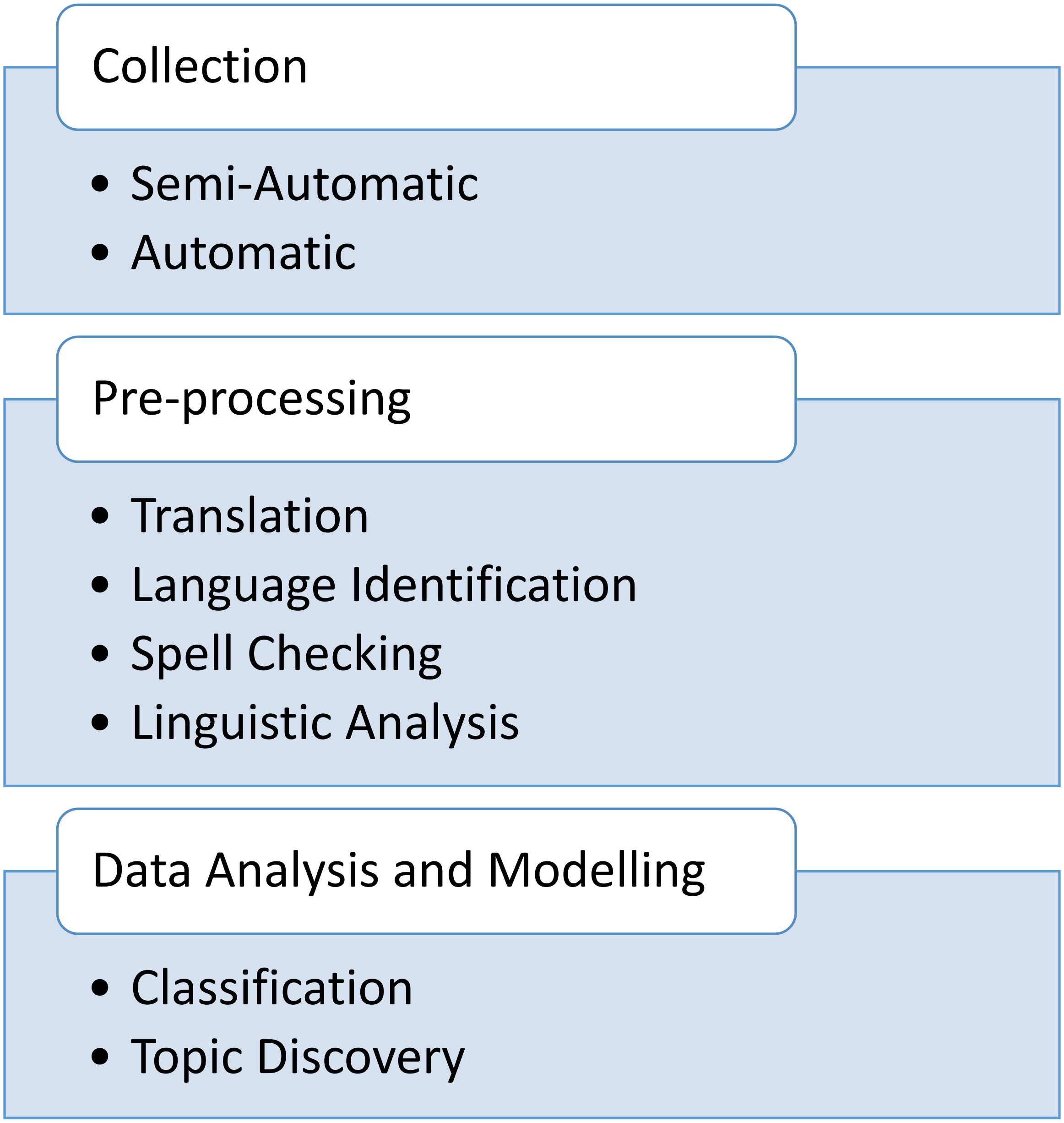
NLP Research Primer



Collection

- Semi-Automatic
- Automatic

Pre-processing

- Translation
- Language Identification
- Spell Checking
- Linguistic Analysis

Data Analysis and Modelling

- Classification
- Topic Discovery

Collection

As research works are now data-driven, there is a need for a databank of Philippine language resources. Towards addressing this concern, students who are interested will develop tools and techniques that can aid automatic collection and categorization of texts. This includes crawling the web for language resources and automatically storing and organizing them based on language. Related work includes clustering the languages and annotating each collected text.

