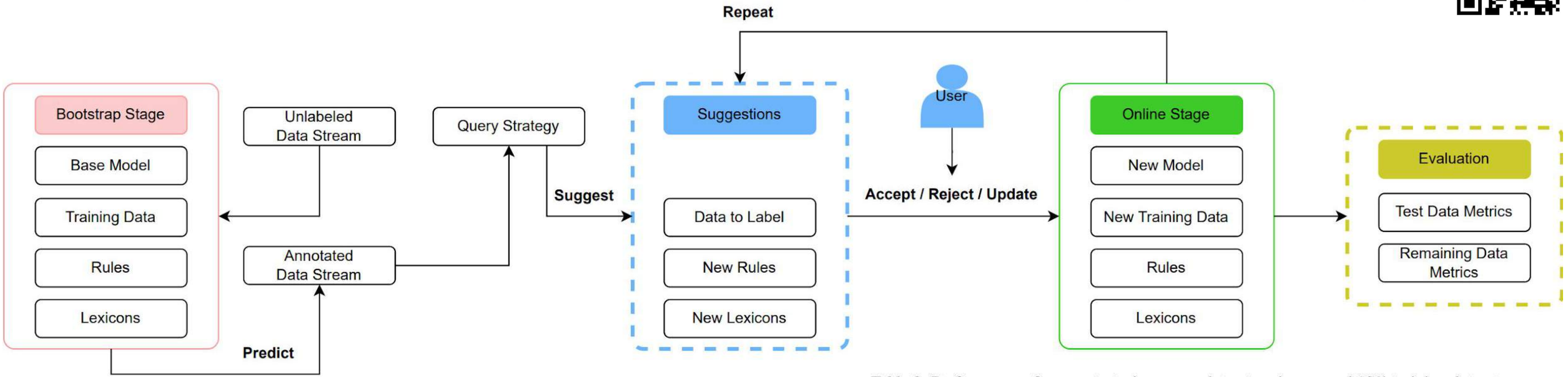
PyTAIL: An Open Source Tool for Interactive and Incremental Learning of NLP Models with Human in the Loop for Online Data

3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS)
6 Dec 2023 @ EMNLP 2023 in Singapore

Shubhanshu Mishra* (shubhanshu.com), Jana Diesner (University of Illinois at Urbana-Champaign),
*Work done while at UIUC

ArXiv: https://arxiv.org/abs/2211.13786
Dataset: https://doi.org/10.5281/zenodo.7236430
Code: https://github.com/socialmediaie/pytail



Problem formulation

- Given a large unlabeled corpus, can we:
  - label it efficiently using fewer human annotations?
  - allow human-in-the-loop injection of rules?
  - update models efficiently to work with new data?
- Proposal:
  - Use active learning for data labeling
  - Use interface to surface and inject prominent rules
  - Use incremental learning algorithms for model
- Highly applicable to social media data:
  - Model should adapt to new and streaming data

PyTAIL Benchmark for Social Media Active Learning

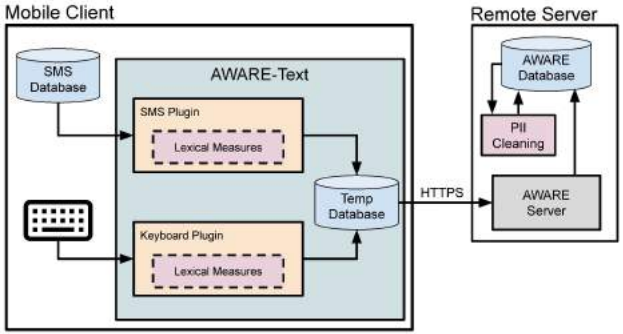
- Tasks for Social Media Text Classification: Abusive, Sentiment, Uncertainty
- 10 tasks, 200K social media posts
- Derived from Social Media IE Multi Task Benchmark – https://doi.org/10.5281/zenodo.5867160

Table 2: Performance of query strategies across datasets using around 10% training dataset.

