Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity

AACL' 23 Tutorial (Bali/Indonesia)

Organizers

Alham Fikri Aji (MBZUAI), Jessica Zosa Forde (Brown University), Alyssa Marie Loo (Brown University), Lintang Sutawika (EleutherAI), Skyler Wang (UC Berkeley, Meta AI), Genta Indra Winata (Bloomberg), Zheng-Xin Yong (Brown University), Ruochen Zhang (Brown University), A. Seza Doğruöz (Universiteit Gent), Yin Lin Tan (Stanford University, NUS), Jan Christian Blaise Cruz (Samsung R&D Institute Philippines)

Introduction

- 1000+ languages in SEA
- ~700 alone is in Indonesia
- Diverse in culture
- yet, underrepresented in NLP research

PS: South-east Asia != South Asia

Agenda

- Opening
- Introduction to Multilingualism
- NLP Resources
  - Indonesia
  - Philippines
  - Singapore
  - Malaysia
- Can Multilingual Generative LLMs code-mix with SEA languages?
- Community Efforts
- Panel
- Q&A

Introduction to Multilingualism

1 November 2023

Multilingualism (Doğruöz et al, 2021)

- Multilinguals: Individuals who speak/communicate with more than one language in daily life.
- Multilingualism: Study of language use by multilingual individuals & within multilingual communities.
- There are millions of multilinguals around the world (e.g., Europe, Africa, India). South East Asian countries host many languages and the area is linguistically highly diverse.

Code-switching (Doğruöz et al, 2021)

- It is common for multilingual speakers to switch between and across languages/dialects they speak.
- Code-switching is commonly observed among multilinguals depending on the context and their conversational partners.
- Due to typological similarities and differences between languages, code-switching can also take many different forms (e.g., single words vs. multi-word phrases).
- In linguistics, there are many theories and classifications about code-switching categories.

Code-switching (Doğruöz et al, 2021)

Example of Code-switching Research in Europe:

- C-S research among bilingual and multilingual children
- C-S research between standard and minority languages
- C-S research in immigrant contexts

Code-switching (Doğruöz et al, 2021)

Examples of Code-switching Research in India:

- Hindi-English C-S is well-known but there are many other languages in India.
- Doğruöz et al (2021) provides a variety of code-switching examples between different languages in India (e.g. Konkani-Kannada, English-Assamese-Bodo, Bengali-English, Malayalam-English, Kannada-English) and contexts (e.g. mass media, educational settings, daily life) based on available literature.
- In addition, there are examples of social & cultural factors (e.g., prestige) which facilitate and/or inhibit code-switching across Indian languages.

Multilingualism in South East Asia (Doğruöz et al, 2021)

- South East Asia (SEA) is a linguistically diverse area hosting many languages and dialects.

- According to Florey (2010), 10% of the world languages are hosted in Indonesia as one of the SEA countries.

- In addition, colonialism brought the European languages (e.g., English, Dutch, Portuguese, Spanish, French) in contact with the local languages as well (Chin & Cavallaro, 2019).

Multilingualism in South East Asia (Doğruöz et al, 2021)

- Linguistic diversity, composition of the communities and policies toward multilingualism are not the same across countries and regions (Chin & Cavallaro, 2019).
- In addition to the local languages, English is also used in education, trade and diplomacy across the region.

Issues about Analyzing Multilingualism

(Doğruöz & Sitaram, 2022)

- Most of the languages spoken in South East Asia are low resource. In addition, the speakers of these languages are multilingual as well.
- There is a need to analyze the attitudes, power & prestige hierarchies in low resource language communities before building up computational tools and resources blindly.
- Not all low resource language speakers may be willing to speak these languages & teach them to their children due to difficulties in employability & low social status associated with these languages in their communities.

Issues about Analyzing Multilingualism
(Doğruöz & Sitaram, 2022)

- It is common for many low resource language (LRL) speakers to switch across languages in daily communication but this is not always taken into account for developing language technologies for these communities and there is an inclination to collect unnatural monolingual data which may not exist in real-life communication.
- Too much cleaning os “LRLs” leads to system failures + low adoption rates by the community members as well as waste of time, energy and resources while building these systems.

Issues about Data
(Doğruöz & Sitaram, 2022)

- Very clean data is good for the accuracy of the systems but they may not represent the real world communication (both monolingual & multilingual).
- It may be possible to improve these flaws for HRLs over time but not so much for LRLs.
- LRLs are not small size HRLs. Applying the same methods for LRLs may not be useful in the same way.

Issues about Data Collection (Doğruöz & Sitaram, 2022)

- Existing HRL data is not representative for LRLs and approximations are not helpful.
- Ideally, LRL data should be collected directly from its speakers. There is not enough research about the validity of synthetic data sets.
- Although they are smaller in size, there may be linguistic data sets collected for linguistic research already.
- Field-work: Ideally, multi-disciplinary teams (e.g. linguists, applied researchers & engineers, members of product & UX teams) should do field-work & prepare reports about language use in LRLs and their communities in advance.

Issues about Data Collection (Doğruöz & Sitaram, 2022)

- Lack of standards in annotation procedures.
- Difficulties with multiple scripts involving LRLs.
- Same words could end up being transcribed for both scripts (e.g. multiple occurrences of the same word).

Issues about Model Building (Doğruöz & Sitaram, 2022)

- Models built with monolingual assumptions may produce errors while processing inherently multilingual LRLs.
- Massive Multilingual Models can process many languages in a single model but they tend to perform worse on LRLs compared to HRLs.
- Wikipedia texts (or random crawled data from the web) do not represent the variation in language use within a LRL community.
- Models that are explainable and easy to debug will also benefit from the feedback provided by the users of LRL communities.

Issues about Evaluation (Doğruöz & Sitaram, 2022)

- Evaluation benchmarks do not always exist for most LRLs.
- Available data sets may not reflect how language is actually used in these communities.
- Over optimization on a small set of benchmarks, performance of NLPs (even on HRLs) could be an overestimate.

Representativeness as a Forgotten Lesson for Multilingual & Code-switched Data Collection & Preparation (Doğruöz et al., 2023)

- Analyses of 68 CSW data sets until 2023.
- Most CSW data (speech & social media) involves English. Most language/dialect pairs (especially LRLs) in CSW is not represented enough.
- Flaws about representativeness in terms of data collection, transcription and annotation.

Doğruöz, A.S., Sitaram, S., Yong, Z. (2023). *Representativeness as a Forgotten Lesson for Multilingual and Code-switched Data Collection and Preparation*, EMNLP’23 Findings.

Issues about Representativeness (Doğruöz et al., 2023)

- CSW data about the same language pairs (e.g., Chinese-English) collected from one region (e.g., Hong Kong) does not represent the CSW for the same language pair in another region (e.g., Singapore) due to sociolinguistic variation.
- We trained ASR models both on ASCEND and SEAME data sets for the same language pairs (Chinese-English) data sets collected in two regions (Singapore, Hong Kong).
- When the models were trained and evaluated on different data sets, there is a substantial performance gap between 25% and 35% in Match Error Rate (MER) and Character Error Rate (CER).

Issues about Representativeness (Doğruöz et al., 2023)

Test Datasets	Pretraining Languages	ASCEND (Train)		SEAME (Train)	
		↓ MER	↓ CER	↓ MER	↓ CER
ASCEND	Chinese (Mandarin)	**26.40**	**22.89**	55.40 (+29.0)	49.26 (+26.37)
	English	**30.33**	**24.17**	61.23 (+30.9)	52.85 (+28.68)
SEAME	Chinese (Mandarin)	65.77 (+33.51)	53.19 (+30.52)	**32.26**	**22.67**
	English	64.39 (+32.65)	54.66 (+32.30)	**31.74**	**22.36**

Table 1: ASR performance trained and evaluated on ASCEND (Lovenia et al., 2021) and SEAME (Lyu et al., 2010) from Hong Kong and Singapore respectively. We indicate the (performance gap) in error rate between models that are trained-and-evaluated on the same datasets (**bold text**) and on different datasets.

Issues about Representativeness **(Doğruöz et al., 2023)**

- Reasons behind these errors:
- Singapore CSW: Mandarin-English but also Hokkien words (especially in informal conversations).
- However, it is not observed in the Hong Kong data set.

Issues about Representativeness (Doğruöz et al., 2023)

- There are sociolinguistic differences across users/speakers in multilingual communities. Not everybody speaks in the same way.
- User/Speaker backgrounds (e.g., age, gender, education, language backgrounds) are not reported or incomplete in current CSW data sets (in NLP).
- If they are not reported, we do not know which community/user group the data sets represents.
- There are differences in CSW based on the background factors (e.g., age, gender, multilingual backgrounds, proficiency levels).

Issues about Representativeness **(Doğruöz et al., 2023)**

- There is often no information about the backgrounds (e.g. language, age, gender) about the data collectors, transcribers or annotators in current CSW data sets and no clear criteria about how they were chosen.
- This leads to errors in the pre-processing of the data sets which lead to more problems in the later stages.
- Example (Diab, 2023): Failed example of speech recognition system in Arabic. The annotators spoke only Moroccan Arabic whereas the data was collected from another Arabic dialect.
- Recruiting annotators without checking their language skills led to a high number of annotation errors and a system failure.
- Same issues hold true for transcribers and data collectors as well.

Issues about Representativeness **(Doğruöz et al., 2023)**

- CSW Data sets based on Speech Data:

  44% studies mentioned the gender of the participants.

  36% mentioned the age.

  52% included some information about the language backgrounds of the speakers (but not in detail).

  Almost none of the studies mentioned any of the socio-demographic factors about the data collectors and/or transcribers.

  Without the socio-demographic factors, it is very hard to make any estimates about representativeness of the data sets or the transcribers.

Issues about Representativeness **(Doğruöz et al., 2023)**

- CSW Data sets based on Social Media Data:
- None of the data sets had any information about socio-demographic backgrounds of the users.
- It is not possible to estimate the representativeness of social media data sets.
- There were no transcribers but annotators.
- However, the socio-demographic information about the annotators were very limited which makes it challenging to estimate the representativeness of the annotators for these data sets.

References

- Doğruöz, A.S. & Sitaram, S., Yong, Z. (2023). Representativeness as a Forgotten Lesson for Multilingual and Code-switched Data Collection & Preparation. EMNLP'23 Findings. Singapore.
- Doğruöz, A.S. & Sitaram, S. (2022). *Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights*. Proceedings of SIGUL at LREC'22. European Language Resources Association.
- Doğruöz, A.S., Sitaram, S., Bullock, B.E., Toribio, A.J. (2021). *A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies*, (ACL-IJNLP'2021), Bangkok, Thailand.
- Nguyen, Doğruöz, Rose, de Jong (2016). Computational Sociolinguistics. Computational Linguistics, MIT Press.
- Chin, N.B. & Cavallaro, F. (2019). *Multilingualism in Southeast Asia: The post-colonial language stories of Hong Kong, Malaysia and Singapore.* In S. Montanari & S. Quay (Eds.) Multidisciplinary Perspectives on Multilingualism. Berlin, Boston: De Gruyter Mouton.
- Florey, M. (2010). Endangered Languages of Austranesia. Oxford University Press.

NLP Resources

1 November 2023

Preamble

Various Scripts

Thai

Vietnamese

Jawi
(Malaysia, Brunei, Indonesia)

Burmese

Lao

Khmr (Cambodian)

Balinese (Indonesian)

Various Language Roots and Influences

Influences through trade and/or colonialism

- French: Laos, Vietnamese
- Dutch: Indonesian (*Kantoor* for office is *Kantor*)
- Spanish: Indonesian (*Sapatos* for shoes is *Sepatu*)
- English: Malay (Bicycle is *Baisikal*)
- Arabic: Malay, Indonesian

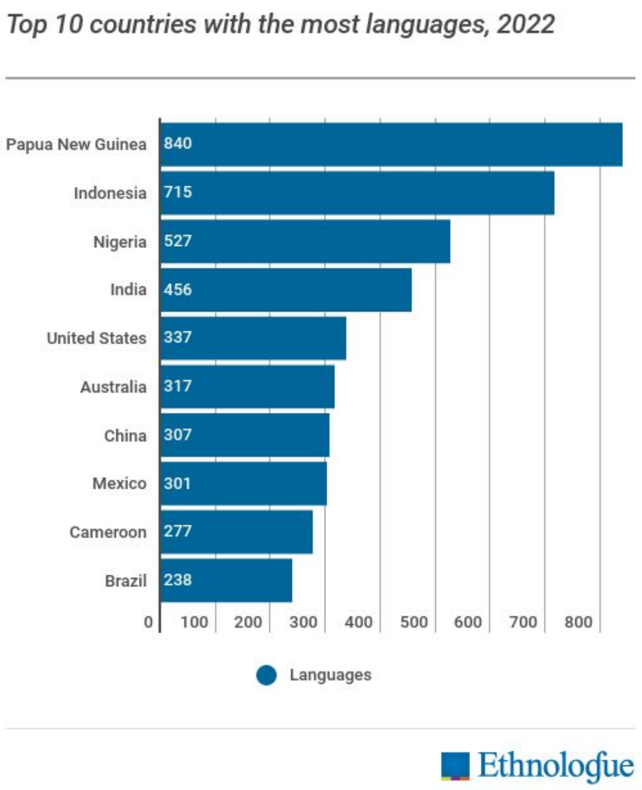
Resource Availability and Language Models

🇮🇩 Indonesia

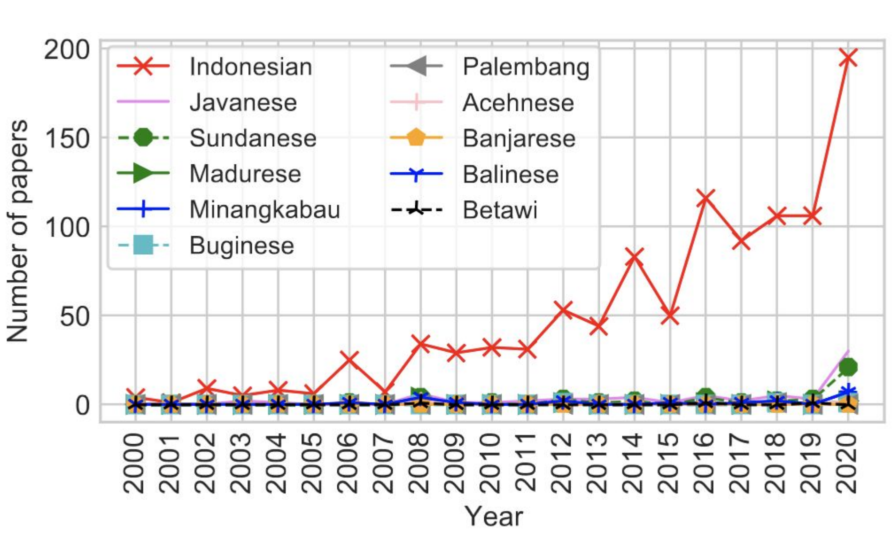
1 November 2023

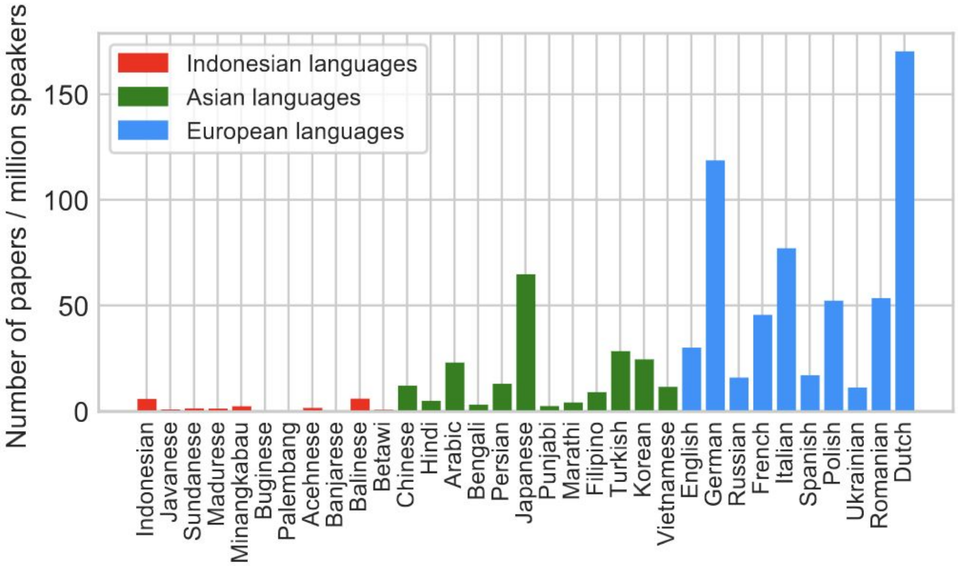
NLP Resources

- The NLP resources for Indonesian languages have a huge progress in the last few years.
- There are more than 700 languages spoken in Indonesia (Aji, et al., 2022)



