Juan Tampubolon

Computational Linguist

Tjuan-PER ☎ +1 (236) 788-4077
juantampubolon ✉ jf.danieltampubolon@gmail.com

Education

University of British Columbia Aug 2024 – Jun 2025
MDS, Master of Data Science in Computational Linguistics *Vancouver, BC, Canada*

University of Michigan Aug 2019 – Apr 2022
BSc, Bachelor of Science in Honors Linguistics *Ann Arbor, MI, USA*

Green River College Sep 2016 – Jun 2019
AS, Associate of Science in Computer Science with High Honors *Auburn, WA, USA*

Skills

Programming and Databases: Python, R, Javascript, Bash, HTML5, CSS, MongoDB, PostGreSQL
Machine Learning and NLP: NLTK, SciKit-Learn, SpaCy, Pytorch, Dplyr, BeautifulSoup, HuggingFace, LangChain
Data Analysis, Visualization, and Tools: Pandas, NumPy, Matplotlib, Altair, ggplot2
Laguages: Indonesian (native), English (native), French (intermediate), Spanish (intermediate)

Experience

NLP Research Engineer Apr 2025 - Jun 2025
Ocarina Studios - Final Project *Vancouver, BC, Canada*

- Led development and refinement of Named Entity Recognition (NER) components within a question generation pipeline, integrating SpaCy noun chunking, WordNet, and NLTK NER to improve semantic coverage beyond standard entity classes.
- Developed comprehensive documentation aligning technical workflows with stakeholder needs, improving transparency, onboarding, and long-term maintainability.
- Conducted pipeline diagnostics to identify and address inefficiencies related to tagging redundancy, entity overlap, and semantic drift in existing classification systems.
- Built evaluation tools measuring cosine similarity, tagging consistency, and tag coverage to monitor model quality and support system scalability.
- Integrated external knowledge databases (Wikidata) to supplement generated questions with relevant background facts and context.
- Collaborated with teammates responsible for classification models, aligning extraction outputs to downstream model needs.
- Served as key liaison coordinating between academic supervisors, technical teammates, and industry stakeholders to align deliverables and technical direction during a two-month research sprint.

Primary Homeroom Teacher Jul 2022 – Jul 2024
Sekolah Victory Plus *Bekasi, West Java, Indonesia*

- Developed immersive learning experiences that increased connection across disciplines, resulting in higher levels of comprehension and engagement observed through classroom participation, discussion, student behavior reports, and student-led initiatives
- Organized out-of-classroom learning activities and expert visitations that enriched students' understanding of different subjects
- Mentored 6th-grade students throughout their end-of-program project culminating in a book which sold over 50 copies
- Successfully coached the Secondary Schools' debate club, resulting in the acquisition of several medals and students' qualification for the following rounds including the Tournament of Champions held at Yale University

Projects

ReDox: Reddit Toxicity Dataset & Web App Mar 2025
Class Project

- Prdouced a web app to allow users to filter, search, and analyze toxic comments, making it an educational tool for online discourse.
- Scraped & collected 1,000+ Reddit comments using Python (BeautifulSoup), storing them in a CSV file.
- Preprocessed, classified data for their toxicity levels (Neutral, Offensive, Abusive, Hate Speech) and identified targets.
- Integrated FastAPI as backend to serve data and built an interactive front-end with HTML, CSS, and JavaScript.

Virtual Assistant with RAG for Cooking Apr 2025
Class Project

- Developed an end-to-end virtual cooking assistant chatbot that incorporates Natural Language Understanding (NLU) and Retrieval Augmented Generation (RAG)
- Built a vector database to optimize search relevance and query speed

Product Review Sentiment Analysis and Detoxification Mar 2025
Class Project

- Developed an LLM-based NLP pipeline for sentiment analysis, toxicity detection, and toxic style transfer, integrating models like LangChain, DetoxLLM, and Granite 3.0-2b-instruct.
- Implemented multilingual support using NLLB-200 and Toucan translation models, enabling processing of non-English text.
- Optimized model performance and evaluation using F1-score, and designed explainability components for transparency.
- Engineered and documented an end-to-end agentic workflow, leveraging sequential chaining and dynamic model routing