The Indonesian Religiolect Corpus: Data Curation for Muslim, Protestant, and Catholic Language Varieties

Dan Sachs
Universitas Gadjah Mada
AMINEF - Fulbright U.S. Student Program
ds1731@georgetown.edu

Abstract

Religiolects—language varieties shaped by religious community identity—are low-resource domains often overlooked within high-resource languages. We present the Indo-Religiolect Corpus,[1] the first large-scale dataset for Indonesian religious language variation, containing 3 million sentences from over 100 institutional websites representing Muslim, Catholic, and Protestant communities. Fine-tuning IndoBERT demonstrates these religiolects are computationally distinguishable: Islamic Indonesian exhibits high distinctiveness (91.73%), while Catholic and Protestant varieties share substantial lexical overlap yet retain detectable shibboleths (86.41% and 86.64%). Our findings indicate a potential for representation collapse: models trained on majority-normative data may default to secular or Muslim-dominant Indonesian, blurring distinct minority voices. We hypothesize that these gaps plausibly translate into downstream fairness risks for applications like content moderation and automated hiring. This corpus offers a template for documenting sub-national varieties, advancing linguistic equity beyond "National Language" benchmarks toward "No Language Variety Left Behind."

1 Introduction: The "Variety Gap" in NLP

The Old Javanese phrase *Bhinneka Tunggal Ika*—"Unity in Diversity"—encapsulates Indonesia's ethos of pluralism. Language is more than a mere tool for communication; it is a powerful lens through which identity, ideology, and belonging are both expressed and constructed. Nowhere is this more evident than in the spaces where religious communities express their faith. In Indonesia, a country famously defined by its vast linguistic and religious diversity, the intersection of language and religion plays a crucial role in navigating both unity and division.

However, the field of Natural Language Processing (NLP) has yet to fully grapple with this reality. While initiatives like Meta's "No Language Left Behind," project (NLLB Team et al., 2022) have made historic strides in covering thousands of distinct languages, they often treat these languages as monolithic entities. This approach is insufficient; the field must advance toward "No Language *Variety* Left Behind," lest we continue to overlook sociolinguistic variation—such as religiolects—nested within larger languages.

While Indonesian itself is generally well-resourced (with models like IndoBERT and large web corpora), the religiolects we study function as low-resource domains: they lack dedicated training data, are absent from standard benchmarks, and remain invisible to current NLP systems trained on generic Indonesian corpora.

Currently, standard Large Language Models (LLMs) flatten linguistic diversity. They treat "Indonesian" as a monolith, ignoring the distinct sociolects specific to religious communities—what Hary and Wein (2013) term "religiolects"—used by millions of speakers. This leads to *representation collapse*: When models are trained on generic web scrapes like CommonCrawl, they default to the majority discourse, effectively erasing minority variations. This bias is not accidental but historical; Fogg (2015) discusses the process of standardizing the Indonesian language, which involved uniting thousands of islands and hundreds of people groups to construct a uniting national identity. The resulting performance gap is systematic: Blasi et al. (2022) show that language technology consistently underperforms for communities with lower digital footprints, a pattern that can extend to smaller language varieties within Indonesian digital spaces.

These religiolects function as *low-resource domains* nested within high-resource languages. The

[1]Data and code are available at https://github.com/dansachs/indo-religiolects

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 426–435

challenges identified in recent policy analysis regarding low-resource contexts—specifically the scarcity of high-quality, representative data (Pava et al., 2024)—apply acutely here. Just as global languages face a "quality gap," so too do these internal varieties.

This paper introduces the Indo-Religiolect Corpus, a curated dataset of ∼3 million sentences designed to capture the distinct linguistic fingerprints of Muslim, Catholic, and Protestant communities. By examining these varieties, this research holds broader implications for sociolinguistics, religious studies, and minority-majority relations, contributing new insights into the role of language in multilingual, religiously plural settings.