NLP Trends for Indonesian





Benchmarks

Recent development on benchmarks

- IndoNLU (Wilie, et al., 2020) NLU
  - 12 Tasks, IndoBERT models
- IndoNLG (Cahyawijaya, et al., 2021) NLG
  - 10 Tasks, IndoGPT, IndoBART model
- IndoLEM (Koto, et al., 2020) NLU
- NusaCrowd (Cahyawijaya, et al., 2023) NLU NLG

New Datasets

Expanding the coverage of languages

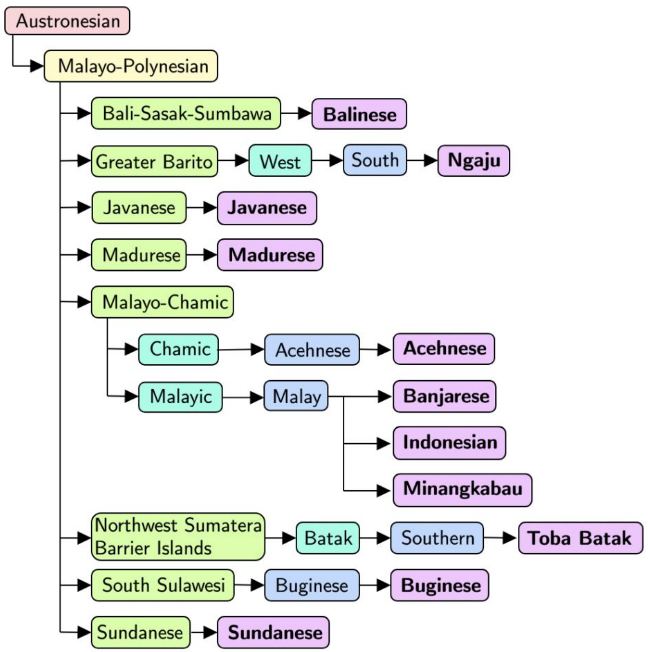
- NusaX (Winata, et al., 2023)
- NusaWrites (Cahyawijaya, et al., 2023)
- IndoNLI (Mahendra, et al., 2021)

New Datasets: NusaX

- **1,000** Samples – **500** train, **100** validation, **400** test
- **2** Tasks
  - **Sentiment Analysis**
  - **Machine Translation**
- **10 local languages** + **English** + **Indonesian**

Language	ISO 639-3	Annotators' Dialect	Example
Acehnese	ace	Banda Aceh	Meureutoh rumoh di Medan keunong ie raya
Balinese	ban	Lowland	Satusan umah ring medan merendem banjir
Toba Batak	bbc	Toba, Humbang	Marratus jabu di medan na hona banji
Banjarese	bjn	Hulu, Kuala	Ratusan rumah di medan tarandam banjir
Buginese	bug	Sidrap	Maddatu bola okko medan nala lempe
Javanese	jav	Matraman	Atusan omah ing medan kebanjiran
Madurese	mad	Situbondo	Ratosan bangko e medan tarendem banjir
Minangkabau	min	Padang, Agam	Ratuihan rumah di medan tarandam banjir
Ngaju	nij	Kapuas, Kahayan	Ratusan huma hong medan lelep awi banjir
Sundanese	sun	Priangan	Ratusan bumi di medan karendem banjir

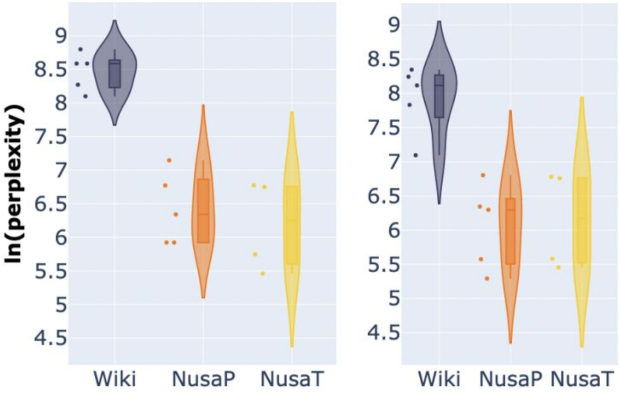
Table 8: Local languages spoken in Indonesia (ID) that are covered in NusaX.



New Datasets: NusaWrites

Effort to collect pretraining corpus for low-resource languages.

Web Crawling is not the best way to collect unlabelled corpus!



(a) **Balanced** setting

(b) **Full** setting

Figure 4: LMs perplexity evaluation of different corpus collection methods. Lower is better. Wiki: Wikipedia, NusaW: NusaParagraph, NusaT: NusaTranslation.

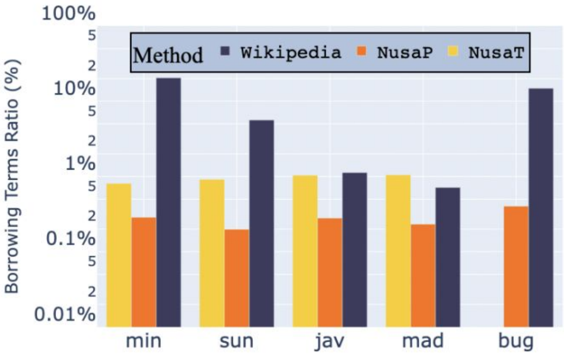


Figure 5: Ratio of loan words per language of different corpus collection methods. Wiki: Wikipedia, NusaW: NusaParagraph, NusaT: NusaTranslation. The ratio is presented in $\log_{10}$ basis.

Language Models

In the area of language models, there are efforts to build language-specific pre-trained language models.

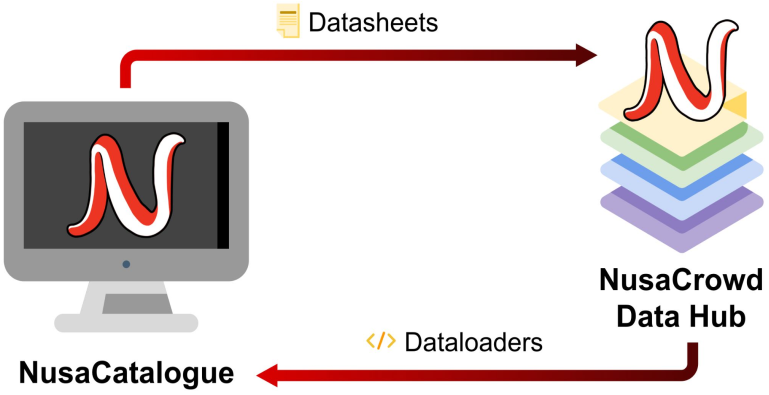
IndoBERT (Wilie, et al., 2020, Koto, et al., 2020)

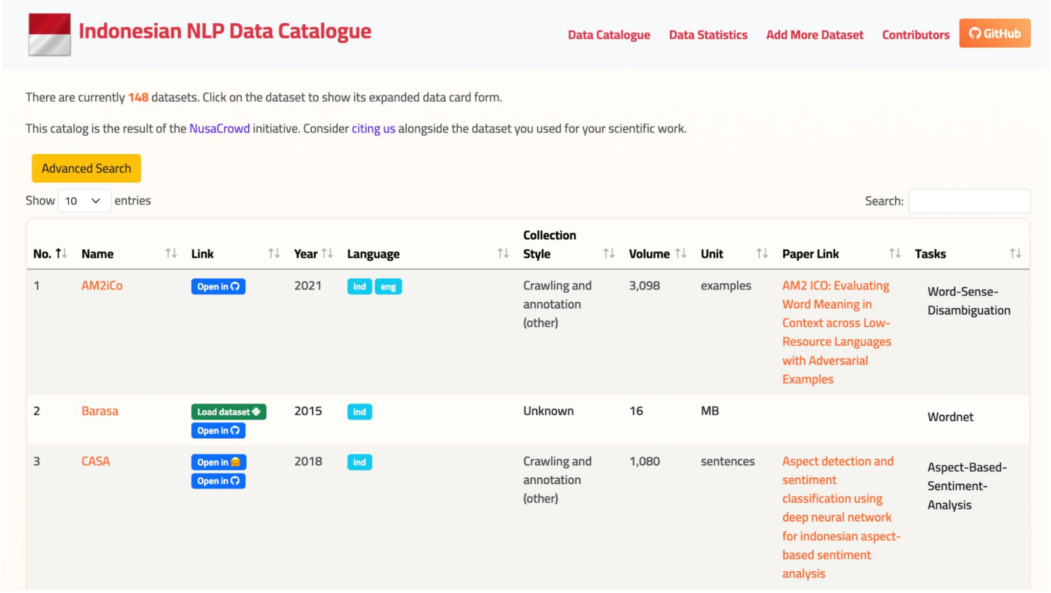
IndoGPT (Cahyawijaya, et al., 2021)

IndoT5 (Coming soon!)

Crowdsource Effort

Community movement to gather resources





Indonesian NLP Data Catalogue

Data Catalogue Data Statistics Add More Dataset Contributors GitHub

There are currently **148** datasets. Click on the dataset to show its expanded data card form.

This catalog is the result of the NusaCrowd initiative. Consider citing us alongside the dataset you used for your scientific work.

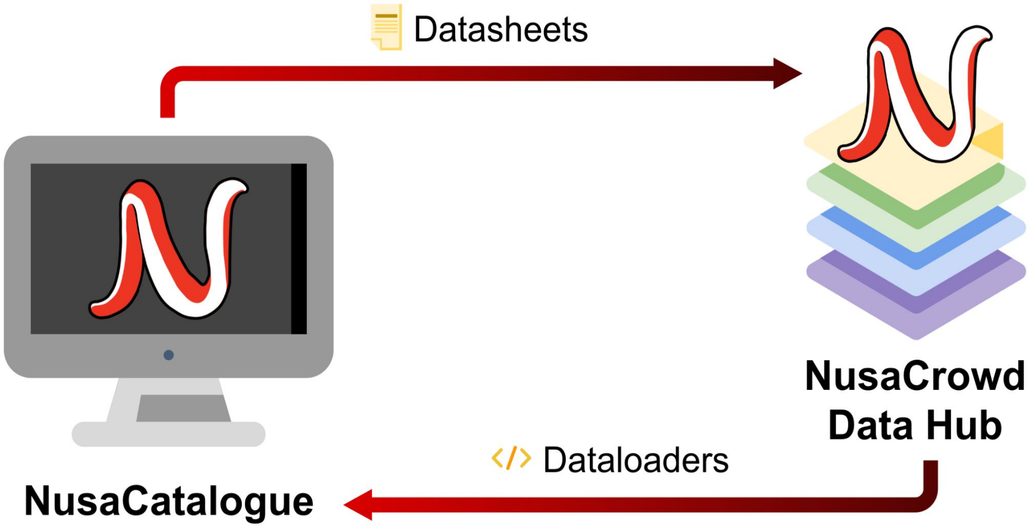
Advanced Search

Show 10 entries Search:

No.	Name	Link	Year	Language	Collection Style	Volume	Unit	Paper Link	Tasks
1	AM2iCo	Open in	2021	ind eng	Crawling and annotation (other)	3,098	examples	AM2 ICO: Evaluating Word Meaning in Context across Low-Resource Languages with Adversarial Examples	Word-Sense-Disambiguation
2	Barasa	Load dataset Open in	2015	ind	Unknown	16	MB		Wordnet
3	CASA	Open in Open in	2018	ind	Crawling and annotation (other)	1,080	sentences	Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis	Aspect-Based-Sentiment-Analysis

NusaCrowd Indonesia

- Open access to 130+ datasheets is provided through NusaCatalogue.
- 110+ dataloader scripts to access the resources are implemented in NusaCrowd Data Hub.
- Published in ACL Findings 2023.



Contribution points

To be considered as co-author, 10 contribution points is required.

- +3 for implementing a dataloader, unless specified otherwise.
- +2 for submitting a datasheet of a public dataset.
- +1 for submitting a datasheet of a private dataset.
- +3 for opening access to a previously private dataset.
  - As a support for the development of local languages datasets:
    - +2 for datasets in Sundanese, Javanese, or Minangkabau. +3 for other local languages.
  - For any dataset that does not achieve a certain minimum standard, 50% contribution score of the dataset will be penalized. This policy affects dataset that is collected without any manual validation.
  - If there is >1 Author for a dataset, main author will be eligible for nominating 1 more author to be granted the same contribution score.

Incentive

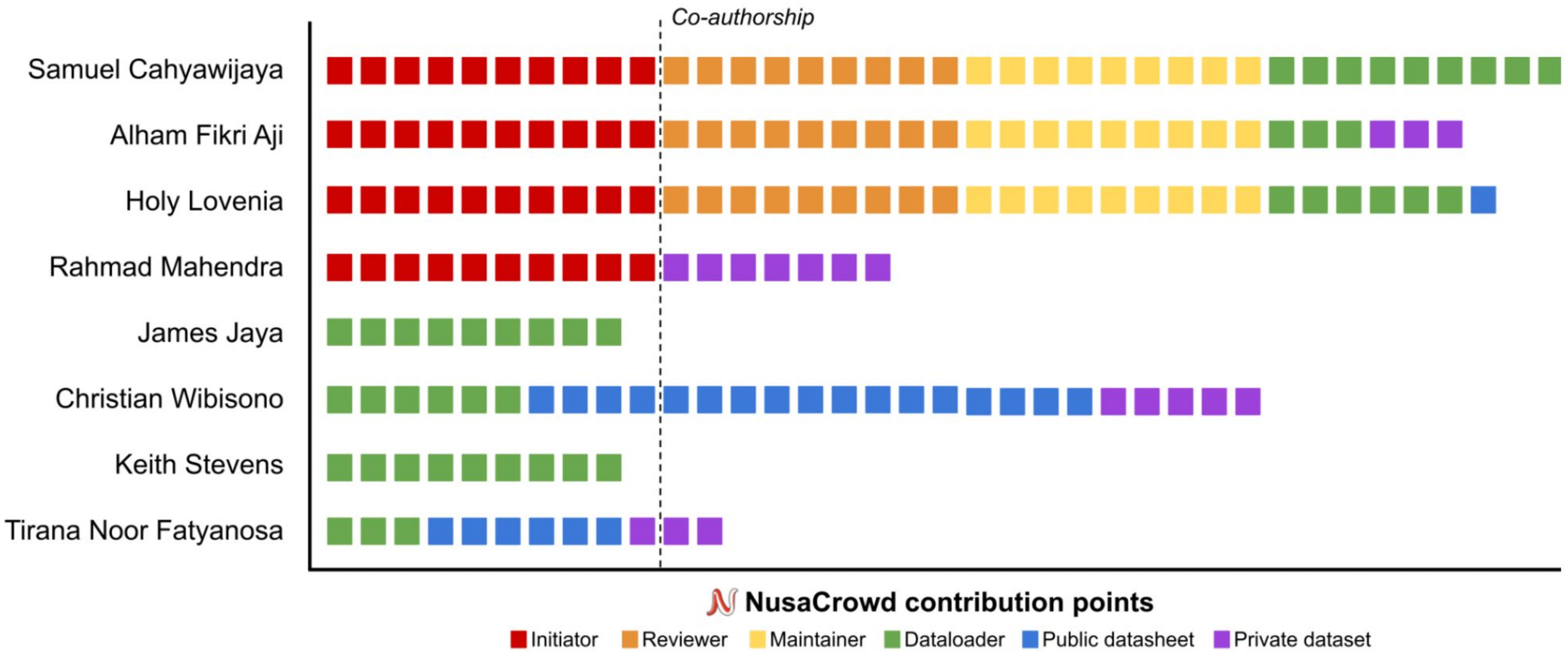
Incentive as a contributor

- New friends & connections 🤗
- Acknowledgement on our paper 🙇‍♂️
- Your name will be listed in our website 🌐
- Hacktoberfest badge*

Incentive as a co-author

- All the above benefits ⬆️
- Free T-Shirt 👕 👉 👉 👉
- Authorship on our research paper 🎓

Contributions



Challenges and Future Work

- **Resource challenges**
  - Resource Availability (Model, Corpus, etc)
  - Accessibility & Standardization
- **Language challenges**
  - Diversity
  - Code-Mixing
- **Societal challenges**
  - Lack of Funding / Resources
  - Technology Disparity

Resource Availability and Language Models

🇸🇬 Singapore

1 November 2023

Languages in Singapore

Official Languages

- English
- Mandarin
- Malay
- Tamil

Spoken by major ethnic groups

Others Languages

Cantonese, Hokkien, Indonesian, Hindi, Thai, Vietnamese, Telugu...

