# A SYSTEMATIC LITERATURE REVIEW OF NATURAL LANGUAGE PROCESSING FOR INDONESIAN REGIONAL LANGUAGES

**Hendri Ahmadian**
*Prodi Teknologi Informasi, Fakultas Sains dan Teknologi UIN Ar-Raniry, Banda Aceh, Indonesia*
*Email korespondensi: hendri@ar-raniry.ac.id*

**Abstract**: This systematic literature review (SLR) investigates the evolution of Natural Language Processing (NLP) for Indonesian regional languages from 2020 to 2025. Analyzing 13 pivotal studies, the research identifies a significant transition from fragmented studies of high-population languages, such as Sundanese and Madurese, toward inclusive, archipelago-wide frameworks covering "low-resource" dialects like Acehnese and Nias. Architecturally, the field has progressed from classical machine learning to Transformer-based Large Language Models (LLMs), including IndoBART and GPT. Furthermore, data provenance has evolved from unstructured social media corpora to standardized multilingual benchmarks like NusaX and NusaCrowd. Despite these advancements, persistent gaps in data standardization and large-scale pretraining resources remain. Future research should prioritize cross-lingual transfer learning and specialized benchmarks to ensure the technological sustainability of Indonesia's diverse linguistic heritage.

**Keywords:** Natural Language Processing, Indonesian Regional Languages, Systematic Literature Review, Large Language Models.

**Abstrak:** Systematic literature review ini mengkaji evolusi Natural Language Processing (NLP) untuk bahasa daerah di Indonesia dari tahun 2020 hingga 2025. Melalui analisis terhadap 13 studi utama, penelitian ini mengidentifikasi transisi signifikan dari studi terfragmentasi pada bahasa berpopulasi tinggi, seperti bahasa Sunda dan Madura, menuju kerangka kerja inklusif yang mencakup dialek minim sumber daya (low-resource) seperti bahasa Aceh dan Nias. Secara arsitektural, bidang ini telah berkembang dari machine learning menuju Large Language Models (LLMs) berbasis Transformer, termasuk IndoBART dan GPT. Selain itu, sumber data telah berevolusi dari korpus media sosial yang tidak terstruktur menjadi tolok ukur (benchmark) multibahasa yang terstandardisasi seperti NusaX dan NusaCrowd. Meskipun terdapat kemajuan, kesenjangan dalam standardisasi data dan sumber daya pratatih berskala besar masih ditemukan. Riset mendatang perlu memprioritaskan cross-lingual transfer learning dan pengembangan benchmark khusus untuk menjamin keberlanjutan teknologi warisan linguistik Indonesia yang beragam.

**Kata kunci:** Natural Language Processing, Bahasa Daerah Indonesia, Systematic Literature Review, Large Language Models.

## 1. Introduction

Research on Natural Language Processing (NLP) continues to demonstrate a hegemony toward high-resource, data-driven languages. While data deficits are often considered a major obstacle for minority languages, the underlying problem lies in the complexity of linguistic diversity and a limited theoretical understanding of the unique characteristics of these languages (Joshi et al., 2020). This phenomenon is particularly critical in Indonesia, which boasts over 700 regional languages but still faces hurdles in literary documentation, minimal integration into formal education, and challenges in standardization (Novitasari et al., 2020). The dominance of Indonesian as the national language has also accelerated a shift in the role of local languages, risking linguistic marginalization that could lead to a monolingual society (Cohn & Ravindranath, 2014). Therefore, the preservation of regional languages can no longer rely solely on traditional linguistic documentation; it demands measurable computational intervention to maintain their relevance in the digital age.

These sociolinguistic dynamics manifest as a significant digital divide in the language technology landscape of the archipelago. Unlike Indonesian, which has achieved technological maturity through curated Large Language Models (LLMs), the majority of regional languages are still classified as low-resource. However, new optimism is emerging through strategic initiatives such as the NusaCrowd (Cahyawijaya et al., 2022), the IndoBERT model (Wilie et al., 2020), and multilingual model optimizations that integrate local language parameters through cross-lingual transfer mechanisms. This transition is also driven by researchers' initiatives to gradually build independent corpora—utilizing local repositories and extracting bilingual dictionaries into machine-processable formats—which serve as a crucial foundation for NLP modeling.

