

Salsabila Zahirah Pranida

📍 Indonesia | 📞 +62 853-2672-5956 | ✉ irasalsabila@gmail.com
🌐 [Salsabila Zahirah Pranida](#) | 🗣 [irasalsabila](#) | 🌐 [Website](#)

PROFILE

Researcher specializing in low-resource and dialectal languages, with expertise across text and speech technologies. Experienced in dataset development, model fine-tuning, and evaluation for multilingual and multicultural contexts. Passionate about advancing language technology that respects cultural and linguistic diversity.

EDUCATION

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) 2024 - 2026
M.Sc in Natural Language Processing Abu Dhabi, UAE

- Research Focus: Computational Social Science, Multilingual and Dialectal, Low-Resource Language
- Thesis: Indodialect—A Multi-Dialect Conversational Dataset for Indonesian Local Languages

Universitas Islam Indonesia 2018 - 2022
S.Kom (Bachelor in Informatics) Indonesia

- Final Project: Customer Satisfaction Sentiment Analysis on Courier Services Using BiLSTM and BiGRU

PROJECTS

IndoDialect—A Multi-Dialect Conversational Dataset for Indonesian Local Languages 2025

- Built a conversational NLP dataset containing 14.2K multi-turn dialogues across 11 dialects and 4 Indonesian local languages.
- Developed and managed a large-scale data annotation and quality assurance pipeline with native speakers for dialect translation and dialogue generation.
- Evaluated multilingual and statistical machine learning models for dialect identification and machine translation in low-resource language settings.

Visiting Project: Persona and Gender Effects in Indonesian Toxicity Classification 2025

- Investigated the impact of evaluator demographics and simulated personas on toxicity classification in Indonesian online discourse.
- Analyzed bias and fairness patterns in large language models using persona-conditioned prompting and the IndoDiscourse dataset.
- Conducted research under Dr. Daniel Herschcovich in collaboration with the University of Copenhagen, supported by the Danish Data Science Academy (DDSA) Visit Grant.

ASR Under Noise: Exploring Robustness for Sundanese and Javanese 🌐 🌐 2025

- Evaluated Whisper-based automatic speech recognition (ASR) models for Javanese and Sundanese under diverse noise and signal-to-noise ratio (SNR) conditions.
- Improved ASR robustness using fine-tuning, SpecAugment, and synthetic noise generation, achieving stronger recognition performance across noisy environments.
- Conducted language-specific error analysis for low-resource speech recognition; accepted to WinNLP 2025 at EMNLP 2025 and awarded *Best Paper Award 2025*

Culturally-Nuanced Story Generation for Reasoning in Low-Resource Languages 🌐 🌐 2024

- Built Javanese and Sundanese Story Cloze datasets for culturally grounded commonsense reasoning in low-resource languages.
- Evaluated native-written, machine-translated, and LLM-generated stories for narrative coherence, cultural fidelity, and reasoning quality.
- Fine-tuned BERT-based models on synthetic training data for story understanding tasks; accepted to MRL 2025 at EMNLP 2025.

- [C.1] Saeed Almheiri*, Bilal* Elbouardi, **Salsabila Zahirah Pranida***, ..., Suma Bhat, Fajri Koto. 2026. **Multilingual Idioms in Sentences and Conversations Across High-, Medium-, and Low-Resource Languages**. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12363–12389, San Diego, California, United States. Association for Computational Linguistics.
- [C.2] Dilara Torunoğlu-Selamet, Dogukan Arslan, **Salsabila Zahirah Pranida**, ..., Zahra Saaberi, Zhuohan Xie. 2026. **A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding**. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, (pp. 9434–9448). European Language Resources Association (ELRA).
- [R.1] Samuel Cahyawijaya, Peerat Limkonchotiwat, **Salsabila Zahirah Pranida**, ..., Cho Chan Myei Oo, Hee Ming Shan. 2026. **Anthropogenic Regional Adaptation in Multimodal Vision-Language Model**. *arXiv preprint arXiv:2604.11490*.
- [R.2] Tyler A. Chang, Catherine Arnett, ..., **Salsabila Zahirah Pranida**, ..., Jiayi Wang, David Ifeoluwa Adelani. 2025. **Global PIQA: Evaluating Commonsense Reasoning Across 100+ Languages and Cultures**. *arXiv preprint arXiv:2510.24081*.
- [C.3] **Salsabila Zahirah Pranida***, Rifo Ahmad Genadi*, Muhammad Cendekia Airlangga*, and Shady Shehata. 2025. **ASR Under Noise: Exploring Robustness for Sundanese and Javanese**. In *Proceedings of the 9th Widening NLP Workshop*, pages 87–99, Suzhou, China. Association for Computational Linguistics. **Best Paper Award at WinNLP 2025**
- [C.4] **Salsabila Zahirah Pranida***, Rifo Ahmad Genadi*, and Fajri Koto. 2025. **Culturally-Nuanced Story Generation for Reasoning in Low-Resource Languages: The Case of Javanese and Sundanese**. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 369–384, Suzhou, China. Association for Computational Linguistics.
- [C.5] Samuel Cahyawijaya, Holy Lovenia, ..., **Salsabila Zahirah Pranida**, ..., Börje F. Karlsson, Peerat Limkonchotiwat. 2025. **Crowdsourcing, Crawl, or Generate? Creating SEA-VL, a Multicultural Vision-Language Dataset for Southeast Asia**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.
- [J.1] **Salsabila Zahirah Pranida** and Arrie Kurniawardhani. 2022. **Sentiment Analysis of Expedition Customer Satisfaction using BiGRU and BiLSTM**. *Indonesian Journal of Artificial Intelligence and Data Mining*, Vol. 5, No. 1, pages 44–53.
- [C.6] **Salsabila Zahirah Pranida***, Muhammad Rizhan Ridha*, Dimas Danu Budi Pratikto*, Ari Farhan Nurihsan*, Salma Aufa Azaliarahma, Dhomas Hatta Fudholi*, and Ari Sujarwo. 2019. **Jendela Jogja: Portal Promosi dan Pemasaran Pariwisata Yogyakarta**. In *Proceedings of the Seminar Nasional Aplikasi Teknologi Informasi (SNATI 2019)*, Yogyakarta, Indonesia.

