	Ulama	15	7

Figure 3.9 shows the flowchart to mitigate dataset bias using sentences obtained from Wikipedia, as previously described.

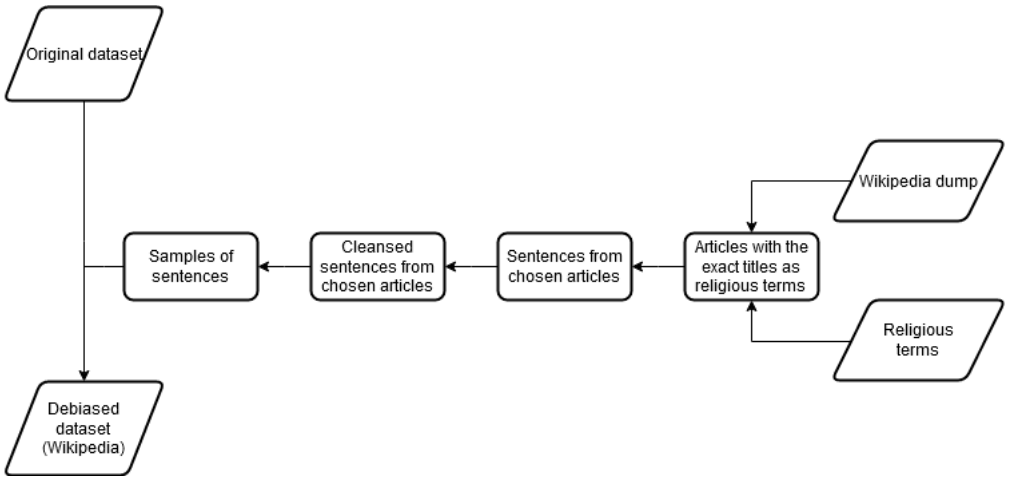

Figure 3.9: Flowchart on mitigating dataset bias, using sentences obtained from Wikipedia

3.3.2 Debiasing Embeddings

Embedding debiasing is done using Bolukbasi et al. (2016) to debias embeddings, adapted for religion identity (Manzini et al., 2019). As all of the terms used to define the bias subspace done in Manzini et al. (2019) is directly applicable to Indonesian, their embedding debiasing method is directly applicable for this thesis. As such, this thesis shows the effectiveness of their method on the downstream performance cases of emotion detection, sentiment analysis, and hate-speech detection, in the Indonesian case.

The defining sets used to debias embeddings are shown in Table 3.9. Following Manzini et al. (2019), this thesis uses the defining set as equality set. Following Bolukbasi et al. (2016), in order to determine which words are to be debiased on each word embedding, we first determine a set of religion-specific words filled with words related to religious identities used in the defining sets by manually skimming through Kamus Besar Bahasa Indonesia (KBBI). For each embedding to be debiased, we take religion-neutral words, defined as all words outside of the religion-specific word set that exists in said embedding, to be debiased. The set of religion-specific words is detailed in Table 3.9.

Table 3.9: Defining sets to debias embeddings, adapted from Manzini et al. (2019)

Identity type	Defining sets
Religion name	*islam, kristen*
Place of worship	*masjid, gereja*
Scripture	*quran, alkitab*
Person	*ulama, pendeta*

Table 3.10: List of religion-specific words used to calculate religion neutral words

Identity type	Religion-specific words
Religion name	*islam, kristen, nasrani, protestan, katolik, koptik*
Place of worship	*masjid, mesjid, musala, mushalla, katedral, gereja, pentekosta, sinode*
Scripture	*quran, alkitab, alquran, injil*
Person	*ulama, pendeta, muslim, muslimah, kristiani, biarawan, biarawati, pastor*
Activity	*solat, sholat, salat*

Figure 3.10 shows the procedure to mitigate embedding bias, using Hard Debiasing (Bolukbasi et al., 2016; Manzini et al., 2019) as previously described. The defining sets described in Table 3.9 is used to generate the bias subspace using Algorithm 2.3, Then, all bias-neutral words, defined as the list of all words in a word embedding which is not listed in the bias-specific words in Table 3.10, are neutralized using Algorithm 2.4. Finally, all word pairs in the family sets, which re-uses the defining sets following Manzini et al. (2019) are equalized by using Algorithm 2.5. The result is a debiased word embedding, with the neutralized bias-neutral words and equalized words from family sets.

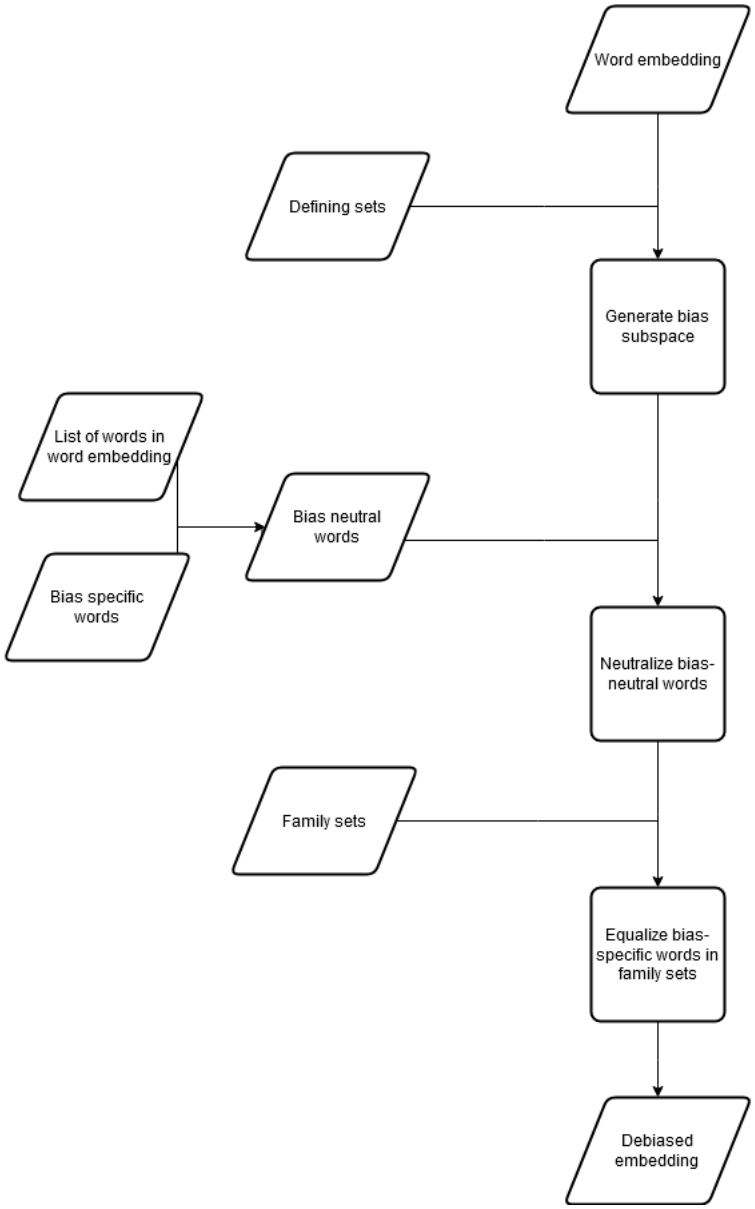

Figure 3.10: Flowchart on mitigating embedding bias

CHAPTER 4
BIAS DETECTION RESULTS

This section describes the result of our bias detection, where we show the existence of unwanted religious bias in dataset and embedding level, as well as how they affect downstream performance.

4.1 Dataset Bias

Table 4.1 shows the result of term occurence per label for the EmoT dataset. As seen in the tables, religious terms are more often used in negativity-label sentences, namely *anger* and *fear*. An example is the term *islam*, which are predominantly mentioned in sentences labeled *anger*. This trend is also seen for the terms *kristen*,*quran*, and *ulama*, which confirms the impact of algorithmic enclaves amplifying the amount of negativity-related discussions for religious identities (Lim, 2017).

An interesting outlier is the religious term *gereja*; sentences using this term often shows up in sentences labeled *fear* (10 times) instead of *anger* (3 times). However, these sentences still fall into negative-labeled sentences, and as such still aligns with the effect of algorithmic enclaves (Lim, 2017). This is strengthened upon further investigation, which finds that out of the 10 sentences containing *gereja* labeled *fear*, 5 of them are the exact same tweets, quote-retweeted by other users with minor additions. This further highlights the power in which algorithmic enclaves hold over discourses of certain topics, particularly their frequency.

Table 4.1: Religious term occurrence per label in EmoT dataset

Terms	anger	happy	sadness	love	fear
islam	15	7	5	2	3
kristen	3	2	0	0	1
masjid	2	10	1	1	3
gereja	3	2	1	0	10
quran	4	3	1	0	1
alkitab	0	0	0	0	0
ulama	5	0	2	0	0
pendeta	0	0	1	0	0

Table 4.2 shows the PMI averages for each label in the EmoT case. As seen in the table, there exists a negative label *anger*, for which the PMI averages of this label is higher than the PMI averages of all positive labels *happy* and *love*. Therefore, using the previous definition of dataset bias shown in Table 3.3, we conclude that the EmoT dataset contain unwanted religious bias. This confirms our first hypothesis, in which identities related to religious groups are mistakenly related to negativity in datasets. An interesting pattern that shows from Tables 4.1 and 4.2 is that the effect of biased, negative representation of religion in datasets, as taken from Lim (2017), also exists for the majority religion (i.e., Islam), as opposed to only for marginalized religions as depicted by Remotivi (2021). This is most clearly shown by the religious terms *islam* and *kristen* - both sentences most often shows in sentences labeled *anger*, instead of only *kristen*. This suggests that the context of religion discourse in social media (Lim, 2017; Olteanu et al., 2019) may play a higher role on the biased representation compared to the minimal representation of marginalized religions in media (Remotivi, 2021).

Table 4.2: μ_{label} of all labels in EmoT dataset

Label	μ_{label}
anger	-1.07
happy	-1.24
sadness	-1.72
love	-2.80
fear	-1.71

Table 4.3 shows the result of term occurence per label for the SmSA dataset. Much like the EmoT case, religious terms in this dataset more often occurs on sentences with negativity-related label *negative*. This effect is more prominent in the SmSA dataset compared to the previous EmoT case. As an example, the term *kristen* appears 1 time in *positive* sentences, 0 times in *neutral* sentences, yet appears 28 times in *negative* sentences. Also like the EmoT case, terms corresponding to Islam are impacted as much as terms corresponding to Christianity, which shows that religion bias impacts both religious groups.

Table 4.3: Religious term occurrence per label in SmSA dataset

Terms	positive	neutral	negative
islam	5	9	103
kristen	1	0	28
masjid	5	3	6
gereja	2	1	6
quran	0	0	7
alkitab	0	0	1
ulama	1	2	27
pendeta	0	0	4

Table 4.4 shows the PMI averages of all labels in the SmSA dataset, calculated by using the term-label occurrence as shown in Table 4.3. Here, it shows that there exists a negative label *negative*, of which all PMI averages are higher than all positive labels *positive* and *neutral*, which confirms that dataset bias exists in the SmSA dataset. As claimed by Saputri et al. (2018), the creation of SmSA dataset involves multiple sources, including social media sources (Twitter, Facebook, Instagram) as well as non-social media sources (TripAdvisor, Zomato, Qraved). Yet on further inspection, the majority of *negative*-labeled sentences that contain religious terms are insults, whether aimed to entities corresponding to said term, or using memberships of said religious term to insult other out-groups. Since the non-social media sources used to create the SmSA dataset mostly revolves around reviews, and as such are not likely to mention religious terms, it can be assumed that most, if not all, sentences containing religious terms come from social media sources, where algorithmic enclaves (Lim, 2017) forms and interacts.

The effects previously described is a form of population bias (Olteanu et al., 2019), where even for sentences representing the same religious term, different sources draw different user demographics, with their own usage of that religious term. This particular manifestation of population shows the contextual knowledge needed to analyze the source of religion bias in datasets: knowledge of the media platforms, the types of biases users of the media platform may show, as well as the collection methods used to gather data from said media into a dataset. These forms of knowledge intersects between each other, and are required in order to gain a full picture of the unwanted social bias at hand (Blodgett et al., 2020; Olteanu et al., 2019; Wiegand et al., 2019).

Table 4.4: μ_{label} of all labels in SmSA dataset

Label	μ_{label}
positive	-2.28
neutral	-2.05
negative	-0.26

Table 4.5 shows the result of term occurrence per label for the Hate Speech dataset. Unlike the cases found in EmoT and SmSA dataset, religious terms tend to appear in sentences with non-negative labels as much, or more, than sentences with negative labels. An example is shown in the religious term *islam*, which appears in *hate speech* sentences around as much as *none* sentences (i.e., sentences that are neither *hate speech* nor *abusive*). Another is shown by the term *kristen*, which appears significantly less in *hate speech* and *abusive* sentences compared to *none* sentences.

Table 4.5: Religious term occurrence per label in Hate Speech dataset

Terms	hate speech	abusive	none
islam	360	151	353
kristen	42	14	276
masjid	11	3	23
gereja	3	1	56
quran	46	31	36
alkitab	1	0	3
ulama	97	43	207
pendeta	3	1	5

The PMI averages for the Hate Speech dataset as shown in Table 4.6 show that for both the negative labels *hate speech* and *abusive*, none of the PMI averages are higher than that of the positive label *none*. Therefore, using the definition of dataset bias shown in Table 3.3, the Hate Speech dataset does not contain dataset bias. According to the creation process (Ibrohim and Budi, 2019), this dataset is created by crawling 2018 Indonesian Twitter data, using terms related to Indonesian hate speech as queries. Some examples of the terms used are politics-related hate speech terms (*cebong*, *bani kotak*, etc.) as well as disparaging terms against certain marginalized identities (*bencong*, *budek*, etc.) Of note, some of these terms intersect with each other due to the nature of intersectioning harms against marginalized identities (Blodgett et al., 2020; Jiang and Fellbaum, 2020; Guo and Caliskan,

2021). Prime examples are the terms *kristen* and *tionghoa*, representing marginalized religion and racial identities respectively, yet are co-opted to be used by various algorithmic enclaves to insult outgroups.

As shown by Wiegand et al. (2019), datasets obtained by random sampling a large population of sentences tend to be fairer against unwanted social biases when compared to datasets obtained by specifically querying for sentences regarding a certain topic. This has two implications: first, it shows that the terms used to create the Hate Speech dataset have enough variety to create a dataset free from religion bias. Second, since Hate Speech share the same Twitter source as EmoT and SmSA, yet only Hate Speech is unbiased, it may show that the methods used to create EmoT and SmSA datasets are not diverse enough to be free of religion bias. Unfortunately, the papers depicting the creation of these datasets (Purwarianti and Crisdayanti, 2019; Saputri et al., 2018) do not contain sufficient information to make proper conclusions as to why their end-results contain religion bias. This highlights the need to properly document dataset creation processes in order to ease the process of bias discovery (Bender and Friedman, 2018).

Table 4.6: μ_{label} of all labels in Hate Speech dataset

Label	μ_{label}
hate speech	-1.35
abusive	-2.13
none	-0.43

4.2 Embedding Bias

Table 4.7 shows the result of word analogies, as previously mentioned in Table 3.4 that all three chosen word analogies are not able to extract the expected semantic relations between religious identities, as expected in Section 3.2.2. In particular, the first and third analogy fails to capture *kafir* and *onta* being co-opted as religious insults, and the second analogy fails to capture the lack of positive representations of marginalized religions on media. This corroborates the findings of Kurniawan (2019) where Indonesian word embeddings struggle to answer hand-crafted semantic analogies, such as '*australia* is to *dolar* as *indonesia* is to *rupiah*'. This may show that word analogies are not an appropriate measurement for measuring embedding bias in Indonesian word embeddings.

Table 4.7: Word analogy results for all embeddings

Word analogy	Twitter	Wiki	Tempo	ConLL
islam : kristen :: kafir : x	nunguin, ngabisin, nungu, bangunin, meluk	dibaptis, pendeta, berdosa, biarawan, efesus	protestan, cruz, pendeta, robert, dibantai	zelot, saduki, 7:21-23, beelzebul, non-katolik
kristen : islam :: kejadian : x	umat, agama, negara, bangsa, pemimpin	dakwah, ulama, sultan, muhammad, syekh	ideologi, kiai, pesantren, muhamadiyah, terorisme	zariah, syari'at, dauliyah, taghayyur, maaliah
islam : kristen :: onta : x	nunguin, nungu, nungguin	aurelius, stanislaus, efesius	cruz, glen, david, carson, pedro	membaptiskan, zelot, 14:13-21, yakobus, 8:9

Table 4.8 shows that for Tempo and Wiki embedding, whose sources (Tempo news articles and Wikipedia articles respectively) are more descriptive in nature, word similarity does not output religiously biased words, instead outputting synonyms or other semantically similar words (e.g: *radikal* is most similar to other political/social stances, *kafir* is most similar to other Islamic terms related to sins). CoNLL embedding show similar effects, although the quality of words returned is low (many typos, existence of symbols inside words). This is the desired outcome of the word similarity test for embedding bias, and may show that religious bias does not manifest for these embeddings.

A different case happens in the Twitter embedding, whose source are one possible source of the algorithmic enclave effect researched by Lim (2017). In this embedding, in line with Lim (2017), certain religion-related words were found to have high similarity to religiously-biased words. As an example, *kafir* is instead closer to other religions, and *radikal* is specifically close to *khilafah* and *komunis*, which are co-opted as political insults. This shows the impact of co-opting religion-neutral concepts as insults (Lim, 2017) on NLP resources that leverage platforms where these co-opting acts happen as corpuses (Olteanu et al., 2019). The word relationships shown by word similarity in the Twitter embedding case, in which words related to religious identities are related to negativity, confirms our second hypothesis for this thesis.

In addition to confirming our prior hypothesis, there are two other key take-

aways regarding the word similarity results shown in Table 4.8. First, it strengthens the finding of Gonen and Goldberg (2019), in that there are forms of embedding bias that can be discovered using bias-by-neighborhood methods but not with bias-by-projection methods. Second, this shows the potential of using word similarity as a bias measurement for bias-by-neighborhood methods. The neighborhood metric used by Gonen and Goldberg (2019) to measure bias-by-neighborhood utilizes the fact that their research concerns binary gender bias between male-presenting and female-presenting identities, which is the trend on existing English research on unwanted bias in machine learning models (Sambasivan et al., 2021). The framing of binary gender bias for neighborhood metric allows Gonen and Goldberg (2019) to assign a cluster for each gender identity, of which the cluster membership of gender-neutral occupations are used to determine occupational gender bias. However, adapting this approach to religion bias is not straightforward, since it is not as straightforward as gender bias, and requires additional contextual information (Sambasivan et al., 2021). The word similarity measurement is an attempt of such adaptation, which replaces the clusters - representing gender stereotypes - by manual inspections utilizing religion bias contexts in Indonesia to discover embedding bias.

Table 4.8: Per-word similarity of certain religion-related terms for all embeddings

Word analogy	Twitter	Wiki	Tempo	ConLL
islam	umat, kristen, ajaran, agama, radikal	muslim, kristen, islamnya, agama dakwah	muslim, keislaman, sekuler, wahabi, khilafah	isalam, isalm, islamdan, islam, agam
kristen	hindu, protestan, beragama, bnudha, muslim	kristiani, katolik, protestan, kekristenan, protestanisme	protestan, nasrani, katolik, koptik, evangelis	katolik, kristiani, protestan, kriten, katholik
radikal	ormas, khilafah, komunis, pancasila, politikus	moderat, progresif, konservatif, anarkis, sentris	fundamentalis, moderat, islamis, ekstremis, radikalisme	redikal, radaikal, radkal, bebas.radikal, antidegeneratif
haram	hukumnya, haramgirls, meluluskan, projek, halal	masjidil, manyarah, makruh, madinah, sunah	haramnya, terlarang, makruh, ilegal, selundupan	diharamkan, mengharamkan, harom, haramnya, diharamkannya

Table 4.9: Per-word similarity of certain religion-related terms for all embeddings (continued)

Word analogy	Twitter	Wiki	Tempo	ConLL
nista	opla, henti2nya, menyemangati, kedudukannya, menjelaskannya	jaba, mandala, ganapati, ler, kedhaton	terkutuk, keji, hina, kemunafikan, biadab	nishta, kasmala, benar/dipenuhi, lobha, kekotoran
kafir	kristen, pribumi, yahudi, islam, radikal	musyrik, munafik, murtad, zalim, fasik	murtad, musyrik, khawarij, beriman, zalim	musyrik, nashroni, kekafiran, musyrikin, kafir/musyrik
halal	bihalal, bpom, syariah, tiens, melimpah	bihalal, kosher, kehalalan, zakat, diharamkan	kehalalan, lppom, labelisasi, halalnya, mui	halal, halalnya, kehalalannya, muisurat, thayiban
syiah	wahabi, puak, feminist, loyalis, perampokan	sunni, wahabi, muslim, syi, salafi	sunni, salafi, najaf, sadr, wahabi	syi'ah, sunni, wahabi, suni, wahhabi

4.3 Impact 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 evaluation for allocation harm, using parity conditions to measure the existence of allocation harm, then proceed to the representation harm results using sentence templates. In all three evaluations, models are referred by the embeddings used as word representation, adding an *lstm_* prefix beforehand (e.g.: *lstm_twitter* means a Bi-LSTM model trained with Twitter embedding).

Table 4.10 shows the accuracy of all models, over both splits (training and validation). 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, in the EmoT case, the low validation split result on all models may suggest overfitting in the

overall dataset. In the Hate Speech case, the accuracy results are low for both training and validation splits for all models, which suggests the difficulty of multi-label learning presented by the dataset.

Table 4.10: Accuracy results on all datasets for each embedding

Data split	Twitter	Wiki	Tempo	ConLL
Training (EmoT)	92.93	82.36	88.47	82.56
Validation (EmoT)	63.86	62.05	65	67.5
Training (SmSA)	98.26	96.35	96.72	95.52
Validation (SmSA)	90.74	90.78	90.74	90.86
Training (Hate Speech)	68.37	69.73	70.02	69.09
Validation (Hate Speech)	68.91	74.6	64.64	69.55

Since measuring allocation harm using parity metrics rely on knowing the distribution of labels in the dataset given certain groups, Table 4.11 shows the label distribution (in negative/positive simplified labels as explained in Section 3.2) for all sentences that contain Islamic and Christianity religious terms, for all datasets.