Target Venue(s):

- Local: PCSC, ICE
- International (SCOPUS): TENCON, IALP, PACLIC

Starting Reference(s):

- Oco, N., Syliongka, L. R., Allman, T., & Roxas, R. E. (2016). Resources for Philippine Languages: Collection, Annotation, and Modeling. In Proceedings of the 30th Pacific Asia Conference on Language, Information, and Computation (pp. 433-438).
- Dita, S., Roxas, R. E., & Inventado, P. (2009). Building online corpora of Philippine languages. In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2 (pp. 646-653).

Pre-Processing

Textual data has been the main resource for numerous software programs. One of the integral considerations is the proper representation and use of high-quality data. To achieve such quality, text pre-processors – or subprograms that modify the raw data to custom fit or provide new data features to a given system – are needed. Currently, there are numerous pre-processors that are available. However, there exists no compilation of tools that are lightweight and flexible to different kind of systems or language domains. Students are to develop pre-processing tools for textual data. These may consist of the following:

- Tokenization
  - Cleaning
  - URLs
  - Special Characters
  - Length Limit
  - Duplicates
  - Stop words
- True-casing (e.g. john -> John)
- Stemming (Root words)
- Part of Speech (POS) Tagging (e.g. “dog” -> noun, “delicious” -> adjective)
- Text Transformation
  - Standard text normalization (e.g. resume -> résumé, canonicalization)
  - Unicode normalization (e.g. ñ -> U+00F1, Å -> U+00C5)
  - Shortcut text normalization (e.g. LOL -> “Laughing Out Loud”, gr8 -> “great”)
  - Spell / grammar check
  - Translation
- Linguistic Analysis
  - Word count, sentence count, phrase count
  - Lexical density, lexical variation
  - Syllable patterns
  - Compound word densities

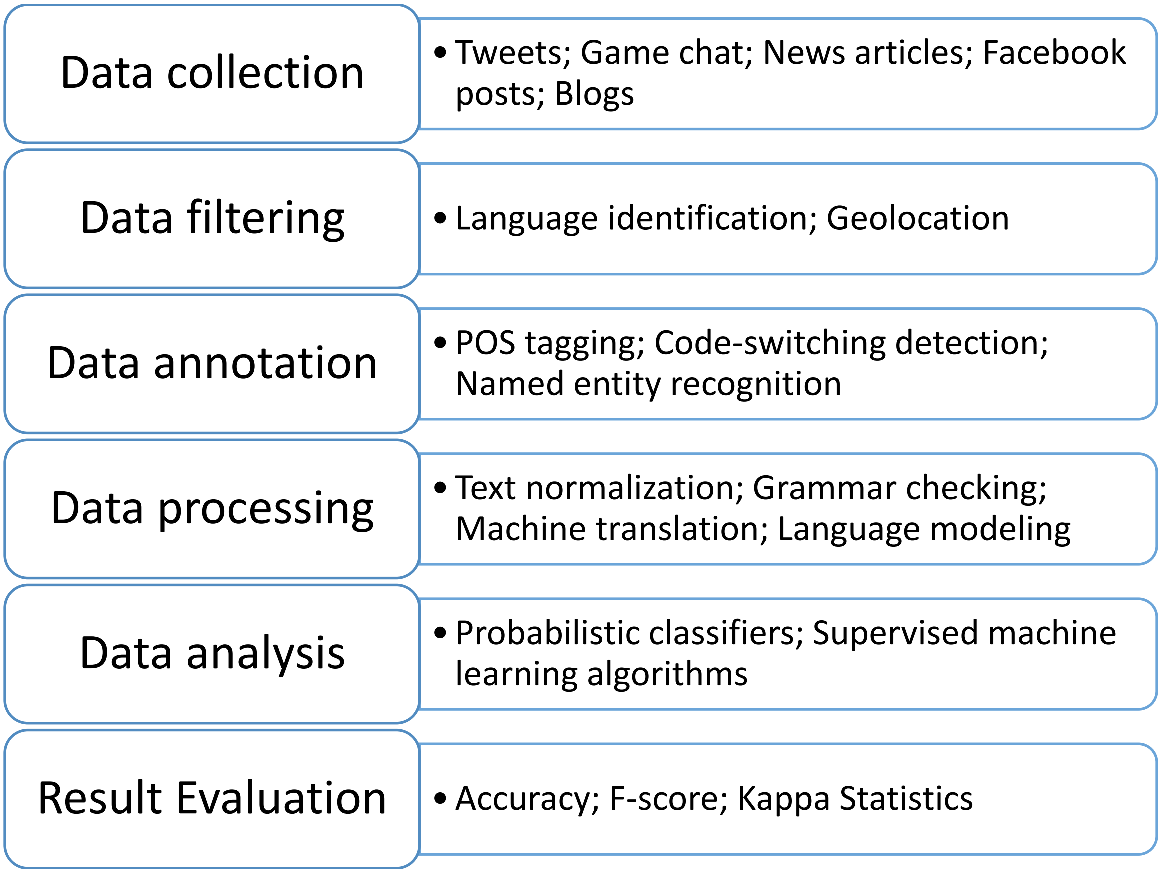
Target Venue(s):

