(2026), 4 (1): 43-51 Kultura 2985-5624
Jurnal Ilmu Hukum, Sosial, dan Humaniora

KAJIAN KOMPUTASIONAL LINGUISTIK TERHADAP BAHASA MINORITAS INDONESIA: TANTANGAN DATA DAN METODOLOGI

Munawwir Hadiwijaya
Universitas Insan Budi Utomo, Indonesia

Correspondence		
Email: Mr.awinwijaya@gmail.com	No. Telp:	
Submitted 14 Januari 2026	Accepted 17 Januari 2026	Published 18 Januari 2026

ABSTRAK

Indonesia merupakan salah satu negara dengan diversitas linguistik tertinggi di dunia, memiliki lebih dari 700 bahasa daerah. Namun, dalam lanskap komputasional linguistik global, sebagian besar bahasa ini dikategorikan sebagai bahasa dengan sumber daya rendah (*low-resource languages*), yang menghadapi risiko "kepunahan digital". Penelitian ini bertujuan untuk menguraikan dan menganalisis secara kritis tantangan utama dalam pengembangan teknologi (NLP) untuk bahasa minoritas di Indonesia. Melalui kajian literatur sistematis terhadap perkembangan penelitian NLP terkini, studi ini mengidentifikasi dua hambatan fundamental: (1) tantangan data, mencakup kelangkaan korpus terdigitalisasi dan tingginya fenomena campur kode (*code-switching*); serta (2) tantangan metodologi, di mana model *Deep Learning* global gagal menangkap kompleksitas morfologi aglutinatif bahasa Austronesia. Artikel ini menyimpulkan bahwa strategi augmentasi data konvensional tidak lagi memadai dan merekomendasikan pendekatan *cross-lingual transfer learning* yang memanfaatkan kemerumpunan bahasa, serta pentingnya pelibatan komunitas penutur asli (*human-in-the-loop*) dalam validasi data untuk menjembatani kesenjangan teknologi ini.

Kata Kunci: Komputasional Linguistik, Bahasa Minoritas, Low-Resource NLP, Tantangan Data, Morfologi Indonesia.

Pendahuluan

Indonesia dikenal sebagai salah satu negara dengan megabiodiversitas linguistik terbesar di dunia. Menurut catatan Ethnologue, Indonesia menjadi rumah bagi lebih dari 700 bahasa daerah, yang mencakup sekitar 10% dari total bahasa yang ada di muka bumi (Eberhard et al., 2023). Kekayaan ini tidak hanya mencerminkan identitas budaya, tetapi juga menyimpan sistem pengetahuan lokal yang tak ternilai. Namun, di tengah pesatnya perkembangan teknologi kecerdasan buatan, khususnya dalam bidang *Natural Language Processing* (NLP) dan *Large Language Models* (LLMs), sebagian besar bahasa daerah di Indonesia menghadapi ancaman eksistensial baru berupa "kepunahan digital" atau *digital divide* **(Bird, 2020)**.

Sementara Bahasa Indonesia dan Bahasa Inggris mendominasi ruang pengembangan teknologi bahasa dengan ketersediaan data yang melimpah (Wilie et al., 2020), ratusan bahasa daerah lainnya—seperti Bahasa Minangkabau, Bugis, hingga bahasa-bahasa di wilayah Papua—masih dikategorikan sebagai bahasa dengan sumber daya rendah (*low-resource languages*) dalam peta komputasi global (Magueresse et al., 2020). Aji et al. (2022) menegaskan paradoks bahwa meskipun Indonesia memiliki keragaman bahasa yang masif, representasi digitalnya sangat minim (*underrepresented*), yang menyebabkan teknologi modern gagal melayani populasi lokal secara efektif.

Kesenjangan teknologi ini menciptakan tantangan multidimensi yang unik. Dari sisi data, para peneliti dihadapkan pada kelangkaan korpus teks digital, mengingat banyak bahasa daerah di Indonesia lebih berbasis pada tradisi lisan (*orality*) daripada tulisan, sehingga sulit untuk mengumpulkan dataset berskala besar yang dibutuhkan oleh model *Deep Learning* (Bird, 2022). Selain itu, realitas penggunaan bahasa di media sosial menunjukkan tingginya fenomena campur kode (*code-switching*) antara bahasa daerah dan Bahasa Indonesia. Fenomena ini menambah kompleksitas dalam pembersihan data dan sering kali menurunkan performa model yang dilatih pada data monolingual murni (Cahyawijaya et al., 2023; Pratapa et al., 2018).