Aligning to the previous dataset bias results for EmoT and SmSA datasets, sentences containing religious terms are mostly labeled as negative, whereas this effect is not seen in the Hate Speech dataset. As an example, for the term *islam*, there are 23 sentences with negativity-related labels in the EmoT dataset compared to 9 with non-negativity-related labels. For the SmSA dataset, there are 103 and 14 sentences with and without negativity-related labels respectively. This is not seen in the Hate Speech dataset - there are 360 sentences labeled *hate speech*, 151 labeled *abusive*, and 353 labeled *none*. As previously shown in Section 4.1, this effect also holds true for Christianity terms.

Table 4.11: Label distribution of all datasets, grouped by religious terms

Dataset Name	Source	Label distribution for sentences with Islamic terms	Label distribution for sentences with Christianity terms
EmoT	Saputri et al, 2018	(-) 40, (+) 22	(-) 19, (+) 3
SmSA	Purwarianti and Crisdayanti, 2019	(-) 133, (+) 24	(-) 36, (+) 3
Hate Speech	Ibrohim and Budi, 2019	(-) 482, (+) 568	(-) 45, (+) 313

The imbalanced label distribution found on datasets, particularly for *negative* label, have certain implications in choosing the appropriate parity metrics to measure allocation harms. To use the SmSA dataset as an example, as shown in Table 4.11, this dataset contains 36 negative sentences with Christianity terms, and 3 positive sentences with them. Since there are few positive sentences but many negative sentences to be used as training data, there is a possibility that Bi-LSTM models trained on this dataset fails to properly generalize positive Christianity-related sentences. Combined with the comparatively higher amounts of negative Christianity-related sentences, Bi-LSTM models trained using SmSA dataset are expected to have low false positive rate (FPR) and high demographic parity (DP) scores. As shown in Table 4.11, the imbalance is also seen for both EmoT and SmSA, for both religion groups.

The label imbalances seen in EmoT and SmSA for both religion groups cause the expected FPR and DP results to be consistent over both religious groups and both datasets, yet does not explain much for allocation harm measurements. Therefore, our main focus in this thesis is on the FNR (false negative rate) parity and TPR (true positive rate) parity.

As shown in Section 4.1, religion bias manifests in the form of religious terms mostly existing in negativity-labeled sentences. Combining the manifestation of religion bias in datasets and the definition of parity metrics as seen in Table 2.3, FPR and TPR can be re-contextualized with relation to religion bias. In particular, FNR measures the amount of non-negative sentences mispredicted as negative, which measures the impact of religion bias on downstream performance. On the other hand, TPR measure the amount of non-negative sentences correctly predicted, which measures the capability of Bi-LSTM models to perform despite existing biases in datasets and embeddings used to develop the model.

Despite not being used to measure allocation harms, this thesis still opts to show

FPR and DP results, for pre-debiasing and all post-debiasing results. This is done primarily to measure whether debiasing also impacts other parity metrics. Since debiasing methods aim to re-balance the imbalanced negative representations found in datasets and embeddings, there may be instances where other parity metrics experience changes. As an example, dataset debiasing adds non-negative-labeled sentences from external sources to existing datasets. These may cause models that learn from the augmented dataset to experience increased FPR scores due to the increased amount of positive-labeled sentences to learn from.

Table4.12 shows the parity metric results for the EmoT dataset pre-debiasing. Here, we note that in the EmoT dataset, FNR tends to be significantly higher and TPR significantly lower for Christian terms for 3 out of 4 models, compared to Islamic terms. As an example, for the Twitter model, TPR scores for sentences containing Islamic and Christianity religious terms are 100 and 66.67% respectively. The FNR scores are its complements - 0 and 33.33% respectively. This shows a performance gap, where the Twitter model performs worse for sentences with Christianity religious terms. These effects can be seen in lstm_twitter, lstm_wiki, and lstm_conll models. Therefore, we conclude that these three models inflict allocation harms against Christianity.

For this case, lstm_tempo exists as an outlier, where the performance gap between Islam and Christianity sentences does not exist. This is seen in the 95.65% and 100% TPR score for Islamic and Christianity sentences respectively. Since all of the Bi-LSTM models shown in the Table 4.12 learns from the same EmoT dataset, yet only the Twitter word embedding was found to contain embedding bias, this finding shows that different embeddings can react differently to the same dataset, and as such can inflict different types of allocation harm. As an example, neither Tempo nor Wiki word embeddings were found to contain embedding bias as analyzed in Table 5.21, yet only lstm_tempo inflict allocation harms against Christianity. This implies that the differences came from the word relationships represented by the embeddings themselves. As such, this shows the importance of analyzing downstream performance results caused by the interaction of multiple NLP resources, as opposed to only consider biases in the resources separately (Blodgett et al., 2020).

Another point of interest comes from the 19 negative sentences containing Christianity terms in the EmoT dataset. Out of these sentences, 8 of them are almost exact copies of the sentence *Tidak sepakat, dan anda sudah menyesatkan orang lain. Dilarang sholat di dalam bangunan yang didirikan untuk kekafiran (salah satunya gereja). Makanya jangan kebanyakan nonton mak lampir. Aku berlindung*

kepada ALLAAH SWT. These 8 sentences have minor differences from each other that allow them to escape initial duplication checks after cleansing. As an example, two such copies end with *dari godaan setan yang terkutuk*, but one of them omits the *SWT* part. These type of posts are tactics commonly done by certain algorithmic enclaves known as 'buzzers', which coordinate spam posts against other out-groups (Lim, 2017). When unchecked, these near-duplicates can worsen model training purposes, which may explain the accuracy drop in validation scores as shown in Table 4.10 if more of these duplicates exist in the dataset. This thesis opts to keep these sentences in the dataset, as well as other types of sentences (if it exists) instead of attempting to remove them, in order to showcase as much of the original religion bias manifestations as possible. This also highlights the importance of documenting the data collection and creation process, in order to better assess the possibilities of biases that may occur on models that learn from this dataset (Bender and Friedman, 2018).

Table 4.12: Parity metric resuls of EmoT dataset, in percentage

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	6.15	0	100	58.46
lstm_twitter - Christianity	0	33.33	66.67	90.91
lstm_wiki - Islam	0	4.35	95.65	66.15
lstm_wiki - Christianity	4.55	66.67	33.33	90.91
lstm_tempo - Islam	1.54	4.35	95.65	64.62
lstm_tempo - Christianity	0	0	100	86.36
lstm_conll - Islam	0	4.35	95.65	66.15
lstm_conll - Christianity	4.55	66.67	33.33	90.91

Table4.13 shows the parity metric results for the SmSA dataset pre-debiasing. In this dataset, models tend to have higher FNR and lower TPR for Islam sentences, showing an initial case of allocation harm against Islam. An example of this is shown in the lstm_tempo case, where it has a 83.33% TPR for Islamic sentences

and 100% TPR for Christianity sentences. Note that even if the both EmoT and SmSA dataset contain dataset bias against both religious groups, the religion group harmed by models that learn from these sentences differ. This is after considering the fact that all of the models that show this case of allocation harm (i.e., all but lstm_twitter) were all trained on unbiased embeddings. This strengthens the prior finding that different embeddings and datasets react differently when combined together, resulting in different types of allocation harms. A short manual inspection is done to the SmSA dataset to show that unlike the EmoT case, no near-duplicate negativity posts were found. This shows that different manifestations of the same dataset bias exists, which can possibly inflict different types of allocation harm.

Another interesting observation from this table can be seen in the lstm_twitter case. For this case, the embedding used to train the Bi-LSTM model was originally found to be biased, yet the model does not inflict allocation harm. The effect can be seen by the 100% TPR and the subsequent 0% FNR score on both Islamic and Christianity sentences. This further highlights the variability of interaction between the biases in datasets and embeddings (if any), as well as the resulting allocation harms caused. In particular, in some cases, the biases that exist in the embedding may interact with the biases in the dataset in such a way that the resulting model does not inflict allocational harms. This highlights the importance of analyzing downstream performance results of an NLP model, on top of analyzing biases in separate components that were used to train such model (Blodgett et al., 2020).

Table 4.13: Parity metric resuls of SmSA dataset, in percentage

Model - Term	FPR	FNR	TPR	DPR
lstm_twitter - Islam	0.75	0	100	84.08
lstm_twitter - Christianity	0	0	100	92.31
lstm_wiki - Islam	3.01	4.17	95.83	82.80
lstm_wiki - Christianity	2.78	0	100	89.74
lstm_tempo - Islam	0.75	16.67	83.33	86.62
lstm_tempo - Christianity	0	0	100	92.31
lstm_conll - Islam	0	4.17	95.83	85.35
lstm_conll - Christianity	0	0	100	92.31

Table 4.14 shows the parity metric results for the Hate Speech dataset pre-debiasing. We note that in Hate Speech dataset, which originally does not constitute dataset bias, does not have significant FNR/TPR difference between Islamic and Christianity terms. However, since there is a significant imbalance towards positive labels in sentences containing Christianity terms in the dataset, models tend to be more overconfident on predicting sentences containing Christianity terms, causing higher FPR rates than sentences containing Islamic terms. For example, in *lstm_wiki*, the FPR rate for Christianity terms is 11%, about five times larger than the 2.27% FPR rate for Islamic terms. However, this is not the main concern of our thesis, since our bias is defined more by association towards negative labels.

Table 4.14: Parity metric results of Hate Speech dataset, in percentage

Model - Term	FPR	FNR	TPR	DPR
lstm_twitter - Islam	3.1	5.24	94.75	47.25
lstm_twitter - Christianity	4.44	2.24	97.77	13.97
lstm_wiki - Islam	2.27	6.64	93.36	48.39
lstm_wiki - Christianity	11.11	2.56	97.44	13.41
lstm_tempo - Islam	3.31	4.55	95.45	46.78
lstm_tempo - Christianity	6.67	0.96	99.04	12.57
lstm_conll - Islam	5.79	5.59	94.41	46.21
lstm_conll - Christianity	13.33	1.6	98.4	12.29

Table 4.15 shows the sentence template results for EmoT dataset pre-debiasing. Here, it shows how representation harms occur in the form of misprediction of sentence templates. As an example, the template *saya anut agama islam*, originally labeled neutral, was mispredicted as negativity-related label *anger*. Similar to this template is the template *sekolah saya ajar agama islam*, mispredicted as negativity-related label *sadness* or *anger* in certain cases. This replicates the finding of religion bias as shown in Figure 1.1.

Unlike allocation harm, representation harm generally harms both religion groups equally. This can be seen in the result of the *saya anut agama islam* and *saya anut agama kristen* sentences, which came from the same sentence template and filled with different religious identities, both being mispredicted as *anger*. The religion groups being inflicted representation harms differ from the allocation harm case, where it was shown in Table 4.12 that only inflicts Christianity. This shows that both allocation and representation harms can differ on whose group it harms.

The result of representation harm shown in Table 4.15 contain several outliers. First, the template *saya cinta agama islam* and *saya cinta agama kristen* are both correctly predicted into a positive label. For this instance, the word *cinta* is often used in sentences declaring love, and as such this implies that the word

cinta massively influences the prediction result of this particular template. Second, there are cases where the sentence *masjid jadi tempat aman bagi seluruh masyarakat indonesia*, as well as the Christianity counterpart *gereja jadi tempat aman bagi seluruh masyarakat indonesia* are correctly predicted into a positive label, and cases where both sentences are mispredicted into a negative label. As shown, lstm_twitter, lstm_wiki, and lstm_conll assign positive labels to these sentences, whereas lstm_tempo assign negative labels to the sentences. Since the predictions are consistent over sentences belonging to both religion groups, this may show that representation harm are mostly consistent over both religion groups, with the variations happening at model level.

Table 4.15: Sentence template results of EmoT dataset

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	anger, 77.89	anger, 70.66	anger, 61.86	anger, 84.26
saya anut agama kristen	anger, 82.98	anger, 80.85	anger, 72.8	anger, 89.74
saya cinta agama islam	love, 99.06	love, 98.54	love, 99.14	love, 96.96
saya cinta agama kristen	love, 98.99	love, 98.29	love, 98.89	love, 96.04
sekolah saya ajar agama islam	happy, 88.17	sadness, 78.37	happy, 72.69	anger, 67.
sekolah saya ajar agama kristen	anger, 83.86	sadness, 77.67	anger, 68.87	anger, 77.07
tenggang rasa antar kaum islam harus jaga	sadness, 86.7	sadness, 70.29	anger, 85.74	anger, 81.59
tenggang rasa antar kaum kristen harus jaga	sadness, 80.68	anger, 69.93	anger, 88.89	anger, 87.64
masjid jadi tempat aman bagi seluruh masyarakat indonesia	happy, 95.46	sadness, 79.24	happy, 99.23	happy, 93.75
gereja jadi tempat aman bagi seluruh masyarakat indonesia	happy, 89.59	sadness, 81.74	happy, 94.12	happy, 87.18
saya tidak setuju dengan ajaran agama islam	anger, 88.17	sadness, 69.14	anger, 80.94	anger, 85.62
saya tidak setuju dengan ajaran agama kristen	anger, 91.03	anger, 72.08	anger, 85.41	anger, 90.44

Table 4.16 shows the sentence template results for SmSA dataset pre-debiasing, which show that Bi-LSTM models trained using SmSA dataset inflict representation harm. The mispredictions given by models trained on the SmSA dataset pre-debiasing are more prominent in this case. As an example, the sentence *saya anut agama islam*, originally a *neutral*-labeled sentence, was mispredicted into a *negative* sentence with around 99% confidence. Other sentences also encounter the same

misprediction with high probability, as shown with the template *tenggang rasa antar kaum islam harus jaga* being assigned a near-perfect 99% misprediction. This is in contrast to mispredictions happening in Table 4.16, where the predictions probability mostly hover between 60-80%. Additionally, there are sentences which were able to be correctly predicted in the EmoT case, yet are mispredicted in the SmSA case. A prime example of this is the sentence *saya cinta agama kristen*, which was correctly predicted into a positive label *love* for all models, yet are consistently mispredicted as *negative* in the SmSA case, with high prediction probability.

The more severe representation harm happening in the SmSA, when combined with the better accuracy scores shown in Table 4.10 for SmSA compared to EmoT, as well as the higher misrepresentation of religion groups in the SmSA dataset as shown in Table 4.4 when compared to the EmoT case, accentuates the interaction between aspects of model training on introducing biases and inflict harms on religion groups. While the corpuses used to create this dataset are more varied due to the inclusion of multiple platforms, the observations done in Table 4.4 show the possibility of most sentences mentioning religion terms to be obtained from the platforms where the limited representations of certain religious identities (Lim, 2017; Remotivi, 2021) take place. This manipulates the population of sentences mentioning religious terms in the platform (Olteanu et al., 2019), and introduces the concept of religion bias, which manifests in the form of religious identities being mostly represented in negativity-related sentences, when compiled into the SmSA dataset (Dixon et al., 2018; Wiegand et al., 2019). These sentences do not contain near-duplicates like the EmoT dataset as discussed in Table 4.12, and as such allows Bi-LSTM models that learn from SmSA dataset to have better performance, as shown in Table 4.10. However, their high accuracy performance, combined with the consistent misrepresentation of religious identities in the SmSA dataset, result in models that learn from this dataset to inflict representation harm with high probability.

Table 4.16: Sentence template results of SmSA dataset

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	negative, 99.99	negative, 99.66	negative, 99.99	negative, 99.66
saya anut agama kristen	negative, 99.99	negative, 99.86	negative, 100	negative, 99.97
saya cinta agama islam	positive, 82.33	positive, 61.56	negative, 92.3	negative, 95.49
saya cinta agama kristen	negative, 94.69	negative, 82.43	negative, 97.32	negative, 99.67
sekolah saya ajar agama islam	negative, 99.9	negative, 99.98	negative, 99.99	negative, 99.88
sekolah saya ajar agama kristen	negative, 99.99	negative, 99.99	negative, 99.99	negative, 99.98
tenggang rasa antar kaum islam harus jaga	negative, 99.36	negative, 96.833	negative, 99.04	negative, 94.1
tenggang rasa antar kaum kristen harus jaga	negative, 99.96	negative, 99.18	negative, 99.58	negative, 98.51
masjid jadi tempat aman bagi seluruh masyarakat indonesia	positive, 69.55	positive, 95.14	positive, 94.96	positive, 54.88
gereja jadi tempat aman bagi seluruh masyarakat indonesia	positive, 71.38	positive, 93.04	positive, 72.56	positive, 43.93
saya tidak setuju dengan ajaran agama islam	positive, 95.48	negative, 96.93	negative, 98.42	positive, 51.85
saya tidak setuju dengan ajaran agama kristen	positive, 92.24	negative, 94.2	negative, 98.51	negative, 73.81

Table 4.17 shows the sentence template results for Hate Speech dataset pre-debiasing, for the *hate speech* label. Since this dataset does not contain dataset bias, it follows that no representation harm occurs in most cases, much like the allocation harm case. However, since there are considerable amount of sentences labeled *hate speech* that contain religious terms, there are specific outliers that are to be discussed for this case.

The most prominent outliers in this table is the sentence *sekolah saya ajar agama islam*, of which 3 out of 4 embeddings output a 20-30% hate speech probability. Considering the fact that this does not happen on the Christianity counterpart *sekolah saya ajar agama kristen*, yet the effect is constant on most embeddings, this may imply that there are specific sentences contained in the dataset which cause the sentence *sekolah saya ajar agama islam* to be mispredicted. As such, this is a dataset outlier happening outside of religion bias effect, and analyzing them is outside the scope of this thesis. However, this does highlight the importance of checking for specific edge cases of downstream performance, even if the overall result does not inflict harm - whether allocation or representational.

The second outlier comes from the lstm_conll model, which often outputs higher probabilities when compared to other models. As such, this particular model can be considered to inflict more representational harm compared to other Bi-LSTM models used in this thesis. This outlier is clearly seen in the prediction result of *saya anut agama islam*, where the model assigns a 29% hate speech probability, in contrast of the 3% predictions given by other models. Other examples are the prediction probability for *saya anut agama kristen*, *saya cinta agama islam*, and *saya cinta agama kristen* to a lesser degree. This further strengthens the variability of datasets and embeddings, with regards to the biases and harms inflicted by the resulting model.

Table 4.17: Sentence template results of Hate Speech dataset (hate speech label)

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0.8	2.44	3.81	29.65
saya anut agama kristen	0.21	1.	0.16	13.48
saya cinta agama islam	3.03	0.83	0.29	16.71
saya cinta agama kristen	1.01	0.45	0.02	6.56
sekolah saya ajar agama islam	27.55	23.77	22.59	9.5
sekolah saya ajar agama kristen	9.10	7.7	1.25	4.22
tenggang rasa antar kaum islam harus jaga	8.64	3.9	4.5	40.75
tenggang rasa antar kaum kristen harus jaga	5.9	1.36	0.19	20.87
masjid jadi tempat aman bagi seluruh masyarakat indonesia	1.29	12.91	9.12	0.5
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.24	12.25	2.89	0.95
saya tidak setuju dengan ajaran agama islam	35.2	13.74	2.04	42.18
saya tidak setuju dengan ajaran agama kristen	11.76	3.65	0.15	20.97

Table 4.18 shows the sentence template results for Hate Speech dataset pre-debiasing, for the *abusive* label, Much like the *hate speech* label, there are no cases of mispredictions happening since the original dataset does not contain bias. Since the distribution of *abusive* sentences containing religious terms are even less than *hate speech*, as seen in Tables 4.5 and 4.6, there are no prominent misprediction outliers in this case.

Table 4.18: Sentence template results of Hate Speech dataset (abusive label)

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0.01	0.02	0.06	1.2
saya anut agama kristen	0	0	0	0.46
saya cinta agama islam	0.03	0.03	0.01	0.56
saya cinta agama kristen	0.01	0	0	0.25
sekolah saya ajar agama islam	0.04	0.05	0.16	0.25
sekolah saya ajar agama kristen	0.01	0	0	0.14
tenggang rasa antar kaum islam harus jaga	0.22	0.06	0.39	9.75
tenggang rasa antar kaum kristen harus jaga	0.07	0	0	5.11
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0.29	0.07	0.04	0
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.1	0.02	0.01	0.01
saya tidak setuju dengan ajaran agama islam	0.04	0.02	0	0.36
saya tidak setuju dengan ajaran agama kristen	0.02	0	0	0.25

4.4 Summary

The findings of both allocation and representation harms in Section 4.3 confirms the impact of dataset and embedding biases on downstream performance, as inquired by the second research question and shown on Figure 3.1 . On the case of allocation harms, represented by the FNR and TPR of different models, the imbalanced nature

of non-negative sentences mentioning certain religious terms on Indonesian NLP datasets causes models to have considerable FNR and TPR differences across Islamic and Christianity mentions. For representation harms, non-negative sentences with mentions of certain religious terms are mis-predicted into negativity-related labels.

Much like the findings of Section 4.1, the allocation and representation harms also impact non-marginalized religion identities (i.e. Islam). In the case of allocation harms, the findings in Table 4.13 shows that allocation harm impacts sentences mentioning Islamic religious identities more than Christianity identities. For representation harms, results in Table 4.15 and Table 4.16 shows that religious identities related to both Islam and Christianity are impacted, as opposed to only marginalized religious identities.

CHAPTER 5
DEBIASING RESULTS

This chapter describes the results of our debiasing methods, where each subchapter describes the result of each method. For methods that use dataset augmentation (i.e dataset debiasing and joint debiasing), the results only show findings from SmSA and Hate Speech dataset, since EmoT is unable to be debiased using our proposed debiasing method. This is because of its dataset creation method (Saputri et al., 2018), where sentences without any explicit emotions are removed prior to creating EmoT dataset.

5.1 Dataset Debiasing with Templates

Table 5.1 shows the impact of dataset debiasing using sentence templates on the SmSA dataset on the term-label occurrence. Since the added sentences on the SmSA dataset for debiasing are non-negative in nature, only the non-negative terms increase quantity. An example is the term *islam*, which in the original dataset mostly exists in sentences labeled as *negative* (103 sentences). Debiasing by sentence templates add multiple *positive* and *neutral* sentences containing *islam*, such that it is now represented by 145 *positive* sentences and 99 *neutral* sentences.

Table 5.1: Religious term occurrence per label in SmSA dataset, after dataset debiasing by templates

Terms	positive	negative	neutral
islam	145	103	99
kristen	141	28	90
masjid	45	6	53
gereja	42	6	51
quran	0	7	20
alkitab	0	1	10
ulama	21	27	12
pendeta	20	4	10

As shown by the PMI scores after dataset debiasing by sentence templates in Table 5.2, the added sentences are able to change the label distribution of sentences containing religious term, so that the majority of them are non-negative. In detail,

the PMI averages for the only negativity-related label *negative* is smaller than the PMI averages for both positivity-related label *positive* and *neutral*. This is a change from the pre-debiasing PMI score seen in Table 4.4, in which the PMI averages for *negative* is higher than both positive labels.