Image from Tang(2016)

Languages in Singapore

Official Languages

- English
- Mandarin
- Malay
- Tamil

Spoken by major ethnic groups

Others Languages

Cantonese, Hokkien, Indonesian, Hindi, Thai, Vietnamese, Telugu…

Singapore Colloquial English (Singlish)

(Figure adapted from https://thesmartlocal.com/read/singaporean-culture-quirks/ by Ranae Cheng)

NLP Resources

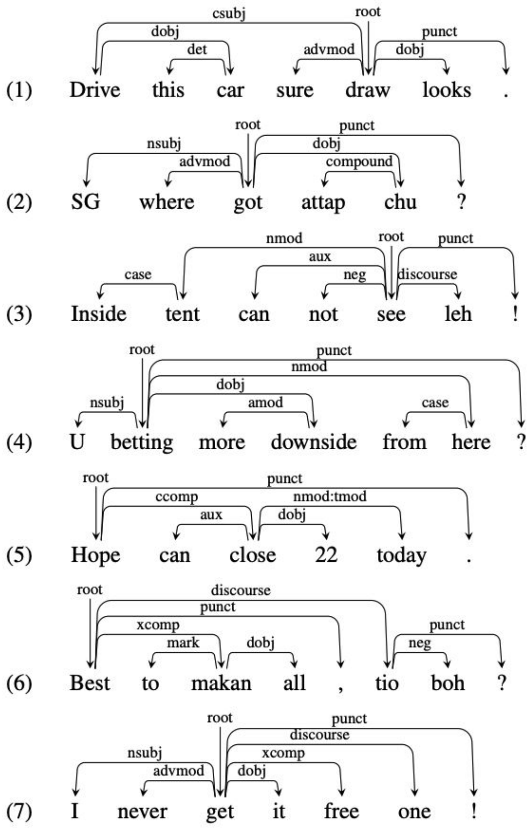
- Mostly developed by people affiliated with NLP labs in local universities or research agencies
  - NUS SMS Corpus (Chen and Kan, 2015)

Source	English SMS	English Contributors	Chinese SMS	Chinese Contributors
MTurk	11,275	75	55	19
ShortTask	280	17	0	0
Zhubajie	0	0	23,789	483
Local	16,701	20	3,544	10
Internet Community	468	4	1,712	3
Total	28,724	116	29,100	515

Number of SMS and Contributors by Source (Chen and Kan, 2011)

NLP Resources

- Mostly developed by people affiliated with NLP labs in local universities or research agencies
  - NUS SMS Corpus (Chen and Kan, 2015)
  - Singlish Treebanks (Wang et al., 2017; Wang et al., 2019)



Unique Singlish Grammars (Wang et al., 2017)

NLP Resources

- Mostly developed by people affiliated with NLP labs in local universities or research agencies
    - NUS SMS Corpus (Chen and Kan, 2015)
    - Singlish Treebanks (Wang et al., 2017; Wang et al., 2019)
    - English Resource Grammar Corpus for Singlish (Chow and Bond, 2022)

NLP Resources

- Mostly developed by people affiliated with NLP labs in local universities or research agencies
  - NUS SMS Corpus (Chen and Kan, 2015)
  - Singlish Treebanks (Wang et al., 2017; Wang et al., 2019)
  - English Resource Grammar Corpus for Singlish (Chow and Bond, 2022)
  - Singlish Message Paraphrasing Corpus (Liu et al, 2022)

Level 1: Lexical Level Normalization
Lexical Variations in User-Generated Content Tackling the common user-generated lexical variations, including lower-case and upper-case (E.g., ‘mrt’ → ‘MRT’, ‘TOdAy’ → ‘TODAY’), spelling typo (e.g., ‘domian’ → ‘domain’, ‘r0bust’ → ‘robust’), single-word abbreviations (e.g., ‘pple’ → ‘people’, ‘coz’ → ‘because’), phonetic substitutions (e.g., ‘tym’ → ‘time’, ‘4U’ → ‘for you’), and other non-standard spellings (e.g., ‘goooood’ → ‘good’). **Lexical Variations in Singlish** Tackling the Singlish lexical variations, such as special short forms (e.g., ‘yck’ → ‘YCK (Yio Chu Kang)’), and colloquial words (e.g.,‘cheapo’ → ‘cheapskate’, ‘gahmen’ → ‘government’). **Non-English Word Borrowing** Replacing the non-English words borrowed from other languages that have a word-to-word mapping. E.g., ‘mei mei’ → ‘sister’ (Mandarin), ‘pa tuo’ → ‘dating’ (Cantonese), ‘ta pau’ → ‘take-away’ (Cantonese), ‘huat’ → ‘to prosper’ (Hokkien), and ‘makan’ → ‘food’ (Malay).
Level 2: Syntactic Level Editing
Missing Pronoun & Copula Recovering the appropriate pronouns, and the necessary verbs (e.g. “m typing a sms” → “I am typing a SMS”, “oh cat so cute” → “oh the cat is so cute”). **Non-Standard Syntax & Grammar** Fixing the non-standard grammar in colloquial Singlish sentences, such as the topic prominence phenomenon (e.g., “A bit late lah, I came there.” → “I came there a bit late.”). **Missing Punctuation** Inserting the punctuation to where it is necessary (e.g., “Is that your book” → “Is that your book?”).
Level 3: Semantic Level Rewriting
Colloquial Wording Some wording is different from colloquial Singlish and English, thus it needs to paraphrase the sentence while retaining the same semantic meaning (e.g. “Call aint going.” → “The call is not coming through.”) **Discourse Particles** Some clausal-final discourse particles indicate much semantic information (e.g., ‘leh’ marks a tentative request, ‘lah’ is a mood marker, and appeals for accommodation). For instance, “U leh, i going back liao.” → “What about you? I am going back.” **Non-English Spans & Code-Switching** Some non-English spans and the code-switching require clause or sentence level translation (e.g. “You sian? Let’s go shopping!” → “Are you feeling bored? Let’s go shopping!”, “makan where?” → “where should we eat?”).

Three sub-tasks of the Singlish message paraphrasing. (Liu et al, 2022)

NLP Resources

- Mostly developed by people affiliated with NLP labs in local universities or research agencies
  - NUS SMS Corpus (Chen and Kan, 2015)
  - Singlish Treebanks (Wang et al., 2017; Wang et al., 2019)
  - English Resource Grammar Corpus for Singlish (Chow and Bond, 2022)
  - Singlish Message Paraphrasing Corpus (Liu et al, 2022)
  - Singlish Benchmark for sentiment analysis and language identification (Gotera et al., 2022)

SENTIMENT ANALYSIS EXAMPLES

Posts	Label
I thought see good offers I share ma	0
bad breath just say dun blame on the durian cookies bro	1
Take 98 from jurong east bus interchange	2

where label 0 and 1 indicates non-Singlish and Singlish, respectively.

SINGLISH IDENTIFICATION EXAMPLES

Posts	Label
Patutnya MySejahtera buat benda macam tu	0
I'm an East Coast resident, and this is what I'd like to see	0
Wah why so many round pulak ni	1
Ministers also not easy la.	1

Task examples. (Gotera et al, 2022)

NLP Resources

- Mostly developed by people affiliated with NLP labs in local universities or research agencies
  - NUS SMS Corpus (Chen and Kan, 2015)
  - Singlish Treebanks (Wang et al., 2017; Wang et al., 2019)
  - English Resource Grammar Corpus for Singlish (Chow and Bond, 2022)
  - Singlish Message Paraphrasing Corpus (Liu et al, 2022)
  - Singlish Benchmark for sentiment analysis and language identification (Gotera et al., 2022)
  - Singapore-centered Online Attacks (SOA) Dataset. (Haber et al., 2023)

Language	n	% of Data
Indonesian	12,212	81.4
Malay	1,635	10.9
Indonesian and English	396	2.6
Singlish	218	1.5
Malay and English	131	0.9
Javanese	92	0.6
English	85	0.6
Sundanese	46	0.3
Javanese and Indonesian	23	0.2
Sundanese and Indonesian	20	0.1
Chinese	11	0.1
Other	121	4.0

Distribution of languages and language combinations for the 15,000 comments gathered from Reddit. Languages or language combinations present in fewer than 10 comments, such as Hokkien Chinese, Arabic and Russian, are combined as 'Other'. (Haber et al., 2023)

Other Modalities

Mostly Audio Resources

- Resource collection led by government: National Speech Corpus (NSC)
  - first large-scale Singapore English corpus spearheaded by the Info-communications and Media Development Authority (IMDA) of Singapore.
  - open speech and transcript data for automatic speech recognition (ASR) research and speech-related applications.
  - 1.2TB in size, available with Singapore Open Data License, access upon registration
- Closed data sources from data company
  - Nexdata: 201 Hours - Singaporean Speaking English Speech Data by Mobile Phone
    - English recordings with transcription from 452 Singaporeans (40% male and 60% female)

Language Models

On Hugging Face Model Hub

- singbert-lite-sg, singbert and singbert-large-sg from user zanelim
  - Pertained with data collected from subreddits- r/singapore and r/malaysia, and forums such as hardwarezone

Community and Industry Efforts

AI Singapore - non-profit organization funded by government

- Released python libraries for general NLP research and for ASEAN languages
  - sgnlp and SeaCoreNLP
- Created BAHASA: a SEA Linguistic and Cultural Evaluation Suite for LLMs (Leong et al., 2023)
- Provided AI Practitioner’s Handbook for providing guidance on AI product development
- Currently developing SEA datasets and LLMs, stayed tuned!

Compute Resources

- Lab compute in Local Universities
- National Compute Environment: National Supercomputing Centre

Challenges and Future Work

Limited Resources

- Language Diversity
  - English vs. other languages
  - unique colloquial languages like Singlish or code-mixing require annotators to have multilingual proficiency
- Resource Scale - Scarce online texts resources -> current datasets need careful manual collection
- Standardized Tasks and Benchmarks
- Limited Computation Resources for LLMs

Resource Availability and Language Models

🇵🇭 The Philippines

1 November 2023

Languages in the Philippines

- Some 130 to 195 languages
  - **Filipino** – “Standardized” version of **Tagalog.** Official national language.
  - Regional Malayo-Polynesian languages (Kapampangan, Cebuano, Ilokano, etc.)
  - **English** – Official language for government and instruction alongside Filipino
  - **Chinese** – Primarily Hokkien and Mandarin, spoken by Filipino-Chinese communities.
  - **Chavacano** – One spanish creole.
  - **Filipino Sign Language** (FSL)

- The Latin Alphabet is used as the medium of writing.
  - There is a recent push towards revitalizing ancient Philippine systems such as **Baybayin**
  - Chinese is widely used in local Filipino-Chinese communities
  - Arabic is also widely used especially in Muslim Mindanao in southern Philippines.

Benchmark Datasets

- There is no single standardized benchmark a la GLUE for Philippine Languages.

- Most datasets in the Philippines are task specific:
    - PagkataoKo (Tighe et al., 2022) – Demographic and Personality-based Social Media
    - NewsPH NLI (Cruz et al., 2021) – Entailment dataset from News domain
    - Fake News Filipino (Cruz et al., 2020) – Classification of Fake News
    - Hatespeech Tagalog (Cabasag et al., 2019) – Classification of Hatespeech on Social Media
    - Dengue Classification (Livelo and Cheng, 2018) – Topic-related classification

Training Corpora

- Language Modeling
  - TLUnified (Cruz et al., 2021) – Unlabeled text dataset for language modeling
  - NewsPH (Cruz et al., 2021) – Text corpus in the news domain
  - WikiText-TL-39 (Cruz and Cheng, 2019) – Earliest crawled unlabeled dataset for Filipino

- Speech and Translation
  - OPUS Project (Tiedemann, 2012) – Contains parallel corpora for Bicolano, Cebuano, Filipino, Hiligaynon, Ilocano, and Pangasinense
  - Filipino-Bisaya Speech (Pascual et al., 2023) – ASR dataset in mixed Filipino and Bisaya

Linguistic Tools and Pretrained Vectors

- End-to-End Tools
  - CalamanCy (Miranda, 2023) – SpaCy-based end-to-end pipeline for Filipino

- Some classical resources are available but only for Filipino
  - FilWordNet (Velasco et al., 2022) – Synthetically-aligned wordnet for Filipino
  - SMTPOST (Nocon and Borra, 2018) – Part of Speech (POS) dataset

- Word vectors are still being trained but usually for specific domains
  - Juris2Vec (Peramo et al., 2021) – Vectors specifically for Philippine law

Pretrained Models

- Only Filipino-based pretrained models exist
  - RoBERTa Tagalog (Cruz and Cheng, 2022) – Currently the best pretrained model for Filipino
  - ELECTRA Tagalog (Cruz and Cheng, 2021)
  - BERT Tagalog (Cruz and Cheng, 2020) – The first pretrained LM in Filipino

- Support for other Philippine languages are lacking
- Likewise, support for Philippine languages in multilingual LLMs are lacking
  - Bactrian-X (Li et al., 2023) – LoRA adapters that include translated Filipino instructions
  - MT5 (Xue et al., 2020) – Includes Filipino (Tagalog) as part of the pretraining corpus
  - XLM-RoBERTa (Conneau et al., 2019)
  - mBERT (Devlin et al., 2018) – Includes Filipino (Tagalog) and Cebuano

Challenges

- Data scarcity is the biggest problem
  - Universal Dependencies Treebank for Tagalog has less than 20k words
  - There is a lack of (non-wikipedia) unlabeled corpora for non-Filipino Philippine languages.
  - Benchmarks are disjoint
- Reliance on Synthetic Data is a problem
  - NewsPH-NLI is synthetically derived from news articles
  - Our only instruction-based dataset is translated via GPT (Bactrian-X)
  - Cebuano Wikipedia is created by translation (Lsjbot, 5.3M pages)
- Most of these problems arise from a lack of resources, funding, and compute

Resource Availability and Language Models

🇲🇾 Malaysia

1 November 2023

Languages in Malaysia

- Standard Malay
- English
- Mandarin Chinese
- Tamil
- Chinese dialects
  - Cantonese, Hokkien, etc.
- South Asian languages
  - Bengali, Hindi, Punjabi, etc.
- Malayic dialects
  - Pahang Malay, Iban, etc.
- Other indigenous languages
  - Dusunic, Aslian, etc.

Image taken from Wikimedia

Languages in Malaysia

- **Standard Malay** — **Official Language**
- English
- Mandarin Chinese
- Tamil

Medium of instruction

- Chinese dialects
  - Cantonese, Hokkien, etc.
- South Asian languages
  - Bengali, Hindi, Punjabi, etc.
- Malayic dialects
  - Pahang Malay, Iban, etc.
- Other indigenous languages
  - Dusunic, Aslian, etc.