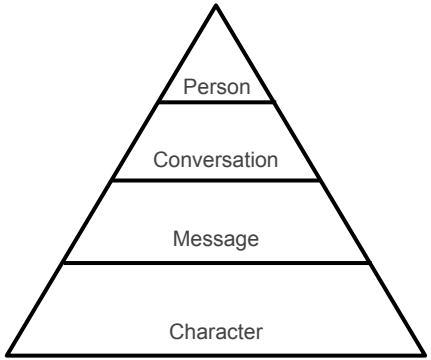
task	dataset	round	N	N_{left}	$\%_{used}$	Full	Rand	E_{top}	E_{prop}	M_{top}	M_{prop}
Test Dataset											
ABUSIVE	Founta	42	41,861	37,661	0.10	0.79	0.77	0.78	0.78	0.79	0.77
	WaseemSRW	14	13,072	11,672	0.11	0.82	0.79	0.78	0.77	0.78	0.76
SENTIMENT	Airline	9	8,725	7,825	0.10	0.82	0.76	0.78	0.79	0.77	0.77
	Clarin	45	44,299	39,799	0.10	0.66	0.63	0.61	0.62	0.63	0.63
	GOP	8	7,121	6,321	0.11	0.67	0.63	0.64	0.63	0.62	0.64
	Healthcare	1	590	490	0.17	0.59	0.64	0.60	0.61	0.60	0.60
	Obama	2	1,777	1,577	0.11	0.63	0.56	0.60	0.58	0.59	0.57
	SemEval	13	12,145	10,845	0.11	0.65	0.59	0.60	0.61	0.58	0.61
UNCERTAINITY	Riloff	2	1,201	1,001	0.17	0.78	0.77	0.76	0.77	0.76	0.79
	Swamy	1	555	455	0.18	0.39	0.39	0.40	0.39	0.34	0.31
Remaining Dataset											
ABUSIVE	Founta	42	41,861	37,661	0.10	NaN	0.77	0.80	0.78	0.81	0.78
	WaseemSRW	14	13,072	11,672	0.11	NaN	0.78	0.79	0.77	0.80	0.76
SENTIMENT	Airline	9	8,725	7,825	0.10	NaN	0.75	0.79	0.79	0.80	0.78
	Clarin	45	44,299	39,799	0.10	NaN	0.62	0.62	0.62	0.64	0.63
	GOP	8	7,121	6,321	0.11	NaN	0.62	0.64	0.62	0.63	0.63
	Healthcare	1	590	490	0.17	NaN	0.53	0.56	0.53	0.47	0.50
	Obama	2	1,777	1,577	0.11	NaN	0.54	0.56	0.57	0.56	0.56
	SemEval	13	12,145	10,845	0.11	NaN	0.61	0.62	0.62	0.63	0.62
UNCERTAINITY	Riloff	2	1,201	1,001	0.17	NaN	0.80	0.82	0.84	0.82	0.81
	Swamy	1	555	455	0.18	NaN	0.37	0.40	0.40	0.33	0.36

AWARE-Text: An Android Package for Mobile Phone Based Text Collection and On-Device Processing

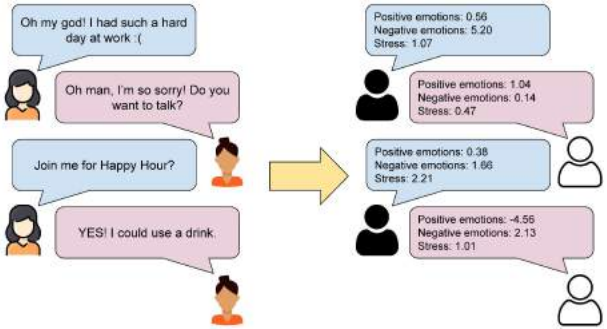
Salvatore Giorgi, Garrick Sherman, Douglas Bellew, Sharath Chandra Guntuku, Lyle Ungar, Brenda Curtis



Pipeline for collecting both SMS and keystroke data



Levels of Data Collection



Privacy preservation via on-device lexical processing

https://github.com/TTRUCurtis/aware-text

NIH National Institute on Drug Abuse Intramural Research Program
Penn University of Pennsylvania

Beyond the Repo: A Case Study on Open Source Integration with GECToR

Sanjna Kashyap, Zhaoyang Xie, Kenneth Steimel, Nitin Madnani
Educational Testing Service

Task

Integrate the open source **GECToR** code and models (developed by Grammarly) into our production NLP pipeline.

Issues Faced

- Not under active development
- Used older versions of Python, Pytorch and AllenNLP
- No versioning or packaging
- Did not fully exploit AllenNLP's high-level abstractions

Lessons Learned

- Projects should explicitly state a purpose.
- Test, test, test!
- Estimation of effort is hard but necessary.
- Always have a contingency plan.