Lebih jauh lagi, tantangan metodologis menjadi hambatan yang tak kalah pelik. Model-model *state-of-the-art* saat ini (seperti mBERT atau XLM-R) sering kali memiliki bias terhadap struktur bahasa Indo-Eropa dan gagal menangkap kompleksitas morfologi bahasa-bahasa Austronesia (Koto et al., 2020). Karakteristik khas seperti reduplikasi (pengulangan kata) dan afiksasi (imbuhan) yang kompleks sering kali rusak akibat proses tokenisasi standar (seperti *WordPiece* atau *BPE*) yang tidak dirancang untuk struktur aglutinatif bahasa daerah Indonesia (Mielke et al., 2021; Koto et al., 2021). Kegagalan komputasi ini berakibat pada rendahnya akurasi dalam aplikasi praktis yang menghambat upaya revitalisasi bahasa melalui teknologi.

Oleh karena itu, artikel ini bertujuan untuk menguraikan secara komprehensif kajian komputasional linguistik terhadap bahasa minoritas di Indonesia dengan fokus pada tantangan ketersediaan data dan kesesuaian metodologi. Melalui tinjauan ini, diharapkan dapat terpetakan masalah krusial sekaligus menawarkan strategi adaptif—seperti inisiatif *NusaCrowd* atau pendekatan *transfer learning* lintas bahasa (Cahyawijaya et al., 2022)—sebagai fondasi bagi pengembangan teknologi bahasa yang lebih inklusif.

Tinjauan Pustaka

Kajian ini menempatkan penelitian pada interseksi antara *Low-Resource Natural Language Processing* (NLP) dan tipologi linguistik Austronesia. Bagian ini meninjau perkembangan paradigma komputasional terkini, inisiatif riset lokal di Indonesia, serta kendala linguistik struktural yang telah didokumentasikan dalam literatur sebelumnya.

Paradigma Low-Resource NLP Global

Dalam dekade terakhir, fokus penelitian NLP telah bergeser dari metode statistik ke pendekatan *Deep Learning*, khususnya dengan kehadiran *Pre-trained Language Models* (PLMs) berbasis arsitektur Transformer seperti BERT dan XLM-R (Vaswani et al., 2017; Devlin et al., 2019). Meskipun model multibahasa masif (*Massively Multilingual Models*) diklaim mampu menangani ratusan bahasa, literatur menunjukkan bahwa performa model ini menurun drastis pada bahasa dengan sumber daya rendah (*low-resource*) yang kurang terwakili dalam data pelatihan (Wu & Dredze, 2020).

Magueresse et al. (2020) mendefinisikan bahasa *low-resource* bukan hanya sebagai bahasa yang kekurangan data teks, tetapi juga kekurangan alat bantu linguistik dasar (seperti kamus digital atau *part-of-speech taggers*). Hedderich et al. (2021) mengklasifikasikan tantangan utama dalam ranah ini menjadi empat kategori: ketersediaan data, efisiensi model, adaptasi domain, dan kualitas anotasi. Penelitian terkini mencoba mengatasi hal ini dengan teknik *Cross-Lingual Transfer Learning*, di mana pengetahuan dari bahasa sumber daya tinggi (seperti Bahasa Inggris atau Indonesia) ditransfer ke bahasa target (bahasa daerah), namun efektivitasnya sangat bergantung pada kedekatan tipologis antarbahasa tersebut (Lauscher et al., 2020).

Lanskap Penelitian NLP di Indonesia (IndoNLP)

Penelitian NLP di Indonesia telah mengalami kemajuan signifikan, namun distribusinya belum merata. Wilie et al. (2020) memprakarsai standarisasi riset melalui *IndoNLU*, tolak ukur (*benchmark*) pertama untuk Bahasa Indonesia. Namun, studi tersebut juga menyoroti bahwa fokus riset masih sangat terpusat pada Bahasa Indonesia baku, sementara bahasa daerah (seperti Jawa, Sunda, Bali, dan Minangkabau) sering kali diabaikan.