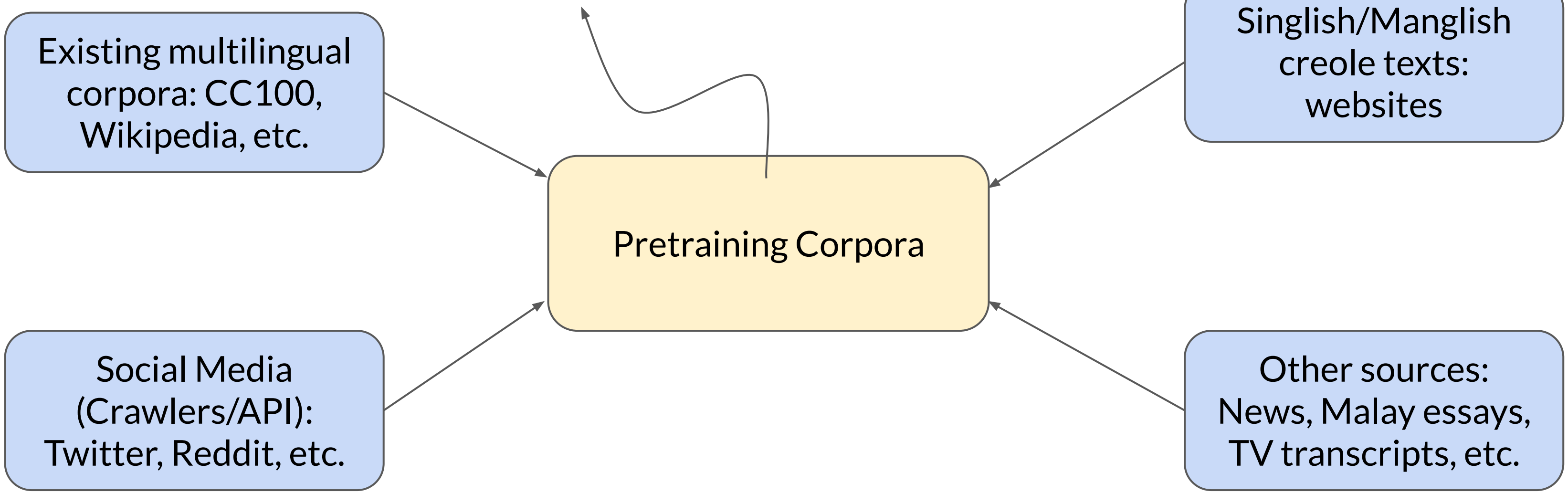
Regional dialects or minority languages

~140 languages

Image taken from Wikimedia

Malay Pretrained LMs (Malaya Suite)

- Created and open-sourced by Mesolitica (Zolkepli, 2018)
- ~15 pretrained language models
  - T5, GPT-2, BERT, BigBird, XLNET, etc.



Malay Dataset (Malaysian-Dataset)

- Created and open-sourced by Mesolitica (Zolkepli, 2018)
- Diverse Datasets covering LM **training**, task **evaluation**, and **lexicon** resources.
  - Instruction-tuning
  - Question-answering
  - Paraphrase generation
  - Bilingual dictionaries
  - ...
- Main Methods:
  - **Translation**: Google Translate or ChatGPT
  - **Synthetic data generation**: LLM (ChatGPT)
  - **Crawlers**: on social media (Twitter, Facebook and Instagram)

However, **no LLM model benchmarking** done on the datasets yet.

Malay Social Media Text: Normalization

- Many Malay NLP research work is done on social media data:
  - Malaysian-dataset (Zolkepli, 2018)
  - Sentiment analysis (Zabha et al., 2019; Samah et al., 2022, inter alia)
  - Public health (Juan, Saee & Mohamad, 2022; Alamoodi et al., 2022)
  - Fake news detection (Rahim & Basri, 2022; Lim et al, 2023; Kong et al., 2023)
  - ...
- Malay social media text is **noisy** (dialects, mixed languages, short forms, grammatical neglects, etc.)

Non-standard terms → **Normalization** → Standard terms

- Better data quality
- Better downstream task performance

Current Method: Rule-based Lexicon Approach

Tweets

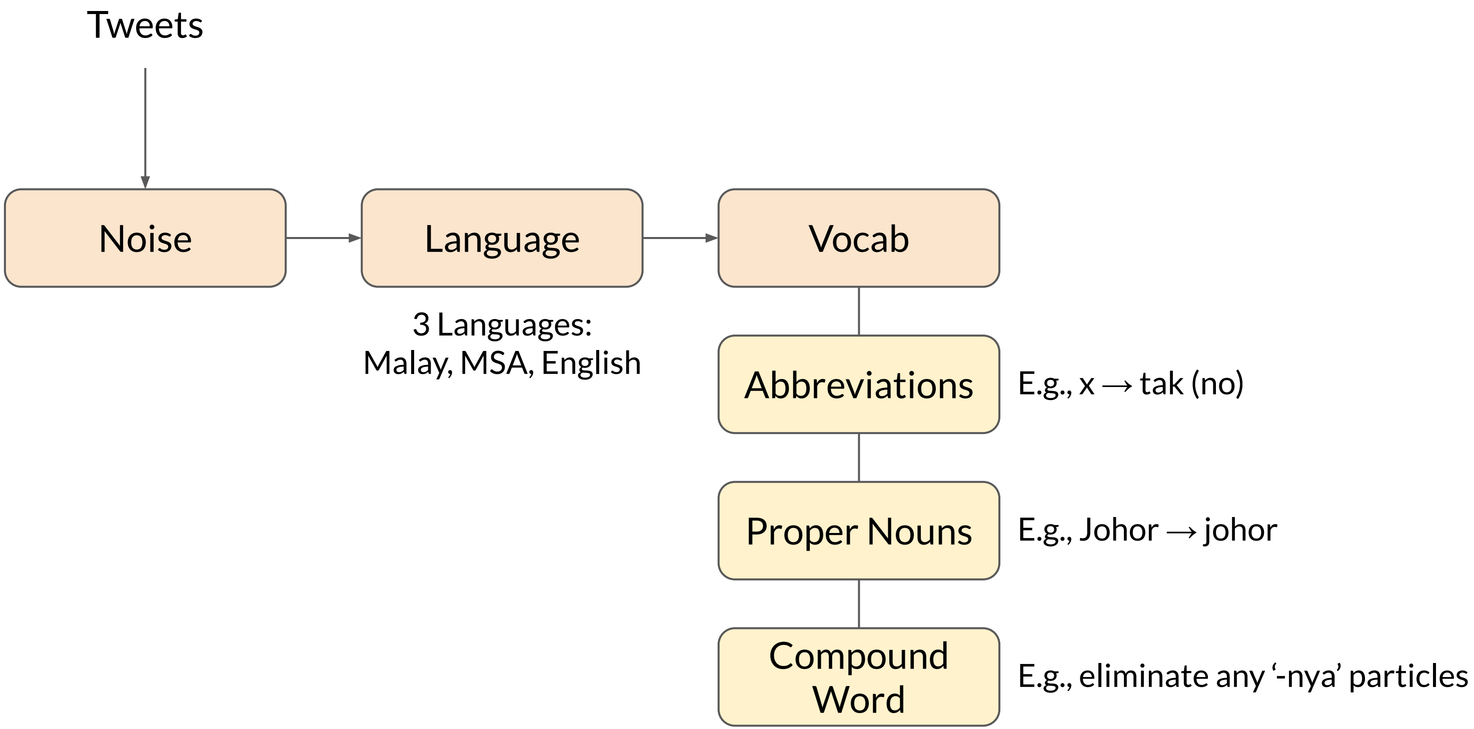
↓

Noise

Filter out non-Latin characters, web links, etc.

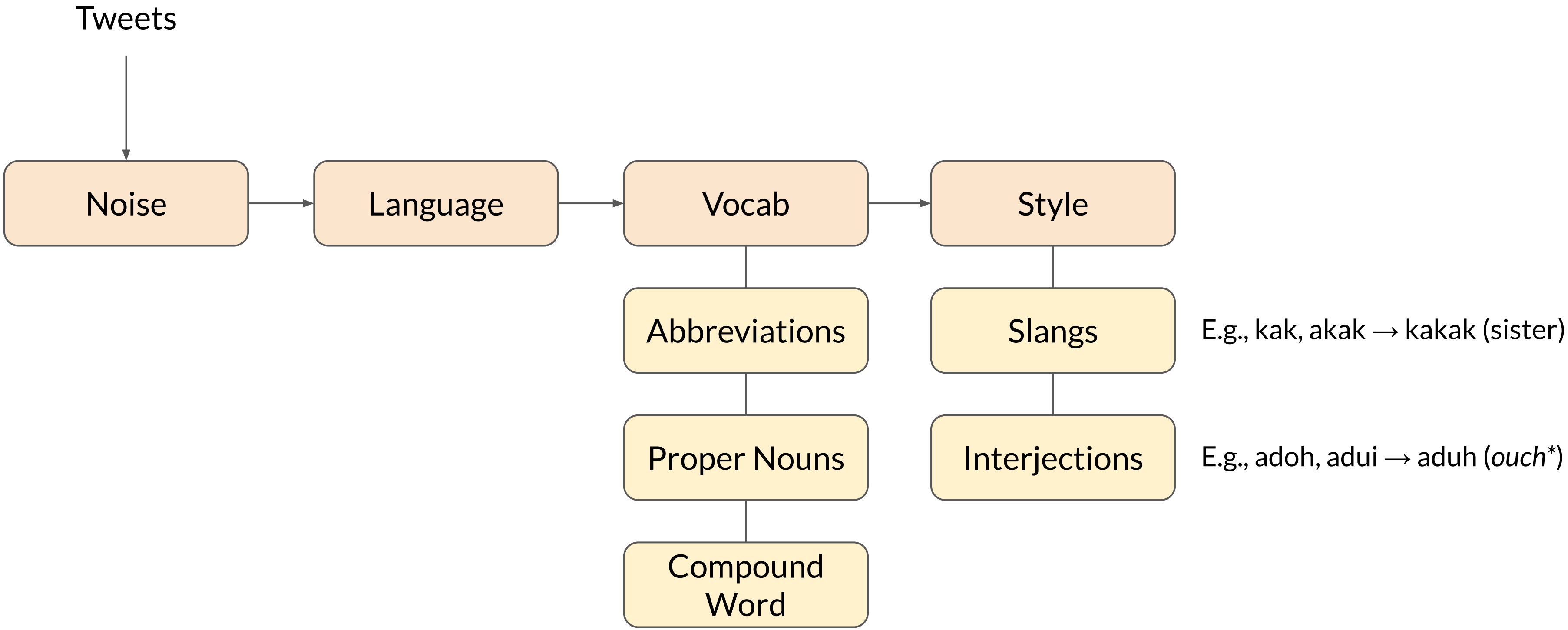
Adapted from [Ariffin, Siti Noor Allia Noor, and Sabrina Tiun. (2020)."Rule-based text normalization for Malay social media texts." IJACSA]

Current Method: Rule-based Lexicon Approach



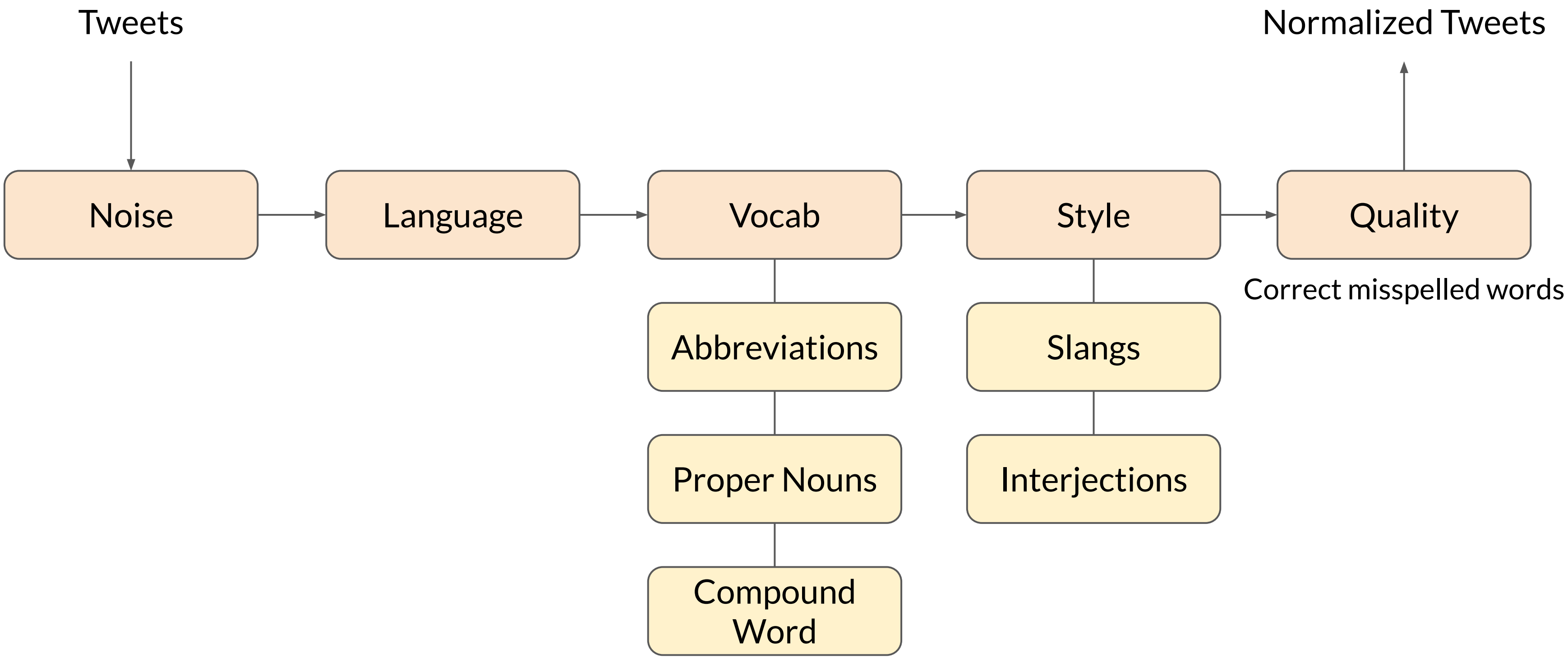
Adapted from [Ariffin, Siti Noor Allia Noor, and Sabrina Tiun. (2020)."Rule-based text normalization for Malay social media texts." IJACSA]

Current Method: Rule-based Lexicon Approach



Adapted from [Ariffin, Siti Noor Allia Noor, and Sabrina Tiun. (2020)."Rule-based text normalization for Malay social media texts." IJACSA]

Current Method: Rule-based Lexicon Approach



Adapted from [Ariffin, Siti Noor Allia Noor, and Sabrina Tiun. (2020)."Rule-based text normalization for Malay social media texts." IJACSA]