2 Related Work

2.1 Religiolects and Sociolinguistics

Sociolinguistics has recognized that religious community membership can motivate stable, socially meaningful language varieties. Hary and Wein (2013) use the term *religiolect* for Jewish-, Christian-, and Muslim-defined varieties that emerge through shared histories, textual traditions, and practices of boundary maintenance. This perspective aligns with classic work on functional distribution and register stratification (e.g., Ferguson (1959)), as well as scholarship on historically religiously-marked contact varieties (e.g., Judeo-Arabic), where linguistic forms become indexical of religious identity rather than merely of topic or setting. Despite this rich descriptive tradition, these religiolects have rarely been operationalized as explicit targets for computational modeling, leaving a gap between sociolinguistic theory and NLP practice.

2.2 Intra-Language Variation and Dialect Identification in NLP

Within NLP, language variation has most often been studied through dialect and variety identification tasks, primarily grounded in geography. The VarDial workshop series catalyzed shared tasks and benchmark datasets for discriminating closely related languages and regional varieties, including Arabic Dialect Identification and German Dialect Identification (Zampieri et al., 2017, 2018). Methodologically, this line of work typically frames "variety" as a supervised classification problem over web or social media text, using lexical and stylistic cues to separate regional standards and dialects (Dunn, 2019). In contrast, demographic and communal varieties (such as those indexed by religious identity) are often less overt in surface form, less consistently tagged, and therefore less represented in existing benchmarks—yet they may be consequential in downstream settings where group membership is socially salient.

2.3 Indonesian NLP and Low-Resource Domains within a High-Resource Language

Indonesian NLP has benefited from rapid recent progress in pretrained models and evaluation resources, including IndoBERT and the IndoLEM benchmark (Koto et al., 2020). In parallel, data initiatives such as NusaCrowd (Cahyawijaya et al., 2023) and NusaX (Winata et al., 2023) have substantially improved coverage for Indonesia's regional languages (e.g., Javanese and Sundanese), complementing multilingual efforts such as NLLB Team et al. (2022). However, these efforts largely operationalize linguistic diversity as either national or regional language coverage. The Indo-Religiolect Corpus targets a different axis: sub-national sociolectal variation within Indonesian itself. By treating Muslim-, Catholic-, and Protestant-associated Indonesian as low-resource domains nested inside a high-resource language, this work highlights a form of underdocumentation that can contribute to representation collapse, where models trained on majority-normative web text default to secular or Muslim-dominant language and blur minority religious voice.

3 Theoretical Framework: Religiolects as Low-Resource Domains

3.1 Defining the Religiolect

This work adopts the framework of Hary and Wein (2013), who define religiolects not merely as a lexicon of religious terms, but as distinct language varieties shaped by the history, texts, and identity markers of a religious community. Crucially, religiolects extend beyond religious contexts: speakers continue using the linguistic patterns of their religious community in secular settings, public-facing websites, and signage. This persistence across contexts distinguishes religiolects from domain-specific jargon used only within worship spaces.

3.2 Historical Foundations

The existence of religious language varieties in Indonesia has deep historical roots. Classical Malay, a precursor to modern Indonesian and other contemporary Malay varieties, was deeply intertwined with Islamic identity, often functioning as a 'Muslim language' across the archipelago (Bausani, 1975). This historical pattern suggests that even minor linguistic differences can carry significant identity markers. The language used by religious communities—internally, in mixed company, and in written texts—serves to distinguish between minority and majority groups. For individual Christians, using Christian-associated linguistic features can signal religious identity without explicit declaration.

The Islamic portion of the corpus exhibits stylistic features traceable to *Kitab Malay*, where literal translation strategies from Arabic often reshape the syntax of Indonesian text (Riddell, 2002). This syntactic influence represents one dimension along which religiolects can diverge beyond simple lexical substitution.