- Local: PCSC
- International (SCOPUS): TENCON, IALP, PACLIC, ACL

Starting Reference(s):

- Go, M. P., & Nocon, N. (2017). Using Stanford part-of-speech tagger for the morphologically rich Filipino language. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (pp. 81-88).
- Imperial, J. M., & Ong, E. (2021). Diverse Linguistic Features for Assessing Reading Difficulty of Educational Filipino Texts. arXiv preprint arXiv:2108.00241.

Supervised Learning: Classification



Data collection
- Tweets; Game chat; News articles; Facebook posts; Blogs

Data filtering
- Language identification; Geolocation

Data annotation
- POS tagging; Code-switching detection; Named entity recognition

Data processing
- Text normalization; Grammar checking; Machine translation; Language modeling

Data analysis
- Probabilistic classifiers; Supervised machine learning algorithms

Result Evaluation
- Accuracy; F-score; Kappa Statistics

Target Venue(s):

- Local: PCSC, NNLPRS
- International (SCOPUS): TENCON, IALP
- Journal (SCOPUS): Philippine Political Science Journal, ACM Transactions on Asian Language Information Processing, Literary and Linguistic Computing

Starting Reference(s):

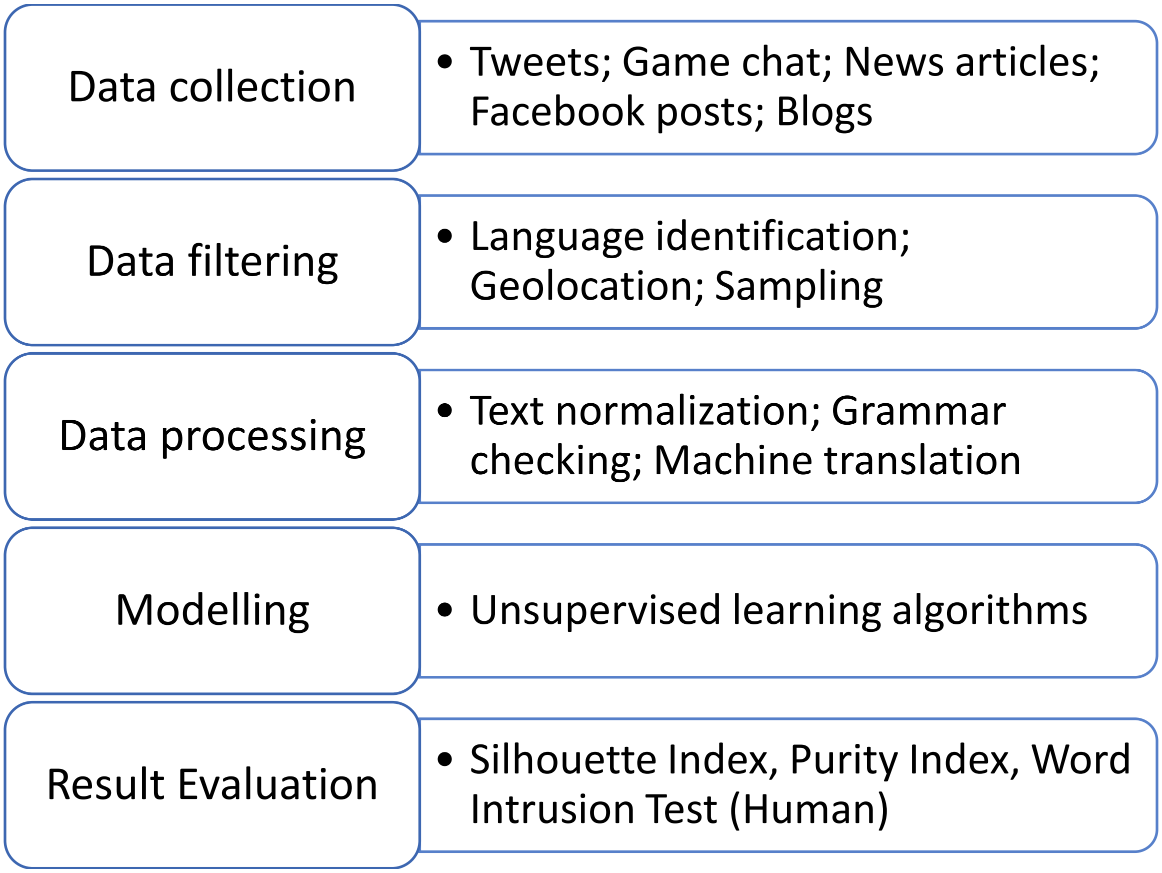
- Maceda, L., Llovido, J., & Satuito, A. (2018). Categorization of earthquake-related tweets using machine learning approaches. In 2018 International Symposium on Computer, Consumer and Control (IS3C) (pp. 229-232). IEEE.

- Maceda, L., Llovido, J., & Satuito, A. (2018). Categorization of earthquake-related tweets using machine learning approaches. In 2018 International Symposium on Computer, Consumer and Control (IS3C) (pp. 229-232). IEEE.

- Imperial, J. M., & Ong, E. (2021). Diverse Linguistic Features for Assessing Reading Difficulty of Educational Filipino Texts. *arXiv preprint* arXiv:2108.00241.