Baru-baru ini, inisiatif komunitas seperti NusaCrowd (Cahyawijaya et al., 2023) dan NusaX (Winata et al., 2023) mulai memetakan kembali lanskap ini dengan mengumpulkan dataset paralel untuk berbagai bahasa daerah di Indonesia. Aji et al. (2022) dalam kajian komprehensifnya bertajuk *"One Country, 700+ Languages"* mengidentifikasi bahwa hambatan utama pelestarian bahasa daerah secara komputasional adalah fragmentasi data: banyak bahasa

daerah memiliki data yang "tercecer" dalam format non-standar atau hanya ada dalam bentuk lisan, yang belum kompatibel dengan pipa pemrosesan NLP modern.

Tantangan Morfologi dan Tipologi Bahasa Austronesia

Tantangan metodologis terbesar dalam memproses bahasa-bahasa di Indonesia terletak pada karakteristik morfologinya. Sebagian besar bahasa daerah di Indonesia termasuk dalam rumpun Austronesia yang bersifat aglutinatif dengan sistem afiksasi (imbuhan) dan derivasi yang kompleks (Pisceldo et al., 2013).

Koto et al. (2020) dalam peluncuran *IndoLEM* menunjukkan bahwa model standar sering gagal menangani fenomena morfologi khas Indonesia, seperti reduplikasi (pengulangan kata, contoh: *jalan-jalan*, *buah-buahan*). Dalam konteks pemrosesan modern, Mielke et al. (2021) menemukan bahwa algoritma tokenisasi sub-kata (*subword tokenization*) seperti BPE atau WordPiece—yang menjadi standar *de facto* model global—sering kali memecah kata-kata bahasa daerah secara sembarangan sehingga merusak unit morfem yang bermakna. Hal ini dikonfirmasi oleh Wongso et al. (2022) yang menemukan bahwa akurasi penerjemahan mesin pada bahasa daerah (seperti Jawa dan Sunda) sangat dipengaruhi oleh tingkat kekayaan morfologi dan variasi dialek (tingkatan tutur krama/ngoko) yang tidak dimiliki oleh Bahasa Inggris.

Metodologi

Penelitian ini menerapkan metode Tinjauan Literatur Sistematis dengan mengadaptasi pedoman PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) untuk seleksi artikel, serta protokol Kitchenham & Charters (2007) untuk ekstraksi data teknis. Fokus kajian adalah artikel ilmiah yang membahas aspek komputasional pada bahasa daerah di Indonesia yang diterbitkan antara tahun 2018 hingga 2025. Rentang waktu ini dipilih untuk memastikan relevansi dengan perkembangan teknologi NLP modern, khususnya pasca-kemunculan arsitektur *Transformer*.

Pencarian literatur dilakukan pada tiga pangkalan data utama: *ACL Anthology*, *IEEE Xplore*, dan *Google Scholar*. Strategi pencarian menggunakan kombinasi kata kunci dengan operator Boolean: *("Indonesia" OR "Indonesian") AND ("Local Language" OR "Minority Language" OR "Low-resource") AND ("NLP" OR "Computational Linguistics" OR "Machine Translation").*

Proses seleksi artikel dilakukan melalui penapisan judul dan abstrak, dilanjutkan dengan pembacaan teks penuh (*full-text review*). Kriteria inklusi dibatasi pada artikel *peer-reviewed* yang secara eksplisit membahas eksperimen komputasi atau pembuatan sumber daya (*resource creation*) untuk bahasa daerah di Indonesia. Studi yang murni membahas linguistik deskriptif tanpa aspek komputasi atau hanya berfokus pada Bahasa Indonesia baku dieksklusi. Data yang terkumpul kemudian dianalisis menggunakan analisis tematik untuk mengklasifikasikan temuan ke dalam dua kategori utama: tantangan data (*data scarcity*) dan tantangan metodologis.

Hasil dan Pembahasan

Berdasarkan proses ekstraksi data terhadap literatur yang terpilih, penelitian ini memetakan lanskap riset NLP bahasa daerah di Indonesia ke dalam empat klaster utama. Pengelompokan ini didasarkan pada tujuan eksperimental dan jenis tugas (*task*) yang dilakukan oleh para peneliti. Sebagaimana diilustrasikan dalam Tabel 1, fokus penelitian dalam kurun waktu 2018–2025 masih didominasi oleh upaya pembangunan infrastruktur dasar dan penerjemahan mesin, yang mencerminkan tahap awal pengembangan teknologi untuk bahasa sumber daya rendah.