3.3 Linguistic Markers and Shibboleths

While some shibboleths are well-documented, others remain understudied. Christians in Indonesia (and elsewhere) use the term *Allah* in daily life, liturgy, and scripture, but two notable differences distinguish Christian from Muslim usage:

Pronunciation: Muslims, depending on organizational affiliation, may pronounce the word /oloh/ or /awloh/, while Christians typically pronounce it /alah/. These pronunciations serve as auditory shibboleths, allowing listeners to infer religious affiliation.

Theological Semantics: A semantic shift has occurred for *Allah* and its partner word *Tuhan*. Muslims say "*Allah* is our *Tuhan*" (Allah is our God), while Christians say "*Tuhan* is our *Allah*" (God is our Allah). For Muslims, *Allah* functions as a personal name while *Tuhan* serves as a general term for deity. For Christians, the relationship inverts: *Allah* is the general term while *Tuhan* can specify their own God (Singgih, 2003). This subtle shift exemplifies natural language change driven by community-specific theological frameworks.

Throughout the Indonesian archipelago, words of Arabic and Persian origin continue to enter the language. Differentiation can occur at the morphological level: Campbell (1996, p. 41) highlights how retention or modification of Arabic suffixes (such as *-at* versus *-ah*) in loanwords serves as a subtle but consistent shibboleth distinguishing Islamic usage from general Indonesian. While Arabic influence dominates, the linguistic landscape remains complex. Van Dam (2010) documents the persistence of Persian loanwords (like *Berkat*, meaning blessing) that have integrated differently across religious communities. Sanskrit loanwords play a similar role, while English loanwords indicate Christian authorship, particularly in Protestant contexts.

3.4 Corpus Examples

To illustrate these distinctions, consider how the three communities discuss natural disasters:

Muslim text:

> *Dalam **khutbah** Jumat ini, **khatib** juga menyinggung **musibah** banjir bandang, dan tanah longsor, yang melanda Aceh dan beberapa provinsi lainnya dalam tiga pekan terakhir.*
>
> "In this Friday **sermon**, the **preacher** also touched upon the **calamity** of flash floods and landslides that struck Aceh and several other provinces in the last three weeks."

Key markers: *khutbah* (Friday sermon), *khatib* (Islamic preacher), *musibah* (Arabic-origin word for calamity/disaster)

Catholic text:

> ***Gereja** juga segera membuka dapur umum untuk memenuhi kebutuhan para korban banjir yang sudah mulai berdatangan.*
>
> "The **Church** also immediately opened a public kitchen to meet the needs of the flood victims who have already started arriving."

Key markers: *Gereja* (Church as institution), focus on institutional charitable action

Protestant text:

> *Kepada **saudara-saudara** kami yang terdampak banjir di Parapat dan sekitarnya, hati kami bersama kalian.*

> "To our **brothers and sisters** [brethren] affected by floods in Parapat and surrounding areas, our hearts are with you."

Key markers: *saudara-saudara* (brothers and sisters/brethren), personal solidarity language

These examples illustrate how identical semantic content—response to a natural disaster—is realized through distinct lexical and stylistic choices that signal the speaker's religious identity.

3.5 Implications for NLP Systems

These linguistic varieties permeate everyday discourse beyond religious contexts, appearing in domestic, educational, and political settings. Every Indonesian speaker exhibits religiolinguistic features that index their background, making awareness of these varieties essential for equitable NLP systems.

Because religiolinguistic features function as identity markers, NLP systems trained exclusively on majority-normative data risk systematically disadvantaging religious minorities in high-stakes applications. A job screening algorithm trained on Muslim-normative Indonesian might flag Christian-marked resumes as linguistically anomalous. Content moderation systems might misclassify minority religious expression as low-quality or foreign content. Machine translation systems might flatten theological nuance by defaulting to majority-normative terminology. Recognition of religiolects as distinct low-resource domains within Indonesian becomes necessary not merely for linguistic documentation, but for building fair and inclusive language technology.