Link to paper

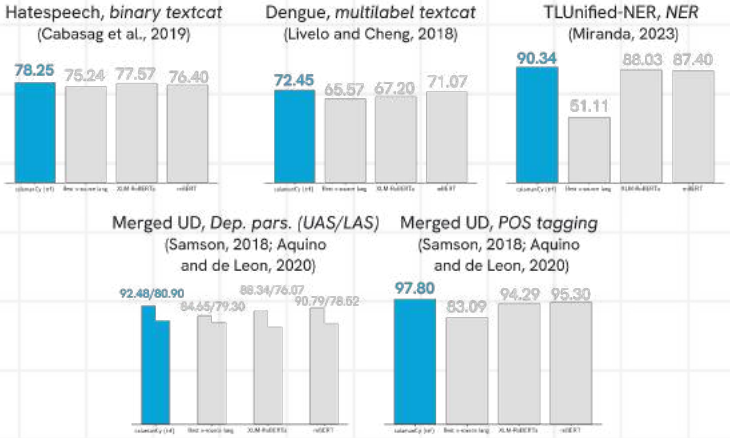
ETS AI LABS

calamanCy: A Tagalog Natural Language Processing Toolkit

Lester James V. Miranda

We created a dedicated NLP toolkit for **Tagalog**, a **low-resource language** from the Philippines.

calamanCy provides **out-of-the-box** pipelines that **outperform** both cross-lingual and multilingual transfer learning techniques on a variety of tasks.



github.com/ljvmiranda921/calamanCy

```python
import calamancy

nlp = calamancy.load("tl_calamancy_md")
# John went to Japan
doc = nlp("Pumunta si Juan sa Japan")

# Get entities
[token.ent_type_ for token in doc]
# Get tags
[(token.pos_, token.tag_) for token in doc]
# Get dep relations
[token.dep_ for token in doc]
```

tl_calamancy_md

Medium-sized pipeline using floret (50k vectors, 200 dims, 77 MB)

tl_calamancy_trf

Transformer-based pipeline using RoBERTa (813 MB)

tl_calamancy_lg

Large-sized pipeline using fasttext (714k vectors, 300 dims, 455 MB)

Work in progress...

Based on an LM pretrained on a more diverse corpus.

Download Paper

The Third Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)

Deepparse : An Extendable, and Fine-Tunable State-Of-The-Art Library for Parsing Multinational Street Addresses

David Beauchemin (david.beauchemin@ift.ulaval.ca), Marouane Yassine

Motivations

- Address parsing is essential to many applications, such as geocoding and record linkage.
- Most applications are confined to academic endeavours or with little availability of free and easy-to-use open-source solutions.

Related work

- Sharma et al. (2018): Parse monolingual address using a feedforward neural network.
- Mokhtari et al. (2019): Parse monolingual address using different RNN architectures.
- OpenVenues (2016): Libpostal Python Library.

Contributions

- We describe an open-source Python library for multinational address parsing.
- We describe its implementation details and natural extensibility due to its fine-tuning possibilities.
- We benchmark it against other open-source libraries.

PyThaiNLP: Thai Natural Language Processing in Python

A free and open-source natural language processing (NLP) library for Thai language.

Features in PyThaiNLP

Tokenizers	**Phonetic Algorithm and Transliteration**	**Embedding**	**Sequence Tagging**
Character Cluster and Syllable Level Word Level Sentence Level	Grapheme-to-Phoneme Soundex Thai-English Transliteration	Word Level Sentence Level	Named-Entity Recognition Part-of-Speech Tagging
Automatic Speech Recognition*	**Co-reference and Entity Linking**	**Spell Checking**	**Machine Translation***

Datasets

VISTEC-TPTH-2020 (Limkonchotiwat et al., 2021)
Task: *Word Tokenization*; Domain: *social media*

SCB-MT-EN-TH* (Lowphansirikul et al., 2020)
Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Thai NER (Phatthiyaphaibun, 2022)
Task: *Named-Entity Recognition*; Domain: *news and Wikipedia articles*

Han-Coref (Phatthiyaphaibun and Limkonchotiwat, 2023)
Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Pre-trained Langauge Models

WangchanBERTa* (Lowphansirikul et al., 2021a)
Thai Pre-trained Language Model

WangchanGLM (Polpanumas et al., 2023)
Multilingual Instruction-Following Model

*: in collaboration with the VISTEC-depa Thailand Artificial Intelligence Research Institute