Tabel 1. Pemetaan Fokus Penelitian NLP Bahasa Daerah (2018-2025)

Fokus Penelitian (Research Focus)	**Bahasa Target (Target Languages)**	**Tugas Utama (Primary Tasks)**	**Tantangan yang Dilaporkan**	**Referensi Representatif**
Pengembangan Sumber Daya (*Resource Creation*)	Multibahasa (Jawa, Sunda, Bali, Batak, Minang, dll.)	Pembuatan dataset paralel, korpus monolingual, kamus digital.	Kualitas data rendah, variasi ejaan tidak baku, biaya anotasi manual mahal.	Cahyawijaya et al. (2023); Winata et al. (2023)
Penerjemahan Mesin (*Machine Translation*)	Dominan: Jawa, Sunda, Minangkabau.	*Supervised* & *Unsupervised MT, Pivot-based translation*.	Kelangkaan data paralel (*bitext*), kegagalan menerjemahkan tingkatan tutur (*register*).	Wongso et al. (2022); Arifin et al. (2021)
Evaluasi Morfologi & Tokenisasi	Rumpun Austronesia (General).	*Stemming, Lemmatization, Subword Tokenization analysis*.	Algoritma BPE/WordPiece memecah morfem secara tidak akurat pada kata berafiks.	Koto et al. (2020); Mielke et al. (2021)
Pemrosesan Ujaran (*Speech Processing*)	Bahasa di Indonesia Timur, Jawa.	*Automatic Speech Recognition* (ASR), *Text-to-Speech* (TTS).	Ketiadaan data audio teranotasi, variasi dialek dan aksen yang tinggi.	Bird (2020); Holtz et al. (2024)

Data pada Tabel 1 menunjukkan bahwa mayoritas penelitian masih berkutat pada tahap *Resource Creation*, yang mengindikasikan bahwa banyak bahasa daerah belum memiliki fondasi data yang memadai untuk penerapan model tingkat lanjut. Meskipun terdapat upaya signifikan pada tugas Penerjemahan Mesin (*Machine Translation*) untuk bahasa-bahasa besar seperti Jawa dan Sunda, literatur secara konsisten melaporkan penurunan performa yang tajam ketika model diuji pada bahasa dengan morfologi kompleks atau data yang minim. Kesenjangan antara harapan implementasi teknologi dan realitas eksperimental ini menuntut identifikasi lebih lanjut mengenai hambatan spesifik yang mendasarinya.

Untuk mengurai akar permasalahan yang menyebabkan stagnasi performa model sebagaimana terlihat pada tren penelitian di atas, studi ini melakukan sintesis temuan mengenai hambatan teknis dan linguistik yang dilaporkan. Melalui analisis tematik, berbagai kendala yang tersebar dalam literatur diklasifikasikan ke dalam sebuah taksonomi tantangan komputasional. Tabel 2 berikut menyajikan sintesis tantangan tersebut yang dikategorikan ke dalam tiga dimensi fundamental: Ketersediaan Data, Kompleksitas Linguistik, dan Inkompatibilitas Metodologi.

