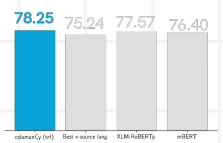
calamanCy: A Tagalog Natural Language Processing Toolkit

Lester James V. Miranda

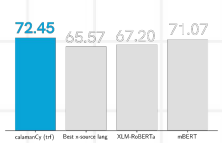
We created a dedicated NLP toolkit for **Tagalog**, a **low-resource language** from the Philippines.

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

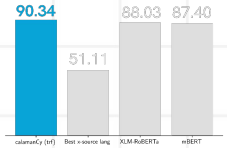
Hatespeech, *binary textcat*
(Cabasag et al., 2019)



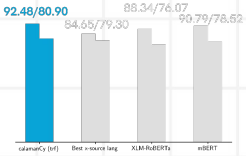
Dengue, *multilabel textcat*
(Livelo and Cheng, 2018)



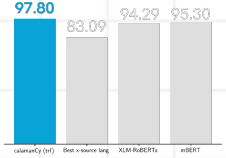
TLUnified-NER, *NER*
(Miranda, 2023)



Merged UD, *Dep. pars. (UAS/LAS)*
(Samson, 2018; Aquino and de Leon, 2020)



Merged UD, *POS tagging*
(Samson, 2018; Aquino and de Leon, 2020)



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)