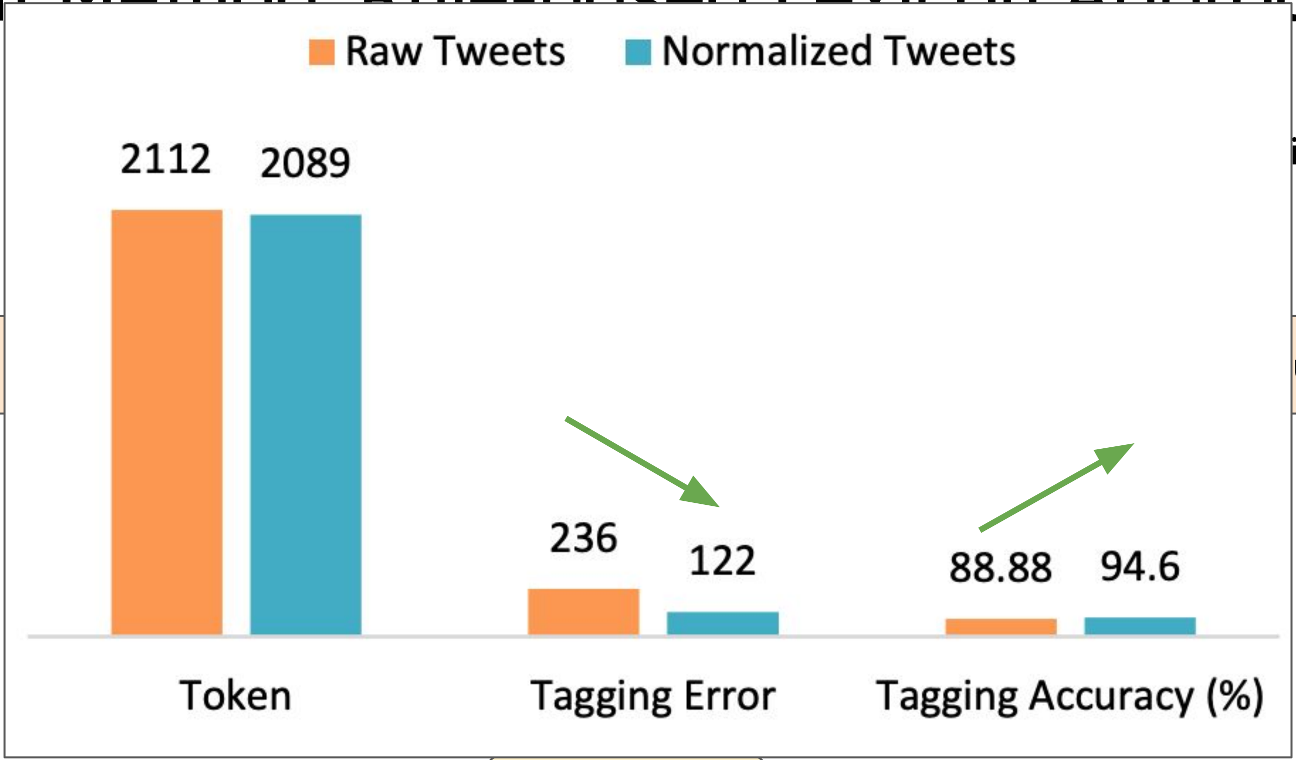
Current Method: Rule-based Lexicon Approach

Tweets

Noise

ized Tweets

uality



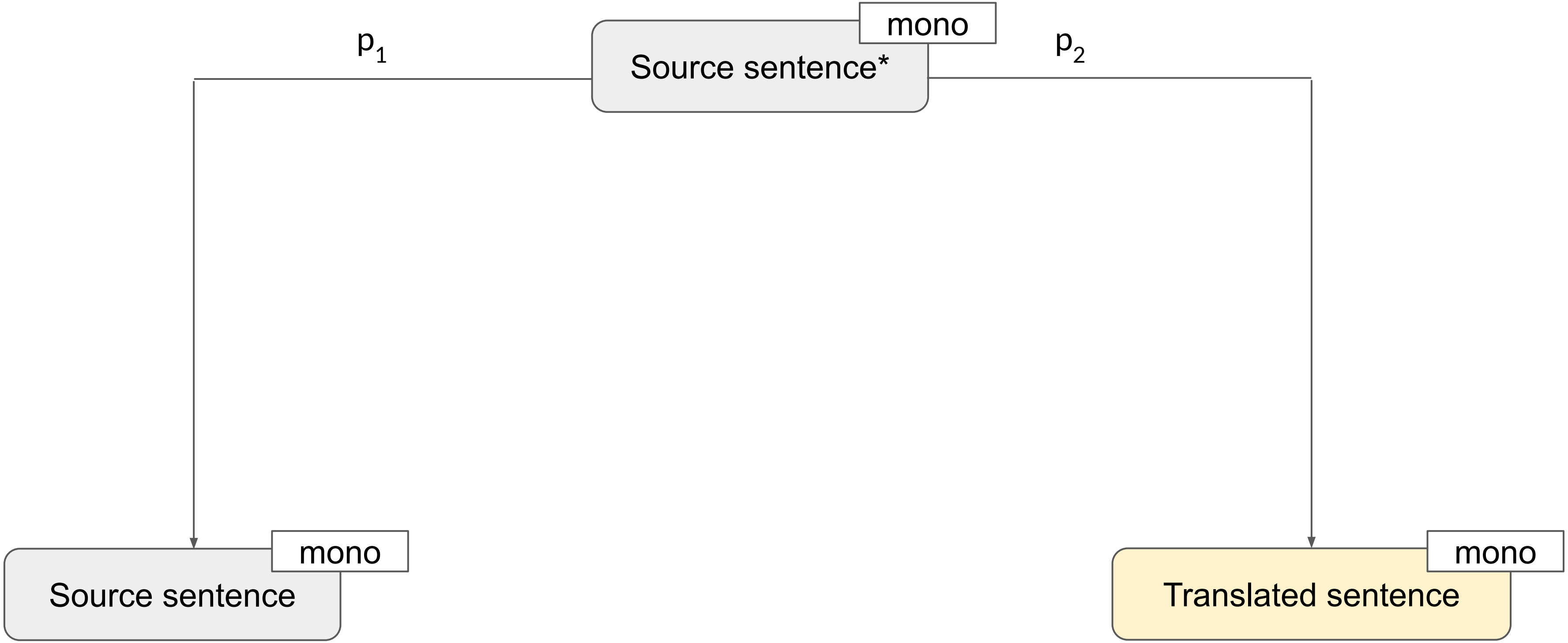
Compound Word

Adapted from [Ariffin, Siti Noor Allia Noor, and Sabrina Tiun. (2020)."Rule-based text normalization for Malay social media texts." IJACSA]

Bahasa Rojak

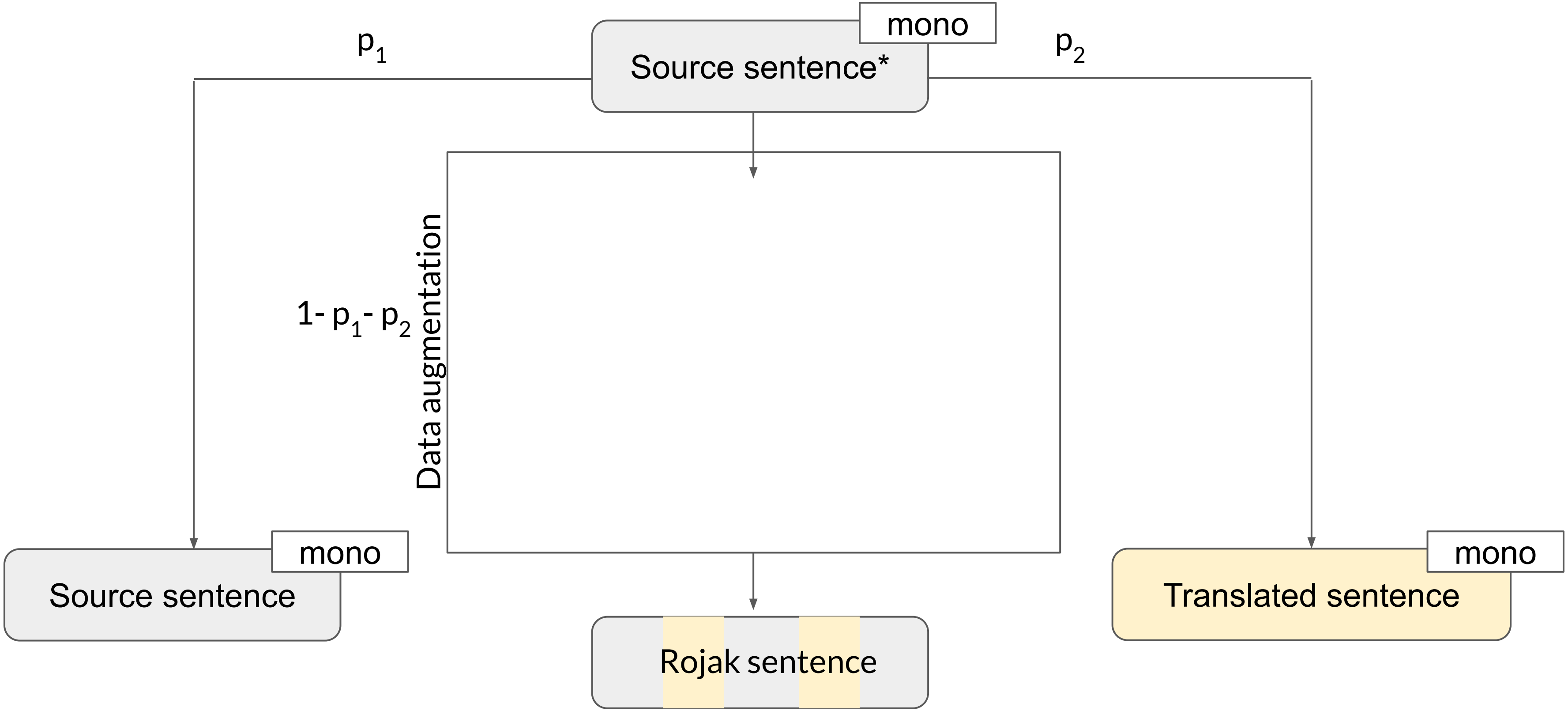
- $\overbrace{\textbf{Bahasa}}^{\text{languages}}\ \overbrace{\textbf{Rojak}}^{\text{mixed}}$ = Malaysian pidgin where speakers mix two or more languages in Malaysia (e.g., Malay, English, Mandarin Chinese, etc.)
- Already observed back in Malacca port during 15th century
    - Epicenter of east-west trade.
    - Polyglot: >80 languages were spoken.

Bahasa Rojak Crawled Corpus (BRCC)



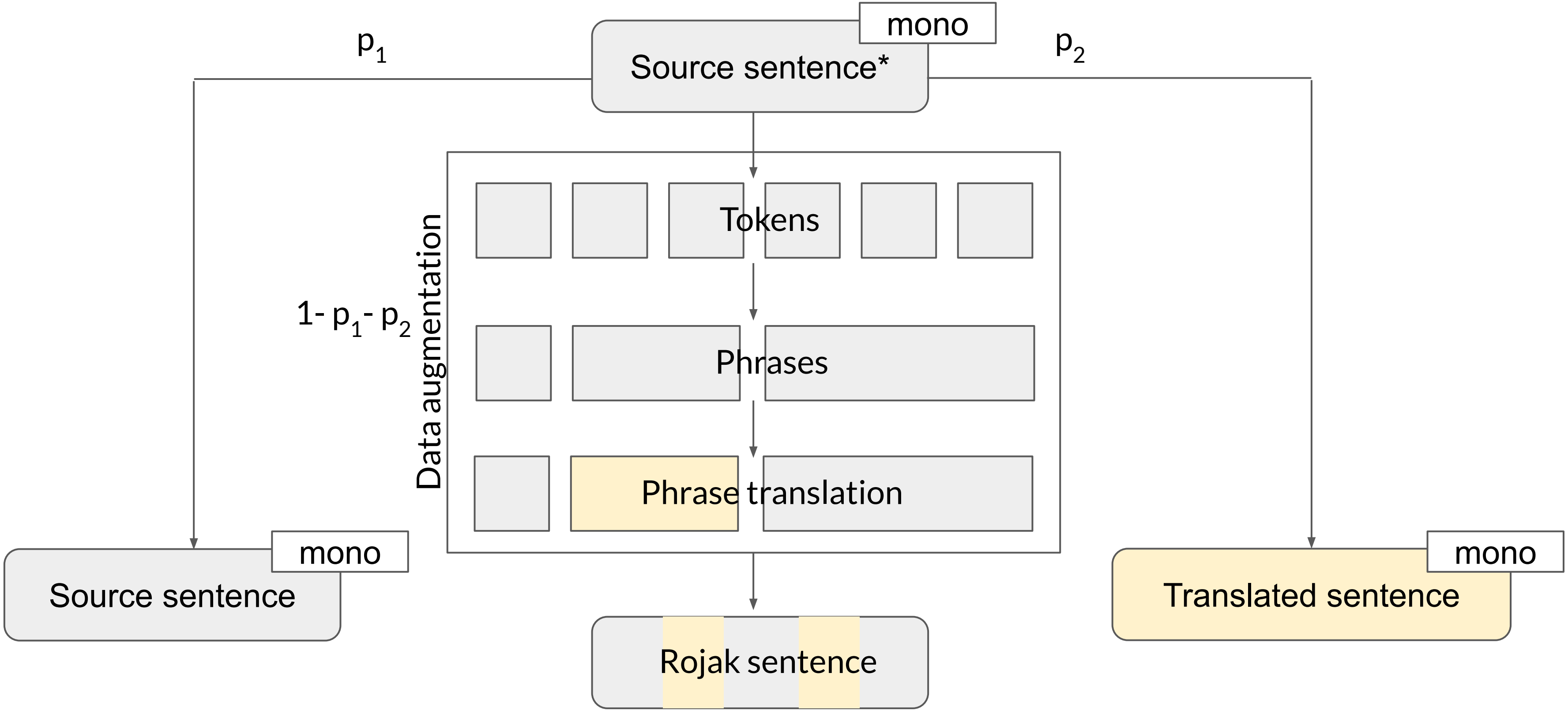
Romadhona, N. P., et al., (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. COLING.

Bahasa Rojak Crawled Corpus (BRCC)



Romadhona, N. P., et al., (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. COLING.

Bahasa Rojak Crawled Corpus (BRCC)



Romadhona, N. P., et al., (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. COLING.

Bahasa Rojak Sentiment Analysis and LM

- **Sentiment Analysis Dataset**: Preprocessing + Data Augmentation (Rojak)
- **Model**: Mixed-XLM (XLM + language embedding layer)

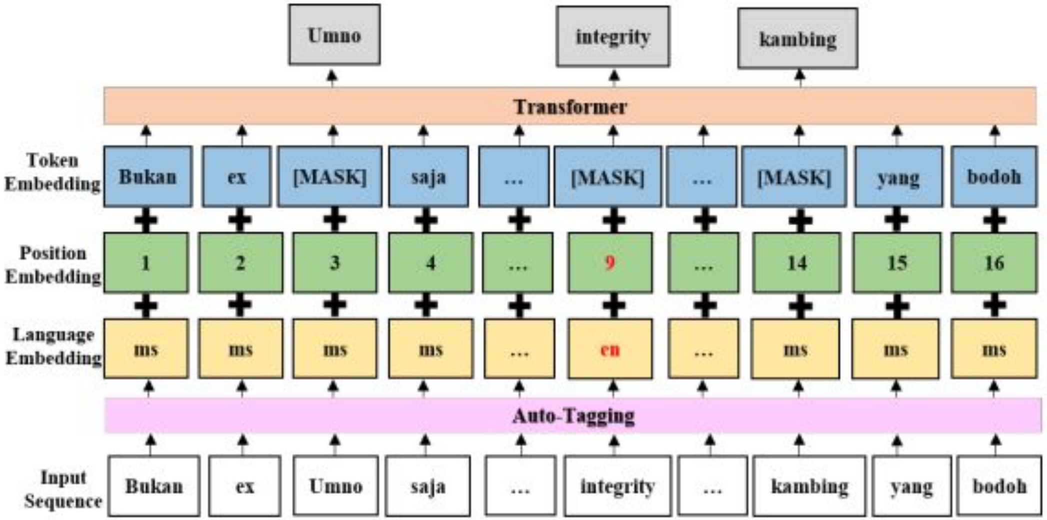


Figure 1: Input representation of Mixed XLM

Romadhona, N. P., et al., (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. COLING.