To ensure the sustainability of this innovation, a comprehensive roadmap is needed to avoid research redundancy and resource inefficiencies. Collaborative synergy in mapping research priorities. To achieve this, a literature review that not only maps studies narratively but also evaluates them through a rigorous methodological framework is essential.

The objective of this study is to map the potential advancements and existing barriers in developing NLP tools for Indonesian local languages. The methodology involved a systematic review of existing literature sourced from major academic databases, including Google Scholar, IEEE Xplore, and Scopus.

## 2. Methods

### 2.1 Selection Criteria and Framework

To ensure a systematic selection of relevant literature, this research utilized the PICOC (Population, Intervention, Comparison, Outcome, and Context) framework to establish stringent eligibility requirements. The PICOC frameworks can be seen in Table 1. Inclusion was limited to peer-reviewed primary studies published between 2020 and 2025 that introduced computational NLP models or datasets specifically for Indonesian regional languages and provided empirical performance data.

Table 1. Methodological Framework for Research Questions

| Element | Description |
|---|---|
| Population (P) | The indigenous vernaculars of the Indonesian archipelago, such as Javanese, Sundanese, and Minangkabau |
| Intervention (I) | The frameworks for computational Natural Language Understanding, including their underlying data and cross-lingual adaptation methods |
| Comparison (C) | Standard benchmarks (such as mBERT and XLM-R) alongside conventional approaches to documentation |
| Outcome (O) | Quantitative assessment measures (F1-score, accuracy, BLEU) and the reachability of relevant data resources |
| Context (C) | Computational safeguarding of local knowledge and the implementation of inclusive technologies for Indonesia’s diverse regions |

To refine the focus of this investigation, four pivotal research questions (RQs) were formulated: First, we categorize the specific Indonesian regional languages currently represented in the literature (RQ1). Second, we examine the prevailing NLP tasks that have garnered the most scholarly attention (RQ2). Third, we evaluate the model architectures and algorithmic frameworks that consistently deliver superior performance (RQ3). Finally, by synthesizing the limitations of existing studies, we identify critical research gaps to outline a strategic roadmap for future exploration (RQ4).

We excluded research focusing solely on standard Bahasa Indonesia, purely theoretical linguistic studies lacking computational validation, and works centered on speech-based modalities. Furthermore, preprints, review papers, and papers with restricted full-text access were omitted to safeguard the methodological quality of the synthesis.

Table 2. Methodology for Study Selection and Eligibility

| Criteria | Inclusion (In-Scope) | Exclusion (Out-of-Scope) |
|---|---|---|
| Language | Regional/Local | Bahasa Indonesia only |
| Modality | Text-based NLP (Textual) | Speech/Audio (ASR, TTS) |
| Study Type | Empirical Research | Review Papers |
| Output | Performance Metrics | Theoretical Frameworks only |

## 2.2 Search Strategy

Our search strategy aimed to bridge the gap between computational linguistics and the sociolinguistic realities of the Indonesian archipelago. We performed an extensive literature harvest across four major digital libraries: Scopus, IEEE Xplore, the ACL Anthology, supplemented by Google Scholar and the Garuda portal to capture local scholarly contributions. The search window (2020–

2025) was chosen to observe the evolution from traditional feature engineering to modern Transformer-based and LLM paradigms.

### 2.3 Selection Rationale and Scoping

These criteria were applied to ensure the resulting corpus directly served the study’s primary goals. By isolating regional languages, the review effectively sidesteps the saturated research field of standard Indonesian, focusing instead on the digital divide affecting low-resource languages. The deliberate exclusion of audio-based signals restricts the scope to text-only NLP, allowing for a focused evaluation of semantics and syntax without the confounding variables of acoustic processing. This empirical focus enables a rigorous benchmarking of model architectures to identify which methodologies best accommodate the unique linguistic structures of Indonesia's regional languages.

## 5. Results and Discussion

The following discourse provides a systematic overview of the literature on Indonesian regional language NLP, with significant findings tabulated in Table 3. By synthesizing this data, the current chapter seeks to resolve the research questions identified in the earlier stages of this study.