4 Corpus Construction

While recent initiatives like NusaCrowd (Cahyawijaya et al., 2023) have made significant strides in cataloging Indonesia's regional languages (e.g., Javanese, Sundanese), there remains a scarcity of resources dedicated to the sociolects that cut across these regional boundaries.

4.1 Dataset Overview

The Indo-Religiolect Corpus contains 3,013,172 sentences distributed across three religious communities:

- **Islam:** 1,455,454 sentences from ∼27 sites (e.g., NU Online)
- **Catholic:** 797,131 sentences from ∼30 sites (e.g., Mirifica, KAS)
- **Protestant:** 760,587 sentences from ∼44 sites (e.g., PGI)

Each entry in the dataset includes the following metadata fields:

- `label`: Religious denomination (Islam, Catholic, Protestant)
- `denomination`: Full denomination name or identifier
- `region`: Geographic region (when available)
- `date`: Publication date
- `title`: Source article title
- `text`: The text to be classified
- `url`: Source URL

4.2 Data Acquisition Pipeline

To mitigate noise inherent in blind web crawling, a custom asynchronous crawler was developed targeting "trusted nodes"—established religious institutional websites. The system utilizes `asyncio` and `aiohttp` for asynchronous concurrency with a domain rotation queue to maximize parallelism while respecting server load.

Ethical scraping constraints were implemented throughout, including compliance with the `robots.txt` protocol (via `urllib.robotparser.RobotFileParser`) and adaptive rate limiting. The system monitors server response times, dynamically adjusting delays and triggering exponential backoff on HTTP 429/503 errors. To ensure high-quality content retrieval, the crawler filters URL traps by rejecting deep paths (>6 segments), repeating segments, and non-content endpoints.

4.3 Text Extraction and Normalization

Content extraction was performed using `trafilatura` with a fallback to `BeautifulSoup`. A critical challenge in scraping religious texts involves handling stylized formatting common in older CMS platforms. A custom "drop cap" fix was implemented to merge styled initial letters back into the text stream (e.g., correcting `H ati` to "Hati").

4.4 NLP Cleaning Pipeline

The raw data underwent a multi-stage cleaning process designed to isolate linguistic style from boilerplate noise:

Language Detection: The pipeline utilized `langdetect` with a fixed seed, strictly filtering for Indonesian (`id`) and Malay (`ms`) while rejecting unrelated pages.

Scripture De-referencing: To ensure models learn the style of each religiolect rather than memorizing scripture, specific citation coordinates were stripped. Bible references (e.g., "Mat 25:46") and Quranic citations (e.g., "QS Al-Baqarah: 183") were removed while preserving surrounding theological commentary.

Heuristic Filtering: Structural filters removed "web noise" by rejecting sentences with high navigation density (>30% keywords), excessive capitalization, or incomplete fragments. Specific rules targeted "Mojibake," navigation keywords, and contact information. Short religious expressions (e.g., "amin," "alhamdulillah") were whitelisted as legitimate content.

4.5 Design Choices and Scope

The corpus construction deliberately prioritized signal quality over raw volume. Given that Indonesia is nearly 90% Muslim, a raw scrape of all Indonesian religious websites would produce massive class imbalance. Balanced undersampling was implemented, discarding majority-class data to create a perfectly balanced three-way split. We chose undersampling (rather than only applying class-weighted losses) to establish a strict baseline benchmark where class size is controlled and each religiolect contributes comparable lexical diversity during training. This design choice optimizes the corpus for benchmark evaluation rather than proportional representation.

The corpus intentionally retains regional language intrusions (Batak, Sundanese, Javanese). In low-resource sociolects, code-switching represents a defining feature rather than noise to be eliminated. This preserves authentic linguistic practices within each religious community.