- Imperial, J. M., & Ong, E. (2020). Exploring hybrid linguistic feature sets to measure Filipino text readability. *In 2020 International Conference on Asian Language Processing* (IALP) (pp. 175-180). IEEE.
- Regalado, R. V. J., Chua, J. L., Co, J. L., & Tiam-Lee, T. J. Z. (2013). Subjectivity Classification of Filipino Text with Features Based on Term Frequency--Inverse Document Frequency. *In 2013 International Conference on Asian Language Processing* (pp. 113-116). IEEE.
- Regalado, R. V. J., & Cheng, C. K. (2012). Feature-based subjectivity classification of Filipino text. *In 2012 International Conference on Asian Language Processing* (pp. 57-60). IEEE.

Unsupervised Learning: Topic Discovery



Data collection

- Tweets; Game chat; News articles; Facebook posts; Blogs

Data filtering

- Language identification; Geolocation; Sampling

Data processing

- Text normalization; Grammar checking; Machine translation

Modelling

- Unsupervised learning algorithms

Result Evaluation

- Silhouette Index, Purity Index, Word Intrusion Test (Human)

Target Venue(s):

- Local: PCSC, NNLPRS
- International (SCOPUS): TENCON, IALP
- Journal (SCOPUS): Philippine Political Science Journal, ACM Transactions on Asian Language Information Processing, Literary and Linguistic Computing

Starting Reference(s):

- Soriano, C. R., Roldan, M. D. G., Cheng, C., & Oco, N. (2016). Social media and civic engagement during calamities: the case of Twitter use during typhoon Yolanda. Philippine Political Science Journal, 37(1), 6-25.
- Maceda, L. L., Llovido, J. L., & Palaoag, T. D. (2017). Corpus analysis of earthquake related Tweets through topic modelling. International Journal of Machine Learning and Computing, 7(6), 194-197.