### RQ1: Regional Language Varieties Under Investigation

From 2020 to 2025, the research environment for Indonesian regional languages underwent a significant geographical and linguistic transformation. Early academic efforts primarily targeted high-density languages with more readily available data, such as Minangkabau, Sundanese, and Madurese. Although these studies provided the essential groundwork for regional NLP by tackling prominent non-national languages, the research was frequently fragmented, emphasizing specific linguistic structures rather than adopting a unified, nationwide outlook.

By the 2023–2025 period, the discipline moved toward a more inclusive and scalable model, extending its reach to numerous low-resource languages. This recent surge in scholarship has integrated underrepresented tongues—such as Acehnese, Balinese, Buginese, Ambon, and Batak—as well as less-documented dialects like Gorontalo and Nias. This progression marks a significant transition from isolated, single-language investigations toward the development of comprehensive, multilingual benchmarks. Collaborative projects like NusaX, NusaWrites, and NusaCrowd have successfully established new industry standards, creating frameworks that protect and support Indonesia's diverse linguistic landscape in a synchronized manner.

### RQ2: Predominant NLP Tasks

The field NLP for Indonesian regional languages has evolved from basic linguistic categorization into a sophisticated domain of generative capabilities. In

2020, research primarily focused on foundational tasks—such as sentiment analysis, emotion classification, and part-of-speech tagging—aimed at identifying the tone and structural properties of regional dialects. However, between 2021 and 2025, the research focus shifted toward more complex applications. Contemporary studies now prioritize generative tasks, including machine translation and automated dialogue summarization, reflecting an increasing demand for technologies capable of human-like interaction and synthesis in regional languages.

Accompanying this functional evolution is a distinct architectural transition from traditional statistical methods to advanced neural networks. While early studies relied heavily on classical machine learning algorithms like Naive Bayes and Support Vector Machines (SVM), the expansion of computational power and data availability has facilitated a move toward deep learning and Large Language Models (LLMs). By the 2023–2025 period, the industry decisively adopted Transformer-based architectures such as mBERT and XLM-R, along with specialized models like IndoBART and cutting-edge generative frameworks like Llama, Mistral, and GPT. This progression signifies a shift from simple classification toward a more nuanced and context-aware era of Indonesian language technology.

**RQ3: Model Architectures and Performance Metrics**

The robustness and reliability of NLP models for Indonesian regional languages are intrinsically linked to the evolution of data provenance. Early research primarily utilized datasets harvested from accessible yet unstructured platforms, such as Twitter and Wikipedia. While these sources provided an essential starting point, they frequently lacked the linguistic precision and formal structure necessary for high-level analytical tasks. In contrast, recent advancements signify a shift toward more specialized and curated resources. Contemporary studies now leverage sophisticated corpora and benchmarks—including the Sunda Corpus, IndoNLG, and the NusaDialogue dataset—offering a more structured and representative foundation for model training in low-resource environments.

To rigorously evaluate model efficacy, researchers employ a suite of standardized metrics tailored to specific NLP objectives. Performance in classification-based tasks, such as sentiment analysis and part-of-speech tagging, is consistently assessed using F1-score and accuracy to ensure a balanced reflection of model validity. As the field has matured into complex generative domains like machine translation and text summarization, evaluation criteria have become increasingly specialized. Researchers now utilize metrics such as BLEU, SacreBLEU, and ROUGE to measure linguistic quality and textual overlap. This systematic evaluative framework facilitates a precise comparison between legacy machine learning methods and contemporary deep learning architectures.

**RQ4: Research Gaps and Future Roadmap**

The synthesis of current literature reveals several persistent challenges that impede the advancement of Indonesian regional language technology, most notably a pervasive lack of standardization. Research conducted between 2020 and 2021 frequently identified a lack of uniform data protocols and a pressing need for more extensive, diverse corpora to enhance model reliability. These foundational hurdles suggest that while progress has been made, the fragmented nature of early datasets continues to obstruct the development of robust linguistic tools. Resolving these discrepancies is vital for transitioning from localized, experimental projects to a more integrated national infrastructure for regional NLP.