The corpus focuses on Indonesia's three largest religious groups. Within these categories, Islam encompasses two main organizational streams with regional variation across islands; Catholicism maintains relatively centralized institutional language despite being a smaller community; Protestantism exhibits the widest denominational variation, though some smaller groups may be underrepresented in this collection.

This corpus supports multiple research directions: text classification by denomination, religiolect analysis of vocabulary and syntactic patterns, and general Indonesian NLP research. Future work could explore whether models can identify religious affiliation through syntactic structure alone via delexicalization techniques.

5 Validation Experiment

To validate the Indo-Religiolect Corpus as capturing distinct low-resource domains, a baseline classification experiment was conducted. The goal was to determine whether a standard pre-trained Indonesian language model could effectively distinguish between the three religious sociolects and to identify patterns of linguistic overlap.

5.1 Experimental Setup

The IndoBERT base model `indolem/indobert-base-uncased` was fine-tuned for three-way classification. IndoBERT is a transformer-based model pre-trained on general Indonesian text and serves as a standard baseline for Indonesian NLP tasks (Koto et al., 2020).

5.2 Training Configuration

To facilitate reproducibility, we report the exact training environment and hyperparameters used for the baseline experiment.

Software and hardware: Training was performed in Python 3.12 using PyTorch 2.2.2 and Hugging Face Transformers 4.57.3 on a single NVIDIA A100-SXM4-40GB GPU. We trained with mixed precision `bf16` (selected for numerical stability on A100 hardware) and enabled gradient checkpointing to reduce memory usage.

Optimization and hyperparameters:

- **Optimizer:** AdamW (`adamw_torch`)
- **Learning rate:** 2×10^{-5} (linear schedule)
- **Warmup:** 500 steps (linear warmup)
- **Weight decay:** 0.01
- **Epochs:** 3

- **Batch size:** 64 (train), 128 (eval)
- **Max sequence length:** 128 tokens

Data balancing: To avoid majority-class dominance, we applied balanced undersampling, capping each class at 666,666 examples (total ≈2 million sentences) prior to splitting. The resulting dataset was split into a training set of 1,799,998 sentences and a held-out test set of 200,000 sentences.

5.3 Results

The fine-tuned model achieved robust overall performance (Table 1), validating both dataset quality and the separability of religiolects. Overall accuracy reached 88.26% with a balanced F1 score of 88.25%. Precision (88.25%) and recall (88.26%) were similarly aligned, demonstrating stable model behavior across all three classes.

Per-class performance (summarized in Table 1 and Figure 2) reveals asymmetric patterns.

Class	Accuracy	Correct	Misclassified as
Islam	91.73%	61,150	Catholic, Protestant
Catholic	86.41%	57,608	Protestant
Protestant	86.64%	57,762	Catholic

Table 1: Per-class performance on the held-out test set. Note the significantly higher accuracy for Islam compared to the bidirectional confusion between Catholic and Protestant classes.

5.4 Analysis

Islamic Distinctiveness: The model achieved highest performance on the Islam category, with remarkably low misclassification rates to either Christian denomination. This suggests the Islamic religiolect possesses highly distinct vocabulary, likely driven by specific Arabic loanwords and theological concepts that clearly separate it from Christian varieties.

Christian Continuum: In contrast, the model struggled to differentiate between Catholic and Protestant texts. The bilateral confusion between these categories (totaling over 11,000 mutually misclassified samples) supports the sociolinguistic hypothesis of a shared "Christian Indonesian" register. Both communities share core theological terms (e.g., *Yesus*, *Kristus*, *Gereja*) that create substantial lexical overlap.