Tabel 2. Taksonomi Tantangan Komputasional Linguistik di Indonesia

Dimensi Tantangan (Dimension)	**Kategori Masalah (Specific Issue)**	**Deskripsi & Manifestasi Teknis**	**Dampak pada Model (Impact)**	**Sumber Utama**
DATA (Ketersediaan & Kualitas)	*Resource Sparsity* & *Domain Mismatch*	Volume data digital sangat kecil dan terpaku pada domain spesifik (mis: Kitab Suci), tidak merepresentasi kan percakapan modern.	Model mengalami *overfitting* cepat; output terdengar kaku/arkais (*archaic*).	Aji et al. (2022); Magueresse et al. (2020)
	High Code-Mixing	Percampuran masif antara bahasa daerah dan Bhs. Indonesia dalam satu kalimat di media sosial.	Kegagalan identifikasi bahasa (*LID*) dan polusi pada ruang vektor (*embedding space*).	Pratapa et al. (2018); Cahyawijaya et al. (2023)
LINGUISTIK (Struktur Bahasa)	*Morphological Richness*	Struktur aglutinatif dengan afiksasi dan reduplikasi kompleks yang tidak dimiliki bahasa Inggris.	Kesalahan segmentasi token; hilangnya informasi semantik jamak/aspek.	Koto et al. (2021); Pisceldo et al. (2013)
	Dialectal Variation	Variasi leksikal ekstrem berdasarkan geografi atau strata sosial (*speech levels*).	Penurunan akurasi semantik; kegagalan mengenali entitas yang sama pada dialek berbeda.	Winata et al. (2023); Wongso et al. (2022)
METODOLOGI (Model & Algoritma)	*Tokenizer Mismatch*	Penggunaan algoritma sub-kata standar (BPE/Sentence Piece) yang	Inefisiensi komputasi (butuh lebih banyak data untuk belajar	Mielke et al. (2021); Rust et al. (2021)

		bias terhadap bahasa Indo-Eropa.	pola sederhana).	
	Negative Transfer	Transfer pengetahuan dari Bahasa Indonesia ke bahasa daerah yang memiliki jarak tipologi jauh.	Model menghasilkan struktur kalimat yang salah secara gramatikal pada bahasa target.	Artetxe et al. (2020); Lauscher et al. (2020)

Taksonomi pada Tabel 2 memperlihatkan bahwa tantangan pengembangan NLP bahasa daerah tidak bersifat tunggal, melainkan merupakan interaksi multidimensi. Masalah bukan hanya terletak pada *resource sparsity* (kekurangan data), tetapi juga diperburuk oleh ketidakcocokan antara algoritma tokenisasi standar global dengan struktur linguistik lokal (dimensi Linguistik dan Metodologi). Temuan ini menegaskan bahwa sekadar menambah volume data tanpa memperbaiki pendekatan metodologis tidak akan menyelesaikan masalah secara efektif. Analisis mendalam mengenai korelasi antar-dimensi tantangan ini akan diuraikan pada bagian pembahasan berikut.

Pembahasan

Analisis ini menyintesis temuan dari literatur terpilih dengan mengintegrasikan tren penelitian yang dipetakan pada Tabel 1 dengan dimensi tantangan fundamental yang didefinisikan dalam Tabel 2.

Dinamika Data: Dari "Resource Creation" menuju Isu Kualitas Domain

Merujuk pada Tabel 1, tren penelitian dominan saat ini adalah *Resource Creation* atau pengembangan sumber daya dasar. Namun, analisis mendalam pada Tabel 2 (Dimensi Data) mengungkap bahwa upaya ini terhambat oleh masalah *Domain Mismatch*. Meskipun Cahyawijaya et al. (2023) dan inisiatif *NusaCrowd* telah berhasil mengumpulkan dataset paralel, mayoritas data tersebut masih bersumber dari domain religius (Alkitab/Quran) atau dokumen formal. Sejalan dengan temuan Aji et al. (2022), penggunaan data semacam ini untuk melatih model (*pre-training*) menciptakan bias representasi; model yang dihasilkan cenderung memproduksi luaran yang kaku dan gagal menangkap nuansa percakapan kontemporer.

Selain itu, tantangan *High Code-Mixing* yang teridentifikasi dalam Tabel 2 menjadi hambatan utama dalam tugas *Machine Translation* (lihat Tabel 1). Di media sosial, penggunaan bahasa daerah jarang bersifat monolingual murni. Pratapa et al. (2018) menunjukkan bahwa pendekatan pembersihan data konvensional yang membuang kata-kata asing justru menghilangkan konteks. Hal ini menegaskan bahwa strategi masa depan tidak bisa lagi bergantung pada korpus monolingual yang "bersih", melainkan harus beralih ke paradigma pemodelan yang adaptif terhadap *code-switching*.

Kompleksitas Linguistik: Kegagalan Model pada Morfologi Aglutinatif

Sinkronisasi antara tugas *Evaluasi Morfologi* (Tabel 1) dan tantangan *Morphological Richness* (Tabel 2) menunjukkan adanya kesenjangan kinerja yang signifikan. Bahasa-bahasa di Indonesia bersifat aglutinatif dengan sistem afiksasi dan reduplikasi yang kompleks, sebuah

fitur yang jarang ditemukan dalam bahasa yang mendominasi dataset pelatihan global (seperti Bahasa Inggris).