To address these limitations, the strategic roadmap for the field proposes a significant methodological shift toward more efficient learning paradigms. Rather than training models from scratch—a process often precluded by the resource constraints of regional languages—modern research advocates for cross-lingual transfer learning, zero-shot prompting via LLMs, and the utilization of multilingual models. This strategic evolution seeks to mitigate data scarcity by leveraging the linguistic knowledge of high-resource languages to support underrepresented dialects. Moving forward, priority must be given to expanding pretraining data and establishing specialized benchmarks, such as the proposed IndoGLUE, to ensure the long-term sustainability and efficacy of Indonesian language technology across the archipelago.

Table 3. Overview of Key Literature regarding NLP for Indonesia's Regional Languages

| No | Work | Database | Regional Language | NLP Task | Model/Architecture | Metric | Dataset Type | Key Findings |
|---|---|---|---|---|---|---|---|---|
| 1 | (Koto & Koto, 2020) | Scopus | Minang | Sentiment Analysis and Machine Translation | Naive Bayes, Support Vector Machine, Logistic Regression, Bi-LSTM, Transformers, mBERT | F1-Score | Twitter and Wikipedia | Lack of standardization and larger corpus |
| 2 | (Putra et al., 2020) | IEEE | Sundanese | Emotion Classification | KNN, Naive Bayes, Support Vector Machine, Logistic Regression | F1-Score | Twitter | Lack of standardization and larger corpus |
| 3 | (Dewi & Ubaidi, 2020) | Google Scholar | Madura | Part of Speech | Brill Tagger | Accuracy | Anthology of Madurese Narratives and Articles | Lack of standardization and larger corpus |
| 4 | (Cahyawijaya et al., 2021) | ACL Anthology | Javanese, Sundanese | summarization, question answering, chit-chat and machine translation | IndoBART, IndoGPT, mBART, mT5 | BLEU | IndoNLG Benchmark | training larger IndoBART and IndoGPT models, using larger pretraining data |
| 5 | (Putri et al., 2021) | Scopus | Javanese, Sundanese, Madurese, Minangkabau, and Musi | Abusive language, Hate speech | Naive Bayes, Support Vector Machine, and Random Forest Decision Tree | F1-Score | Twitter | Lack of standardization and larger corpus |
| 6 | (Wongso et al., 2022) | Scopus | Sundanese | Emotion Classification | Transformers | Macro-F1, accuracy | Sunda Corpus | Create a benchmark like Sundanese GLUE |

| No | Work | Database | Regional Language | NLP Task | Model/Architecture | Metric | Dataset Type | Key Findings |
|---|---|---|---|---|---|---|---|---|
| 7 | (Nugraha & Romadhony, 2023) | ACL Anthology | Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak | Machine Learning Task | Support Vector Machine, Naive Bayes Classifier, Decision Tree, Rocchio Classification, Logistic Regression, and Random Forest | macro-F1, accuracy | NusaX Benchmark | Machine learning approaches |
| 8 | (Sulistyo et al., 2023) | Scopus | Madurese | Machine Translation | Long Shor Term Memory | BLEU | Madura Corpus | Machine learning approaches |
| 9 | (Winata et al., 2023) | ACL Anthology | Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak | sentiment analysis and machine translation | Naive Bayes, SVM Logistic Regeression, IndoNLU, IndoLEM, m-BERT, XLM-R, IndoGPT, IndoBART,mBART, mT5 | Macro-F1, SacreBLEU | NusaX Benchmark | Cross-lingual transfer |
| 10 | (Cahyawijaya et al., 2023) | ACL Anthology | Ambon, Batak, Betawi, Bima, Buginese, Javanese, Madurese, Makassarese, Minangkabau, Palembang / Musi, Rejang, and Sundanese. | Sentiment analysis, emotion classification and machine translation | Naive Bayes, SVM Logistic Regeression, IndoLEM, IndoNLU, m-BERT, XLM-R, BLOOMZ, mT0 | macro-F1, SacreBLEU, ChrF++ | NusaWrites Benchmark | Cross-lingual transfer, zero shot prompting via LLM |
| 11 | (Sujaini & Putra, 2024) | Scopus | Javanese, Batak, Sundanese, Bugis, | Detecting local language | Naive Bayes and KNN | F1-score | NusaCrowd | Alternative machine learning approaches |