However, the fact that the model correctly classified the majority of Christian texts indicates that

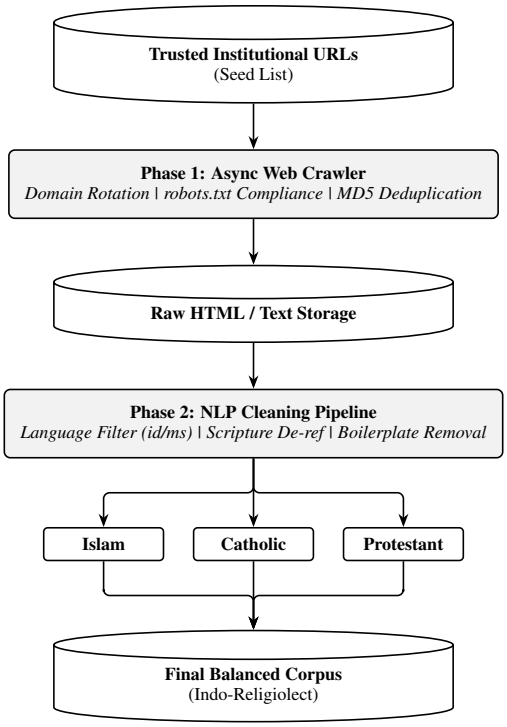

Figure 1: Data collection and preprocessing pipeline for the Indo-Religiolect Corpus. Phase 1 performs ethical, asynchronous crawling over trusted institutional seeds; Phase 2 applies language filtering and text-cleaning heuristics before producing denomination-labeled streams and a final corpus.

subtle shibboleths—such as the *Allah*/*Tuhan* distinction and denomination-specific styles discussed in Section 2—are preserved in the corpus and computationally detectable. The approximately 5% performance gap between Islam and the Christian denominations quantifies this "Christian Continuum" effect, where shared lexical items create a natural ceiling on classification accuracy that does not exist for the Muslim variety.

6 Limitations

This work represents an initial step toward capturing Indonesian religiolects, but several limitations warrant acknowledgment and suggest directions for future research.

Exclusion of Smaller Communities: The corpus captures Indonesia's three largest religious communities (Muslim, Protestant, Catholic) but does not yet include Hindu, Buddhist, Confucian, and others like Indigenous faith communities, which collectively comprise approximately 3% of the population. This exclusion stems from methodological challenges. Smaller religious groups may lack the centralized institutional websites that enabled

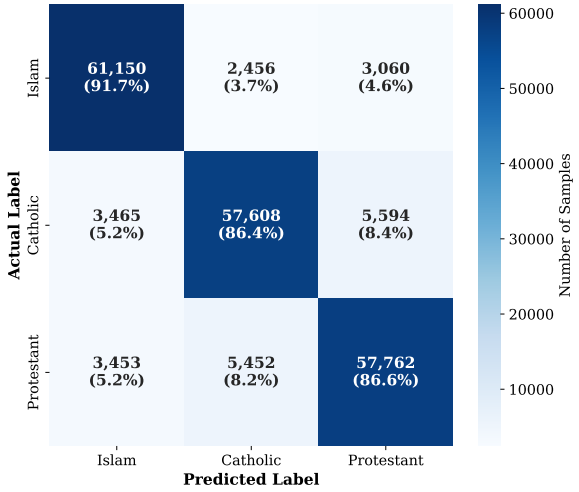

Figure 2: Confusion matrix of IndoBERT fine-tuned on the Indo-Religiolect test set ($N = 200,000$). The model exhibits high separability for Islam (91.7%) alongside bilateral confusion between Catholic and Protestant categories, illustrating the shared "Christian Continuum" register.

the "trusted node" approach used here. Effectively identifying, locating, and aggregating high-quality corpora for groups with more dispersed digital presence remains an open research question requiring different collection strategies.

Internal Denominational Diversity: The "Protestant" label collapses substantial theological and sociolinguistic heterogeneity into a single class. Indonesian Protestantism spans mainline and Reformed traditions (including Lutheran-identified communities such as HKBP) as well as Charismatic and Pentecostal movements, which differ in worship style, institutional histories, and transnational ties. These differences can plausibly surface in written language: older mainline institutions may retain more formalized ecclesial registers and translation conventions shaped by earlier Dutch- or Kitab Malay–influenced textual practices, while Charismatic/Pentecostal communities often employ more contemporary, affective rhetoric and may more readily introduce calques and discourse patterns aligned with globalized (often American) worship media.