- Ancheta, J. R., Sy, C., Maceda, L., Oco, N., & Roxas, R. (2017). Computer-assisted thematic analysis of typhoon Fung-Wong tweets. In TENCON 2017-2017 IEEE Region 10 Conference (pp. 723-726). IEEE.
- Austero, L. D., Sy, C. Y., & Canon, M. J. P. (2018). Discovering Themes from Online News Articles on the 2018 Mt. Mayon Eruption. In 2018 International Symposium on Computer, Consumer and Control (IS3C) (pp. 242-245). IEEE.
- Imperial, J. M., Cruz, A. D. L., Malaay, E., & Roxas, R. E. (2022). Cross-Textual Analysis of COVID-19 Tweets: On Themes and Trends Over Time. In Proceedings of Sixth International Congress on Information and Communication Technology (pp. 813-821). Springer, Singapore.

Resources

Corpora

Philippine Languages Online Corpora (https://github.com/imperialite/Philippine-Languages-Online-Corpora) - This is a repository of various language resources collected over the past several years. The contents of the dataset are described below.

- Tweets – 2013 Election, COVID-19 PH Tweets, Typhoon Yolanda
- WordNets – Filipino WordNet
- Dictionaries – Filipino dictionary
- Tagged data – Tagged
- Language models – Religious text in different languages
- Multilingual corpora – Religious articles; Parallel corpus
- English and Filipino monolingual corpora – Wikipedia articles

NLPinas (https://www.nlpinas.org.ph/) – An online repository maintained by NLPinas, a non-profit group for Natural Language Processing research and development in the Philippines.

NLP-Phil (https://sites.google.com/site/roxasreo/home/nlp-phil) – Launched by the Computing Society of the Philippines Special Interest Group on Natural Language Processing (CSP SIG-NLP), NLP-Phil hopes to provide a space for NLP researchers in the Philippines and/or those working on Philippine languages to share, and also access links to research papers, open language resources, and codes.

Preprocessing Tools

Filipino Linguistic Extractors (https://github.com/imperialite/filipino-linguistic-extractors). This repository contains scripts for extracting linguistic features from Filipino texts. The scripts were created for Joseph Imperial's MSCS thesis in readability assessment of children's books. The complete list of linguistic features including the formulas and descriptions are uploaded with this repo.

Filipino Part-of Speech Taggers. Extracting part-of-speech tags may be an important step in many NLP tasks. The following repositories contain code and instructions for implementing POS tagging for Filipino texts. Conveniently, both repositories follow the same tagset information which can be found at http://goo.gl/dY0qFe.

1. Based on LSTM (by Jan Christian Blaise Cruz, DLSU Machine Learning Group) - https://github.com/jcblaisecruz02/filipino-pos
2. Filipino POS Tagger integrated at Stanford CoreNLP (by Go and Nocon, PACLIC 2017) - https://github.com/matthewgo/FilipinoStanfordPOSTagger

Models

HuggingFace Pretrained Models for Tagalog (https://huggingface.co/jcblaise) – This HuggingFace repository contains a wide range of neural language models pretrained using various architectures (BERT, GPT-2, RoBERTa, ELECTRA) for Tagalog. Maintained by Jan Christian Blaise Cruz, DLSU Machine Learning Group.

Online Tools

Here are other tools that you may find useful for doing NLP research:

- Twint (https://github.com/twintproject/twint) – Need Twitter data for your research? Use Twint. Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles **without** using Twitter's API.

- LanguageTool (https://languagetool.org/dev) – Need to add a preprocessing or filtering step such as spellchecking or language recognition? Use LanguageTool as an API and integrate it on your Java code.

- Weka (https://www.cs.waikato.ac.nz/ml/weka/) - Need an easy application to run your experiments without programming? Use Weka. Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.