| No | Work | Database | Regional Language | NLP Task | Model/Architecture | Metric | Dataset Type | Key Findings |
|---|---|---|---|---|---|---|---|---|
| | | | Malay, Dayak, Madurese, and Minang | | | | | |
| 12 | (Wongso et al., 2025) | ACL Anthology | Javanese, Sundanese, Acehnese, Minangkabau, Banjarese Balinese, Gorontalo, Banyumasan, Buginese, Nias, Tetum | Emotion Classification, Sentiment Analysis, Aspect-based Sentiment Analysis, Textual Entailment, Topic Modeling, Rhetorical Mode, Named Entity Recognition, Part-of-Speech Tagging and Span Extraction | Naive Bayes, SVM Logistic Regeression, IndoNLU, IndoLEM, m-BERT, XLM-R, NusaBERT | macro-F1, accuracy | Nusa Translation, Nusa Paragraph, Nusa X, IndoNLU | Cross-lingual transfer |
| 13 | (Purwarianti et al., 2025) | ACL Anthology | Minangkabau, Balinese, and Buginese | Dialogue-Summarization | IndoNLU-IndoBART, IndoNLU-IndoGPT, mT5, T5, Llama, Merak, Mistral, Wizard-Vicuna, bloom, GPT, zephyr | ROUGEL, ROUGE-2 | NusaDialogue Dataset | Cross-lingual transfer |

## 5. Conclusions and Future Works

### 5.1. Conclusions

The systematic analysis of NLP research for Indonesian regional languages from 2020 to 2025 reveals a field in the midst of a significant technological and linguistic transformation. The research landscape has matured from isolated studies of high-population languages—specifically Minangkabau, Sundanese, and Madurese—to a more inclusive, archipelago-wide focus that encompasses low-resource and niche dialects such as Acehnese, Buginese, Gorontalo, and Nias.

Technologically, the field has transitioned from traditional statistical methods used for basic classification tasks like sentiment analysis, toward sophisticated deep learning and generative architectures. The adoption of Transformers and LLMs like IndoBART, Llama, and GPT has enabled more complex NLP applications, including machine translation, automated summarization, and dialogue-based tasks. Furthermore, the shift from unstructured social media data to curated benchmarks such as NusaX, NusaCrowd, and NusaWrites indicates a critical move toward standardization. While early research was hindered by fragmented datasets and a lack of uniform protocols, current trends emphasize collaborative, multi-language frameworks that prioritize both algorithmic efficacy and linguistic robustness.

### 5.2. Future Works

Despite the rapid advancements documented in this review, several persistent gaps remain that define the strategic roadmap for future research. To ensure the long-term sustainability of Indonesian language technology, the following directions are recommended:

- Expansion of Multilingual Pretraining Data. Future efforts should prioritize the expansion of pretraining data for underrepresented languages. While existing benchmarks provide a foundation, there is a pressing need for larger, high-quality corpora sourced from literary archives and formal documents to improve the performance of models beyond simple zero-shot prompting.
- Development of task-specific benchmarks. There is a clear call for the creation of specialized benchmarks similar to the proposed IndoGLUE. These benchmarks should be designed to evaluate model performance across a wider variety of regional linguistic nuances, ensuring that technology remains accurate across different dialects and speech levels.
- Methodological innovation via transfer learning: To mitigate the scarcity of regional data, future research should move away from training from scratch. Instead, exploring cross-lingual transfer learning and Parameter-Efficient Fine-Tuning (PEFT) will be vital in leveraging the linguistic similarities within the Austronesian language family to support low-resource tongues.

- Socio-Linguistic sensitivity. Future models must be evaluated not only on technical metrics like BLEU or F1-score but also on their ability to handle culturally specific elements of Indonesian regional languages, such as formal and informal speech registers. This will ensure that the developed technology is culturally respectful and applicable to real-world social contexts.

## Daftar Kepustakaan

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G. I., Wilie, B., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Koto, F., Santoso, J., Moeljadi, D., Hudi, C. W. and F., Parmonangan, I. H., Alfina, I., Wicaksono, M. S., Putra, I. F., Oenang, S. R. and Y., Septiandri, A. A., … Purwarianti, A. (2022). *NusaCrowd: Open Source Initiative for Indonesian NLP Resources*.