EMNLP 2023

PyThaiNLP/PyThaiNLP

NOMIC

GPT4All: An Ecosystem of Open Source Compressed Language Models

Yuvanesh Anand, Zach Nussbaum, Adam Treat, Aaron Miller, Richard Guo,
Ben Schmidt, Planet Earth, Brandon Duderstadt*, Andriy Mulyar*

TL;DR

A technical report and case study of how an open source gpt-3.5-turbo clone became the 3rd fastest growing github repository of all time

55,000+
Github Stars

40,000+
Chat Client Monthly Active Users

66,000+
Python Package Downloads/Month

25,000+
GPT4All Discord Members

nanoT5: A PyTorch Framework for Pre-training and Fine-tuning T5-style Models with Limited Resources

850+ Stars

Steps to test new Transformer idea:

< ------ Reproduce the baseline
- T5-base pre-training on C4 + T5-instruct fine-tuning on SNI
- We document software + hardware + configs + training curves

< ------ Modify the code with your idea
- Exposed training loop + minimalistic T5 implementation
- Written and optimised in PyTorch

< ------ Test it
- Training starts within minutes and takes 16 hours on 1 x A100
- Supports multi-GPU training, weights from HF Hub, other datasets

"nanoT5 is the ideal template to start your LLM research"
~Happy User

PyThaiNLP: Thai Natural Language Processing in Python

A free and open-source natural language processing (NLP) library for Thai language.

Features in PyThaiNLP

Tokenizers	**Phonetic Algorithm and Transliteration**	**Embedding**	**Sequence Tagging**
Character Cluster and Syllable Level Word Level Sentence Level	Grapheme-to-Phoneme Soundex Thai-English Transliteration	Word Level Sentence Level	Named-Entity Recognition Part-of-Speech Tagging
Automatic Speech Recognition*	**Co-reference and Entity Linking**	**Spell Checking**	**Machine Translation***

Datasets

VISTEC-TPTH-2020 (Limkonchotiwat et al., 2021)
Task: *Word Tokenization*; Domain: *social media*

SCB-MT-EN-TH* (Lowphansirikul et al., 2020)
Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Thai NER (Phatthiyaphaibun, 2022)
Task: *Named-Entity Recognition*; Domain: *news and Wikipedia articles*

Han-Coref (Phatthiyaphaibun and Limkonchotiwat, 2023)
Task: *Coreference Resolution*; Domain: *news and Wikipedia articles*

Pre-trained Langauge Models

WangchanBERTa* (Lowphansirikul et al., 2021a)
Thai Pre-trained Language Model

WangchanGLM (Polpanumas et al., 2023)
Multilingual Instruction-Following Model

*: in collaboration with the VISTEC-depa Thailand Artificial Intelligence Research Institute

EMNLP 2023

PyThaiNLP/PyThaiNLP

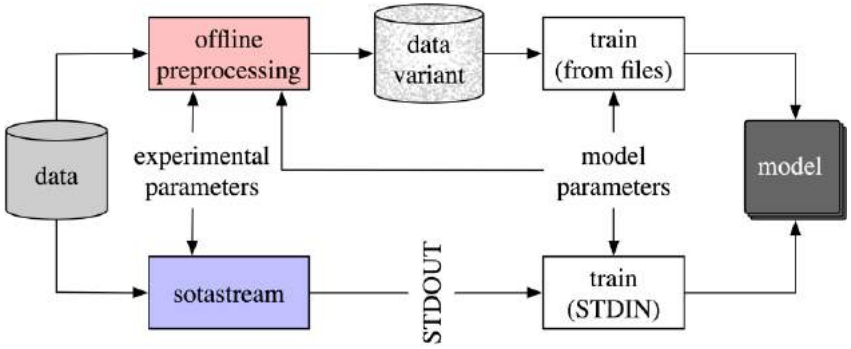
SOTASTREAM: A Streaming Approach to Machine Translation Training

Matt Post, Thamme Gowda, Roman Grundkiewicz, Huda Khayrallah, Rohit Jain, Marcin Junczys-Dowmunt

- Problem: standard off-line data preparation is expensive
  - data tensorized ahead of time
  - takes up time and disk space
  - ties the prepared dataset to a model configuration (e.g., vocabulary)
- Solution: generate data dynamically, on-the-fly!
- SOTASTREAM
  - ✓ Just as accurate
  - ✓ Just as fast
  - ✓ Saves disk space
  - ✓ More flexible

</>

- **pip install sotastream**
- https://github.com/marian-nmt/sotastream
- MIT License



Use Cases

Mixing multiple streams of data	Data augmentation for robustness	Filtering bad data examples
Subword tokenization sampling	Training document-context models	Alignments and other data types
Data collection tools: e.g., mtdata	Generating datasets for offline use	...