Aggregating such subgroups into a single "Protestant" category therefore encourages the model to learn a "median" representation of Protestant-associated Indonesian that may smooth over denomination-specific shibboleths and community identity markers. This matters both for sociolinguistic validity and for model use: high performance on an aggregated label does not imply uniform coverage across denominations, and errors may be concentrated in underrepresented subtraditions. Future work should evaluate finer-grained denominational splits (where ethically and practically feasible) and/or hierarchical taxonomies (e.g., mainline vs. charismatic) to test whether distinct Protestant sub-registers are separable and to better reflect the Indonesian Christian landscape.

Register Limitations: The current corpus consists entirely of edited, institutional web text, and thus disproportionately reflects *Bahasa Baku* (standard/formal Indonesian) rather than the range of registers used in everyday interaction. Indonesian sociolinguistics commonly describes a sharp functional distinction between formal standard usage and more colloquial styles (often glossed as *bahasa gaul*), which can be understood through the lens of diglossia-like register compartmentalization (Ferguson, 1959; Sneddon, 2003). Because institutional websites are public-facing and reputationally sensitive, they often enforce "sanitized" stylistic norms (standardized morphology, careful orthography, reduced slang), which can attenuate precisely the interactional cues where religious identity may be most salient.

This register bias matters for generalization. Religious identity is frequently indexed through informal choices—including discourse particles, address terms, stance-taking, and code-switching patterns with regional languages (e.g., Javanese, Sundanese, Batak)—that may appear more strongly in sermons, testimonies, small-group communication, or social media than in official news posts. As a result, models trained and validated on formal web text may perform well in the institutional register while degrading on informal or spoken data, where orthography is variable and register boundaries relax. Future work should therefore test cross-register robustness by adding sermon transcripts and conversational or social media data (with appropriate ethical safeguards) and by evaluating whether religiolect signals persist after controlling for register.

Temporal Scope: The corpus reflects contemporary language use (primarily 2020–2025) and cannot capture historical evolution of religiolects or predict future changes. As Indonesia's religious communities continue to develop and interact, these language varieties may shift, merge, or diverge in ways not captured by this snapshot.

Ethical Considerations

While this corpus enables important research on linguistic diversity and model fairness, it also carries potential for misuse. Classification models trained on this data could theoretically be deployed for religious profiling or discrimination. The research community must remain vigilant about dual-use concerns and work to ensure such tools are used to reduce bias rather than enable it.

This work aligns with emerging research on measuring social harm in language models, such as recent empirical studies on religious bias (Sadhu et al., 2024), by extending such scrutiny to the specific context of Indonesian religious communities.

7 Conclusion

This paper presents the first large-scale corpus and benchmark for Indonesian religiolects, demonstrating that these varieties function as low-resource domains nested within a high-resource language. The Indo-Religiolect Corpus contains 3 million sentences across Muslim, Catholic, and Protestant communities, validated through classification experiments showing both the distinctiveness of these varieties and patterns of overlap that align with sociolinguistic theory.

The validation experiment reveals that while religiolects are computationally distinguishable (88.26% accuracy), they exhibit asymmetric separability: Islamic Indonesian is highly distinct, while Catholic and Protestant varieties share substantial lexical overlap yet retain detectable shibboleths. This finding has direct implications for NLP practitioners.