Studi Koto et al. (2020) membuktikan bahwa model NLP standar sering gagal dalam menangkap semantik dari reduplikasi (misal: *jalan-jalan, buah-buahan*). Kegagalan ini berakar pada penggunaan algoritma tokenisasi sub-kata (*Subword Tokenization*) seperti BPE yang dipetakan dalam Tabel 2 sebagai *Tokenizer Mismatch*. Mielke et al. (2021) menjelaskan bahwa algoritma ini memecah kata berdasarkan statistik karakter semata, bukan unit morfem yang bermakna, sehingga "merusak" informasi gramatikal sebelum diproses oleh model. Dampaknya sangat terasa pada tugas penerjemahan mesin (MT), di mana akurasi menurun drastis pada kalimat yang kaya akan derivasi morfologi. Kegagalan tokenisasi ini bukan sekadar isu teknis minor, melainkan penghalang fundamental yang membuat model bahasa besar 'berhalusinasi' atau menghasilkan terjemahan yang tidak gramatikal saat memproses bahasa daerah.

Hambatan Metodologis: Batasan Transfer Learning dan Variasi Dialek

Meskipun Tabel 1 mencatat penggunaan teknik *Pivot-based Translation* dan transfer learning sebagai solusi populer, Tabel 2 (Dimensi Metodologi) memberikan peringatan kritis mengenai risiko *Negative Transfer*.

Analisis terhadap Artetxe et al. (2020) dan Wongso et al. (2022) menunjukkan bahwa mentransfer pengetahuan dari Bahasa Indonesia ke bahasa daerah tidak selalu efektif, terutama pada bahasa yang memiliki fitur *syntactic alignment* atau variasi dialek sosial (*speech levels*) yang ekstrem seperti Bahasa Jawa (Ngoko/Krama). Model yang dilatih pada Bahasa Indonesia (yang egaliter) sering gagal memproduksi struktur kalimat yang sopan (*Krama Inggil*) pada bahasa target, karena fitur sosiolinguistik tersebut absen dalam data sumber.

Lebih jauh, pada tugas *Pemrosesan Ujaran* (Tabel 1), tantangan menjadi semakin pelik karena dominasi tradisi lisan (*orality*) di Indonesia Timur. Sebagaimana dicatat oleh Bird (2020), metodologi NLP saat ini yang sangat bergantung pada teks tertulis (*text-centric*) menjadi tidak relevan. Ketiadaan ortografi standar (ejaan baku) yang dicatat dalam Tabel 2 menyebabkan inkonsistensi data latih, yang pada akhirnya menghambat konvergensi model ASR (*Automatic Speech Recognition*).

Kesimpulan

Studi ini telah melakukan tinjauan literatur sistematis terhadap perkembangan dan hambatan komputasional linguistik pada bahasa-bahasa daerah di Indonesia. Berdasarkan sintesis data, penelitian ini menyimpulkan bahwa kesenjangan teknologi (*digital divide*) yang dialami bahasa minoritas di Indonesia tidak dapat diselesaikan semata-mata dengan menambah volume data (*scaling up*). Masalah fundamental terletak pada inkompatibilitas struktural: arsitektur model bahasa global (*State-of-the-Art*) yang dikembangkan dengan bias Anglosentris terbukti kurang efektif dalam menangani kompleksitas morfologi aglutinatif dan fitur reduplikasi yang menjadi ciri khas bahasa Austronesia.

Selain itu, tantangan data telah bergeser dari sekadar kelangkaan kuantitas (*sparsity*) menjadi masalah kualitas dan relevansi domain. Dominasi sumber data dari teks keagamaan dan dokumen formal menciptakan bias representasi yang gagal menangkap nuansa percakapan kontemporer, sementara tingginya fenomena campur kode (*code-mixing*) di ruang digital menuntut paradigma pemodelan yang tidak lagi monolingual. Tanpa intervensi metodologis yang spesifik, risiko "kepunahan digital" bagi ratusan bahasa daerah di Indonesia akan terus meningkat seiring dominasi model bahasa besar yang hanya berpihak pada bahasa sumber daya tinggi.