Model	Product Review		Movie Review		Stock Market		Avg.	
	acc	f1	acc	f1	acc	f1	acc	f1
mBERT (CM)	0.652	0.603	0.661	0.653	0.563	0.496	0.625	0.584
mBERT (EN-MS-CM)	0.653	0.651	0.631	0.576	0.571	0.562	0.618	0.596
XLM (EN-MS)	0.658	0.592	0.764	0.751	0.672	0.568	0.698	0.637
Mixed XLM (CM)	0.703	0.671	0.761	0.746	0.690	0.581	0.718	0.666
Mixed XLM (EN-MS-CM)	**0.718**	**0.696**	**0.812**	**0.803**	**0.706**	**0.615**	**0.745**	**0.705**

Table 9: Results of different models on SentiBahasaRojak

Model	Product Review		Movie Review		Stock Market		Avg.	
	acc	f1	acc	f1	acc	f1	acc	f1
mBERT (CM)	0.801	0.793	0.691	0.663	**0.717**	**0.773**	0.736	0.743
mBERT (EN-MS-CM)	0.803	0.794	0.755	0.745	0.702	0.771	0.753	0.770
XLM (EN-MS)	0.813	0.812	0.701	0.689	0.675	0.746	0.730	0.749
Mixed XLM (CM)	0.807	0.804	0.792	0.771	0.643	0.712	0.747	0.762
Mixed XLM (EN-MS-CM)	**0.823**	**0.826**	**0.813**	**0.787**	0.677	0.743	**0.771**	**0.785**

Table 10: Results of different models on English

Model	Product Review		Movie Review		Stock Market		Avg.	
	acc	f1	acc	f1	acc	f1	acc	f1
mBERT (CM)	0.813	0.802	0.782	0.756	0.690	0.756	0.762	0.771
mBERT (EN-MS-CM)	0.815	0.743	0.780	0.782	0.683	0.765	0.759	0.763
XLM (EN-MS)	0.823	0.802	0.751	0.744	0.683	0.742	0.752	0.763
Mixed XLM (CM)	0.824	0.805	0.783	0.764	0.661	0.736	0.756	0.768
Mixed XLM (EN-MS-CM)	**0.828**	**0.818**	**0.805**	**0.785**	**0.696**	**0.769**	**0.776**	**0.791**

Table 11: Results of different models on Malay

Romadhona, N. P., et al., (2022). BRCC and SentiBahasaRojak: The First Bahasa Rojak Corpus for Pretraining and Sentiment Analysis Dataset. COLING.

Malay NLP **Challenges**

Standardized Benchmarking

LLM-centric NLP

Malay NLP **Challenges**

Lack of large-scale, human-curated, (generative task) data

Many work is still rule-based and uses traditional ML

Lack of standardized evaluation practice

Standardized Benchmarking

LLM-centric NLP

Missing from LLM-centric NLP discourse

Linguistic diversity of Malay dialects

Resource constrained

Can Multilingual Generative LLMs code-mix with SEA languages?

1 November 2023

Can we **generate code-mixed data with LLMs**?

Code-mixing (or code-switching) is common is SEA but **code-mixed data are hard to collect.**

Can we **use multilingual LLMs to generate code-mixed data for SEA languages**?

Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages

Prompting LLMs

CM

Assuming that you are an English-Malay bilingual speaker, how would you write a natural English-Malay code-mixed sentence about traffic?

(ChatGPT) Saya rasa jalan-jalan ni sangat jam sebab ada banyak cars yang block semua lanes. (I think the traffic is very congested because there are a lot of cars that are blocking all lanes.)

In this sentence, I have mixed English and Malay words together to form a code-mixed sentence. The sentence starts in Malay, with ... **<truncated>** 3

(Davinci-003) "Kerja yang mengambil masa yang lama seperti menunggu lama di laluan raya membuatkan saya bosan!" 0

(Davinci-002) The traffic is really bad today. 0

(a) Template: Assume to be bilingual speaker

CM

Imitate the speaking style of a person who can speak Singlish in one sentence about family.

(ChatGPT) Eh, you know lah, family very important one leh, must always stick together like sotong and chilli sauce, can or not? 3

(Davinci-003) "My family lah, we so close, can go anywhere together one!" 3

(Davinci-002) My family is the best! 0

(b) Template: Imitate speaking style

[Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages]

Methodology

5 topics: AI, family, food, traffic, weather

7 languages: Chinese, Indonesian, Malay, Tagalog, Tamil, Vietnamese, *Singlish*

6 prompt templates

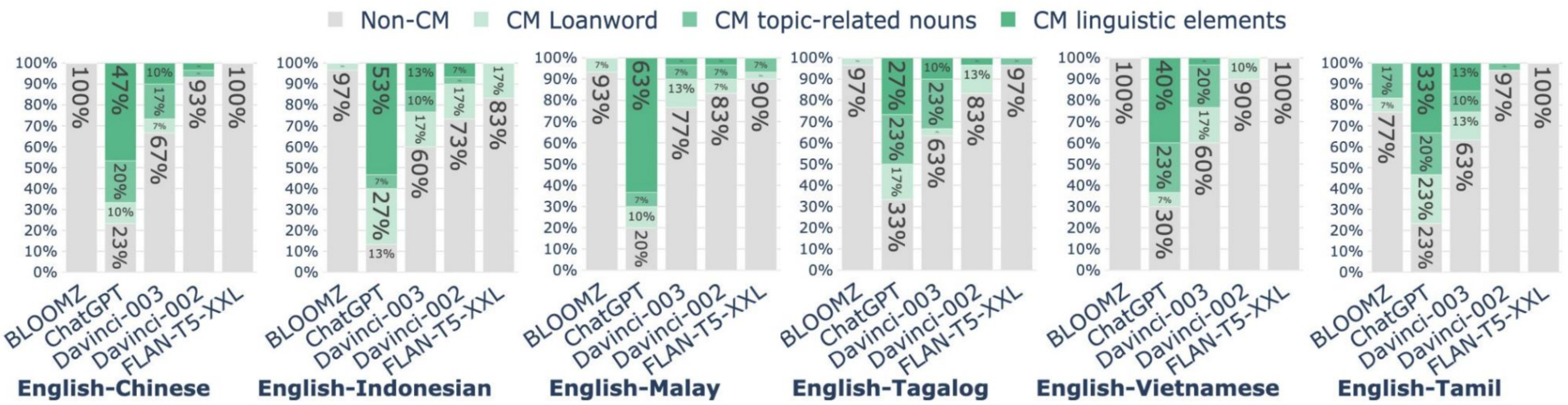
5 Models: ChatGPT, InstructGPT (davinci-002, -003), Flan-T5-XXL, BLOOMZ

Evaluation:

- **4 Code-Mixed level:** No CM, Loanwords, Topic-Related Nouns, Linguistic elements.
- **Fluency Analysis**:
  - Unnatural
  - Might be spoken by someone else
  - Annotators can see themselves speaking them

Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages

Results - Can LLMs generated Code-mixed text?



See our paper ([Yong et al., 2023](#)) for breakdown analysis according to topics, languages and templates

Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages

Results - Can LLMs generated Code-mixed text?

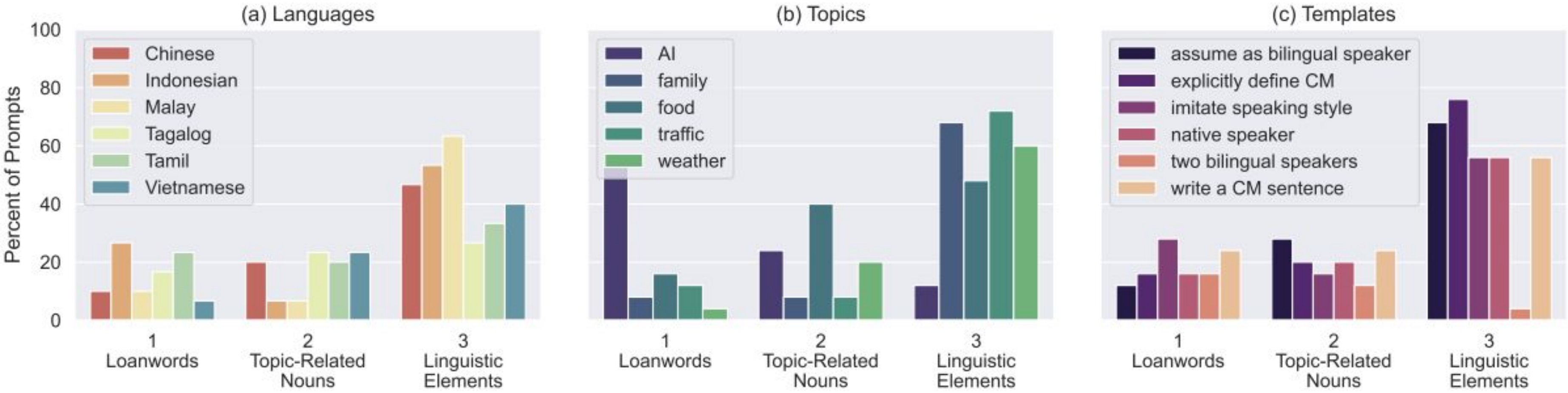
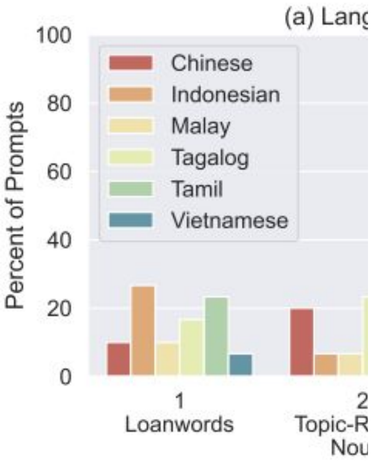


Figure 4: Analysis of code-mixed data generated by ChatGPT.

Results - -mixed text?



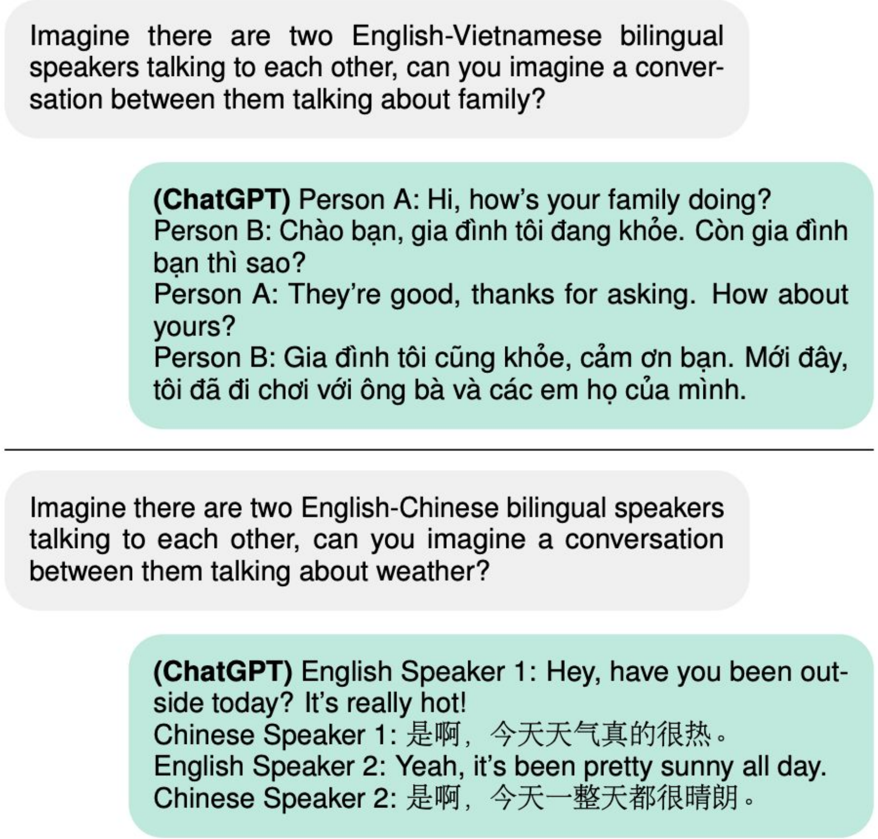
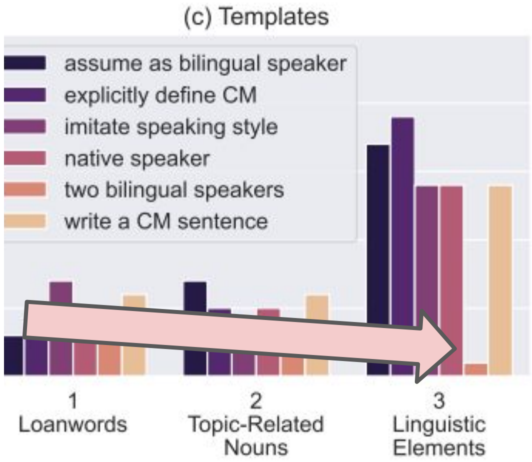


Figure 5: Failure cases when we prompt ChatGPT to imagine two bilingual speakers conversing about a specific topic.



atGPT.

Results - Can LLMs generated Code-mixed text?

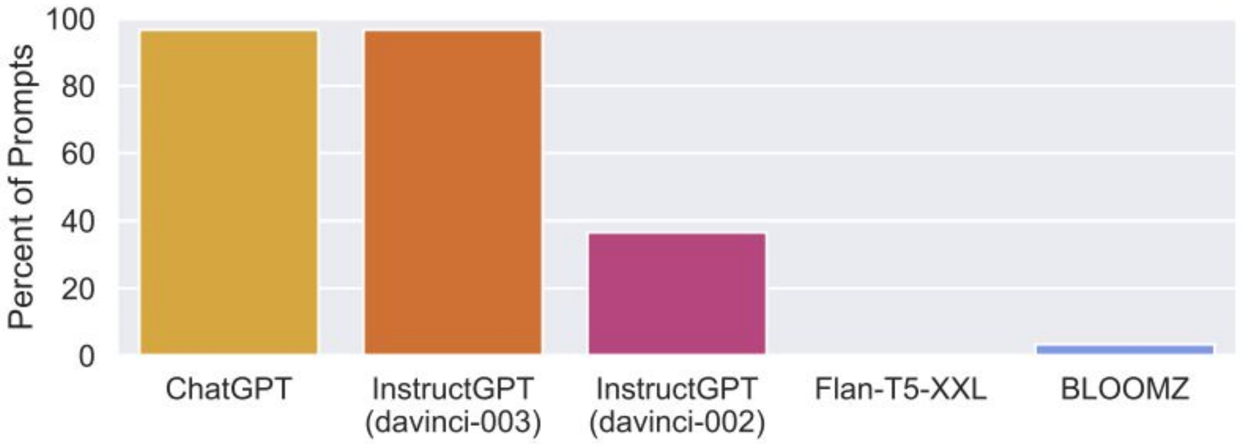


Figure 7: Models' performance in generating Singlish expressions through zero-shot prompting.

[Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages]

Fluency Analysis (ChatGPT)

Huge variance in fluency.

Singlish is mostly fluent, but had some semantic accuracies

- ChatGPT generate incorrect expression *"sotong and chilli sauce"* to describe close family bonds, where "sotong" is a Malay word for "squid."

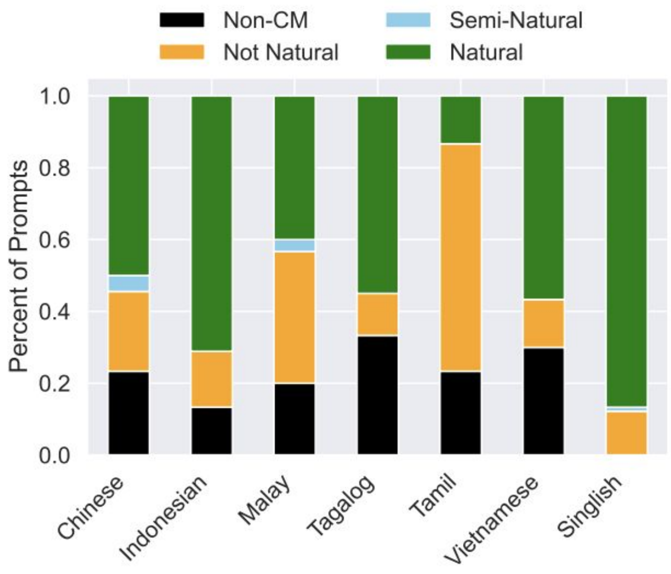


Figure 6: Naturalness of code-mixed (CM) text generated from ChatGPT. "Semi-Natural" indicates that the annotators see someone else from other region/culture speaking it, whereas "Natural" indicates that the annotators see themselves speaking it.

[Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages]

Failure Cases - Fail to follow instructions and give correct explanations

Imagine that you are a Malaysian speaker. How would you talk about traffic using both English and Chinese in a single sentence?

(ChatGPT) Assuming that you are asking for a sentence that combines both English and Chinese in a natural and commonly used way by Malaysian speakers, here is an example:
“今天的 traffic 很 teruk，我要赶紧走去 KLCC for meeting。”
This sentence incorporates the English word “traffic” and the Chinese word “teruk” (which means “bad” or “terrible” in English) to describe the situation of the traffic being very bad. The speaker then switches back to English to express their intention to hurry to KLCC for a meeting. This kind of code-switching is common in Malaysia, where speakers often mix languages in their daily conversations.

Figure 9: ChatGPT’s failure to follow the prompt. Underlined texts show its language misclassification.

Code-mixing refers to a phenomenon of combining two or more languages in a single sentence. Can you generate a code-mixed English-Malay sentence about weather?

(ChatGPT) “*The weather hari ni sangat panas, I rasa macam nak minum air sejuk je.*” (Translation: “The weather today is very hot, I feel like drinking cold water.”)
In this sentence, “hari ni” and “sangat panas” are in Malay, while “I” and “nak minum air sejuk je” are in English.

Figure 10: Underlined texts show wrong explanations about the code-mixed text. We italicize the *code-mixed sentences* to make it explicit to the reader.

Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages

Takeaways (Yong et al., 2023)

- ChatGPT has shown relative success in generating code-mixed texts, but we advise researchers to **exercise heavy caution**
  - Singlish: we find that syntactically-sound responses may contain semantic inaccuracies that are difficult for non-native speakers to detect.
- **Multilingual ≠ Code-Mix Compatible**
- We need to **incorporate code-mixed output recognition and generation capabilities in LLMs**.
- We cannot confidently identify ***how*** **ChatGPT code-mix** so due to the **lack of transparency**

Yong et al. (2023). Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages

Community efforts for NLP Development in SEA

1 November 2023

Aya: No Languages Left Behind (**Model** + Data)

Text Generation

Tulis satu karangan yang membincangkan cara-cara memupuk amalan gaya hidup yang sihat.

Langkah pertama untuk mengamalkan gaya hidup yang sihat adalah dengan menjaga kebersihan dan kesihatan diri …

Sentiment analysis

Review: We came here on a Saturday night. It wasn't as packed as I thought it would be [...] What is the sentiment [...]

It is a positive review.

Question answering

人为什么会流鼻血？

引起鼻出血的原因大致可分为3类：局部因素、全身因素和特发性因素。包括鼻部干燥、鼻部创伤 …

…

(Up to 101 languages, different tasks)

Aya: No Languages Left Behind (Model + **Data**)

- **Multilingual instruction-style** dataset covering **101 languages.**
- Fully **open-source (Apache license)**
- Led by Cohere For AI
  - Involve researchers from MIT, Berkeley, Apple, etc.
  - Contributors from all around the world
- **Looking for data contributors in our last 7-week sprint**

Prompt-Completion Dataset Contributions

- Release **high-quality, diverse, prompt-completion datasets.**

Long completions

Variety of tasks

Text-to-text format

Contribute Your Language (Task 2) → Rate Model Performance (Task 1) → Review User Feedback (Task 3)

Task 2: **Contribute Your Language**

Prompt

Apakah tanda, gejala dan kesan demam denggi?

Completion

Biasanya, individu yang dijangkiti virus denggi tidak simptomatik (80%) atau hanya mengalami gejala ringan seperti demam yang tidak rumit. Yang lain mempunyai penyakit yang lebih teruk (5%), dan dalam jumlah yang kecil ia mengancam nyawa. Tempoh inkubasi (masa antara pendedahan dan permulaan gejala) antara 3 hingga 14 hari, tetapi selalunya ia adalah 4 hingga 7 hari. Oleh itu, pelancong yang pulang dari kawasan endemik tidak mungkin mengalami demam denggi jika gejala bermula lebih dari 14 beberapa hari selepas tiba di rumah. Kanak-kanak sering mengalami gejala yang serupa dengan selesema biasa dan gastroenteritis (muntah dan cirit-birit) serta mempunyai risiko komplikasi yang lebih besar, walaupun gejala awal umumnya sederhana tetapi termasuk demam panas.

Task 1: Rate Model Performance

Prompt

Look out for: ✅ Correct Grammar ✅ Reasonable Length ✅ Clear Instructions

Apakah resipi untuk kek batik? Tunjukkan langkah-langkah secara terperinci.

Edit Distance[Prompt]: 0

Completion

Look out for: ✅ Clear Answer ✅ Full Sentences with Correct Answer ✅ Reasonable Length

Kek Batik ialah sejenis hidangan pencuci mulut yang berasal dari Malaysia. Di Malaysia ia dipanggil kek batik kerana ia mempunyai corak seperti batik daripada biskut yang telah dihancurkan.

Bahan: 1 cawan (100gm) Milo, 1 bungkus (300gm) biskut marie, ½ buku (125gm) butter/mentega, 2 sudu besar serbuk koko, ¾ tin (375gm) susu pekat atau krimer manis, Chocolate rice

Cara penyediaan:
1. Pecah-pecahkan setiap keping biskut Marie kepada 3-4 bahagian.
2. Alas loyang 8x8 inci dengan kertas parchment atau aluminium foil. Kertas hendaklah menutup dasar dan juga sisi

Edit Distance[Completion]: 191

Task 3: Review User Feedback

Review from Task 1

Read the Original Prompt

Jawab teka-teki yang lawak ini: Mee apa paling berat dalam dunia?

Read the Original Completion

Jawapan: Mee wantan (One tan).

Read the Edited Prompt Show Edits

Jawab teka-teki ~~yang~~ lawak ~~ini~~berikut : Mee apa paling berat dalam dunia?

Edit Distance[Prompt]: 0

Read the Edited Completion Show Edits

Jawapan: Mee wantan (One tan).

Edit Distance[Completion]: 0

Final Review (Task 3)

* Required

Is the edited prompt an improvement over the original?

- 1 - Much worse than Original
- 2 - Worse than original
- 3 - No noticeable difference compared to the original
- 4 - Better than original
- 5 - Much better than original (selected)

* Required

Is the edited completion an improvement over the original?

- 1 - Much worse than Original
- 2 - Worse than original
- 3 - No noticeable difference compared to the original (selected)
- 4 - Better than original
- 5 - Much better than original

(Optional)

Further improve the completion

Jawapan: Mee wantan. Ini kerana sebutan wantan mirip "one" tan. Unit tan ialah unit ukuran berat yang bersamaan dengan 2240 lb. Maka ia dikatakan sangat berat.

Aya: **Prizes**

500 Contribution Points

+ Digital Certificate and Badge
+ Aya limited edition stickers

1. 🏅 Digital Certificate and Badge for 500 Contribution points
2. 🌠 Aya limited edition stickers
3. 📘 C4AI Exploring the Unknown, Together Notebook

1,000 Contribution Pts

+ Digital Certificate & Badge
+ Aya limited edition stickers

1. 🏅 Digital Certificate and Badge for 1,000 Contribution points
2. 🌠 Aya limited edition stickers
3. 📘 C4AI Exploring the Unknown, Together Notebook
4. 👕 **Limited edition Aya t-shirt**

5,000 Contribution Pts

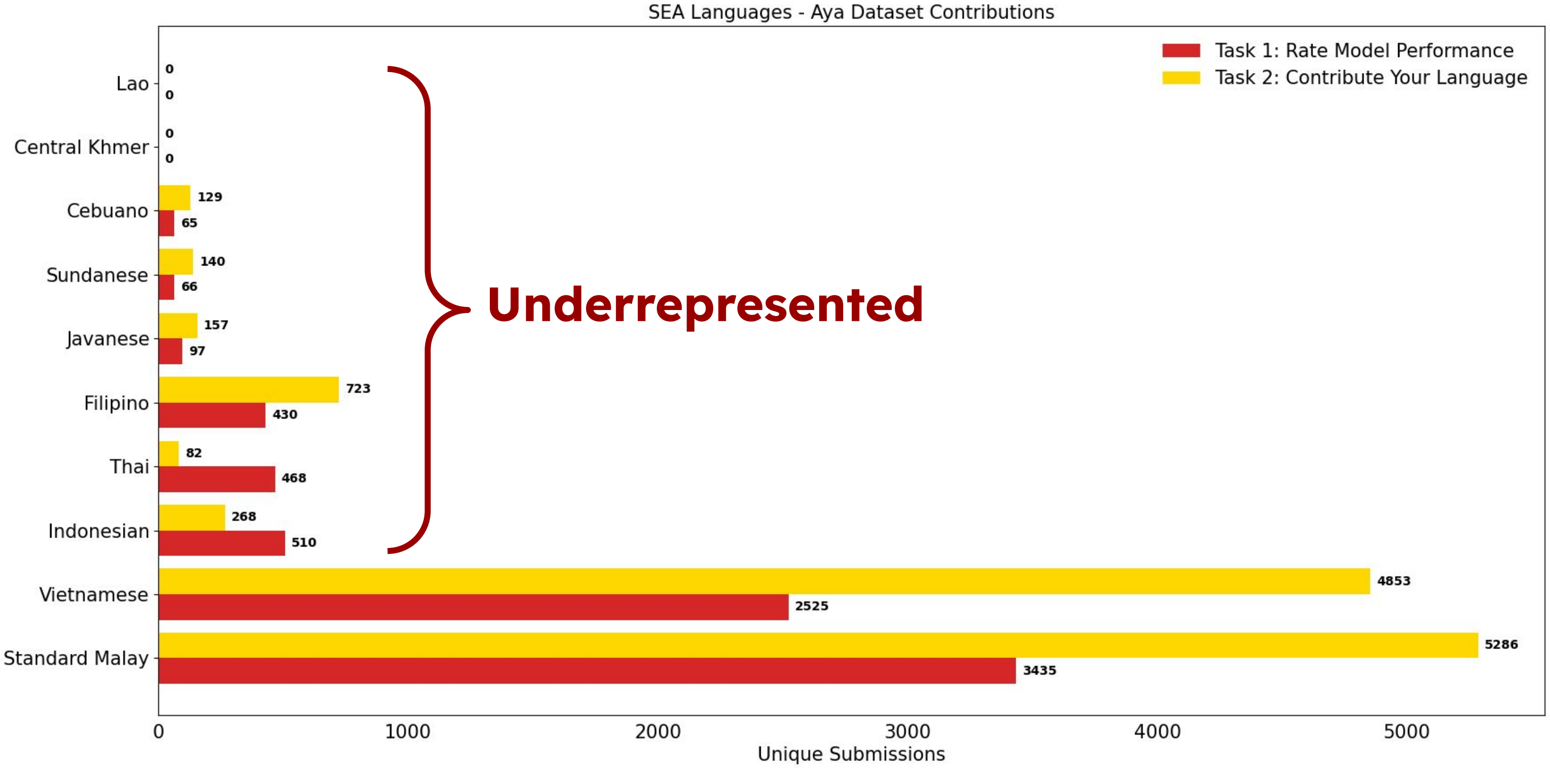
+ Digital Certificate & Badge
+ Aya limited edition stickers

1. 🏅 Digital Certificate and Badge for 5,000 Contribution points
2. 🌠 Aya limited edition stickers
3. 📘 C4AI Exploring the Unknown, Together Notebook
4. 👕 Limited edition Aya t-shirt
5. ✨ **Special edition Aya sweater**

Aya: **Prizes**

1. 🎖️ Digital Certificate and Badge for 5,000 Contribution points
2. 🌠 Aya limited edition stickers
3. 📘 C4AI Exploring the Unknown, Together Notebook
4. 👕 Limited edition Aya t-shirt
5. ✨ Special edition Aya sweater
6. 🧢 **C4AI Baseball Cap**
7. 🥳 **Virtual party exclusive invite**
8. 📑 Recognition as an **Aya Quality Champion in the paper acknowledgments.**
9. ✏️ **Eligibility to be considered for authorship on Aya dataset paper** based upon overall leadership in discord, attending meetings, and contributions beyond just UI
10. 💸 The Top 10 on this leaderboard will also receive a credit grant of **$500 to the Cohere API**

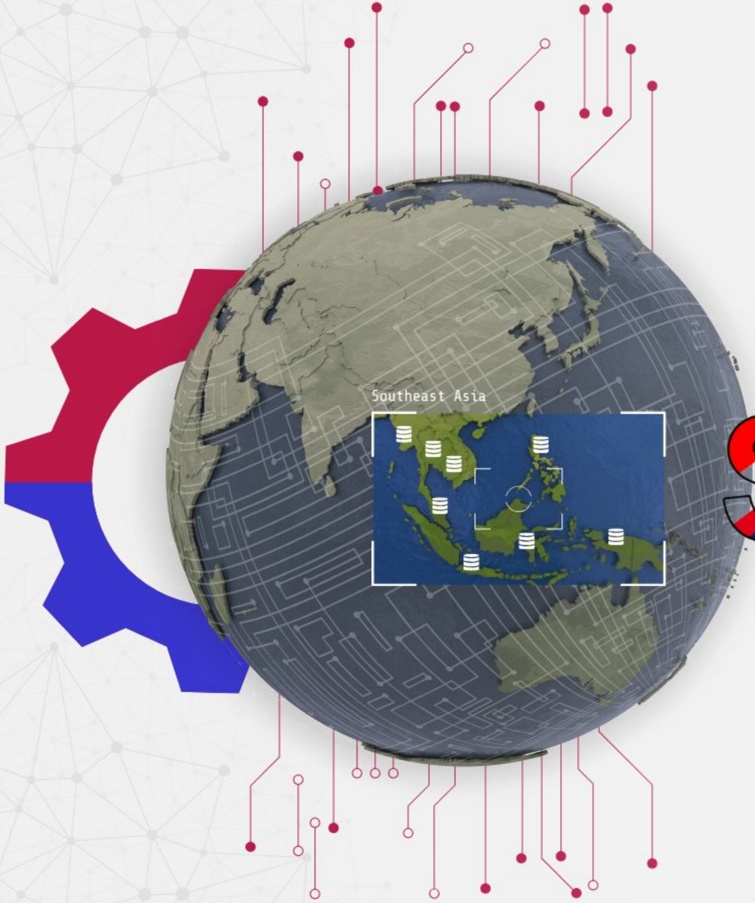
Aya: **SEA Languages**



Aya: **SEA Languages**

- Looking for **contributors** in next 7 weeks
    - Indonesian
    - Lao
    - Javanese
    - Sundanese
    - Malay
    - Thai
    - Vietnamese
    - Filipino
    - Cebuano
    - Khmer

Join Aya!
https://sites.google.com/cohere.com/aya-en/home



FROM THE CREATORS OF NUSACROWD

SEACrowd

Consolidating **Open Datasets** of
Southeast Asian Languages & Culture

introducing our our team.

HOLY LOVENIA

AI SINGAPORE

SAMUEL CAHYAWIJAYA

HKUST

RUOCHEN ZHANG

BROWN UNIVERSITY

FAJRI KOTO

MBZUAI

ZHENG-XIN YONG

BROWN UNIVERSITY

GENTA I. WINATA

BLOOMBERG

JAN CHRISTIAN BLAISE CRUZ

SAMSUNG R&D INSTITUTE PHILIPPINES

SEBASTIAN RUDER

GOOGLE

ALHAM F. AJI

MBZUAI

AYU PURWARIANTI

ITB & PROSA.AI

WILLIAM TJHI

AI SINGAPORE

YOU

SOMEONE EXCITED TO BE A PART OF SOMETHING BIG.

SEACrowd

SEACrowd aims to **centralize** & **standardize** all **publicly available AI datasets** in **SEA indigenous languages** and/or for **SEA cultures**.