Current language models trained on generic Indonesian corpora risk systematically misunderstanding or misrepresenting religious minorities. Although we do not evaluate downstream tasks in this work, the observed separability of religiolects suggests plausible mechanisms by which representation collapse *may* produce disparate outcomes in deployed NLP pipelines, analogous to documented cases where dialectal or identity-linked markers spuriously correlate with toxicity or other labels (Dixon et al., 2018; Sap et al., 2019; Blodgett et al., 2020):

- **Content moderation systems** may over-flag Christian-marked Indonesian as anomalous or toxic if religiolectal cues are treated as out-of-distribution or are spuriously associated with policy-violating categories (Dixon et al., 2018; Sap et al., 2019).
- **Machine translation systems** may flatten religious nuance by normalizing minority theological vocabulary toward majority-normative equivalents, reducing fidelity for community-specific terms.
- **Sentiment analysis tools** may misinterpret affect and stance when religiolectal expressions (including scripture-adjacent phrasing) are underrepresented in training data.
- **Automated hiring systems** could penalize applicants whose writing contains minority religiolect markers if screening models reward majority-normative phrasing or treat minority registers as low quality (Blodgett et al., 2020).

To build equitable AI technology, future NLP work must move beyond monolithic "National Language" benchmarks to include sub-national varieties—whether regional, socioeconomic, or religious. The methodology presented here—trusted node scraping, balanced undersampling, and validation through classification—offers a template for capturing other low-resource domains nested within high-resource languages.

Finally, this work opens critical directions for future research. Can models identify religious affiliation through syntactic structure alone, independent of obvious lexical markers? Delexicalization studies could reveal whether religiolects exhibit distinctive grammatical patterns beyond vocabulary choices. Additionally, expanding to spoken registers, smaller faith communities, and finer-grained denominational distinctions would deepen understanding of how language and religious identity interweave in Indonesia's plural society.

Acknowledgments

During this research, Dan Sachs was a Fulbright U.S. Student Researcher (AMINEF) hosted by the Indonesian Consortium of Religious Studies at Universitas Gadjah Mada, Yogyakarta, Indonesia (2024–2025). The author extends his gratitude to Dr. Zainal Abidin Bagir and Dr. Leonard Chrysostomos Epafras for their generous support.

References

Alessandro Bausani. 1975. Is classical malay a "muslim language"? *Boletin de la Asociacion Española de Orientalistas*.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, and 28 others. 2023. Nusacrowd: Open source initiative for indonesian nlp resources. *Preprint*, arXiv:2212.09648.

Stuart Campbell. 1996. The distribution of -at and -ah endings in malay loanwords from arabic. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 152(1):23–44.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 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.

Jonathan Dunn. 2019. Modeling global syntactic variation in english using dialect classification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 42–53. Association for Computational Linguistics.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Kevin W. Fogg. 2015. The standardisation of the indonesian language and its consequences for islamic communities. *Journal of Southeast Asian Studies*, 46(1):86–110.

Benjamin Hary and Martin J. Wein. 2013. Religiolinguistics: On jewish-, christian- and muslim-defined languages. *International Journal of the Sociology of Language*, 2013(220):85–108.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770. International Committee on Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

J. N. Pava, C. Meinhardt, H. B. U. Zaman, and T. Friedman. 2024. Mind the (language) gap: Mapping the challenges of llm development in low-resource language contexts. Stanford HAI Policy Brief.

Peter G. Riddell. 2002. Literal translation, sacred scripture and kitab malay. *Studia Islamika*, 9(1).

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2024. Social bias in large language models for bangla: An empirical study on gender and religious bias. *Preprint*, arXiv:2407.03536.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678. Association for Computational Linguistics.

Emanuel Gerrit Singgih. 2003. *Doing Theology in Indonesia: Sketches for an Indonesian Contextual Theology*. ATESEA.

James N. Sneddon. 2003. *The Indonesian Language: Its History and Role in Modern Society*. University of New South Wales Press.

Nikolaos Van Dam. 2010. Arabic loanwords in indonesian revisited. *Bijdragen tot de Taal-, Land- en Volkenkunde*, 166(2/3):218–243.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. *Preprint*, arXiv:2205.15960.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank

Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17. Association for Computational Linguistics.