Authors marked with * contributed equally.

WORK EXPERIENCE

PT Dua Empat Tujuh [🌐]

Aug 2022 - Jul 2024

Data Scientist

Indonesia

- Enhanced benchmarking methodologies using ROUGE and BLEU metrics to assess media analytics functions, including document summarization and keyword extraction.
- Refined revenue and expense forecasting models for a telecommunications company, improving financial projections, cost estimation, and budgeting accuracy.
- Delivered a rule-based Named Entity Recognition (NER) system for Indonesians to identify entities such as organizations, landmarks, persons, and time references.

PT Dua Empat Tujuh [🌐]

Jun 2021 - Jun 2022

Data Scientist Intern

Indonesia

- Developed and fine-tuned a sentiment analysis model for a Media Analytics product using BiLSTM, GRU, BiGRU, and FastText methodologies.
- Engineered an API to detect sentiment in ambiguous text, incorporating POS tagging techniques for enhanced accuracy.
- Streamlined the dataset collection process by web scraping over 20,000 tweets for sentiment analysis using Python.
- Established and standardized dataset tagging guidelines to ensure consistency and precision in analysis.

ACHIEVEMENTS & AWARDS

OpenAI Mini-Grant Recipient2025

Dept. of NLP, MBZUAI

- Awarded a \$1,500 internal mini-grant from MBZUAI’s NLP Department to support thesis work on multilingual and dialectal language modeling.

Catalyst Grant Program Recipient2025

Cohere

- Awarded a \$1,000 Catalyst Grant Program by Cohere to support thesis work on multilingual and dialectal language modeling.

DDSA Visit Grant RecipientJun – Aug 2025

Danish Data Science Academy

- Awarded a 15,000 DKK DDSA Visit Grant to conduct collaborative research at the University of Copenhagen, working under the supervision of Dr. Daniel Herschovich.

Bangkit 2021 Scholarship - Machine Learning PathJan 2021

Google, Tokopedia, Gojek, and Traveloka

- Selected as an awardee for the Bangkit 2021 program, focusing on AI and machine learning applications.
- Graduated as a Distinct Graduate, recognized among the top 10% of participants.
- Collaborated on a team project, **Elevate!**, focusing on product input classification and data processing using BiGRU, achieving 84.7% accuracy.

SKILLS

Machine Learning & Deep Learning: PyTorch, TensorFlow, HuggingFace Transformers, Scikit-learn, Keras

Natural Language Processing: Text & Speech Processing, LLM Fine-tuning, Multilingual NLP, Dataset Construction, Evaluation & Benchmarking

Languages: English (Fluent), Indonesian (Native), Javanese (Native)

VOLUNTEER EXPERIENCE

VolunteerJan 2025

COLING 2025 – The 31st International Conference on Computational Linguistics

- Supported the registration desk, assisted during poster sessions, and served as a floating volunteer throughout the conference.
- Helped coordinate attendee logistics and provided guidance to participants, contributing to the smooth execution of the event.

Course TesterOct 2022 - Jul 2024

DeepLearning.AI

- Reviewed new courses from DeepLearning.AI, providing feedback to improve content quality and user experience.
- Identified and corrected typos in course labs and recommended enhancements to increase clarity and effectiveness.