It follows from the dataset bias definition in Table 3.3 that the addition of sentences from sentence templates successfully debiases the SmSA dataset, since the existence of unwanted religious bias in datasets are represented by sentences containing religious term being mostly negativity-related sentiments or emotions, due to the effect of algorithmic enclave (Lim, 2017) and minimal marginal representation (Remotivi, 2021).

Table 5.2: μ_{label} of all labels in SmSA dataset, after dataset debiasing by templates

Label	μ_{label}
positive	-0.86
negative	-1.85
neutral	-0.9

Table 5.3 shows the impact of dataset debiasing using sentence templates on the Hate Speech dataset on the term-label occurrence. Since all sentences obtained from sentence templates are labeled *none* for the purpose of debiasing Hate Speech, only the occurences of the *none* label increases. As an example, for the term *islam*, the amount of *hate speech* and *abusive* (360 and 151 sentences respectively) yet the *none*-labeled sentences increase to 583.

Table 5.3: Religious term occurrence per label in Hate Speech dataset, after dataset debiasing by templates

Terms	hate speech	abusive	none
islam	360	151	583
kristen	42	14	506
masjid	13	6	113
gereja	3	1	146
quran	46	31	56
alkitab	1	0	13
ulama	97	43	237
pendeta	3	1	35

Since only *none*-labeled sentences are added to the Hate Speech dataset for debiasing purposes, the overall PMI averages in this dataset do not change. This

is seen in Table 5.4, where the PMI averages of both negative labels *hate speech* and *abusive* are considerably lower than the positive label *none*, much like the pre-debiasing case shown in Table 4.6. This shows that on top of being able to mitigate dataset bias in biased datasets, debiasing dataset using sentence templates do not introduce additional bias in datasets that were originally unbiased.

Table 5.4: μ_{label} of all labels in Hate Speech dataset, after dataset debiasing by templates

Label	μ_{label}
hate speech	-2.12
abusive	-2.93
none	-0.23

As shown on Table 5.5, for the SmSa dataset case, dataset debiasing with sentence templates manage to maintain accuracy scores for both training and validation splits, with approximately equal values. An example of this is seen in the lstm_twitter case, which achieved 98.26% training accuracy and 90.74% validation accuracy. After dataset debiasing using sentence templates, the training and validation accuracy metrics changes into 98.44% and 91.43% respectively. Other models experience the same effect. Since dataset debiasing by sentence templates augments 10 copies of 57 unique sentences into datasets, there is a possibility of the same sentence templates being included in both the training and validation sets for a given cross validation fold. Therefore, while dataset debiasing by sentence templates has shown to maintain or increase overall accuracy scores, it is unclear on whether the increases are caused by the model being able to generalize better regarding religious terms, or simply because of the model memorizing sentence templates.

The same cannot be said for the Hate Speech dataset, where the impact of dataset debiasing with templates are more varied. The most extreme case is on the lstm_twitter model, where the accuracy post-debiasing decreases by 3-4%. In particular, before debiasing, the training and validation accuracy scores are 68.37% and 68.91%. After debiasing, the accuracy scores decrease into 65.41% and 64.54% respectively. Interestingly, for the other 3 models, the accuracy score for the validation split increases, which imply better performance for these models. This may imply that for the specific lstm_twitter case, the accuracy decrease happen as a reaction to the embedding bias that exists in the Twitter embedding.

Table 5.5: Accuracy results on all datasets for each embedding, after dataset debiasing by sentence templates

Data split	Twitter	Wiki	Tempo	ConLL
Training (SmSA)	98.44	96.11	97.02	95.9
Validation (SmSA)	91.43	90.88	90.57	89.67
Training (Hate Speech)	65.41	70.6	65.57	68.74
Validation (Hate Speech)	64.54	76.03	66.01	71.12

Table 5.6 parity metric results after dataset debiasing by sentence templates, for the SmsA dataset case. In this table, it shows that for Islamic related sentences, the FNR metrics for all models hovers around 0-1%, thereby mitigating allocation harm. As an example, lstm_tempo obtains 16.67% and 0% FNR for Islamic and Christianity-related sentences, showing allocation harm against Islam. After dataset debiasing, the FNR scores change into 0.28% and 0% respectively, mitigating the allocation harm previously shown by the model. However, for all models, all FNR scores for Christianity-related terms are maintained at 0%, which is the same FNR score pre-debiasing. Pre-debiasing, there were only 3 positive Christianity sentences in the SmSA dataset, and as such the 0% FNR obtained can be attributed to chance. However, debiasing by sentence templates adds 360 positive Christianity sentences into the dataset, yet the FNR score is still 0% after debiasing. This implies that the FNR metric performance obtained after debiasing by sentence templates may be attributed to the model remembering sentences that were being added as part of the debiasing process, instead of the model actually learning to generalize better by the added sentences after debiasing.

An interesting impact of dataset debiasing in this dataset is in the massive reduction of DP over all models. However, this effect is caused by the nature of dataset augmentation itself, instead of the method being able to properly mitigate allocation harms. The reasoning is as follows: note that the TPR of all embeddings maintains or increases post-debiasing. Since the dataset debiasing method adds non-negative sentences (which are labelled positive for allocation harm case), this means that most augmented sentences are correctly labeled as positive. This mas-

sively increases the true positive count, which in turn increases the denominator of the DP formula, therefore lowering the DP metric for all models and terms.

Table 5.6: Parity metric results of SmSA dataset, in percentage, after dataset debiasing by templates

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	0	0.85	99.15	27.93
lstm_twitter - Christianity	0	0	100	10.03
lstm_wiki - Islam	1.50	0.56	99.44	27.31
lstm_wiki - Christianity	2.78	0	100	9.75
lstm_tempo - Islam	3.01	0.28	99.72	26.69
lstm_tempo - Christianity	5.56	0	100	9.47
lstm_conll - Islam	2.26	0	100	26.69
lstm_conll - Christianity	5.56	0	100	9.47

As shown in Table 5.7, there are little changes in FNR, TPR, and DP after dataset debiasing by templates. This shows that debiasing by sentence templates does not introduce additional forms of allocation harm to a dataset that does not originally contain unwanted bias. However, the effect of models having high FPR for sentences containing Christianity terms previously seen before debiasing, as seen in Table 4.14 worsens, as shown by the increase in FPR. A primary example is seen in the lstm_twitter example, where for Christianity terms, the FPR score is 4% pre-debiasing and 15% after debiasing by sentence templates. This shows that debiasing datasets that did not contain bias can possibly worsen model performance. As such, it is important to check whether a dataset contains bias or not before deciding on whether to debias said dataset.

Table 5.7: Parity metric results of Hate Speech dataset, in percentage, after dataset debiasing by templates

Model - Term	FPR	FNR	TPR	DPR
lstm_twitter - Islam	2.27	3.47	96.53	45.36
lstm_twitter - Christianity	15.56	0.58	99.42	10.26
lstm_wiki - Islam	2.48	6.78	93.22	47.11
lstm_wiki - Christianity	11.11	1.74	98.26	11.79
lstm_tempo - Islam	1.65	6.61	93.39	47.38
lstm_tempo - Christianity	11.11	1.16	98.84	11.28
lstm_conll - Islam	5.17	2.48	97.52	43.53
lstm_conll - Christianity	24.44	0.58	99.42	9.23

The result of representation harm using sentence templates is shown in Table 5.8 for the SmSA case. As shown in the table, dataset debiasing by sentence templates manage to correct most mispredicted sentence templates, both for templates previously seen in the augmentation process and new templates. However, there may be hints of overfitting from this method, as seen in the results for templates used for augmentation (*saya anut agama [agama]* and *saya cinta agama [agama]*). For these templates, the prediction results are very close to its original label (*neutral*, with near-perfect probability). As an example, for the model lstm_twitter, the sentence *saya anut agama islam* obtains a (negative, 99.99) prediction before debiasing. After debiasing, this sentence is predicted as *neutral*, with 99% probability. Additionally, the template *saya tidak setuju dengan ajaran agama [agama]* is consistently mispredicted into non-negative labels, often with high probability. This strengthens the findings of allocation harms as shown in Table 5.6, in that the mitigation results happens because the model remembers augmented sentences instead of learning to generalize better.

A notable exception for the results of dataset debiasing by templates can be seen in the *lstm_twitter case*. For this model, the template *tenggang rasa antar kaum*

[agama] harus dijaga, unseen in the augmentation process, fails to be mitigated post-debiasing, instead staying at a *negative* label prediction. Since the Twitter embedding was originally found to contain embedding bias, this misprediction likely happens because of the existing embedding bias. This also supports that the mitigation effect done by dataset debiasing using sentence templates is mostly impacted by overfitting, instead of the model actually learning to generalize better.

Additionally, there are cases where dataset debiasing by sentence templates changes the non-negative predicted label of a sentence to another non-negative label. As an example, consider the sentence *saya cinta agama islam*, which was originally labeled as *neutral* in the dataset augmentation process but predicted as *positive* pre-debiasing. In the *lstm_twitter* and *lstm_conll* case, dataset debiasing by sentence templates changes the label of said sentence into *neutral*. While this does maintain the prediction of said sentence to a non-negative label, the label change is influenced by overfitting in the dataset. This supports the previous point where overfitting may influence the strong performance of dataset debiasing by sentence templates on mitigating representational harms.

Table 5.8: Sentence template results of SmSA dataset, after dataset debiasing by sentence templates

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	neutral, 99.99	neutral, 99.99	neutral, 99.99	neutral, 99.99
saya anut agama kristen	neutral, 99.99	neutral, 99.95	neutral, 99.98	neutral, 99.99
saya cinta agama islam	neutral, 99.98	positive, 68.23	positive, 99.91	neutral, 81.36
saya cinta agama kristen	neutral, 99.98	positive, 86.67	positive, 99.98	neutral, 84.5
sekolah saya ajar agama islam	neutral, 100	neutral, 99.99	neutral, 99.99	neutral, 99.99
sekolah saya ajar agama kristen	neutral, 100	neutral, 99.99	neutral, 99.99	neutral, 100
tenggang rasa antar kaum islam harus jaga	negative, 91.37	positive, 99.09	positive, 70.98	positive, 80.84
tenggang rasa antar kaum kristen harus jaga	negative, 94.24	positive, 99.19	positive, 92.48	positive, 87.84
masjid jadi tempat aman bagi seluruh masyarakat indonesia	positive, 98.5	positive, 99.14	positive, 99.97	positive, 90.46
gereja jadi tempat aman bagi seluruh masyarakat indonesia	positive, 99.46	positive, 99.46	positive, 99.98	positive, 98.5
saya tidak setuju dengan ajaran agama islam	neutral, 99.16	positive, 48.61	positive, 99.95	neutral, 87.06
saya tidak setuju dengan ajaran agama kristen	neutral, 97.87	positive, 70.73	positive, 99.99	neutral, 87.63

Table 5.9 shows the impact of dataset debiasing using sentence templates on models that learn from the Hate Speech dataset, for the *hate speech* label. Here, it shows that the debiasing method manages to maintain correct predictions, and therefore not introduce additional representation harm to the models. As an example, for the sentence *saya anut agama islam*, which is non-hate speech sentence in nature, the lstm_twitter trained before debiasing outputs a 0.8% *hate speech* label

probability. This is maintained after dataset debiasing using sentence templates, with 0.006% probability. However, there are edge cases where dataset debiasing by sentence templates increase *hate speech* probability by 10%. The most notable increase is the sentence *saya cinta agama islam* for the lstm_wiki case, where it was originally predicted as 0.83% hate speech before debiasing and 9.5% hate speech probability after debiasing by sentence templates. While the increases are minor, they still nevertheless show the possibility of debiasing introducing other forms of bias on datasets that do not originally contain them.

When compared to the pre-debiasing sentence templates result at Table 4.17, dataset debiasing by sentence templates manage solve prior issues seen in this dataset for representation harm cases. First, the lstm_conll model, trained on CoNLL word embedding, was found to prescribe relatively higher *hate speech* probabilities when compared to other models. The most notable case is the sentence *tenggang rasa antar kaum islam harus jaga*, which was given a 40% *hate speech* probability before debiasing. Dataset debiasing using sentence templates managed to mitigate this by reducing the probability into 26.9%.

Additionally, for sentences originally mispredicted as having high probability of hate speech (*saya tidak setuju dengan ajaran agama islam*, dataset debiasing using sentence templates manages to correctly reduce the hate speech probability for this sentence, improving the overall model performance. A prominent example is seen in the lstm_twitter case, where this sentence was given a 35.2% *hate speech* probability. After dataset debiasing using sentence templates, this probability is reduced down to 4.89%, showing mitigation effects.

Table 5.9: Sentence template results of Hate Speech dataset (hate speech label), after dataset debiasing by sentence templates

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0	2.79	7.17	1.9
saya anut agama kristen	0	0.26	0.52	0.15
saya cinta agama islam	2.45	9.5	2.72	3.48
saya cinta agama kristen	0.17	1.18	0.19	0.23
sekolah saya ajar agama islam	0.58	1.56	4.93	7.04
sekolah saya ajar agama kristen	0.01	0.07	0.31	0.52
tenggang rasa antar kaum islam harus jaga	0.51	5.36	3.11	26.90
tenggang rasa antar kaum kristen harus jaga	0.03	0.48	0.24	5.16
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0.49	3.31	2	0.02
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.15	1.61	1.43	0
saya tidak setuju dengan ajaran agama islam	4.9	7.50	5.17	14.41
saya tidak setuju dengan ajaran agama kristen	0.12	0.62	0.51	2.02

Table 5.10 shows the impact of dataset debiasing using sentence templates on models that learn from the Hate Speech dataset, for the *abusive* label. Much like the *hate speech* case, dataset debiasing using sentence templates does not introduce additional representation harms for this label. This is shown by the predictions maintaining the low prediction scores previously given before debiasing. As an ex-

ample, the sentence *saya anut agama islam* is correctly predicted as 0-1% *abusive*, both before dataset debiasing by sentence templates and after.

Table 5.10: Sentence template results of Hate Speech dataset (abusive label), after dataset debiasing by sentence templates

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0	0.02	0.16	0.09
saya anut agama kristen	0	0	0.03	0.01
saya cinta agama islam	0.06	0.1	0.57	0.2
saya cinta agama kristen	0.01	0	0.05	0.03
sekolah saya ajar agama islam	0.03	0.01	0.05	0.15
sekolah saya ajar agama kristen	0	0	0	0
tenggang rasa antar kaum islam harus jaga	0.02	0.03	0.85	2.04
tenggang rasa antar kaum kristen harus jaga	0	0	0.04	0.38
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0.07	0.05	0.04	0
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.04	0.02	0.02	0
saya tidak setuju dengan ajaran agama islam	0.16	0.14	0.07	0.28
saya tidak setuju dengan ajaran agama kristen	0.01	0	0	0.09

The results of Section 5.1 shows the impact of dataset debiasing using sentence templates on dataset bias, as inquired by the third research question of this thesis, as well as the impact of dataset debiasing using sentence templates on allocation

and representation harms as inquired by the fifth research question. As shown in Table 5.2 and tab 5.4, dataset debiasing using sentence templates manages to mitigate biases in datasets without introducing additional biases on originally-unbiased datasets.

The results of dataset debiasing using sentence templates on downstream performance are less straightforward, which showcases the weaknesses of this particular debiasing method. As shown in Table 5.6, dataset debiasing using sentence templates was successful of mitigating allocation harms inflicted on Islamic-related sentences. However, as explained on the table, the 100% FNR prediction on all 363 positive Christianity sentences, casts doubt on the true effectiveness of dataset debiasing using sentence templates. As argued before, there is a possibility that the Bi-LSTM models merely memorizes all sentences added to debias the SmSA dataset, instead of utilizing the added sentences to decouple religious identities from negativity, which is the main goal of debiasing.

The result of dataset debiasing using sentence templates show that the method is able to mitigate representation harm by correcting the mispredictions done on most sentences, regardless of whether they were also used to debias the dataset. Therefore, the impact of duplicated sentences does not impact representation harm mitigation as much as the previous allocation harm case. However, the sentences *tenggang rasa antar kaum islam harus jaga* and *tenggang rasa antar kaum kristen harus jaga* fails to be mitigated in the lstm_twitter case. Since this model uses the Twitter embedding, which was found earlier to contain biases, this shows the possibility that the limited variability of the sentences introduced by dataset debiasing using sentence templates can not manage the embedding bias. As such, this builds the case of joint debiasing, which is done in later sections. Additionally, there are cases where debiasing on the Hate Speech dataset causes an increase in misprediction rate, which may show possibility of debiasing introducing other forms of bias on datasets that do not originally contain them. As such, ideally, debiasing datasets should be avoided if the original dataset does not contain bias.

5.2 Dataset Debiasing using Wikipedia

Table 5.11 shows the impact of dataset debiasing using Wikipedia sentences on the SmSA dataset for term-label occurrences. Much like the sentence template case seen in Table 5.1, dataset debiasing using Wikipedia sentences manages to debias SmSA dataset by changing the label distribution for sentences with religious terms.

Table 5.11: Religious term occurrence per label in SmSA dataset, after dataset debiasing by Wikipedia

Terms	positive	negative	neutral
islam	5	103	89
kristen	1	28	191
masjid	5	6	121
gereja	2	6	140
quran	0	7	98
alkitab	0	1	62
ulama	1	27	18
pendeta	0	4	0

Table 5.12 shows the impact of dataset debiasing using Wikipedia sentences on the SmSA dataset for PMI scores. As shown by the table, the negative label *negative* does not have higher PMI average than the label *neutral*. As such, this debiases the dataset, aligning to the dataset bias definition shown in Table 3.3. Since all of the augmented sentences are labelled as *neutral*, without any *positive* sentences, the PMI average for *negative* is still higher than *positive*. However, since the dataset bias definition requires a negativity-related label to out-represent all other positivity-related labels in the dataset, this finding is not enough to determine the existence of dataset bias, after debiasing by Wikipedia sentences.

Table 5.12: μ_{label} of all labels in SmSA dataset, after dataset debiasing by Wikipedia

Label	μ_{label}
positive	-3.62
negative	-1.92
neutral	-0.42

Table 5.13 shows the impact of dataset debiasing using Wikipedia sentences on the Hate Speech dataset, for term-label occurrences. Since the Wikipedia sentences added to debias datasets are neutral in sentiment, all of them are labeled as *none* in this dataset. Therefore, only the *none* label experience increases in term occurrences.

Table 5.13: Religious term occurrence per label in Hate Speech dataset, after dataset debiasing by Wikipedia

Terms	hate speech	abusive	none
islam	360	151	433
kristen	42	14	467
masjid	13	6	145
gereja	3	1	195
quran	46	31	134
alkitab	1	0	65
ulama	97	43	223
pendeta	3	1	5

Table 5.14 shows the impact of dataset debiasing using Wikipedia sentences on the Hate Speech dataset, for PMI averages. As shown, the positive-related label *none* has higher PMI averages than both negativity labels *hate speech* and *abusive*, which shows that this dataset does not contain religion bias. Therefore, much like dataset debiasing using sentence templates, this shows that debiasing dataset using Wikipedia sentences do not introduce additional bias to originally-unbiased datasets.

Table 5.14: μ_{label} of all labels in Hate Speech dataset, after dataset debiasing by Wikipedia

Label	μ_{label}
hate speech	-2.22
abusive	-2.81
none	-0.24

Table 5.15 shows the impact of dataset debiasing using Wikipedia sentences on accuracy scores. For models trained on the SmSA dataset, dataset debiasing by Wikipedia tends to perform better on the validation split compared to both pre-debiasing and dataset debiasing by sentence templates. As an example, for lstm_wiki, the validation accuracy metrics are 90.78% before debiasing, 90.88% after dataset debiasing using sentence templates, and 91.32% after dataset debiasing using Wikipedia sentences.

Since validation splits were not seen in the training process, this implies the models are able to generalize better compared to the two previously-mentioned states (pre-debiasing and dataset debiasing using sentence templates). The generalization capability is likely contributed by the high variety of sentences contained in Wikipedia articles, compared to the repetitions done in the sentence template case.

A notable exception to this case is the *lstm_twitter* case, where the accuracy score for validation score decreases into 90.39%, when compared to pre-debiasing (90.74%) and dataset debiasing using sentence templates (91.43%). The biased Twitter embedding may play a role in this effect, where the bias in the embedding cause certain sentences to be wrongly identified.

On the Hate Speech dataset, the performance of dataset debiasing by Wikipedia tends to be better than dataset debiasing by sentence templates, and are at least comparable to pre-debiasing accuracy scores, on both splits. As an example, for the lstm_tempo model, the accuracy metric results after dataset debiasing by Wikipedia sentences are 69.24% and 67.27%, for training and validation splits respectively. The accuracy scores obtained here are better than after dataset debiasing by sentence templates (65.57% and 66.01% for training and validation splits respectively) and are comparable to pre-debiasing (70.02% training, 64.64% validation).

One possible explanation for this effect is the nature of the Hate Speech dataset itself, which originally contain no dataset bias. Since there is no need to change the label distribution of religious terms in the Hate Speech dataset, the increased variability of sentences obtained from Wikipedia instead acts as data points that increase the generalization capability of Bi-LSTM models, and therefore increase their accuracy scores. This is in contrast of the duplicated sentences from sentence templates, which adds little to no new information for the Bi-LSTM models to learn, and therefore lowers overall performance. A notable exception to this effect is the *lstm_wiki* case, where the accuracy of dataset debiasing by Wikipedia on the validation split (70.23%) are noticeably lower than both base (74.6%) and dataset debiasing by sentence templates (76.03%).

Table 5.15: Accuracy results on all datasets for each embedding, after dataset debiasing by Wikipedia sentences

Data split	Twitter	Wiki	Tempo	ConLL
Training (SmSA)	98.64	95.78	97.11	95.31
Validation (SmSA)	90.39	91.32	91.16	91.43
Training (Hate Speech)	71.7	70.79	69.24	69.6
Validation (Hate Speech)	68.76	70.23	67.27	72.34

The results of parity metrics to detect allocation harms is presented at Table 5.16. As shown by the results of FNR and TPR, dataset debiasing by Wikipedia sentences manage to mitigate allocation harms inflicted against Islamic sentences by making the results of said metrics more equal across both religious groups. However, compared to dataset debiasing by sentence templates, the results of dataset debiasing by Wikipedia sentences contain less extreme values (i.e., 0% FNR and 100% TPR, as shown in the results after dataset debiasing using sentence templates).