Cahyawijaya, S., Lovenia, H., Koto, F., Adhista, D., Dave, E., Oktavianti, S., Akbar, S., Lee, J., Shadieq, N., Cenggoro, T. W., Linuwih, H., Wilie, B., Muridan, G., Winata, G., Moeljadi, D., Aji, A. F., Purwarianti, A., & Fung, P. (2023). NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 921–945. https://doi.org/https://doi.org/10.18653/v1/2023.ijcnlp-main.60

Cahyawijaya, S., Winata, G. I., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). *IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation*. http://arxiv.org/abs/2104.08200

Cohn, A. C., & Ravindranath, M. (2014). Local languages in Indonesia: Language maintenance or language shift. *Linguistik Indonesia*, *32*(2), 131–148.

Dewi, N. P., & Ubaidi, U. (2020). Pos tagging Bahasa Madura dengan menggunakan algoritma Brill tagger. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(6), 1121–1128.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Koto, F., & Koto, I. (2020). Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation. In M. Le Nguyen, M. C. Luong, & S. Song (Eds.), *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 138–148). Association for Computational Linguistics. https://aclanthology.org/2020.paclic-1.17/

Novitasari, S., Tjandra, A., Sakti, S., & Nakamura, S. (2020). Cross-Lingual Machine Speech Chain for Javanese, Sundanese, Balinese, and Bataks Speech Recognition and Synthesis. In D. Beermann, L. Besacier, S. Sakti, & C. Soria (Eds.), *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 131–138). European Language Resources association. https://aclanthology.org/2020.sltu-1.18/

Nugraha, A. B., & Romadhony, A. (2023). Identification of 10 Regional Indonesian Languages Using Machine Learning. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, *7*(4), 2203–2214.

Purwarianti, A., Adhista, D., Baptiso, A., Mahfuzh, M., Sabila, Y., Adila, A., Cahyawijaya, S., & Aji, A. F. (2025). NusaDialogue: Dialogue Summarization and Generation for Underrepresented and Extremely Low-Resource Languages. *Proceedings of the Second Workshop in South East Asian Language Processing*, 82–100. https://aclanthology.org/2025.sealp-1.8/

Putra, O. V., Wasmanson, F. M., Harmini, T., & Utama, S. N. (2020). Sundanese Twitter Dataset for Emotion Classification. *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, 391–395. https://doi.org/10.1109/CENIM51130.2020.9297929

Putri, S. D. A., Ibrohim, M. O., & Budi, I. (2021). Abusive Language and Hate Speech Detection for Indonesian-Local Language in Social Media Text. In P. Meesad, Dr. S. Sodsee, W. Jitsakul, & S. Tangwannawit (Eds.), *Recent Advances in Information and Communication Technology 2021* (pp. 88–98). Springer International Publishing.

Sujaini, H., & Putra, A. B. (2024). Analysis of language identification algorithms for regional Indonesian languages. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *13*(2), 1741.

Sulistyo, D. A., Wibawa, A. P., Prasetya, D. D., & Ahda, F. A. (2023). LSTM-based machine translation for Madurese-Indonesian. *Journal of Applied Data Sciences*, *4*(3), 189–199.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *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*, 843–857. https://doi.org/https://doi.org/10.18653/v1/2020.aacl-main.85

Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Fung, P., Baldwin, T., Lau, J. H., Sennrich, R., & Ruder, S. (2023). NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 815–834. https://doi.org/https://doi.org/10.18653/v1/2023.eacl-main.57

Wongso, W., Lucky, H., & Suhartono, D. (2022). Pre-trained transformer-based language models for Sundanese. *Journal of Big Data*, *9*(1), 39. https://doi.org/10.1186/s40537-022-00590-7

Wongso, W., Setiawan, D. S., Limcorn, S., & Joyoadikusumo, A. (2025). NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural. In D. Wijaya, A. F. Aji, C. Vania, G. I. Winata, & A. Purwarianti (Eds.), *Proceedings of the Second Workshop in South East Asian Language Processing* (pp. 10–26). Association for Computational Linguistics. https://aclanthology.org/2025.sealp-1.2/