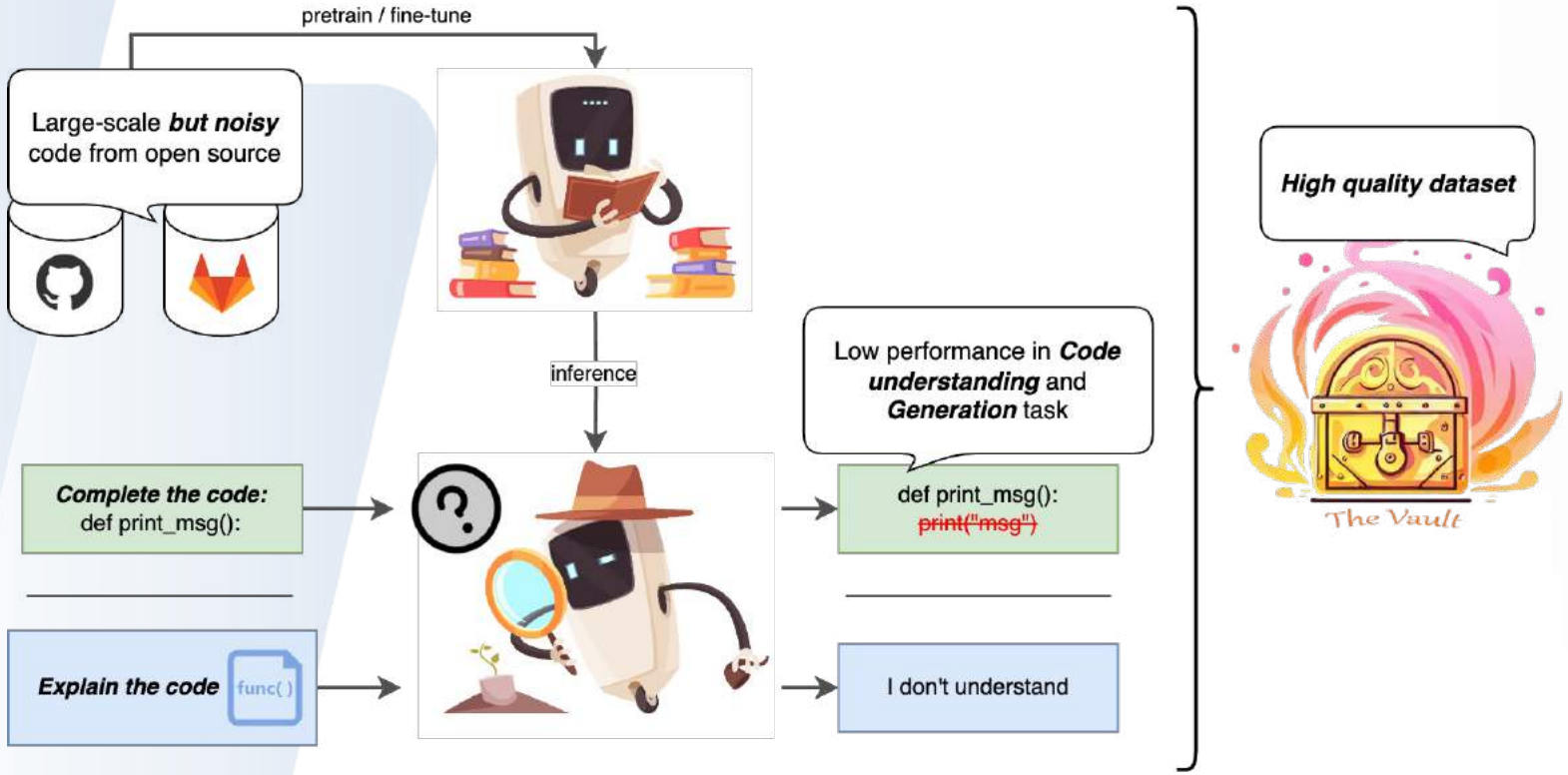
The Vault: A Comprehensive Multilingual Dataset for Advancing Code Understanding and Generation

EMNLP 2023

FPT Software AICenter

TL;DR

- TheVault is the ***largest corpus*** containing a ***multilingual code-text dataset.***
- CodeLLMs show a ***superior in performance*** when fine-tuned on The Vault for a wide range of tasks.



Check our repo