An example of this is seen in the lstm_tempo model, which originally has 16.67% FNR for Islamic sentences and 0% for Christianity sentences, showing an initial allocation harm against Islamic sentences. After dataset debiasing using sentence templates, the values change into 0.28% and 0% for Islamic and Christianity sentences respectively. This mitigates the initial allocation harm, but the 0% FNR on Christianity sentences may indicate that the model memorizes the added sentences, which increases performance but not general model capability. After dataset debiasing using sentence templates, the FNR values change into 2.05% for Islamic sentences, and 1.21% for Christianity sentences. This mitigates the allocation harm by closing the FNR gap between Islamic and Christianity sentences, yet also introduces less extreme values as seen in the 1.21% FNR for Christianity. The result shows that the amount of unique sentences obtained from Wikipedia adds variability to the SmSA dataset while simultaneously debiasing them.

The added variability makes learning from this augmented dataset a harder problem compared to previous debiasing methods, which allows the model to generalize better. This is supported by two findings: first, the FPR results of dataset debiasing by Wikipedia sentences all increase over all models and groups of religious terms. This shows that there are more sentences containing various religious groups that are mispredicted as positive. Since the goal of debiasing done in this thesis is to decouple religious terms from negativity (i.e., sentences are not automatically labeled as negative just because a religious term exists), this shows that dataset debiasing by Wikipedia sentences are a step towards that goal. Second, the increase in accuracy scores for most validation splits shown in Table 5.15 show that all LSTM models are able to generalize better to a set of sentences unseen in the training process.

Table 5.16: Parity metric results of SmSA dataset, in percentage, after dataset debiasing by Wikipedia

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	3.76	2.4	97.6	31.76
lstm_twitter - Christianity	2.78	1.22	98.78	10.68
lstm_wiki - Islam	2.26	1.03	98.97	31.29
lstm_wiki - Christianity	2.78	0.3	99.7	9.86
lstm_tempo - Islam	0.75	2.05	97.95	32.47
lstm_tempo - Christianity	2.78	1.22	98.78	10.68
lstm_conll - Islam	7.52	1.71	98.29	30.12
lstm_conll - Christianity	11.11	0.3	99.7	9.04

As shown in Table 5.17, dataset debiasing using Wikipedia sentences does not introduce additional biases in the form of allocation harm, for either religious groups. Instead, there are improvements in the FNR and TPR metrics for the Hate Speech dataset. As an example, consider the lstm_conll model, which obtains 5.59% and 1.59% FNR scores for Islamic and Christianity sentences, respectively. After dataset debiasing by sentence templates, these scores are 2.47% and 0.57% respectively, showing an overall improvement in both FNR cases. After dataset debiasing by Wikipedia sentences, the FNR scores are now 2.26% and 0.62% for Islamic and Christianity-related sentences respectively. This improves the FNR when compared to pre-debiasing results, and is comparable to the results of dataset debiasing using sentence templates.

These improvements are at least equal, and oftentimes better, than dataset debiasing by sentence templates and pre-debiasing results for all models. This, much like the case shown in the SmSA case after dataset debiasing using Wikipedia sentences, shows the capability of unique sentences obtained from Wikipedia to act as additional data points for the models to learn from.

A side effect of dataset debiasing by Wikipedia is that the overconfidence shown

in high FPR for sentences with Christianity terms are maintained or reduced when compared to dataset debiasing using sentence templates and pre-debiasing. The two exceptions are lstm_twitter and lstm_conll, whose 11% FPR and 22% for Christianity sentences are higher than pre-debiasing scores (although lower than the FPR scores after dataset debiasing using sentence templates).

Table 5.17: Parity metric results of Hate Speech dataset, in percentage, after dataset debiasing by Wikipedia

Model - Term	FPR	FNR	TPR	DPR
lstm_twitter - Islam	4.96	2.39	97.62	36.25
lstm_twitter - Christianity	11.11	0.47	99.53	6.29
lstm_wiki - Islam	3.1	6.79	93.21	39.73
lstm_wiki - Christianity	11.11	1.41	98.59	7.16
lstm_tempo - Islam	1.86	4.76	95.24	38.9
lstm_tempo - Christianity	2.2	0.63	99.37	7.02
lstm_conll - Islam	5.58	2.26	97.74	35.95
lstm_conll - Christianity	22.22	0.63	99.37	5.7

The result of representation harm using Wikipedia sentences is shown in Table 5.18 for the SmSA case. Much like the dataset debiasing by sentence templates, dataset debiasing by Wikipedia sentences manage to correct mispredictions from sentence templates. The mitigation effect is less aggressive than dataset debiasing by sentence templates. As an example, for the *lstm_tempo* case, the sentence *saya cinta agama islam* went from (negative, 92.3017) before debiasing to (positive, 99.9109) after dataset debiasing by sentence templates, but only to (negative, 56.3298) after dataset debiasing by Wikipedia. This shows that the debiasing process relies less on remembering the sentences used to debias, and instead more on better generalization. The increased generalization capability given by the unique sentences is shown to be able to lower the negative impacts of embedding bias to some extent, as seen in the case of *tenggang rasa antar kaum islam harus jaga*

and *tenggang rasa antar kaum kristen harus jaga* for the lstm_twitter case. After dataset debiasing using sentence templates, these sentences fail to be mitigated at 91.37% *negative* and 94.24% *negative*. The misprediction for *tenggang rasa antar kaum islam harus jaga* lowers into 80% *negative* after dataset debiasing by Wikipedia, whereas the misprediction for *tenggang rasa antar kaum kristen harus jaga* is properly mitigated at 72.62% *neutral*.

However, for dataset debiasing by Wikipedia, the amount of sentences that fail to be mitigated are more than the dataset debiasing by sentence templates method. This is most evidently seen on the lstm_twitter case, where this dataset debiasing method fails to mitigate cases where other models succeed. An example is the sentence *saya cinta agama islam*, which is still being mispredicted as *negative* with 97.56% probability, when compared to other models that either reduces the probability or outright changes the label into a positivity-related label.

This shows that there is still room for improvement for the dataset debiasing using Wikipedia method. Specifically for the lstm_twitter case, note the fact that sentences failed to be mitigated by this model after dataset debiasing by Wikipedia but able to be mitigated by other models also happen in the dataset debiasing using sentence template case. Since findings in Table 4.8 show that the Twitter embedding contains embedding bias, this may show the persistence of embedding bias preventing representation harm mitigation, despite the debiasing done to the dataset.

Table 5.18: Sentence template results of SmSA dataset, after dataset debiasing by Wikipedia

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	negative, 75.96	negative, 98.38	negative, 94.18	neutral, 51.63
saya anut agama kristen	neutral, 90.14	negative, 86.49	negative, 56.03	neutral, 94.79
saya cinta agama islam	negative, 97.56	positive, 56.14	negative, 56.33	negative, 57.67
saya cinta agama kristen	negative, 66.09	positive, 62.9	neutral, 85.63	neutral, 47.75
sekolah saya ajar agama islam	negative, 85.4	negative, 87.8	negative, 99.78	neutral, 69.55
sekolah saya ajar agama kristen	neutral, 66.58	negative, 56.03	negative, 97.67	neutral, 97.9
tenggang rasa antar kaum islam harus jaga	negative, 80	neutral, 60.04	neutral, 59.51	neutral, 67.12
tenggang rasa antar kaum kristen harus jaga	neutral, 72.63	neutral, 86.41	neutral, 96.82	neutral, 95.16
masjid jadi tempat aman bagi seluruh masyarakat indonesia	positive, 87.41	neutral, 72.26	neutral, 76.43	neutral, 97.82
gereja jadi tempat aman bagi seluruh masyarakat indonesia	positive, 98.2	neutral, 69.39	neutral, 74.82	neutral, 96.51
saya tidak setuju dengan ajaran agama islam	negative, 99.86	negative, 86.74	negative, 99.32	negative, 55.03
saya tidak setuju dengan ajaran agama kristen	negative, 98.89	negative, 73.55	negative, 96.85	neutral, 79.61

The same trend for SmSA is seen again in the Hate Speech case, as shown in Table 5.19. For most cases, dataset debiasing using Wikipedia sentences does not introduce additional representation harm to either religious group, and can reduce misprediction when it does occur, when compared to the pre-debiasing scores. An example is the sentence *saya anut agama islam*, for the lstm_conll case. This sentence was mispredicted into 29.65% *hate speech* probability, but was corrected into

7.04% probability after dataset debiasing using Wikipedia.

However, there are edge cases where sentences fail to be mitigated after dataset debiasing by Wikipedia sentences. As an example, the sentence *sekolah saya ajar agama islam* was predicted as 22% hate speech by the *lstm_tempo* model, but worsens into 25% hate speech after dataset debiasing by Wikipedia. This shows that there are certain impacts from the embedding itself that need to be debiased, which are unable to be solved by the dataset debiasing method alone.

Table 5.19: Sentence template results of Hate Speech dataset (hate speech label), after dataset debiasing by Wikipedia

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	2.16	9.01	15.82	7.05
saya anut agama kristen	0.32	0.21	1.03	0.84
saya cinta agama islam	4.26	3.14	4.25	16.64
saya cinta agama kristen	0.53	0.11	0.66	1.92
sekolah saya ajar agama islam	6.05	8.62	25.17	10.24
sekolah saya ajar agama kristen	0.96	0.22	7.88	0.85
tenggang rasa antar kaum islam harus jaga	1.81	47.98	8.95	45.32
tenggang rasa antar kaum kristen harus jaga	0.41	9.33	1.17	11.7
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0.1	42.53	1.12	0.16
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.18	56.04	0.5	0.06
saya tidak setuju dengan ajaran agama islam	3.1	38.49	60.15	43.5
saya tidak setuju dengan ajaran agama kristen	0.53	1.4	27.93	9.51

Table 5.20 shows the impact of dataset debiasing using Wikipedia sentences on Hate Speech dataset, for the *abusive* label. Much like other cases, dataset debiasing using Wikipedia sentences maintain the already-low misprediction from pre-debiasing scores. Also like dataset debiasing using sentence templates, dataset debiasing using Wikipedia manages to mitigate the mispredictions done by the

lstm_conll model for certain sentences. As an example, the sentence *tenggang rasa antar kaum islam harus jaga* has 9.74% *abusive* probability before debiasing, and 0.77% *abusive* probability after debiasing.

Table 5.20: Sentence template results of Hate Speech dataset (abusive label), after dataset debiasing by Wikipedia

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0.11	0.08	0.27	0.08
saya anut agama kristen	0.06	0	0	0.01
saya cinta agama islam	0.23	0.2	0.59	0.25
saya cinta agama kristen	0.1	0	0.02	0.04
sekolah saya ajar agama islam	0.14	0.13	0.2	0.22
sekolah saya ajar agama kristen	0.05	0	0.01	0.02
tenggang rasa antar kaum islam harus jaga	0.04	0.75	0.16	0.77
tenggang rasa antar kaum kristen harus jaga	0.02	0.08	0.02	0.17
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0.02	0.19	0	0
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.04	0.22	0	0
saya tidak setuju dengan ajaran agama islam	0.11	0.44	2.19	0.22
saya tidak setuju dengan ajaran agama kristen	0.05	0.02	0.08	0.05

The results of Section 5.2 shows the impact of dataset debiasing using Wikipedia sentences on dataset bias, as inquired by the third research question of this thesis, as

well as the impact of dataset debiasing using Wikipedia sentences on allocation and representation harms as inquired by the fifth research question. Much like the case of dataset debiasing using sentence templates, dataset debiasing using Wikipedia sentences manages to mitigate biases in datasets without introducing additional biases on originally-unbiased datasets. Since the Wikipedia sentences used for augmentation are neutral in label, only the *neutral* label exhibit changes after dataset debiasing.

The impact of dataset debiasing using Wikipedia sentences on allocation harms show that the increased variety of Wikipedia sentences allow Bi-LSTM models to generalize better, when compared to dataset debiasing using sentence templates. This is best seen in the results of allocation harms on Tables 5.16 and 5.17, where it reduces the effect of allocation harms on the augmented SmSA dataset by reducing the FNR and TPR gap over both religious identities, without introducing extreme values (i.e, 0% FNR and 100% TPR).

The impact of dataset debiasing using Wikipedia sentences on representational harms support the findings on allocation harms. In particular, it manages to mitigate sentences by relying on increased generalizing capabilities compared to dataset debiasing by sentence templates, as seen by the successful mitigation of *tenggang rasa antar kaum kristen harus jaga*. However, there are sentences that fail to be mitigated using this debiasing method, which is more than debiasing by sentence templates. Additionally, the debiasing effects are relatively weaker than dataset debiasing by sentence templates. Some sentences that were originally mispredicted prior to debiasing, which was mitigated into the correct non-negative label after debiasing by sentence templates, still maintains the same mispredicted label after dataset debiasing by Wikipedia sentences, albeit with considerable decrease in label probability. This shows that there is still room for improvement on the mitigation effect.

5.3 Embedding Debiasing

Table 5.21 shows the impact of embedding debiasing on cosine similarity results, where it managed to mitigate biases in the Twitter embedding by lowering the similarity of marginalized religious terms to religious insults (e.g: *kafir* is no longer close to Christianity-related terms). However, few biases remain, although both cases were not considered in the main problem formulation. The term *kafir* is still close to *pribumi*, a racial-sensitive insult, which highlights the importance of accomodating for intersectionality on debiasing endeavors. Another can be seen in

the word *syiah*, itself a marginalized branch of Islam often used as insults towards other Muslims in Indonesia, is close to *wahabi*, another marginalized branch of Islam. This may suggest considering certain branches of Islam as religious minorities to be considered in religion bias, as well as considering other forms of bias (e.g: racial bias) and possibility of intersectioning harms.

Table 5.21: Per-word similarity of certain religion-related terms for all embeddings, after embedding debiasing

Word similarity	Twitter	Wiki	Tempo	ConLL
radikal	ormas, khilafah, politikus, politisi, teroris	moderat, progresif, konservatif, sentris, revolusioner	fundamentalis, ekstremis, moderat, islamis, militan	redikal, radaikal, radikal, bebas.radikal, radikal-radikal
haram	jantan, dorang, depa, haramgirls, haaa	masjidil, manyarah, najis, makruh, haramnya	haramnya, terlarang, ilegal, laknat, selundupan	diharamkan, mengharamkan, haramnya, harom, diharamkannya
kafir	heran, munafik, goblok, pribumi, setan	musyrik, murtad, munafik, berdosa, zalim	murtad, musyrik, khawarij, beriman, bidah	musyrik, mashroni, kekafiran, musyrikin, kafir/musrik
halal	bihalal, tiens, bpom, herbal, syariah	kosher, bihalal, kehalalan, qurban, diharamkan	kehalalan, halalnya, labelisasi, lppom, bihalal	hahal, thayiban, muisurat, kehalalannya, halalnya
syiah	puak, wahabi, feminist, offshore, neves	sunni, wahabi, nasrani, syi, salafi	sunni, najaf, salafi, sadr, hazara	syi'ah, sunni, wahabi, wahhabi, ahlussunnah

The impact of embedding debiasing on accuracy is shown in Table 5.22. When

compared to pre-debiasing accuracy scores, embedding debiasing increases accuracy on the training split, but decreases accuracy on the validation split. An example of this is seen in the lstm_wiki model, which obtains a 96.35% training accuracy and 90.78% validation accuracy pre-debiasing. After embedding debiasing, the accuracy scores change into 95.78% and 91.32% for training and validation respectively.

Compared to the dataset debiasing method done in the SmSA and Hate Speech datasets, the accuracy on the training split is higher than both dataset debiasing methods, but vice versa on the validation split. Since embedding debiasing changes the relation of all word vectors with regards to certain religious terms, the decrease on the validation split may represent the bias-variance tradeoff, where less biased embeddings are traded for lower overall model performance. Combined with the higher accuracy on the training set, embedding debiasing may also work to improve model performance in general, even for models that utilize unbiased embeddings.

A notable edge case of embedding debiasing shows up in the *lstm_tempo* case on Hate Speech dataset, where it manages to improve accuracy scores compared to both dataset debiasing methods and pre-debiasing, on both training and validation splits. Another edge case from the same dataset is the *lstm_twitter* case, where it outperforms dataset debiasing using sentence templates on both splits, although it is still outperformed by dataset debiasing using Wikipedia. This shows that there are points of variability that exist in the embeddings themselves, which reacts differently to the same debiasing method. Additionally, these cases show that debiasing datasets that originally contain no unwanted bias may instead harm model performances.

Table 5.22: Accuracy results on all datasets for each embedding, after embedding debiasing

Data split	Twitter	Wiki	Tempo	ConLL
Training (EmoT)	98.18	96.65	92.56	94.01
Validation (EmoT)	63.75	62.95	62.61	66.48
Training (SmSA)	98.68	97.83	98.31	97.45
Validation (SmSA)	89.4	89.27	89.23	89.2
Training (Hate Speech)	70.64	72.59	69.43	71.61
Validation (Hate Speech)	67.03	69.02	72.44	66.78

As shown in Table 5.23, embedding debiasing generally fails to mitigate allocation harms in the EmoT dataset, and in some cases worsen existing allocation harm. As an example, the FNR score for lstm_twitter on Christianity-related sentences is 33%, compared to 8.69% for Islam-related sentences. Since the FNR gap between Islamic and Christianity sentences still exist after embedding debiasing, it shows that embedding debiasing fails to mitigate allocation harm in this case. In fact, the score is worse than pre-debiasing, which has the exact 33% FNR for Christianity sentences, and 0% FNR for Islam-related sentences.

However, there are edge cases where the performance of embedding debiasing mitigates allocation harm. As an example, Table 5.23 shows that for *lstm_conll*, embedding debiasing massively improves FN (both reduced to 0%) and TP (both increased to 100%). Considering the fact that embedding debiasing does not manipulate the original label distribution of sentences in the dataset, these findings shows that embedding debiasing, as a debiasing method, is capable of mitigating allocation harms in certain situations. Additionally, these findings show that different embeddings may react differently to the same debiasing method.

Table 5.23: Parity metric results of EmoT dataset, in percentage, after embedding debiasing

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	4.76	8.7	91.3	64.62
lstm_twitter - Christianity	10.53	33.33	66.67	81.82
lstm_wiki - Islam	0	8.7	91.3	67.69
lstm_wiki - Christianity	5.26	33.33	66.67	86.37
lstm_tempo - Islam	2.38	4.35	95.65	64.62
lstm_tempo - Christianity	0	33.33	66.67	90.91
lstm_conll - Islam	0	0	100	64.62
lstm_conll - Christianity	5.26	0	100	81.82

Much like the EmoT case, Table 5.24 shows that embedding debiasing generally fails to mitigate allocation harms in the SmSA case, and in some cases worsen existing allocation harm. A prominent example for this effect is seen in Table 5.24, where the FN metric results for *lstm_wiki* and *lstm_conll* does not change after embedding debiasing, still showing allocation harms against Islamic terms. This shows that in cases where the datasets also contain unwanted bias, embedding debiasing alone is insufficient to mitigate allocation harms.

However, for some models, embedding debiasing manages to mitigate allocation harms. In the case shown in Table 5.24, for *lstm_tempo*, embedding debiasing improves FN (from 16.67% pre-debiasing into 4.67%) and TP (from 83.33% pre-debiasing into 95.83%) for the Islamic terms case. This strengthens the findings in the EmoT results as shown in Table 5.23, which highlights the variability at model level as a result of dataset-embedding interaction.

Table 5.24: Parity metric results of SmSA dataset, in percentage, after embedding debiasing

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	0.75	0	100	84.08
lstm_twitter - Christianity	0	0	100	92.31
lstm_wiki - Islam	2.26	4.17	95.83	83.44
lstm_wiki - Christianity	0	0	100	92.31
lstm_tempo - Islam	1.5	4.17	95.83	84.08
lstm_tempo - Christianity	0	0	100	92.31
lstm_conll - Islam	1.5	4.17	95.83	84.08
lstm_conll - Christianity	0	0	100	92.31

Table 5.25 shows the impact of embedding debiasing on Hate Speech dataset, which originally does not contain unwanted religious bias. In this case, embedding debiasing manages to improve the overall performance (in terms of lowering FN and increasing TP) on 3 out of 4 models, excluding *lstm_wiki*. As an example, for the *lstm_twitter* case, embedding debiasing improves TP for both sentences with Islamic terms (from 94.75% into 97.55%) and Christianity terms (from 97.76% into 99.04%).

The prior findings confirm both effects of embedding debiasing previously shown in Tables 5.23 and 5.24. In particular, the overall improvements of model performance using parity metrics shown in Table 5.25 shows that embedding debiasing works well in environments with unbiased datasets. Additionally, the decreased performance of *lstm_wiki* as opposed to improvements in other cases of embedding debiasing confirms that different embeddings may react differently to the same embedding debiasing method.

Table 5.25: Parity metric results of Hate Speech dataset, in percentage, after embedding debiasing

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	3.1	2.45	97.55	45.74
lstm_twitter - Christianity	17.78	0.96	99.04	11.17
lstm_wiki - Islam	1.03	8.04	91.96	49.72
lstm_wiki - Christianity	6.67	4.79	95.21	15.92
lstm_tempo - Islam	6.4	3.15	96.85	44.6
lstm_tempo - Christianity	11.111	1.92	98.08	12.85
lstm_conll - Islam	5.17	1.92	98.08	44.51
lstm_conll - Christianity	15.56	0.96	99.04	11.45

The insufficiency of embedding debiasing on biased datasets are shown again on Tables5.26 and 5.27, for EmoT and SmsA datasets respectively. On the EmoT case, embedding debiasing fails to correct misprediction of most templates. Additionally, there are cases where embedding debiasing mispredicts sentences already correctly predicted before debiasing. This is shown in the *saya cinta agama kristen* sentence, which was correctly predicted as *love* prior to debiasing, yet is mispredicted into *fear* and *sadness* labels for lstm_twitter and lstm_tempo respectively.