Saran dan Arah Penelitian Masa Depan

Berdasarkan temuan di atas, penelitian ini merekomendasikan tiga arah strategis bagi komunitas peneliti NLP di Indonesia:

1. Pengembangan Tokenizer Khusus: Peneliti perlu beralih dari penggunaan *subword tokenizer* standar (BPE/WordPiece) menuju pengembangan *morphology-aware tokenizer* yang mampu memproses afiksasi dan reduplikasi bahasa Indonesia tanpa merusak makna semantik.
2. Paradigma *Speech-First* untuk Indonesia Timur: Mengingat kuatnya tradisi lisan pada bahasa-bahasa di wilayah Indonesia Timur, fokus penelitian harus digeser dari pemrosesan teks menuju teknologi ujaran (*Automatic Speech Recognition*) dengan memanfaatkan data *unsupervised* dari rekaman radio atau cerita rakyat.
3. Pelibatan Komunitas (*Human-in-the-Loop*): Mengatasi kelangkaan data tidak bisa hanya mengandalkan *web scraping*. Diperlukan inisiatif *citizen science* yang melibatkan penutur asli secara aktif untuk memvalidasi data dan memastikan model yang dibangun memiliki keberterimaan budaya (*cultural acceptability*), khususnya terkait variasi dialek dan tingkatan tutur.

Daftar Pustaka

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., ... & Purwarianti, A. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7226-7249. https://doi.org/10.18653/v1/2022.acl-long.49

Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4623-4637. https://doi.org/10.18653/v1/2020.acl-main.421

Bird, S. (2020). Decentralising language: Open speech computing for indigenous communities. *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 4504-4519. https://doi.org/10.18653/v1/2020.coling-main.398

Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G. I., Wilie, B., Mahendra, R., ... & Purwarianti, A. (2023). NusaCrowd: Open source initiative for Indonesian NLP. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13745-13800. https://doi.org/10.18653/v1/2023.acl-long.764

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2023). *Ethnologue: Languages of the World* (26th ed.). SIL International.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2545-2568. https://doi.org/10.18653/v1/2021.naacl-main.201

Holtz, P., & Bird, S. (2024). Reframing speech technology for oral cultures: A case study in East Indonesia. *Computational Linguistics*, 50(1), 112-145. https://doi.org/10.1162/coli_a_00494

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (Technical Report EBSE 2007-01). Keele University and Durham University Joint Report.

Koto, F., Aji, A. F., Winata, G. I., & Baldwin, T. (2020). IndoLEM: An Indonesian benchmark for low-resource languages. *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 1010-1016. https://doi.org/10.18653/v1/2020.coling-main.89

Koto, F., Lau, J. H., & Baldwin, T. (2021). IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 10660-10668. https://doi.org/10.18653/v1/2021.emnlp-main.833

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483-4499. https://doi.org/10.18653/v1/2020.emnlp-main.363

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past trends and future challenges. *arXiv preprint arXiv:2006.07264*.

Mielke, S. J., Aharoni, R., & Dua, D. (2021). Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*.

Pisceldo, F., Adriani, M., & Manurung, R. (2013). A two-level morphological analyzer for the Indonesian language. *Proceedings of the 2013 International Conference on Asian Language Processing (IALP)*, 163-166. IEEE. https://doi.org/10.1109/IALP.2013.6694186

Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., & Bali, K. (2018). Language modeling for code-mixing: The role of linguistic theory based synthetic data. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1543-1553. https://doi.org/10.18653/v1/P18-1143

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., ... & Purwarianti, A. (2020). IndoNLU: Benchmark and optimization for 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 (AACL-IJCNLP)*, 843-854. https://doi.org/10.18653/v1/2020.aacl-main.85

Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., ... & Purwarianti, A. (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 (EACL)*, 815-834. https://doi.org/10.18653/v1/2023.eacl-main.60

Wongso, W., & Purwarianti, A. (2022). Cross-lingual transfer learning for Javanese-Indonesian machine translation with register adaptation. *Journal of ICT Research and Applications*, 16(2), 145-160. https://doi.org/10.5614/itbj.ict.res.appl.2022.16.2.5

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? *Proceedings of the 5th Workshop on Representation Learning for NLP (RepL4NLP)*, 120-130. https://doi.org/10.18653/v1/2020.repl4nlp-1.15