SEACrowd

SEACrowd aims to **centralize** & **standardize** all **publicly available AI datasets** in **SEA indigenous languages** and/or for **SEA cultures**.

To greatly increase the **accessibility** of SEA datasets

To **understand** the AI landscape in SEA

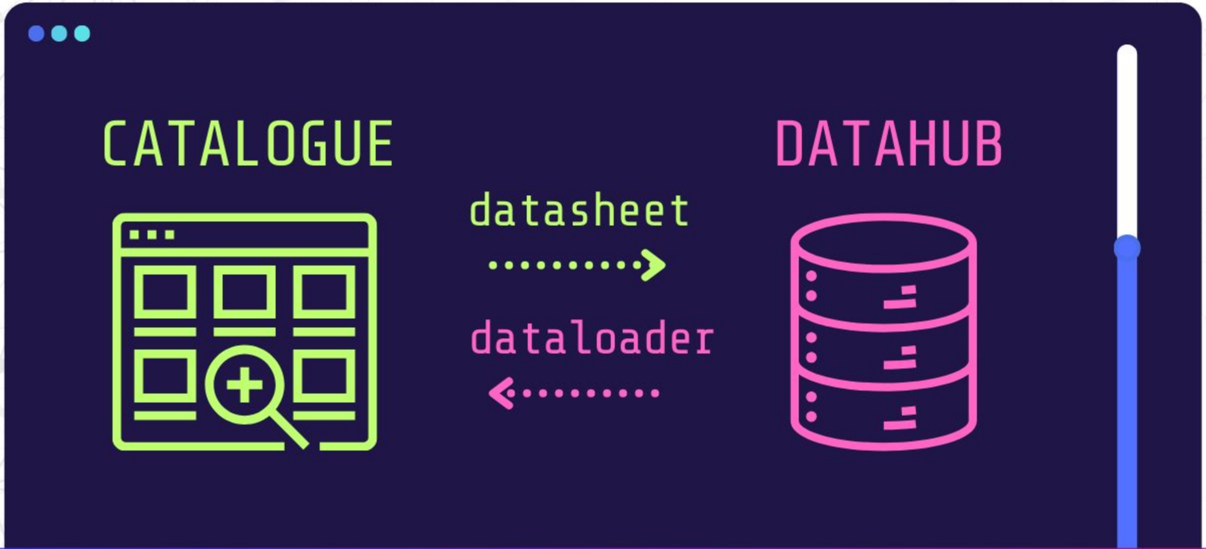
Promote **research** in SEA languages and cultures

Build AI models that **represent** SEA

SEACrowd aims to **centralize** & **standardize** all **publicly available AI datasets** in **SEA indigenous languages** and/or for **SEA cultures**.

Framework

Open access to the datasheets collected is provided through SEACatalogue, and the dataloaders that are used to retrieve the resources are implemented in SEACrowd Data Hub.



able AI
s.

Catalogue, and the
ed in SEACrowd Data Hub.

Predecessor

NUSACrowd in Indonesia

130+
Datasheets

Open access to 130+ datasheets is provided through NusaCatalogue

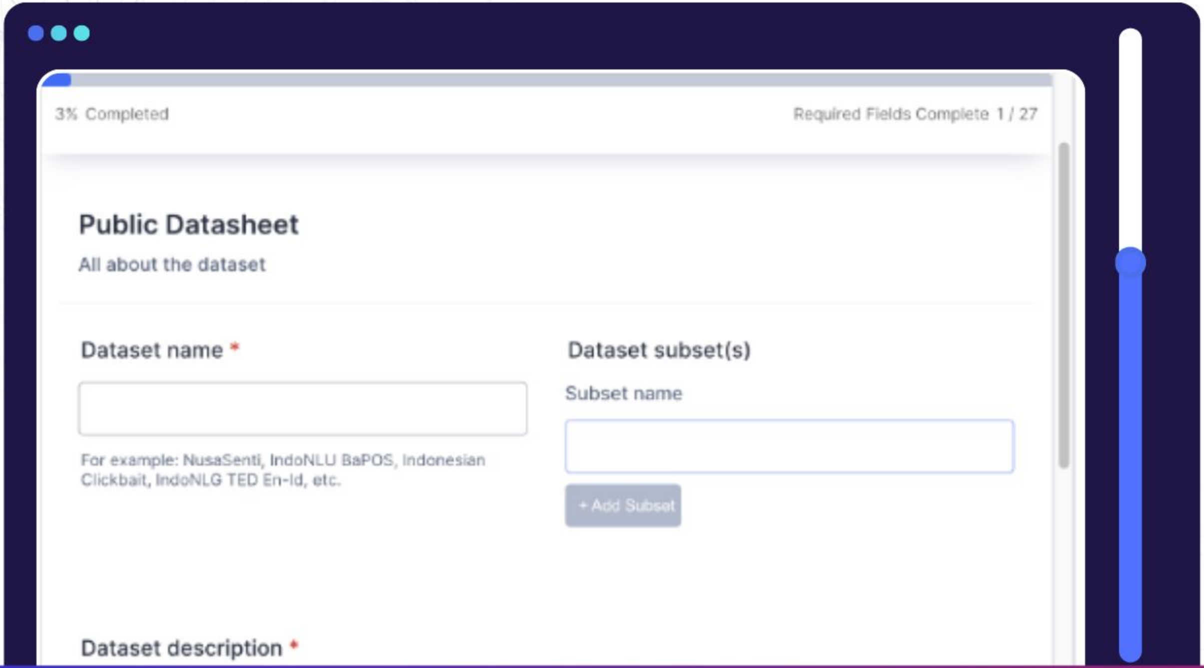
110+
Dataloaders

110+ dataloaders to access the resources implemented in NusaCrowd Data Hub

Published in ACL Findings 2023

01 Submitting Metadata for Existing Public Datasets

Approved datasheets will show up and indexed in **SEACrowd Catalogue**, which is under construction.



Metadata includes:

- Language dialect/register
- License
- Size
- Domain
- Data annotation protocol

02 Building DataLoader

From the **approved datasheets** from the previous task, you can help us build **HuggingFace's dataset dataloader** to ensure that all datasets in SEACrowd are standardised in terms of formatting.

```
class NusaXSenti(datasets.GeneratorBasedBuilder):
    """NusaX-Senti is a 3-labels (positive, neutral, negative) sentiment analysis dataset for 10 Indonesian local lang

    BUILDER_CONFIGS = (
        [nusantara_config_constructor(lang, "source", _SOURCE_VERSION) for lang in LANGUAGES_MAP]
        + [nusantara_config_constructor(lang, "nusantara_text", _NUSANTARA_VERSION) for lang in LANGUAGES_MAP]
        + [nusantara_config_constructor("", "source", _SOURCE_VERSION), nusantara_config_constructor("", "nusantara_te
    )

    DEFAULT_CONFIG_NAME = "nusax_senti_ind_source"

    def _info(self) -> datasets.DatasetInfo:
        if self.config.schema == "source":
            features = datasets.Features(
                {
                    "id": datasets.Value("string"),
                    "text": datasets.Value("string"),
                    "label": datasets.Value("string"),
                }
            )
        elif self.config.schema == "nusantara_text":
            features = schemas.text_features(["negative", "neutral", "positive"])
```

Predefined schema depending on task, everything is standardized.

nding

03 Identifying Private AI Datasets of SEA Languages, Cultures, and/or Regions

Unfortunately, some prior AI research on SEA languages is still hidden behind closed data.

In this task, you will search for prior research publications that did not make their data open. Our team will contact the reported work to negotiate the opening of their data with us.

03 Identifying Private AI Datasets of SEA Languages, Cultures, and/or Regions

nding

04 Opening Your Private AI Dataset of SEA

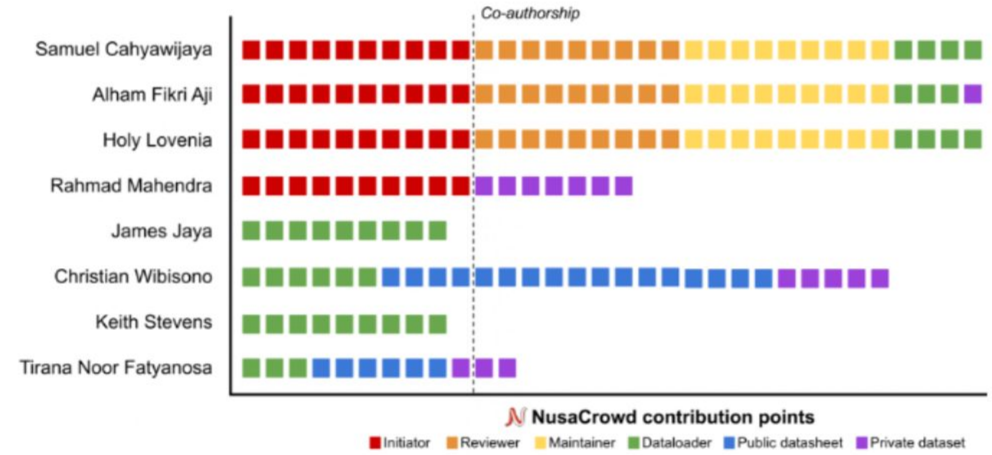
If you have previous work with closed data (or have been contacted by us), consider releasing your data and logging it with us. The data will still be owned by you and tied to your previous work, as we are simply creating a catalog of it.

More points on rare languages/high-quality dataset!

Why Contribute to SEACrowd.

Why Contribute to SEACrowd.

Contributions will be counted through a point system for fairness



Why Contribute to SEACrowd.

Contributions will be counted through a point system for fairness

Contributors up to certain threshold will be rewarded with merchandises

Why Contribute to SEACrowd.

Contributions will be counted through a point system for fairness

Contributors up to certain threshold will be rewarded with merchandises

Contributors up to certain threshold will be the co-author of the upcoming paper

Why Contribute to SEACrowd.

vill be
a point
ness

Contributors up to certain threshold will be rewarded with merchandises

Contributors up to certain threshold will be the co-author of the upcoming paper

Releasing and listing your dataset will give your work more recognition and exposure

Of course,

Meet Friends & Collaborators

doing something great.

https://github.com/SEACrowd

There's still a lot more to do, and SEACrowd could use ***your*** help.

Together, we stand on the verge of a breakthrough for Southeast Asia.
We **invite** you to contribute and be a part of this exciting journey.

OPEN CALL FOR CONTRIBUTIONS

1 NOVEMBER 2023 - 31 MARCH 2024

SEACrowd

SEACrowd aims to **centralize & standardize all publicly available AI datasets in SEA indigenous languages and/or for SEA cultures.**

A follow up of NusaCrowd, a similar movement for Indonesian

SEACrowd

Task 1: Submitting **metadata** of **existing, public** dataset of SEA languages

Dataset name *
For example: NusaSenti, IndoNLU BaPOS, Indonesian Clickbait, IndoNLG TED En-Id, etc.
Your answer

Dataset URL *
Direct link to the dataset repository.
Your answer

Dataset's HuggingFace URL
Link to the HuggingFace version of the dataset (if present). For example: https://huggingface.co/datasets/indonlu.
Your answer

Metadata includes:

- Language dialect and register
- License
- Size
- Domain
- Data annotation protocol
- and more

SEACrowd

Task 2: Working on DataLoader for approved dataset from Task 1. Predefined schema depending on task, everything is standardized

```
class NusaXSenti(datasets.GeneratorBasedBuilder):
    """NusaX-Senti is a 3-labels (positive, neutral, negative) sentiment analysis dataset for 10 Indonesian local lang

    BUILDER_CONFIGS = (
        [nusantara_config_constructor(lang, "source", _SOURCE_VERSION) for lang in LANGUAGES_MAP]
        + [nusantara_config_constructor(lang, "nusantara_text", _NUSANTARA_VERSION) for lang in LANGUAGES_MAP]
        + [nusantara_config_constructor("", "source", _SOURCE_VERSION), nusantara_config_constructor("", "nusantara_te
    )

    DEFAULT_CONFIG_NAME = "nusax_senti_ind_source"

    def _info(self) -> datasets.DatasetInfo:
        if self.config.schema == "source":
            features = datasets.Features(
                {
                    "id": datasets.Value("string"),
                    "text": datasets.Value("string"),
                    "label": datasets.Value("string"),
                }
            )
        elif self.config.schema == "nusantara_text":
            features = schemas.text_features(["negative", "neutral", "positive"])
```

SEACrowd

Task 3: Listing private SEA datasets

Explore existing work that do not open their data

List to us:
- Contact for data owner
- Data language and task
- Data size
- and so on

We will then contact the original owner!

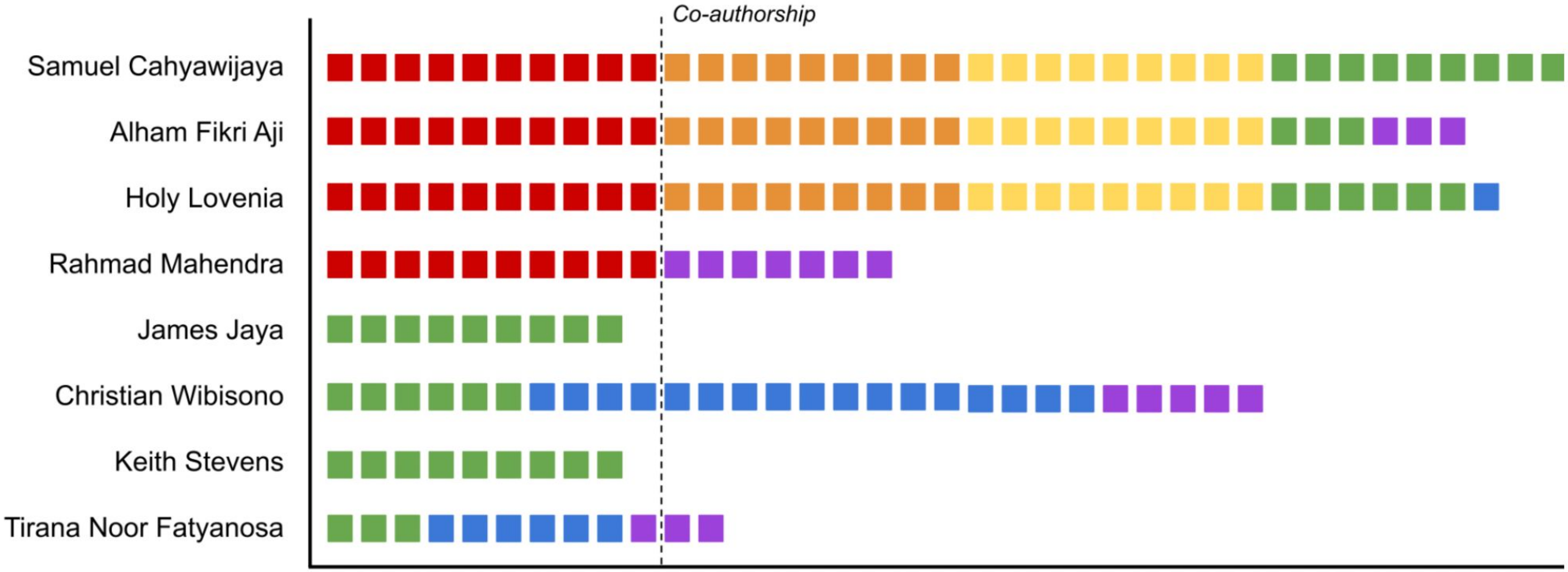
SEACrowd

Task 4: Opening up your private SEA dataset

- If you have closed NLP dataset, consider releasing it!
- We will also contact you, assuming your data is submitted through contribution 3.
- **More points** on **rare** languages/**high-quality** dataset!

SEACrowd

Contribute and earn points. Co-author our paper with enough contribution!



SEACrowd

Want to contribute more? Contact us here in AACL!

SEACrowd

Why contributing to SEACrowd?

- Contribution will be counted through a system point
- Contributors up to certain threshold will be rewarded with SWAG
- Contributors up to certain threshold will be the co-author of the paper
- Releasing and listing your dataset will (hopefully) gives your work more recognition.