Table 5.26: Sentence template results of EmoT dataset, after embedding debiasing

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	happy, 93.94	happy, 76.65	sadness, 83.28	anger, 81.3
saya anut agama kristen	anger, 91.94	fear, 70.72	sadness, 84.97	anger, 85.22
saya cinta agama islam	love, 93.75	love, 99.22	love, 97.43	love, 84.94
saya cinta agama kristen	fear, 82.34	love, 98.07	sadness, 93.82	love, 81.02
sekolah saya ajar agama islam	sadness, 94.22	sadness, 84.48	sadness, 98.11	sadness, 86.53
sekolah saya ajar agama kristen	sadness, 86.46	sadness, 87.62	sadness, 97.73	sadness, 85.48
tenggang rasa antar kaum islam harus jaga	anger, 99.43	anger, 99.49	anger, 99.23	anger, 89
tenggang rasa antar kaum kristen harus jaga	anger, 99.87	anger, 99.60	anger, 99.49	anger, 92.64
masjid jadi tempat aman bagi seluruh masyarakat indonesia	happy, 97.08	happy, 96.23	happy, 99.55	happy, 99.07
gereja jadi tempat aman bagi seluruh masyarakat indonesia	happy, 95.1142	happy, 90.9947	happy, 99.1701	happy, 99.258
saya tidak setuju dengan ajaran agama islam	sadness, 73.61	anger, 74.36	sadness, 90.3	anger, 85.95
saya tidak setuju dengan ajaran agama kristen	anger, 88.08	anger, 79.92	sadness, 90.48	anger, 88.96

On the SmSA case, the case is more varied over templates, where certain templates are harder to correct than others. As an example, the sentence templates *saya anut agama [agama]* and *sekolah saya ajar agama [agama]* consistently fails to be mitigated over all models. On the other hand, longer sentence templates such as *tenggang rasa antar kaum [agama] harus jaga* and *[tempat ibadah] jadi tempat aman bagi seluruh masyarakat indonesia* are mitigated, although the effect of

mitigation varies per model.

An observation of embedding debiasing on the SmSA case shows that there are sentences that are mitigated after embedding debiasing but not after dataset debiasing, and vice versa. An example is shown on the *lstm_twitter* case, which was able to correct the misprediction of *tenggang rasa antar kaum islam harus jaga* into a positive prediction after embedding debiasing, but not after both dataset debiasing methods. On the other hand, the sentence *saya tidak setuju dengan ajaran agama islam* was mispredicted into a positive sentence on 2 out of 4 models after embedding debiasing, but is always corrected into a negative sentence after dataset debiasing by Wikipedia.

Notably, this effect happens even on embeddings that do not contain embedding bias. This is seen in the *lstm_wiki* case, where the sentence *saya anut agama islam* has better mitigation results after embedding debiasing, when compared to dataset debiasing by Wikipedia (from 98% into 87%). Since the findings on Table 5.18 shows that dataset debiasing by Wikipedia has edge cases that fails to be mitigated, there is a possibility that combining embedding debiasing and dataset debiasing can cover the weaknesses of each other.

Table 5.27: Sentence template results of SmSA dataset, after embedding debiasing

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	negative, 99.99	negative, 87.92	negative, 99.88	negative, 99.66
saya anut agama kristen	negative, 100	negative, 99.56	negative, 99.99	negative, 99.95
saya cinta agama islam	positive, 99.77	positive, 69.93	positive, 99.9	positive, 77.3
saya cinta agama kristen	negative, 93.14	negative, 84.42	positive, 50.53	negative, 92.78
sekolah saya ajar agama islam	negative, 99.999	negative, 99.96	negative, 99.89	negative, 99.95
sekolah saya ajar agama kristen	negative, 100	negative, 99.99	negative, 99.99	negative, 99.99
tenggang rasa antar kaum islam harus jaga	positive, 91.21	neutral, 72.92	positive, 56	neutral, 39.5
tenggang rasa antar kaum kristen harus jaga	negative, 89.49	negative, 46.87	negative, 97.3	negative, 84.94
masjid jadi tempat aman bagi seluruh masyarakat indonesia	positive, 71.51	neutral, 50.43	positive, 89.68	positive, 84.98
gereja jadi tempat aman bagi seluruh masyarakat indonesia	neutral, 85	positive, 53.22	positive, 47.91	positive, 52.1
saya tidak setuju dengan ajaran agama islam	positive, 78.79	negative, 96.5	positive, 73.05	negative, 80.47
saya tidak setuju dengan ajaran agama kristen	negative, 92.05	negative, 99.6	positive, 54.34	negative, 94.59

Table 5.28 shows that for certain sentences, embedding debiasing introduces representation harms on models that learn from originally-unbiased embeddings, but consistently reduces representation harms on models that learn from originally-biased embeddings (i.e., *lstm_twitter*). This is in contrast of dataset debiasing using Wikipedia, where said debiasing method worsens the misprediction on those sentences over all models. As an example, for the *lstm_twitter* case, the sentence

saya anut agama islam has 0.8% *hate speech* probability before debiasing, 2% after dataset debiasing by Wikipedia, and 0% probability after embedding debiasing. However, for the same sentence, *lstm_wiki* achieves 2.4% probability before debiasing, 9% after dataset debiasing by Wikipedia, and 27% after embedding debiasing. This happens in contrast of the results of Table 5.27, where embedding debiasing can correct certain mispredictions even if the embedding itself is unbiased.

Table 5.28: Sentence template results of Hate Speech dataset (hate speech label), after embedding debiasing

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0	27.63	0.43	11.57
saya anut agama kristen	0	8.71	0.16	0.78
saya cinta agama islam	0.17	7.99	2.13	0.69
saya cinta agama kristen	0.02	2.51	0.86	0.07
sekolah saya ajar agama islam	1.09	34.94	0.82	6.63
sekolah saya ajar agama kristen	0.12	9.73	0.42	0.58
tenggang rasa antar kaum islam harus jaga	18.51	73.88	32.71	1.24
tenggang rasa antar kaum kristen harus jaga	1.13	25.17	17.1	0.14
masjid jadi tempat aman bagi seluruh masyarakat indonesia	3.98	14.31	0.33	0.04
gereja jadi tempat aman bagi seluruh masyarakat indonesia	17.27	16.15	0.56	0.04
saya tidak setuju dengan ajaran agama islam	2.16	39.04	16.11	9.1
saya tidak setuju dengan ajaran agama kristen	0.47	17.37	14.68	1.61

The impact of embedding debiasing for the *abusive* label of Hate Speech dataset can be seen on Table 5.29. Here, it shows that embedding debiasing generally does not introduce additional mispredictions for the *abusive* label on most models. Since negativity-related mentions of religious identities that exist in this dataset are mostly *hate speech* related, it follows that the pre-debiasing models rarely mispredict them,

due to the label representation. As such, embedding debiasing is able to maintain or lower the misprediction rates, since it does not change the label distributions of the original Hate Speech dataset.

The lstm_tempo model is an anomaly to the embedding debiasing results, showing a misprediction increase in *tenggang rasa antar kaum islam harus jaga* (from 0.39% *abusive* into 4.26% *abusive*) and *saya tidak setuju dengan ajaran agama islam* (from 0% to 2.19% *abusive*). While the resulting misprediction after embedding debiasing is still relatively low, it still highlights the possibility of variance in model level on debiasing impacts.

Table 5.29: Sentence template results of Hate Speech dataset (abusive label), after embedding debiasing

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0	0.79	0.16	0.83
saya anut agama kristen	0	0.27	0.03	0.07
saya cinta agama islam	0.02	0.12	0.97	0.03
saya cinta agama kristen	0	0.05	0.2	0
sekolah saya ajar agama islam	0	0.24	0.22	0.23
sekolah saya ajar agama kristen	0	0.09	0.05	0.03
tenggang rasa antar kaum islam harus jaga	0.09	2.77	4.26	0.24
tenggang rasa antar kaum kristen harus jaga	0	0.68	0.27	0.03
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0	0.02	0	0
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0	0	0	0
saya tidak setuju dengan ajaran agama islam	0.03	0.91	2.43	0.22
saya tidak setuju dengan ajaran agama kristen	0.01	0.35	0.68	0.05

The results of Section 5.3 shows the impact of embedding debiasing on embedding bias, as inquired by the fourth research question of this thesis, as well as the impact of embedding debiasing on allocation and representation harms as inquired by the fifth research question. As shown in Table 5.21, embedding debiasing managed to mitigate embedding bias by lowering the similarity of marginalized re-

ligious terms to religious insults. However, other types of unwanted bias emerge, particularly for marginalized religious identities not considered in the debiasing process, as well as for certain racial identities.

The prevalence of biased datasets massively hampers the effectiveness of embedding debiasing on mitigating allocation and representational harms. In particular, for the case of EmoT and SmSA datasets, which were found to contain biases, embedding debiasing fail to mitigate allocation and representational harms caused by models that learn from these datasets. However, for the sentence templates used to measure representational harm, there are cases where embedding debiasing can mitigate the misprediction but not both dataset debiasing methods, and vice versa. Combined with the fact that embedding debiasing tend to improve overall model performance in the form of accuracy, as well as how it works well on models that learn from unbiased datasets, this may show the merit of joint debiasing (Ghai et al., 2022). By combining both debiasing methods, there is a possibility that the joint debiasing method inherits the best of both debiasing methods

5.4 Joint Debiasing with Sentence Templates

Table 5.30 shows the impact of joint debiasing using sentence templates on accuracy scores. A common trend in joint debiasing using sentence templates over both SmSA and Hate Speech datasets is that it tends to lower accuracy scores in validation splits when compared to pre-debiasing scores. This agrees to the findings of Ghai et al. (2022), where increasing amounts of debiasing treatments lower overall model performance, agreeing to the bias-variance tradeoff. Debiasing methods increases the variability of the learning problem to be solved by the Bi-LSTM models, which subsequently reduces the biases the Bi-LSTM models can learn from. This increases the learning difficulty of the models, and therefore lowers performance.

In the SmSA dataset, joint debiasing using sentence templates has better accuracy on the training splits when compared to dataset debiasing methods, and are roughly comparable to the results of embedding debiasing. However, on the validation split, the accuracy scores are lower compared to dataset debiasing methods, although they are still higher compared to embedding debiasing. Since embedding debiasing was previously shown on Table 5.22 to perform better than all dataset debiasing methods on the training split, but worse on the validation split, this shows that joint debiasing using sentence templaes is able to utilize the best of both methods on their overall performance.

In the Hate Speech dataset, joint debiasing by sentence templates tend to out-

perform dataset debiasing by sentence templates on both splits, which supports both how debiasing datasets unbiased datasets can result in lower performance and joint debiasing being able to utilize the improvements of embedding debiasing. However, joint debiasing by sentence templates tend to perform worser than dataset debiasing by Wikipedia on both splits. This is likely caused by the nature of dataset debiasing by Wikipedia sentences when compared to dataset debiasing by sentence templates, as explained in Table 5.15. An exception happens in the *lstm_tempo* case, where joint debiasing by template outperforms dataset debiasing by Wikipedia on the validation split, and *lstm_conll* where the better performance instead happens on the training split. This again shows the variability in the embeddings themselves, and how they would react to the same embedding debiasing method.

Table 5.30: Accuracy results on all datasets for each embedding, after joint debiasing by sentence templates

Data split	Twitter	Wiki	Tempo	ConLL
Training (SmSA)	98.65	98.17	98.11	97.32
Validation (SmSA)	90.14	90.02	88.82	88.84
Training (Hate Speech)	67.58	67.82	68.65	71.4
Validation (Hate Speech)	65.97	68.22	71.39	68.17

As shown in Table 5.31, joint debiasing by sentence template outperforms both dataset debiasing methods as well as embedding debiasing in terms of mitigating allocation harms. This is shown by the reduction of FNR, of which joint debiasing by sentence template reaches the lowest when compared to both dataset debiasing methods, embedding debiasing, and pre-debiasing scores, as well as the increase of TPR, of which joint debiasing reaches the highest. Additionally, the increased FPR for *lstm_twitter* on both sentences containing Islam and Christianity terms, when compared to the FPR of other models, show that the embedding debiasing does impact the previously-biased Twitter embedding. When compared to the results of embedding debiasing as shown in Table 5.23 and Table 5.24, this confirms the findings that embedding debiasing alone is not enough to properly debias models that also learn from biased datasets.

Table 5.31: Parity metric results of SmSA dataset, in percentage, after joint debiasing by sentence templates

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	5.26	0	100	25.87
lstm_twitter - Christianity	2.78	0	100	9.749
lstm_wiki - Islam	0.75	0.28	99.72	27.31
lstm_wiki - Christianity	2.78	0	100	9.75
lstm_tempo - Islam	3	0.28	99.72	26.69
lstm_tempo - Christianity	0	0	100	10.03
lstm_conll - Islam	0.75	0.56	99.44	27.52
lstm_conll - Christianity	0	0	100	10.03

Joint debiasing by sentence templates tends to perform well on the Hate Speech dataset as seen in Table 5.32, where it manages to consecutively lowers FNR and increases TPR for sentences of both religious groups when compared to pre-debiasing scores. However, the FNR and TPR scores are worser (i.e., higher FNR and lower TPR) when compared to both debiasing methods as well as embedding debiasing. This is likely influenced by the nature of Hate Speech dataset, which does not contain dataset bias. The unbiased nature of this dataset means that the results of parity metrics such as FNR and TPR are purely influenced by model performance, unlike the case of biased datasets where the results of said parity metrics can be influenced by the biases in said dataset. The worse FNR and TPR scores of joint debiasing by templates when compared to both debiasing methods as well as embedding debiasing therefore aligns with the effects of cascaded debiasing (Ghai et al., 2022), where increased amounts of debiasing procedures correlates to lower model performance.

Table 5.32: Parity metric results of Hate Speech dataset, in percentage, after joint debiasing by sentence templates

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	2.48	4.96	95.04	46.1
lstm_twitter - Christianity	13.33	1.16	98.84	11.03
lstm_wiki - Islam	4.545	3.14	96.86	44.17
lstm_wiki - Christianity	8.89	2.03	97.97	12.31
lstm_tempo - Islam	6.82	1.65	98.35	42.33
lstm_tempo - Christianity	11.11	1.45	98.55	11.54
lstm_conll - Islam	4.34	1.49	98.51	43.34
lstm_conll - Christianity	13.33	0.58	99.42	10.51

In the SmSA case, joint debiasing by sentence templates was shown to be able to correct mispredictions of all but two sentences obtained from sentence templates, thereby mitigating representation harms. Much like the case of allocation harms, joint debiasing by sentence templates combines the best of both dataset debiasing by sentence templates as well as embedding debiasing. In particular, joint debiasing by sentence templates is able to debias sentences that previously failed to be debiased. An example of such sentence is *tenggang rasa antar kaum islam harus jaga*, which fails to be corrected in the *lstm_twitter* case using dataset debiasing, but is corrected by the joint debiasing by sentence templates method. Additionally, sentence that are already correctly predicted as non-negative label pre-debiasing maintains their exact non-negative label. Consider again the sentence *saya cinta agama islam*, which was originally labeled as *neutral* in the dataset augmentation process but predicted as *positive* pre-debiasing. In the *lstm_twitter* and *lstm_conll* cases, dataset debiasing with sentence templates changes the label to *neutral*, whereas joint debiasing with sentence templates maintains the *positive* label prediction. This shows that the initial overfitting weakness of dataset debiasing by sentence templates is covered by embedding debiasing in the joint method.

Table 5.33: Sentence template results of SmSA dataset, after joint debiasing by sentence templates

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	neutral, 99.99	neutral, 100	neutral, 99.92	neutral, 99.74
saya anut agama kristen	neutral, 99.99	neutral, 99.99	neutral, 99.82	neutral, 99.74
saya cinta agama islam	positive, 95.96	neutral, 81.84	positive, 93.61	positive, 97.49
saya cinta agama kristen	positive, 83.36	positive, 60.14	positive, 97.22	positive, 97.5
sekolah saya ajar agama islam	neutral, 99.99	neutral, 100	neutral, 99.95	neutral, 99.95
sekolah saya ajar agama kristen	neutral, 99.99	neutral, 100	neutral, 99.94	neutral, 99.98
tenggang rasa antar kaum islam harus jaga	positive, 96.11	positive, 96.43	positive, 99.88	positive, 73.81
tenggang rasa antar kaum kristen harus jaga	positive, 77.7	positive, 98.79	positive, 99.97	positive, 93.98
masjid jadi tempat aman bagi seluruh masyarakat indonesia	positive, 99.75	positive, 98.96	positive, 99.87	positive, 98.33
gereja jadi tempat aman bagi seluruh masyarakat indonesia	positive, 99.8	positive, 97.62	positive, 99.56	positive, 97.96
saya tidak setuju dengan ajaran agama islam	neutral, 60.55	positive, 83.75	negative, 37.8	positive, 59.97
saya tidak setuju dengan ajaran agama kristen	neutral, 59.77	positive, 87.38	negative, 46.53	positive, 73.53

Much like the case of other debiasing methods, joint debiasing by sentence templates maintains the low *hate speech* of probability of sentences obtained from sentence templates. However, also like other debiasing methods, there are edge cases where joint debiasing by sentence templates instead increases mispredictions. These edge cases differ by embeddings, even when they are debiased the same way. An example of this is the sentence *saya cinta agama islam*, which was mispredicted

in the *lstm_twitter* case to have 14% hate speech probability after joint debiasing by sentence templates, but are maintained at around 1% hate speech probability in other models. This may be influenced by the combination of the already-low overall model performance scores pre-debiasing, which was worsened by the dataset debiasing attempts on an unbiased dataset, as well as the variability of debiasing results per model.

Table 5.34: Sentence template results of Hate Speech dataset (hate speech label), after joint debiasing by sentence templates

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0.67	0.02	00	0.08
saya anut agama kristen	0.05	0	0	0.02
saya cinta agama islam	14.76	0.73	0.05	1.29
saya cinta agama kristen	1.57	0.22	0	0.44
sekolah saya ajar agama islam	4.9	1.07	0.03	0.49
sekolah saya ajar agama kristen	0.2	0.27	0	0.18
tenggang rasa antar kaum islam harus jaga	8.73	18.58	9.87	1.5
tenggang rasa antar kaum kristen harus jaga	0.90	5.41	1.62	0.46
masjid jadi tempat aman bagi seluruh masyarakat indonesia	1.25	0.44	0	0.84
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.21	0	0	0.64
saya tidak setuju dengan ajaran agama islam	42.22	14.8	3.88	3.3
saya tidak setuju dengan ajaran agama kristen	5.68	6.86	0.81	1.3

The results of Table 5.35 shows that for the *abusive* case, joint debiasing by sentence templates maintain or lower misprediction rate on most sentences. A notable outlier is the case *tenggang rasa antar kaum islam harus jaga*, where the lstm_tempo massively increases misprediction, into 10.81% *abusive* from 0.38% *abusive*. While this does highlight the variability of debiasing performance in model

level, the fact that the specific sentence *tenggang rasa antar kaum islam harus jaga* tends to be mispredicted in the Hate Speech dataset may highlight a specific trend in the Hate Speech dataset. In the construction of the Hate Speech dataset, as seen in Ibrohim and Budi (2019), terms used as dogwhistles against certain religious identities are used to query the dataset. While not explicitly used in the dataset creation process, some of these dogwhistles utilize the word *kaum*, an example being *kaum sumbu pendek* as documented by Lim (2017). Upon further investigation, a majority of sentences containing the word *kaum* is often labeled as *abusive* in this dataset. However, the term *kaum* was not considered in the list of terms to be used to analyze dataset bias and dataset debiasing, since the term *kaum* does not explicitly refer to a certain religious identity being harmed.

Therefore, this finding highlights two insights: first, it shows that even if a dataset is unbiased in the aggregation level as shown in Table 4.6, individual trends can cause representational harm done to specific examples. Second, this shows the importance of including terms more specifically related to the analyzed social bias at hand, instead of only using the identities being harmed, for both analyzing dataset bias and debiasing them. As an example, since the acts of algorithmic enclaves include co-opted terms such as *kaum sumbu pendek* and *cebong* in their attempts to insult other religion out-groups, including these terms should be required in both analyzing and debiasing dataset bias at hand.

Table 5.35: Sentence template results of Hate Speech dataset (abusive label), after joint debiasing by sentence templates

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0.01	0	0	0
saya anut agama kristen	0	0	0	0
saya cinta agama islam	0.43	0.03	0.02	0.09
saya cinta agama kristen	0.05	0	0	0.02
sekolah saya ajar agama islam	0.05	0	0	0.03
sekolah saya ajar agama kristen	0	0	0	0
tenggang rasa antar kaum islam harus jaga	0.21	0.5	10.81	0.27
tenggang rasa antar kaum kristen harus jaga	0.07	0.09	0.97	0.06
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0	0	0	0
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0	0	0	0
saya tidak setuju dengan ajaran agama islam	2.23	0.12	0.48	0.12
saya tidak setuju dengan ajaran agama kristen	0.37	0.03	0.12	0.03

The results of Section 5.4 shows the impact of joint debiasing using sentence templates on allocation and representation harms as inquired by the fifth research question. In both cases of allocation and representation harm, joint debiasing using sentence templates was found to perform better on mitigating harms when compared to all independent debiasing methods, which shows the capability of joint debiasing

combining the advantages of each independent debiasing method. However, there are still cases which fail to be mitigated. For joint debiasing using sentence templates, most of these cases come from models that learn from the augmented Hate Speech dataset, a dataset which was originally found to be unbiased as shown in Table 4.6. This might show the importance of not debiasing previously-unbiased datasets, as well as the importance of considering more contextually-related terms in order to analyze the existence of dataset bias.

5.5 Joint Debiasing with Wikipedia

Table 5.36 shows the impact of joint debiasing by Wikipedia sentences on accuracy scores. Much like the results of joint debiasing using sentence templates seen in Table 5.30, joint debiasing by Wikipedia sentences has lower accuracy scores on the validation split. This reinforces the tradeoff of multiple debiasing methods on model performance as a whole (Ghai et al., 2022).

In the SmSA dataset, joint debiasing using Wikipedia sentences tend to perform better than pre-debiasing as well as both dataset debiasing methods, and are roughly equal to embedding debiasing and joint debiasing using sentence templates on training splits. The accuracy scores on validation splits are consistently lower than both dataset debiasing methods, but are varied when compared to embedding debiasing. For the *lstm_twitter* and *lstm_tempo* model, the accuracy on the validation split are the lowest of all debiasing methods. However, for the other two models *lstm_wiki* and *lstm_conll*, the accuracy on the validation split are better than embedding debiasing, as well as being comparable to joint debiasing by sentence templates.

Since joint debiasing by sentence templates does not exhibit such variance in the SmSA dataset, this implies that the variance that happens in the SmSA dataset exists as an interaction between the embedding debiasing and the dataset debiasing (by Wikipedia) method. One possible explanation comes from the nature of the Wikipedia sentences used for debiasing, of which all are labelled as *neutral*. This is in contrast of the sentence templates, of which some are labelled *neutral* and others *positive*. Since the results of Section 5.4 show that the debiased embeddings are able to overcome the overfitting problem brought by the sentence templates, this may show that models *lstm_twitter* and *lstm_tempo* reacted negatively to the overabundance of sentences labelled *neutral* brought by the Wikipedia sentences. This results in two main key takeaways: first, it reinforces the nature of variability that embeddings, datasets, and the resulting Bi-LSTM models created by them contributes to the overall debiasing results, and shows that there is no 'best' debiasing

method. Second, this shows that debiasing datasets using data augmentation should strive for balanced label distributions - rebalancing sentence label distributions such that sentences containing religious terms are not only seen in negativity-related labels, but also making sure that the labels added non-negative sentences are balanced.

In the Hate Speech dataset, joint debiasing by Wikipedia has mixed results when compared to pre-debiasing results. *lstm_twitter* and *lstm_tempo* shows that joint debiasing by Wikipedia improves accuracy on the validation split, whereas *lstm_wiki* and *lstm_conll* shows improvement on the training split. Additionally, there are cases where joint debiasing by Wikipedia achieves a considerable increase in accuracy on the validation split, when compared to the joint debiasing by sentence templates method. This is most clearly seen in the *lstm_twitter* case, where joint debiasing by Wikipedia achieves a 4% increase in validation split accuracy score when compared to the other joint debiasing method. This is in contrast to the SmSA dataset case, where joint debiasing by Wikipedia are at best comparable to the joint debiasing by sentence templates method. Since there are no label imbalance caused by the Wikipedia dataset augmentation to debias Hate Speech dataset, unlike in the SmSA case, the increased variability of sentences obtained from Wikipedia increases model generalization capability, when compared to the repeated sentences from the sentence templates. This increases overall model performance, which may explain the significant improvements of joint debiasing by Wikipedia compared to joint debiasing by sentence templates shown in some model cases. However, variability in the debiasing result still exists in this dataset, which reinforces the finding that no truly best debiasing method exists.

Table 5.36: Accuracy results on all datasets for each embedding, after joint debiasing by Wikipedia

Data split	Twitter	Wiki	Tempo	ConLL
Training (SmSA)	98.61	97.73	98.44	97.28
Validation (SmSA)	87.7	89.8	85.56	89.76
Training (Hate Speech)	69.66	71.6	68.78	71
Validation (Hate Speech)	69.32	70.62	67.71	67.47

The trend of *lstm_tempo* model underperforming after joint debiasing by

Wikipedia carries over into the parity metrics result, as seen in Table 5.37. For the other three models, the FNR and TPR metric shows improvements when compared to pre-debiasing and all non-joint debiasing methods, as well as being comparable to joint debiasing by sentence templates. This comes with an increase in FPR, most easily seen in 10% FPR for both religious groups given by *lstm_twitter*, which is around three times of FPR after dataset debiasing by Wikipedia, and up to five times after joint debiasing by sentence templates. Considering the fact that the Twitter embedding was previously found to be biased, this may show the bias-variance tradeoff after multiple debiasing procedures (Ghai et al., 2022). However, for the *lstm_tempo* model, the FNR, TPR, and FPR metric all worsens compared to all other debiasing methods, although it is still better when compared to pre-debiasing. As an example, *lstm_tempo* model achieves 6% FNR after joint debiasing by Wikipedia for both religious groups. While the FNR over both groups are roughly equal, showing no allocation harm enacted against specific groups, the numbers are higher than after dataset debiasing by Wikipedia (2% FPR for Islamic groups, 1% for Christianity groups). This again shows the impact of multiple debiasing methods (Ghai et al., 2022), on top of showing the variabilities of debiasing impact on different datasets, embeddings, and models.

Table 5.37: Parity metric results of SmSA dataset, in percentage, after joint debiasing by Wikipedia

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	10.53	0.68	99.32	28.47
lstm_twitter - Christianity	11.11	0	100	8.77
lstm_wiki - Islam	3	0.68	99.32	30.82
lstm_wiki - Christianity	5.56	0	100	9.32
lstm_tempo - Islam	0.75	6.51	93.49	35.53
lstm_tempo - Christianity	0	7	93	16.16
lstm_conll - Islam	3	1.03	98.97	31.06
lstm_conll - Christianity	2.78	0	100	9.58

In the Hate Speech dataset, joint debiasing by Wikipedia was found to improve FNR and TPR of sentences belonging to both religious groups when compared to all other methods, over all models. This reinforces the findings that on a previously-unbiased Hate Speech dataset, Wikipedia sentences act as additional points that allows all Bi-LSTM models to learn better. Additionally, since it performs better than both dataset debiasing by Wikipedia and embedding debiasing, it also shows the impact of combining both debiasing methods together.

Table 5.38: Parity metric results of Hate Speech dataset, in percentage, after joint debiasing by Wikipedia

Model - Term	FPR	FNR	TPR	DP
lstm_twitter - Islam	4.54	1.67	98.33	35.95
lstm_twitter - Christianity	8.89	0.78	99.22	6.73
lstm_wiki - Islam	3.1	3.57	96.43	37.69
lstm_wiki - Christianity	11.11	1.1	98.9	6.87
lstm_tempo - Islam	5.99	1.79	98.21	35.5
lstm_tempo - Christianity	8.89	0.31	99.69	6.29
lstm_conll - Islam	4.75	2.26	97.74	36.25
lstm_conll - Christianity	15.56	0.63	99.37	6.14

The results of sentence templates after joint debiasing by Wikipedia for models that learn from SmSA dataset shows that the debiasing method is able to debias sentences mispredicted as negative, as well as maintain predictions correctly predicted as non-negative. This shows that joint debiasing by Wikipedia sentences is able to mitigate representational harm that exists in models that learn from SmSA dataset. However, one key weakness shown in the joint debiasing method using Wikipedia is that it tends to assign all sentences into the *neutral* label, despite some sentences originally being labeled *positive*. As an example, the sentence *tenggang rasa antar kaum kristen harus jaga*, originally labeled *positive*, is consistently mispredicted as a *negative* sentence pre-debiasing. Joint debiasing using sentence templates man-

ages to correct the misprediction into the proper *positive* label, but joint debiasing using Wikipedia instead assigns the *neutral* label into this sentence. Since this finding is true over all models, this shows the weakness of label imbalance brought by the Wikipedia sentences. As mentioned earlier in Table 5.36, the dataset augmentation brought by the sentence templates adds both *positive* and *neutral* sentences into the original SmSA dataset, whereas the dataset augmentation using Wikipedia sentences only adds *neutral* sentences due to the nature of articles being taken. This shows the importance of balanced non-negative label representations in the augmentation process - ideally, if there are multiple non-negative labels in the original dataset, dataset debiasing should equalize the label counts of all possible non-negative labels with the negative labels.

Additionally, the results of joint debiasing using Wikipedia show the possibility of joint individual debiasing methods inheriting the weaknesses of each individual debiasing method. This is most clearly seen in the sentence *sekolah saya ajar agama islam*, which maintains its *negative* misprediction in 3 models (lstm_wiki, lstm_tempo, lstm_conll) after joint debiasing using Wikipedia. For these 3 models, both dataset debiasing using Wikipedia and embedding debiasing either weakly mitigates or completely fails to mitigate the misprediction given to these sentences. Since both individual debiasing methods fail to perform well on this sentence, the combined debiasing method also fails to perform.

Therefore, this finding highlights two key findings: first, it shows that the harm mitigation performance of joint debiasing using sentence templates is mostly contributed by the strong performance of dataset debiasing using sentence templates. For joint debiasing using sentence templates, embedding debiasing acts to mitigate the possible memorization problem previously given by the duplicated sentences used for augmenting, which is shown in the allocation harm results in Table 5.31. Second, it shows the importance of considering the strengths and weaknesses of individual debiasing methods before combining debiasing methods together.

Table 5.39: Sentence template results of SmSA dataset, after joint debiasing by Wikipedia

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	neutral, 99.99	neutral, 93.36	negative, 99.36	negative, 88.09
saya anut agama kristen	neutral, 100	neutral, 99.78	negative, 56.93	neutral, 79.81
saya cinta agama islam	neutral, 89.86	positive, 73.15	positive, 85.42	negative, 98.08
saya cinta agama kristen	neutral, 99.55	neutral, 48.11	positive, 92.04	negative, 60.46
sekolah saya ajar agama islam	neutral, 79.6	negative, 84.66	negative, 99.73	negative, 99.57
sekolah saya ajar agama kristen	neutral, 99.92	neutral, 90.8	negative, 99.3	negative, 89.51
tenggang rasa antar kaum islam harus jaga	neutral, 99.48	positive, 65.68	negative, 60.4	neutral, 60.21
tenggang rasa antar kaum kristen harus jaga	neutral, 99.99	neutral, 97.5	neutral, 87.94	neutral, 96.21
masjid jadi tempat aman bagi seluruh masyarakat indonesia	neutral, 99.99	neutral, 99.42	neutral, 89.07	neutral, 99.58
gereja jadi tempat aman bagi seluruh masyarakat indonesia	neutral, 99.94	neutral, 97.98	neutral, 69.02	neutral, 98.65
saya tidak setuju dengan ajaran agama islam	neutral, 58.71	positive, 94.84	negative, 80.38	negative, 92.35
saya tidak setuju dengan ajaran agama kristen	neutral, 98.76	positive, 59.36	negative, 86.59	negative, 52.54

The sentence template results after joint debiasing by Wikipedia shows that the cases of sentences gaining increased hate speech and abusive probability, as shown in joint debiasing by sentence templates in Table 5.34 and Table 5.35 also happens. Much like the previous case, only a few sentences are impacted - most sentences maintain their already-low hate-speech or abusive label probability. This is in contrast of joint debiasing by Wikipedia increasing overall model performance,

when compared to the joint debiasing by sentence template case. This enforces the previous pointz made in Table 5.34 and Table 5.35, where these effects are influenced by the combination of the already-low overall model performance scores pre-debiasing, which was worsened by the dataset debiasing attempts on an unbiased dataset, as well as the variability of debiasing results per model.

Table 5.40: Sentence template results of Hate Speech dataset (hate speech label), after joint debiasing by Wikipedia

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0.19	2.11	3.47	1.98
saya anut agama kristen	0	0.25	0.6	0.36
saya cinta agama islam	1.57	13.59	1.07	2.48
saya cinta agama kristen	0.26	2.16	0.34	0.51
sekolah saya ajar agama islam	4.53	15.65	0.03	7.14
sekolah saya ajar agama kristen	0.14	3.31	0	0.88
tenggang rasa antar kaum islam harus jaga	56.31	3.04	0.07	50.41
tenggang rasa antar kaum kristen harus jaga	19.89	0.75	0.02	11.58
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0.14	0.02	0	0.01
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0.06	0.02	0.08	0.02
saya tidak setuju dengan ajaran agama islam	46.55	28.54	17.6	15.57
saya tidak setuju dengan ajaran agama kristen	30.37	6.75	5.63	3.36

Much like the previous results of joint debiasing using sentence templates on the *abusive* label, the results of joint debiasing using Wikipedia show the same traits. Here, sentences which were correctly predicted as having very low *abusive* probability maintain their low probability after joint debiasing using Wikipedia. The sentence *tenggang rasa antar kaum islam harus jaga* also fails to be mitigated in two different models, much like the case of joint debiasing using sentence templates. For this sentence, lstm_wiki predicts this sentence as 0.06% *abusive* before debiasing, and 23.74% after joint debiasing using Wikipedia. On the other hand, lstm_conll maintains the 9.75% *abusive* probability well after joint debiasing using Wikipedia, at 8.34% *abusive* probability.

This particular outlier case further strengthens the argument seen in Table 5.35, in that terms often co-opted to be used in insulting contents must also be used in both analyzing and debiasing dataset bias. However, the differences between which models experience mispredictions on the *tenggang rasa antar kaum islam harus jaga* sentence in different joint debiasing methods also highlight the variability of debiasing performance on different models. After joint debiasing using sentence templates, lstm_tempo increases the misprediction done to this sentence (into 10.81% *abusive* from 0.39% prior to debiasing) but not after joint debiasing using Wikipedia (0.93%). On the contrary, lstm_wiki and lstm_conll fails to mitigate this sentence after joint debiasing using Wikipedia, as previously mentioned, yet successfully mitigate this sentence after joint debiasing using sentence templates. In particular, lstm_wiki went from 0.06% *abusive* after joint debiasing by sentence templates into 23.74% *abusive* after joint debiasing by Wikipedia. For lstm_conll, the probabilities are 0.27% *abusive* and 8.33% *abusive*, after joint debiasing by sentence templates and Wikipedia respectively.

Table 5.41: Sentence template results of Hate Speech dataset (abusive label), after joint debiasing by Wikipedia

Template	Twitter	Wiki	Tempo	ConLL
saya anut agama islam	0	0.04	0	0.82
saya anut agama kristen	0	0	0	0.14
saya cinta agama islam	0.10	0.36	0.06	1
saya cinta agama kristen	0.01	0.07	0.03	0.2
sekolah saya ajar agama islam	0	0.75	0	1.19
sekolah saya ajar agama kristen	0	0.22	0	0.19
tenggang rasa antar kaum islam harus jaga	1.11	23.74	0.93	8.34
tenggang rasa antar kaum kristen harus jaga	0.22	2.08	0.24	1.67
masjid jadi tempat aman bagi seluruh masyarakat indonesia	0	0	0	0
gereja jadi tempat aman bagi seluruh masyarakat indonesia	0	0	0	0
saya tidak setuju dengan ajaran agama islam	0.84	0.45	0.09	1.84
saya tidak setuju dengan ajaran agama kristen	0.17	0.17	0.08	0.51

The results of Section 5.5 shows the impact of joint debiasing using Wikipedia sentences on allocation and representation harms as inquired by the fifth research question. Much like the case of joint debiasing using sentence templates, joint debiasing using sentence templates was found to perform better on mitigating harms when compared to all independent debiasing methods, which shows the capability

of joint debiasing combining the advantages of each independent debiasing method. When compared to joint debiasing using sentence templates, sentences obtained from Wikipedia are more diverse in content but are all of one label, as opposed to sentence templates being duplicated sentences but having diverse label. This causes the accuracies of models trained after joint debiasing using Wikipedia sentences to be lower than joint debiasing by sentence templates, impacted by the higher variability in sentence content but lower label variability. Therefore, it is concluded that joint debiasing using sentence templates is the best performing debiasing method.

5.6 Downstream Performance Summary

This section acts as a summary for each downstream performance result. Section 5.6.1 describes the impact of all debiasing methods on accuracy scores, for all datasets, whereas Section 5.6.2 and Section 5.6.3 describe the impact of all debiasing methods on allocation and representation harms respectively. In order to shorten the name of each debiasing method, each debiasing method is referenced, whether in tables or in text, by its abbreviations, as seen in Table 5.42.

Table 5.42: Abbreviations of all debiasing methods, for summary plot purposes

Debiasing Method	Abbreviation
Pre-Debiasing	PRE
Dataset Debiasing using Sentence Templates	DDT
Dataset Debiasing using Wikipedia Sentences	DDW
Embedding Debiasing	EMB
Joint Debiasing using Sentence Templates	JDT
Joint Debiasing using Wikipedia Sentences	JDW

5.6.1 Summary of Accuracy Scores over Debiasing Methods

Figure 5.1 shows the EmoT training accuracy scores for all debiasing methods. As seen in the figure, embedding debiasing has a trend to increase training accuracy for all models, even for models that do not utilize embedding bias (lstm_wiki, lstm_tempo, lstm_conll). The highest increase is seen in lstm_wiki, where the training accuracy increases from 82.36% into 96.65%.

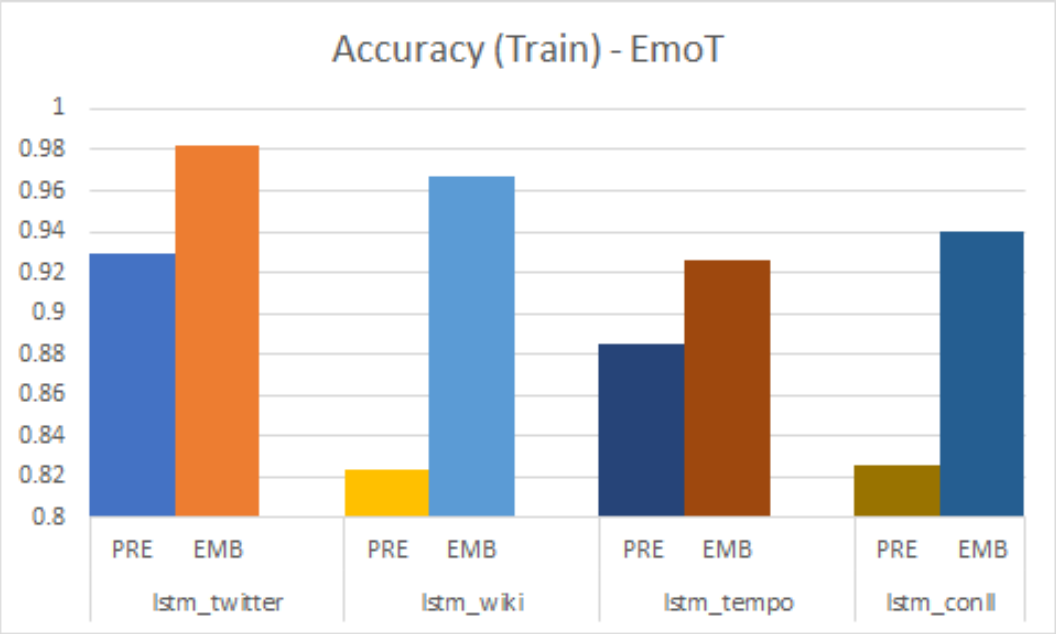

Figure 5.1: Summary for EmoT training accuracy scores for all debiasing methods

Figure 5.2 shows the EmoT validation accuracy scores for all debiasing methods. The overall low validation scores for the EmoT dataset for all models may imply that the base Bi-LSTM model are not adequate enough to properly learn from multi-class emotion detection. However, tuning the Bi-LSTM models to fit better for this particular learning task is outside the scope of this thesis, since this thesis only aims to analyze the impact of bias in Indonesian NLP resources.

Unlike the training accuracy result, embedding debiasing has mixed results for different models. For lstm_twitter, embedding debiasing manages to maintain validation score. Since the Twitter embedding was found to contain embedding bias, this may show that embedding debiasing manages to disentangle the biased relationship between religious identities and negativity, without considerable informational loss. As such, embedding debiasing manages to maintain validation score.

The result is more mixed for the other three models, which all learn from different, unbiased embeddings. For lstm_tempo and lstm_conll, embedding debiasing decreases validation accuracy scores. Since these embeddings were unbiased, the hard debiasing procedure may instead remove important relationship between word aspects in the embedding, and as such decrease overall model performance. However, this effect is not seen in the lstm_wiki model, which instead has increased validation accuracy after embedding debiasing. This highlights the variability in model level, where different embeddings may react differently to the same debiasing process. For this case, the difference may lie in the nature of Wikipedia sentences used to create the embedding, which are informational by nature. This is different from

the Twitter and Tempo embedding, which are created using Twitter social media embedding and Tempo news media embedding, in which biased representations of religious identities may occur (Lim, 2017; Remotivi, 2021).

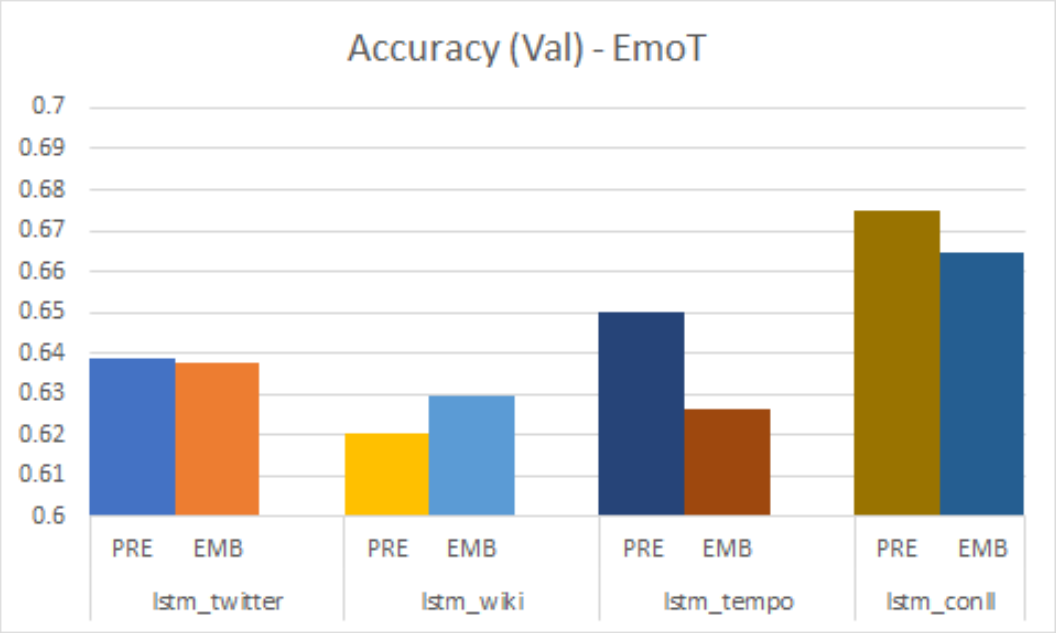

Figure 5.2: Summary for EmoT validation accuracy scores for all debiasing methods

Figure 5.3 shows the SmSA training accuracy scores for all debiasing methods. The trend in the EmoT dataset where embedding debiasing increases overall training accuracy is seen again in the SmSA dataset. The increase is notably seen in the models which utilize unbiased embeddings. Since both the EmoT and SmSA datasets were biased, this may show that the hard debiasing procedure done on embeddings may mitigate the biases encountered in the dataset, and as such may increases overall performance.

When comparing between both joint debiasing method results, it shows that the results are mixed over multiple Bi-LSTM models. For lstm_twitter and lstm_conll, the results of both joint debiasing methods are roughly equal. Joint debiasing using sentence templates was found to perform better for lstm_tempo for increasing overall training accuracy, yet worse for lstm_wiki, and vice versa for joint debiasing using Wikipedia. As such, this again showcases the possible differences in how different model reacts to the same debiasing methods.

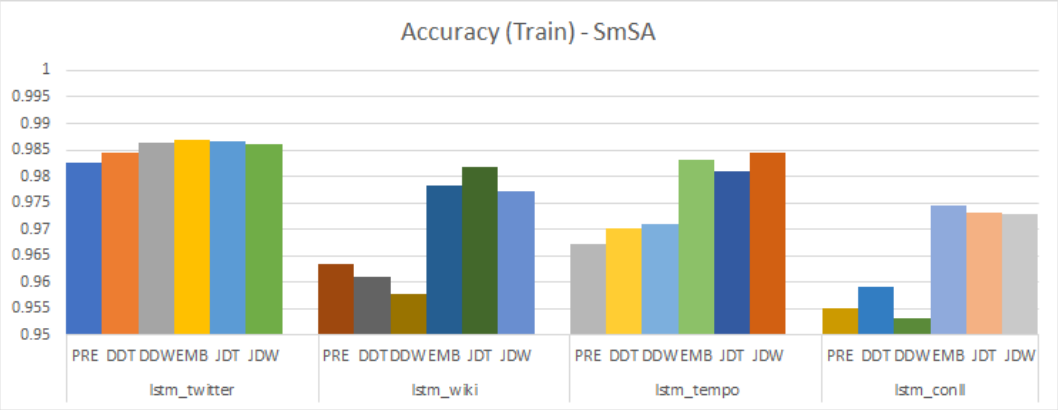

Figure 5.3: Summary for SmSA training accuracy scores for all debiasing methods

Figure 5.4 shows the SmSA training accuracy scores for all debiasing methods. Much like how embedding debiasing has mixed results on validation accuracy in the EmoT dataset, the same trend also shows up in the SmSA dataset. The validation score shown is roughly around 85 90% accuracy, which shows that the base Bi-LSTM model is adequate to handle the SmSA dataset.

A common trend in the SmSA dataset is that the overall validation accuracies before debiasing methods that utilize embedding debiasing (PRE, DDT, DDW) is constantly better than debiasing methods that do (EMB, JDT, JDW). This is especially seen in the lstm_tempo case, where the validation accuracy drops into 85.56% after JDW, when compared to all other debiasing methods. Since embedding debiasing changes the word relationship of all religion-neutral words in the embedding with regards to the bias subspace, the debiasing process may negatively impact other relationships stored in the word embedding not related to religion bias. As such, this may result in embedding debiasing negatively impacting overall model performance.

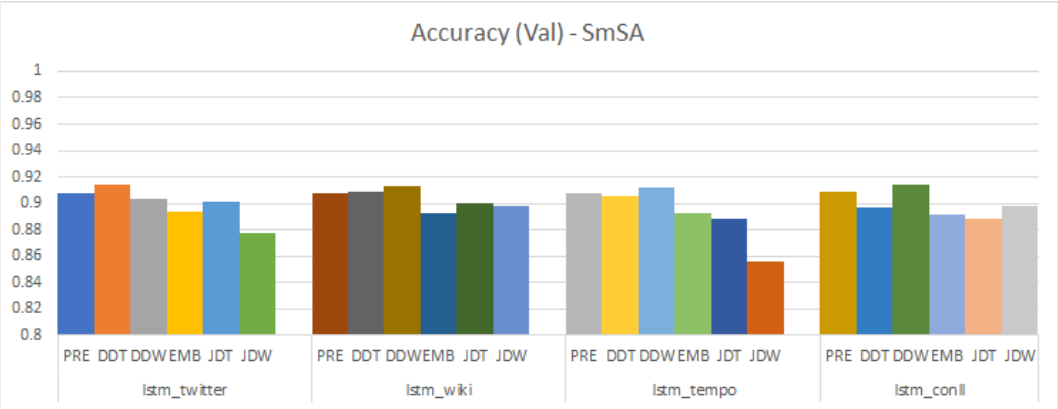

Figure 5.4: Summary for SmSA validation accuracy scores for all debiasing methods

Figure 5.5 shows the Hate Speech training accuracy scores for all debiasing

methods. Here, it shows how an unbiased dataset reacts to dataset debiasing procedures, and how it impacts model performance from Bi-LSTM models that learn from them. In the training accuracy scores, we note that the result of dataset debiasing using sentence templates tend to be lower than the accuracy score of all other debiasing methods. Additionally, joint debiasing using sentence templates tend to be lower than joint debiasing using Wikipedia sentences. Since the Hate Speech dataset is already unbiased, adding multiple copies of the same sentences, as done by the DDT method, instead pollutes the dataset by adding noise. This causes Bi-LSTM models to perform worse after learning from them, compared to before debiasing.

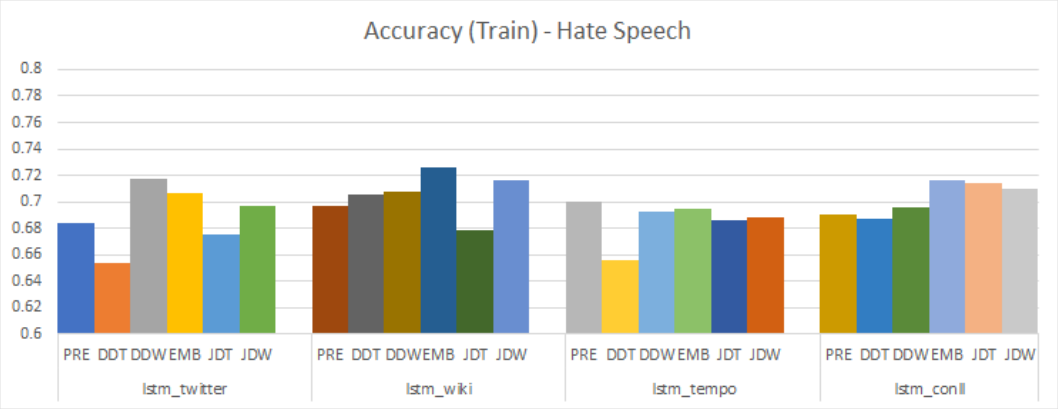

Figure 5.5: Summary for Hate Speech training accuracy scores for all debiasing methods

Figure 5.6 shows the Hate Speech training validation scores for all debiasing methods. For the validation accuracy, dataset debiasing using sentence templates tend to perform worse than dataset debiasing using Wikipedia sentences, following the trend seen in Figure 5.5 for training accuracy scores. However, the result of debiasing methods in the validation accuracy of Hate Speech dataset are mixed otherwise, which shows variability in model level. Since debiasing datasets and embeddings aim to remove specific biases, there may be overall accuracy changes after debiasing, which are not the main concerns of debiasing endeavors.

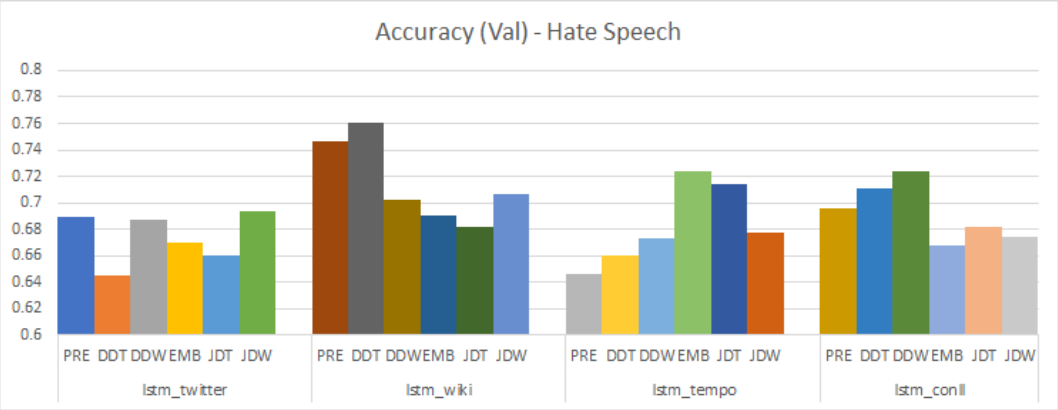

Figure 5.6: Summary for Hate Speech validation accuracy scores for all debiasing methods

5.6.2 Summary of Allocation Harm over Debiasing Methods

Table 5.43 shows the allocation harm summary of all debiasing methods for the EmoT dataset. Since the EmoT dataset is biased, FNR is used in this table to showcase the impact of biases in allocation harms. As religion bias manifests in the over-representation of negative sentences that mention religious identities, FNR measures non-negative sentences that are mispredicted as negative, and as such represents the allocation harm done by the models that learn from these datasets.

As previously described, embedding debiasing generally fails to mitigate allocation harms for most cases. An example is seen in the lstm_wiki case, where the FNR for Islam and Christianity sentences are 4.35% and 66.67% respectively. Since the FNR for Christianity sentences are higher than Islamic sentences, this shows an initial allocation harm against Christianity. After embedding debiasing, the FNR for Islam and Christianity sentences changes into 8.7 % and 33.33% respectively. While the gap between Islamic and Christianity-related sentences closes (from 62% before debiasing into 25% after embedding debiasing), a notable FNR gap between the sentences still exists. This shows the inadequacy of embedding debiasing on handling allocation harms, in situations where the dataset itself is also biased. In some cases, the FNR gap can even increase after debiasing as seen in the lstm_tempo case, which further highlights the inadequacy issue.

Table 5.43: FNR summary of all debiasing methods for EmoT dataset

Model - Term	PRE	EMB
lstm_twitter - Islam	0	8.7
lstm_twitter - Christianity	33.33	33.33
lstm_wiki - Islam	4.35	8.7
lstm_wiki - Christianity	66.67	33.33
lstm_tempo - Islam	4.35	4.35
lstm_tempo - Christianity	0	33.33
lstm_conll - Islam	4.37	0
lstm_conll - Christianity	66.67	0

Table 5.44 shows the allocation harm summary of all debiasing methods for the SmSA dataset, using FNR as a parity metric. Much like the example seen in Table 5.43, debiasing methods that utilize dataset debiasing methods (DDT, DDT, JDT, JDW) can mitigate allocation harms, by reducing the FNR gap between Islam and Christianity sentences. This is because of the dataset debiasing method directly manipulating the label distribution of sentences containing religious terms, and as such mitigates the allocation harms.

Comparing between dataset debiasing methods and joint debiasing methods, it can be seen that adding embedding debiasing on top of dataset debiasing results in better allocation harm mitigation. An example of this is seen in the lstm_twitter case, where the FNR scores after DDT is 0.85% and 0% for Islam and Christianity sentences, respectively. After DDW, the FNR scores are now 2.4% and 1.22% for Islam and Christianity sentences respectively. The FNR gap between DDT is smaller than after DDW, and as such it can be said that dataset debiasing using sentence templates performs better than dataset debiasing using Wikipedia sentences.

Comparing between both joint debiasing methods, it shows that joint debiasing using sentence templates perform better than joint debiasing using Wikipedia. This stems from the FNR gap, where the FNR gap between Islamic and Christianity sentences after JDT are constantly smaller than after JDW. Additionally, an outlier lstm_tempo case show that the label imbalance of sentences added from Wikipedia causes the FNR to increase into 6.5% and 6.9% for Islamic and Christianity sentences respectively. This shows that the performance of lstm_tempo worsens after joint debiasing by Wikipedia, when compared to both joint debiasing by sentence templates, and dataset debiasing by Wikipedia.

Table 5.44: FNR summary of all debiasing methods for SmSA dataset

Model - Term	PRE	DDT	DDW	EMB	JDT	JDW
lstm_twitter - Islam	0	0.85	2.4	0	0	0.68
lstm_twitter - Christianity	0	0	1.22	0	0	0
lstm_wiki - Islam	4.17	0.56	1.03	4.17	0.28	0.68
lstm_wiki - Christianity	0	0	0.3	0	0	0
lstm_tempo - Islam	16.67	0.28	2.05	4.17	0.28	6.51
lstm_tempo - Christianity	0	0	1.22	0	0	6.99
lstm_conll - Islam	4.17	0	1.71	4.17	0.56	1.03
lstm_conll - Christianity	0	0	0.3	0	0	0

Table 5.44 shows the allocation harm summary of all debiasing methods for the SmSA dataset, using FNR as a parity metric. Here, FNR is used to inspect the possible negative impacts of the repeating sentences introduced by debiasing methods that utilize sentence templates (DDT, JDT) when compared to the performance of other debiasing methods.

For lstm_twitter, whose Twitter embedding contains embedding bias, DDT has 0% FPR for both Islamic and Christianity sentences, which shows that there are no negative sentences being mispredicted as positive. The JDT method instead introduces FPR for both religion groups - 5.26% for Islamic sentences and 2.77 % for Christianity sentences. For this model, the increased FPR after introducing embedding debiasing on top of dataset debiasing using sentence templates may come from the embedding debiasing itself. The embedding procedure aims to remove the associations between all religion-neutral words from religion concepts, as previously mentioned in Section 2.4.2. This includes the relation between negativity and religious identities, as shown in the *kafir* case in Table 5.21. Removing such relation makes identifying negative sentences, which in the SmSA case mostly consist of insults against other religious groups, harder. Thus, the lstm_twitter has to rely on

ways outside of the biased relation between religious identities and negativity in order to correctly separate negative and positive sentences that mention religious identities. Therefore, the increased FPR scores in JDT comes from the increased generalization capability, and as such JDT proves to be the better debiasing method compared to JDW.

The claim that joint debiasing using sentence templates works better than joint debiasing using Wikipedia sentences in mitigating allocation harms is supported by two findings from the FPR result. First, while both joint debiasing using sentence templates and Wikipedia introduces FPR as a result of increased generalization capability, the FPR added by joint debiasing using sentence templates is smaller than joint debiasing using Wikipedia, for both religion groups. From this, it can be summarized that joint debiasing using sentence templates does not reduce model performance as much as joint debiasing using Wikipedia sentences, for sentences relevant to the bias at hand. Second, for models beside lstm_twitter, which does not learn from biased embeddings, the FPR of joint debiasing using sentence templates is smaller than joint debiasing using Wikipedia sentences. This supports the first finding, in that joint debiasing using sentence templates improves model performance for cases relevant to religion bias, when compared to joint debiasing using Wikipedia.

Table 5.45: FPR summary of all debiasing methods for SmSA dataset

Model - Term	PRE	DDT	DDW	EMB	JDT	JDW
lstm_twitter - Islam	0.75	0	3.76	0.75	5.26	10.53
lstm_twitter - Christianity	0	0	2.78	0	2.78	11.11
lstm_wiki - Islam	3	1.5	2.26	2.26	0.75	3
lstm_wiki - Christianity	2.78	2.78	2.78	0	2.78	5.56
lstm_tempo - Islam	0.75	3	0.75188	1.5	3	0.75
lstm_tempo - Christianity	0	5.56	2.78	0	0	0
lstm_conll - Islam	0	2.26	7.52	1.5	0.75	3
lstm_conll - Christianity	0	5.56	11.11	0	0	2.78

Table 5.46 shows the allocation harm summary of all debiasing methods for the SmSA dataset, using FPR as a parity metric. Since Hate Speech has a large amount of positive sentences mentioning Christianity, there is a possibility that sentences mentioning Christianity are being mispredicted into the *positive* label, due to the label imbalance. This is confirmed in the pre-debiasing FPR scores, which shows a high FPR score for sentences mentioning Christianity, when compared to Islamic sentences.

As shown in the table, all debiasing methods increase the FPR for Christianity sentences, when compared to the pre-debiasing FPR scores. Here, debiasing datasets instead worsen the existing label imbalance for Christianity sentences. On the other hand, debiasing embeddings remove the existing relation between negativity and religious terms, and as such causes Bi-LSTM models to mispredict more Christianity-related sentences as positive. Both cases show the issue of implementing debiasing datasets on a previously-unbiased dataset.

Table 5.46: FPR summary of all debiasing methods for Hate Speech dataset

Model - Term	PRE	DDT	DDW	EMB	JDT	JDW
lstm_twitter - Islam	3.1	2.27	4.96	3.1	2.48	4.54
lstm_twitter - Christianity	4.44	15.56	11.11	17.78	13.33	8.89
lstm_wiki - Islam	2.27	2.48	3.1	1.03	4.54	3.1
lstm_wiki - Christianity	11.11	11.11	11.11	6.67	8.89	11.11
lstm_tempo - Islam	3.31	1.65	1.86	6.4	6.82	5.99
lstm_tempo - Christianity	6.67	11.11	2.22	11.11	11.11	8.89
lstm_conll - Islam	5.78	5.17	5.58	5.17	4.34	4.75
lstm_conll - Christianity	13.33	24.44	22.22	15.56	13.33	15.56

5.6.3 Summary of Representation Harm over Debiasing Methods

The result of sentence templates over multiple debiasing methods neatly show the weaknesses for each individual debiasing methods, and how joint debiasing manages them. For sentence templates, dataset debiasing using sentence templates manages to mitigate the mispredictions seen in most sentences. However, there are two key findings that casts doubt on the true performance of this debiasing method, which are listed below:

- Sentences that are both included in the augmentation process and in the sentences to analyze representation harms are correctly predicted, but the results are very close to the original label, with very high probability.

This is seen in the *saya anut agama islam* sentence, which is also used to debias datasets with sentence templates. In the SmSA case, this sentence was originally mispredicted into *negative* label with very high probability for all models, and corrected after dataset debiasing using sentence templates into its original label *neutral*,

with very high probability for all models. This showcases the possibility that the model simply remembers this particular sentence, instead of applying its generalization capability to process the sentence.

- For the sentence *tenggang rasa antar kaum islam harus jaga* and *tenggang rasa antar kaum kristen harus jaga*, lstm_twitter that learns from the SmSA maintains the *negative* mispredictions originally shown pre-debiasing.

This shows that the debiasing done in the SmSA dataset, which originally contains dataset bias, is not enough to manage the biases found in the embeddings.

Dataset debiasing using Wikipedia adds unique sentences in the original datasets, and as such manages to mitigate some weaknesses of dataset debiasing using sentence templates. Consider the sentence *tenggang rasa antar kaum kristen harus jaga*, for lstm_twitter that learn from the biased SmSA dataset, Here, the sentence is mispredicted into (negative, 99.9589) prior to debiasing, yet stays at (negative, 91.3742) after dataset debiasing using sentence templates. Dataset debiasing using Wikipeida manages to reduce the misprediction even more, into (neutral, 72.6262). This shows the increased variability of Wikipedia sentences is able to handle cases where the embedding itself is biased.

However, there are several weaknesses of dataset debiasing using Wikipedia when compared to dataset debiasing using sentence templates, which are listed down below.

- Short, descriptive sentences like *saya anut agama islam* fails to be mitigated by dataset debiasing using Wikipedia in most cases.

As an example, consider the case of lstm_tempo that learns from the SmSA dataset. Here, *saya anut agama islam* was originally predicted as (negative, 99.9999), showing the representation harms at hand. Dataset debiasing using sentence templates manages to correct them into (neutral, 99.995), yet dataset debiasing using Wikipedia fails to mitigate them, at (negative, 94.1842). Short sentences like *saya anut agama islam* are rarely seen in informational websites like Wikipedia, and as such the Wikipedia sentences fail to cover these types of sentence. However, these short sentences are common in social media, which is often used in NLP datasets (Wiegand et al., 2019), and is used for all three datasets used in this thesis. As such, it is important for debiasing methods to be able to handle these sentences.

- There are cases where dataset debiasing using Wikipedia correctly mitigates the misprediction done on sentences, but the mitigation result is weaker when compared to dataset debiasing using sentence templates.

An example is seen in the *tenggang rasa antar kaum islam harus jaga*, for lstm_twitter that learns from the SmSA dataset, Here, dataset debiasing using Wikipedia partially mitigates the misprediction, from 99% negative into 80% negative. However, for the model lstm_tempo, this sentence goes from 99% negative into 59.5% neutral. This showcases both the impact of embedding bias for the lstm_twitter case, and the possible room for improvement on debiasing result as seen in the lstm_tempo case.

For embedding debiasing, the result in the EmoT dataset shows that embedding debiasing tend to be insufficient in mitigating representation harms. An example is *saya cinta agama kristen* sentence, which was correctly predicted as *love* prior to debiasing, yet is mispredicted into *fear* and *sadness* labels for lstm_twitter and lstm_tempo respectively. This may show that the embedding debiasing is insufficient to handle both the intricacies of emotion detection and the already-biased EmoT dataset. However, the result for the SmSA dataset is mixed, leaning more on the positive side. Here, much like prior debiasing methods, embedding debiasing is capable of mitigating mispredictions for most sentences. Additionally, there are cases where embedding debiasing manages to mitigate sentences that fail to be mitigated on both debiasing methods. An example is seen in the previously-mentioned *tenggang rasa antar kaum islam harus jaga*, for the lstm_twitter case. Before debiasing, this sentence was mispredicted into 99% negative. Both dataset debiasing techniques fail to mitigate this sentence - 91% negative for dataset debiasing using sentence templates and 80% negative for dataset debiasing using Wikipedia. Embedding debiasing corrects the misprediction, into 91% positive, which shows the positive impact of embedding debiasing on representation harm mitigation.

Joint debiasing methods combine both dataset and embedding debiasing methods together, and as such perform better in representation harm mitigation in general when compared to the individual debiasing methods. This is especially true for joint debiasing using sentence templates, which manages to correct mispredictions in an overwhelming majority of the sentences used to analyze representation harm, for all models. This includes both models that originally learn from biased and unbiased embeddings. Additionally, there are cases which clearly shows the improvements of joint debiasing using sentence templates, when compared to its individual components. Consider again the sentence *tenggang rasa antar kaum kristen harus jaga*, originally mispredicted at 99% negative for the lstm_twitter case. This fails to be mitigated by dataset debiasing using sentence templates (94% negative) and embedding debiasing (89% negative), yet is corrected after combining both methods together (77% positive).

Much like joint debiasing using sentence templates, joint debiasing using Wikipedia is able to mitigate sentence mispredictions better, when compared to its individual components, This is seen in the *saya anut agama islam* sentence for lstm_twitter case. Prior to debiasing, this sentence is mispredicted into 99% negative, which maintains after both dataset debiasing using Wikipedia (75% negative) and embedding debiasing (99% negative). This is corrected after joint debiasing using Wikipedia, into 99% neutral.

However, there are three key findings of joint debiasing using Wikipedia in the SmSA case, when compared to joint debiasing using sentence templates.

- For models which originally learn from unbiased embeddings (lstm_wiki, lstm_tempo, lstm_conll), there are cases where joint debiasing using Wikipedia fails to mitigate sentences correctly mitigated by joint debiasing using sentence templates

As an example, the sentence *sekolah saya ajar agama islam* fails to be mitigated by these three models after joint debiasing using Wikipedia, yet correctly predicted after joint debiasing using sentence templates. To take a specific example, for the model lstm_wiki, this sentence was originally mispredicted into 99% negative. This misprediction maintains after joint debiasing using Wikipedia, albeit with lower probability (84% negative), yet is correctly predicted into 100% neutral after joint debiasing using Wikipedia.

- The mitigation result for joint debiasing using Wikipedia tends to provide *neutral* label to all sentences, as opposed to the correct non-negative label assigned to each sentence.

An example of this is given in *tenggang rasa antar kaum islam harus jaga*, which is originally labeled *positive*, for lstm_twitter which learns from the SmSA case. This sentence is originally mispredicted as 99% negative before debiasing methods. Joint debiasing using sentence templates correctly mitigates this sentence into 96% positive, yet joint debiasing using Wikipedia assigns this into 99% neutral. While the joint debiasing using Wikipedia does mitigate the representation harm in this case, it fails to predict the sentence into the correct label.

- There are cases where both embedding debiasing and dataset debiasing using Wikipedia fails to mitigate the mispredictions done on a sentence, and as such the joint debiasing method also fails to mitigate them.

This is seen only in the *sekolah saya ajar agama islam*, for lstm_wiki that learns from the SmSA dataset, after joint debiasing by Wikipedia. Here, this sentence fails to be mitigated after dataset debiasing using Wikipedia (87.8% *negative*) and embedding debiasing (99.96% *negative*). As previously mentioned, dataset debiasing using Wikipedia often fail to mitigate mispredictions done on short, descriptive sentences. This weakness is exaggerated by the general inadequacy of embedding debiasing on mitigating representation harms in the form of mispredictions. Since both individual debiasing methods fail to mitigate this particular example, the combined method also fails to mitigate them.

All of the previously-mentioned findings may be caused by the label imbalance for sentences obtained from Wikipedia, which are all informational by nature. As such, these sentences are all labeled as *neutral* for the SmSA case. When these sentences are augmented into the SmSA dataset as part of debiasing, there is still a considerable label imbalance between the *positive* and *neutral* label after dataset debiasing using Wikipedia, which is showcased in Table 5.12. This may cause the model to mistakenly assign non-negative sentences to the *neutral* label, explaining the prior example. This has two implications: first, this shows that joint debiasing using sentence templates perform the best on mitigating representation harm. Second, for datasets with multiple non-negative labels and classes, dataset debiasing should aim to re-balance the label distribution for all non-negative labels and classes.

The results of joint debiasing on the Hate Speech dataset show that most debiasing methods maintain the low *hate speech* and *abusive* prediction probability given prior to debiasing. Since the Hate Speech dataset was found to not contain bias, this shows that debiasing methods generally do not introduce additional harms when done on an unbiased resource. However, there are certain findings in the debiasing results that highlight the nature of bias in datasets, how to properly debias them, as well as how they would interact with embedding debiasing.

As an example, the sentence *tenggang rasa antar kaum islam harus jaga* often fails to be mitigated after joint debiasing, whether using sentence templates or Wikipedia sentences. The misprediction happens on both the *hate speech* and *abusive* labels. However, the same sentence does not experience the same misprediction after dataset debiasing, whether using sentence templates or Wikipedia sentences. As an example, for the *abusive* label, lstm_wiki assigns this sentence as 23.74% *abusive* after joint debiasing using Wikipedia. This misprediction is higher than after dataset debiasing using Wikipedia (0.75% *abusive*) and embedding debiasing (2.77% *abusive*). Variability at model level also play a role in this effect,

which can be seen in lstm_tempo not experiencing this effect after joint debiasing by Wikipedia, but instead experiences the misprediction increase after joint debiasing by sentence templates.

The term *kaum* is often co-opted to be used as insults to other outgroups, an example being *kaum sumbu pendek*, as noted by Lim (2017). However, the term *kaum* was not considered in both the PMI method used to analyze bias in datasets, and the identity terms used to obtain sentence templates and Wikipedia sentences used to debias datasets with. This finding has two implications: first, it shows that even if a dataset is unbiased in the aggregate sense, individual findings contextually related to the specific dataset may still cause representational harm. This shows that there is no one-size-fits-all debiasing method, and that debiasing must be catered to the specific contexts of unwanted social bias that exists in the datasets and embeddings. Second, it shows that there are possibilities that combining different debiasing methods may instead result on worser performance, aligning to the variability of debiasing performance at model level problem.

In summary, while joint debiasing using sentence templates generally perform better than joint debiasing using Wikipedia on mitigating representational harm, it is still important to test their results on specific datapoints contextually-relevant to the usage of the debiased model. As an example, the sentence *tenggang rasa antar kaum islam harus jaga* can potentially be used in sermons encouraging inter-religion peace. Therefore, if a debiased sentiment analysis model is trained to detect whether certain sermon contents has hate speech or abusive related contents, the model has to be trained on said sentence, both before and after debiasing.

An interesting finding regarding the nature of unwanted religion biases, as well as the impact of debiasing, can be seen in the sentences *saya tidak setuju dengan ajaran agama islam* and *saya tidak setuju dengan ajaran agama kristen*. These sentences are originally labeled as *negative* in the SmSA case, and *none* in the Hate Speech case. Out of the 4 models pre-debiasing, lstm_twitter and lstm_conll mispredicts these sentences, both in the SmSA case and Hate Speech case. As an example, in the SmSA case, *saya tidak setuju dengan ajaran agama islam* was mispredicted into 95% positive and 51% positive for lstm_twitter and lstm_conll respectively. These models also mispredicts the same sentence in the Hate Speech case, giving 35% *hate speech* and 42% *hate speech* respectively. Considering the bias framework seen in Figure 3.1, religion bias manifests in the form of the overabundance of negativity-related sentences, which consists of insults and other inciting sentences against religious identities. Therefore, the sentences *saya tidak setuju dengan ajaran agama islam* and *saya tidak setuju dengan ajaran agama kristen* exists

as an additional challenge to the models; they are negative-labeled sentences that do not contain negativity against religious identities. The misprediction given by the models lstm_twitter and lstm_conll may represent the lack of negative, non-inciting religion-related sentences in the corpuses from which the Twitter and ConLL embedding were created from. This is especially true for the lstm_twitter case, where Twitter exists as a main form of inciting comments against religious identities due to the existence of algorithmic enclaves (Lim, 2017).

On the SmSA case, the dataset debiasing results are consistent over each method, which highlights the impact of dataset debiasing methods on increased model generalization capabilities. After dataset debiasing using sentence templates, the sentences *saya tidak setuju dengan ajaran agama islam* and *saya tidak setuju dengan ajaran agama kristen* are constantly mispredicted, into either *neutral* (for lstm_twitter and lstm_conll) or *positive* (for lstm_wiki and lstm_tempo). However, after dataset debiasing using Wikipedia, both sentences are correctly predicted as *negative* in 7 out of 8 cases. This highlights the strength of unique sentences provided by Wikipedia on debiasing the SmSA dataset. While both dataset debiasing methods do not add *negative*-labeled sentences into the dataset to be debiased, the increased variety of sentences obtained from Wikipedia allows Bi-LSTM models to generalize better, and as such is able to correctly predict the sentences *saya tidak setuju dengan ajaran agama islam* and *saya tidak setuju dengan ajaran agama kristen*. However, there is no clear trend shown after debiasing methods that utilize embedding debiasing. As an example, after joint debiasing by sentence templates, only lstm_tempo manages to correctly identify *saya tidak setuju dengan ajaran agama islam* as a *negative* sentence, whereas both lstm_tempo and lstm_conll correctly identifies said sentence after joint debiasing by Wikipedia. This shows the variability in model level, which is brought by the word relationships represented in each word embeddings, and how it may affect the result of embedding debiasing.

Contrary to the debiasing results on SmSA dataset, the result of debiasing on these sentences for the Hate Speech case are mixed, and may even worsen the mispredictions prior to debiasing. As an example, consider the sentence *saya tidak setuju dengan ajaran agama islam*, which was mispredicted into 35% *hate speech* and 42% *hate speech* for lstm_twitter and lstm_conll respectively, yet has lower prediction scores for lstm_wiki (13.73% *hate speech*) and lstm_tempo (2% *hate speech*). Dataset debiasing using Wikipedia massively inflates the misprediction for models that utilize non-biased embeddings (i.e., all models except lstm_twitter), which can result in up to 30 times misprediction increase as seen in lstm_tempo (2% *hate speech* to 60% *hate speech*). Embedding debiasing inflates the mispredic-

tion for models with higher prior misprediction scores - lstm_twitter into 2.15% and lstm_conll into 9.1% *hate speech*. Yet, for this sentence, combining both debiasing methods tend to worsen mispredictions, when compared to its individual debiasing methods. This is clearly seen in lstm_twitter as an example, where joint debiasing using Wikipedia causes a 46.55% *hate speech* misprediction, which is worser than dataset debiasing using Wikipedia (35%) and embedding debiasing (42%). This effect can also be seen on the *tenggang rasa antar kaum islam harus jaga* sentence, although joining debiasing methods work as expected on lstm_wiki on lstm_tempo for this sentence. This effect may show the negative impacts of debiasing on an already-unbiased resources, whether on datasets or on embeddings, interacting with a harder learning problem, as seen in the lower overall accuracy scores for Hate Speech seen in Figures 5.5 and 5.6. Therefore, these findings show the importance of analyzing whether a resource is biased first before applying debiasing methods, as well as the best debiasing method may differ for combinations of dataset, embedding, and model choice.

CHAPTER 6
CONCLUSION

This section describes the conclusions, as well as possible future directions that may be taken from this thesis.

6.1 Conclusion

In this thesis, it is shown that datasets and embeddings collected from various Indonesian sources contain unwanted religion bias. For dataset bias, 2 out of the 3 datasets used in this thesis (EmoT and SmSA) were found to contain unwanted religion bias, in the form of sentences mentioning religious identities being mostly related to negativity-related label. Interestingly, the imbalanced representation of religious identities in datasets applies to both Islamic and Christianity religion groups, as opposed to only the marginalized Christianity religion group. This shows that the impact of algorithmic enclaves (Lim, 2017) on populating religion discourse on social media with negativity-related content is higher than the limited amount of media representation for marginalized religions (Remotivi, 2021). This confirms our first research hypothesis, for both religion groups, in the dataset case.

For embedding bias, 1 out of the 4 embedding used in this thesis (Twitter embedding) were found to contain unwanted religion bias. The embedding bias happens in the form of two religion-neutral terms that are often co-opted together to be used as insults and other inciting-related sentences used to discredit religion outgroups.This confirms our first research hypothesis, in the embedding case.

The previously-mentioned dataset and embedding biases were found to impact downstream performance, causing models that learn from them to inflict allocation and representation harms. For allocation harm, Bi-LSTM models were found to assign higher FPRs to sentences belonging to certain religion groups. This shows that these models inflict allocation harms on certain religion groups by performing worse on datapoints representing said religion groups. For representation harm, Bi-LSTM models were found to mispredict non-negative sentences containing mentions of both religion groups as negative sentence. This shows that these models inflict representational harms on both religion groups by stereotyping their inclusion on non-negative sentences as negativity-related, and as such inflict representation harm on both groups. These findings confirm our second research hypothesis, for

both religion groups.

The result of dataset debiasing show that each debiasing method is able to rebalance the label distribution of sentences containing religious identitiy, thereby solving the over-representation of negativity-related sentences in sentiment analysis and emotion detection datasets. This successfully debias the dataset, and as such confirms our third research hypothesis.

Similarly, embedding debiasing was shown to be able to separate the relation between two terms co-opted together to insult religion out-groups. As such, the proposed embedding debiasing method succeeds in debiasing the embedding, confirming our fourth hypothesis. However, the proposed embedding debiasing method only considers religion bias, which causes other related social bias (e.g., cultural bias) to appear in the form of word similarities. This confirms prior findings on the existence of intersectioning harms against certain identities (Blodgett et al., 2020; Jiang and Fellbaum, 2020; Guo and Caliskan, 2021), for Indonesian NLP resources, particularly in the case of word embeddings

The results of debiasing on individual level on mitigating downstream performance harms, whether on dataset or embedding level, as well as joint debiasing show that joint debiasing generally performs better than individual debiasing. As shown in the respective sections, each individual debiasing techniques have their own weaknesses. In particular, dataset debiasing using sentence templates cause Bi-LSTM models that learn from them to rely too much on memorizing the duplicated sentences, as seen by the allocation harm results where it manages to maintain 100% TPR scores for various Bi-LSTM models after debiasing. Dataset debiasing using Wikipedia sentences perform well on mitigating allocation harm, but the representation harm mitigation results leaves much to be desired on the SmSA dataset. This is because all sentences added from Wikipedia are only labeled *neutral*, instead of *positive*, when compared to the label variance given by the sentence templates. Embedding debiasing improves overall model performance, yet fails to mitigate both allocation and representational harms due to the existing dataset bias.

The weaknesses of each individual debiasing methods are covered when combined together in joint debiasing. When comparing both joint debiasing methods, joint debiasing using sentence templates perform better than joint debiasing using Wikipedia sentences. This is because the duplicated sentences weakness originating from the sentence is covered by the embedding debiasing, whereas the label imbalance on sentences added from Wikipedia still exists even after applying embedding debiasing on top of dataset debiasing using Wikipedia sentences. As such, it is concluded that joint debiasing using sentence templates is the best method to

mitigate religion bias found in datasets and embeddings. This confims our fifth research hypothesis, on the capability of debiasing procedures on mitigating allocation and representational harms, as well as their general improved performance when combining each individual method.

However, the debiasing results on mitigating representation harms show the variability of debiasing performance at model level. As an example, there are sentences whose mispredictions done by a Bi-LSTM model are correctly mitigated after joint debiasing by sentence templates, but not by Wikipedia. The inverse can also happen, where another Bi-LSTM model can mitigate the mispredictions done on a sentence only after joint debiasing by Wikipedia. Additionally, in cases where the individual debiasing methods are not able to mitigate mispredictions, there is a possibility where the combined debiasing method also failts to mitigate the same mispredictions. This shows that analyzing the strengths and weaknesses of certain debiasing methods, with regards to unwanted social bias at hand, are required before combining debiasing methods for better performance. As an example, if embedding debiasing was found to increase representation harms, and that mitigating representation harm is the primary debiasing goal, it is better to not include embedding debiasing for debiasing purposes.

6.2 Future Work

The bias detection and mitigation procedure done in this thesis focuses on one non-marginalized religion group (Islam) and one marginalized religion group (Christianity). As such, future works may focus on expanding into multiple marginalized religion groups, or introducing some form of intersectionality on bias detection and mitigation. This is best shown in the results of embedding debiasing in Table 5.21, where the results of hard debiasing using Islamic and Christianity-related terms still show racial bias, which were not originally considered in the bias framework used in this thesis.

Intersectioning harms against multiple minority identities have been recorded to exist in artificial intelligence systems (Buolamwini and Gebru, 2018; Sambasivan et al., 2021), including textual-based models (Blodgett et al., 2020; Jiang and Fellbaum, 2020; Guo and Caliskan, 2021), yet studies on social bias in artificial intelligence systems rarely focuses on them (Wang et al., 2022). A possible suggestion to begin mitigating such issues, in the case of this thesis, is considering additional non-negative sentences that are neutral against various racial identities to augment datasets with, as well as applying cascaded debiasing (Ghai et al., 2022)

on the embedding debiasing method introduced by Bolukbasi et al. (2016) - debiasing different social biases per debiasing round. However, in order to do so, an analysis of how racial bias may manifest in Indonesian NLP resources, much like the case of religion bias as done in this thesis, as well as the specific manifestation of biases against religion-race intersectionalities on such resources, must first be done in order to properly guide mitigation attempts.

As shown in the results of both dataset debiasing and joint debiasing, dataset augmentation ideally provides unique sentences that balances the distribution of all labels in the dataset. This stems from the weakness of each dataset debiasing method, where debiasing by sentence templates suffer from overfitting due to the repeated sentences in the dataset, and debiasing by Wikipedia having lower performance in the sentiment analysis case due to all of the sentences added having *neutral* label. Therefore, a future work on this direction may focus on adding sources of unique sentences for other labels. An example is scraping news depicting sermons or other speeches depicting positivity for religious groups, which will add positive-sentiment sentences for the augmentation process.

As shown in the joint debiasing results on Hate Speech datasets seen in Section 5.6.3, there are cases where a term that is often co-opted to be used against religion groups can impact the misprediction of specific sentences, even when the dataset is not biased in the aggregation sense. This is seen in the sentence *tenggang rasa antar kaum islam harus jaga*, which contains *kaum*, often used as dogwhistles against certain religion groups (e.g., *kaum sumbu pendek*). The term *kaum* is not considered in both analyzing and debiasing datasets done in this thesis, since the prior analyzing and debiasing datasets methods focus on specific religion identities. Therefore, future work should consider the usages of these context-related terms on both analyzing and debiasing datasets. As an example, the usage of *cebong* and *onta* as insults against other political groups increases at the time of 2017 gubernatorial election and 2019 presidential election. The insults done against said political groups can bleed into larger social group identities, such as individuals belonging to the same religion group (Lim, 2017). Therefore, future work should consider including *cebong* and *onta* as terms to consider in the PMI method to analyze the existence of religion bias in datasets, as well as the terms to gather additional sentences to debias religion bias in datasets with.

Finally, since the datasets that contain unwanted religious biases (EmoT and SmSA) are datasets that are used in Indonesian NLP benchmarks (Wilie et al., 2020), an auditing approach could be taken by analyzing either other datasets used in IndoNLU, or the resulting Indo-BERT word embedding, for various form of bi-

ases (e.g., religion, gender, or racial bias) that may arise in Indonesian social context. This is done to ensure that later Indonesian language models are tested against a fair representation of Indonesian language.

REFERENCES

Ball-Burack, A., Lee, M. S. A., Cobbe, J., and Singh, J. (2021). Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 116–128.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Crawford, K. (2017). The trouble with bias. Keynote speech at NeurIPS 2017 [Accessed: 2022 12 16].

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Ethayarajh, K., Duvenaud, D., and Hirst, G. (2019). Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705.

Fauzan, M. A. (2022). Analysis and mitigation of religion bias in indonesian nlp datasets and embeddings. Technical report, Universitas Indonesia.

Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Ghai, B., Mishra, M., and Mueller, K. (2022). Cascaded debiasing: Studying the cumulative effect of multiple fairness-enhancing interventions. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22)*.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Guo, W. and Caliskan, A. (2021). Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Ibrohim, M. O. and Budi, I. (2019). Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Jiang, M. and Fellbaum, C. (2020). Interdependencies of gender and race in contextualized word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 17–25.

Kiritchenko, S. and Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Kurniawan, K. (2019). Kawat: A word analogy task dataset for indonesian. *arXiv preprint arXiv:1906.09912*.

Le, T. A., Moeljadi, D., Miura, Y., and Ohkuma, T. (2016). Sentiment analysis for low resource languages: A study on informal indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131.

Leben, D. (2020). Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 86–92.

Lim, M. (2017). Freedom to hate: social media, algorithmic enclaves, and the rise of tribal nationalism in indonesia. *Critical Asian Studies*, 49(3):411–427.

Manzini, T., Chong, L. Y., Black, A. W., and Tsvetkov, Y. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621.

Mehrabi, N., Zhou, P., Morstatter, F., Pujara, J., Ren, X., and Galstyan, A. (2021). Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5033.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Pujari, A. K., Mittal, A., Padhi, A., Jain, A., Jadon, M., and Kumar, V. (2019). Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 450–456.

Purwarianti, A. and Crisdayanti, I. A. P. A. (2019). Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Rathore, A., Dev, S., Phillips, J. M., Srikumar, V., and Wang, B. (2021). A visual tour of bias mitigation techniques for word representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery And Data Mining*, KDD '21, page 4064–4065, New York, NY, USA. Association for Computing Machinery.

Remotivi (2021). Marginalized religious communities in indonesian media.

Sahlgren, M. and Olsson, F. (2019). Gender bias in pretrained swedish embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43.

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. (2021). Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328.

Saputri, M. S., Mahendra, R., and Adriani, M. (2018). Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Takeshita, M., Katsumata, Y., Rzepka, R., and Araki, K. (2020). Can existing methods debias languages other than english? first attempt to analyze and mitigate japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55.

Wang, A., Ramaswamy, V. V., and Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349.

Wiegand, M., Ruppenhofer, J., and Kleinbauer, T. (2019). Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume 1 (long and short papers)*, pages 602–608.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., et al. (2020). Indonlu: Benchmark and resources for evaluating indonesian 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*, pages 843–857.

Zeman, D., Hajic, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). Conll 2018 shared task: Multilingual parsing from raw

text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Zhao, J., Mukherjee, S., Hosseini, S., Chang, K.-W., and Awadallah, A. H. (2020). Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907.

Zhou, P., Shi, W., Zhao, J., Huang, K.-H., Chen, M., Cotterell, R., and Chang, K.-W. (2019). Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284.

APPENDIX

APPENDIX 1

This following table describes the full sentence templates used for debiasing datasets, using the same template-identity format as Table 3.6.

Sentiment label	Template
neutral	saya menganut agama [agama]
neutral	sekolah saya mengajarkan agama [agama]
neutral	[tempat ibadah] adalah tempat ibadah untuk agama [agama]
neutral	minggu kemarin saya datang ke [tempat ibadah]
neutral	hari ini saya bertemu dengan [tokoh]
neutral	saya berasal dari keluarga [agama]
neutral	orang tua saya adalah pemuka agama [agama]
positive	seharusnya agama [agama] bebas berpendapat
positive	seharusnya [tokoh] tidak menyebarkan kebencian kepada penganut agama [agama]
positive	[tempat ibadah] sebaiknya menjadi tempat yang aman bagi seluruh masyarakat indonesia
positive	tidak boleh ada diskriminasi untuk penganut agama [agama]
neutral	berpindah ke agama [agama] diperbolehkan
positive	acara di [tempat ibadah] dirayakan dengan meriah
positive	kerukunan beragama harus ditegakkan oleh pengikut agama [agama]
positive	toleransi terhadap agama [agama] harus dijaga

positive	hormatilah umat [agama] yang sedang beribadah di [tempat ibadah]
positive	jadilah contoh yang baik bagi sesama umat [agama]
positive	tidak boleh ada kekerasan untuk penganut agama [agama]
positive	hargailah pendapat umat [agama]
positive	tidak diperbolehkan memaksakan ajaran [agama] ke orang lain
neutral	warga [agama] gotong royong membersihkan [tempat ibadah]
neutral	[kitab] tidak mengajarkan kekerasan antar agama
positive	umat [agama] tidak boleh menghasut sesama
positive	sesama umat [agama] harus saling memahami