NLP Topic Map

Collection

- Semi-automatic
- Automatic

Pre-processing

- Annotators
- Translation
- Language identification
- Spell checking

Processing: Social computing

- Sentiment analysis
- Classification
- Topic discovery
- Information extraction

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, NNLPRS
- International (SCOPUS): TENCON, IALP, PACLIC, ACM Transactions on Asian Language Processing

Starting Reference(s):

Authors	Oco, Nathaniel; Syliongka, Leif Romeritch; Allman, Tod; Roxas, Rachel Edita
Title	Resources for Philippine Languages: Collection, Annotation, and Modeling
Publication	The 30th Pacific Asia Conference on Language, Information and Computation
Pages	433-438
Year	2016
Publisher	Institute for the Study of Language and Information at Kyung Hee University

Authors	Dita, Shirley N; Roxas, Rachel EO; Inventado, Paul;
Title	Building Online Corpora of Philippine Languages
Publication	The 23rd Pacific Asia Conference on Language, Information and Computation
Pages	646-653
Year	2009
Publisher	City University of Hong Kong

Authors	Oco, Nathaniel; Ilao, Joel; Roxas, Rachel Edita; Syliongka, Leif Romeritch;
Title	Measuring language similarity using trigrams: Limitations of language identification
Publication	2013 International Conference on Recent Trends in Information Technology (ICRTIT)
Pages	478-481
Year	2013
Publisher	IEEE

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. In order 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)
- Feature Extraction (Affixes)
- Stemming (Root words)
- 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

Target Venue(s):

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

Starting Reference(s):

Authors	Nocon, Nicco; Borra, Allan;
Title	SMTPOST: Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging
Publication	Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation
Pages	391-396
Year	2016
Publisher	Institute for the Study of Language and Information at Kyung Hee University

Authors	Nocon, Nicco; Oco, Nathaniel; Ilao, Joel; Roxas, Rachel Edita;
Title	Philippine component of the network-based ASEAN language translation public service

Publication	2014 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)
Year	2014
Publisher	IEEE

Authors	Oco, Nathaniel; Roxas, Rachel Edita;
Title	Pattern matching refinements to dictionary-based code-switching point detection
Publication	Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation
Pages	07-10
Year	2012

Authors	Oco, Nathaniel; Borra, Allan;
Title	A grammar checker for Tagalog using LanguageTool
Publication	Asian Language Resources collocated with IJCNLP 2011
Pages	2-9
Year	2011

Annotators

To enable computers to understand human language as text, it is fed with large number of data to teach it that certain words are *common nouns*, some words are *PERSON* entity type, and certain sequence of text is considered a *noun phrase*. These annotations are then used by higher level systems such as classifiers, grammar checkers, information extractors as generalized representation of the words. Currently, there are numerous approaches available but only very few have been applied to local language domains. Students are to develop annotator tools for textual data. These may consist of the following:

- Part-of-Speech (POS) Tagging
- Named Entity Recognition
- Constituency Parsing (note: this must have POS tagging as prerequisite)
- Dependency Parsing (note: this must have constituency parsing as prerequisite)

Target Venue(s):

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

Starting Reference(s):

Authors	Nocon, Nicco; Borra, Allan;
Title	SMTPOST: Using Statistical Machine Translation Approach in Filipino Part-of-Speech Tagging
Publication	Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation
Pages	391-396
Year	2016
Publisher	Institute for the Study of Language and Information at Kyung Hee University

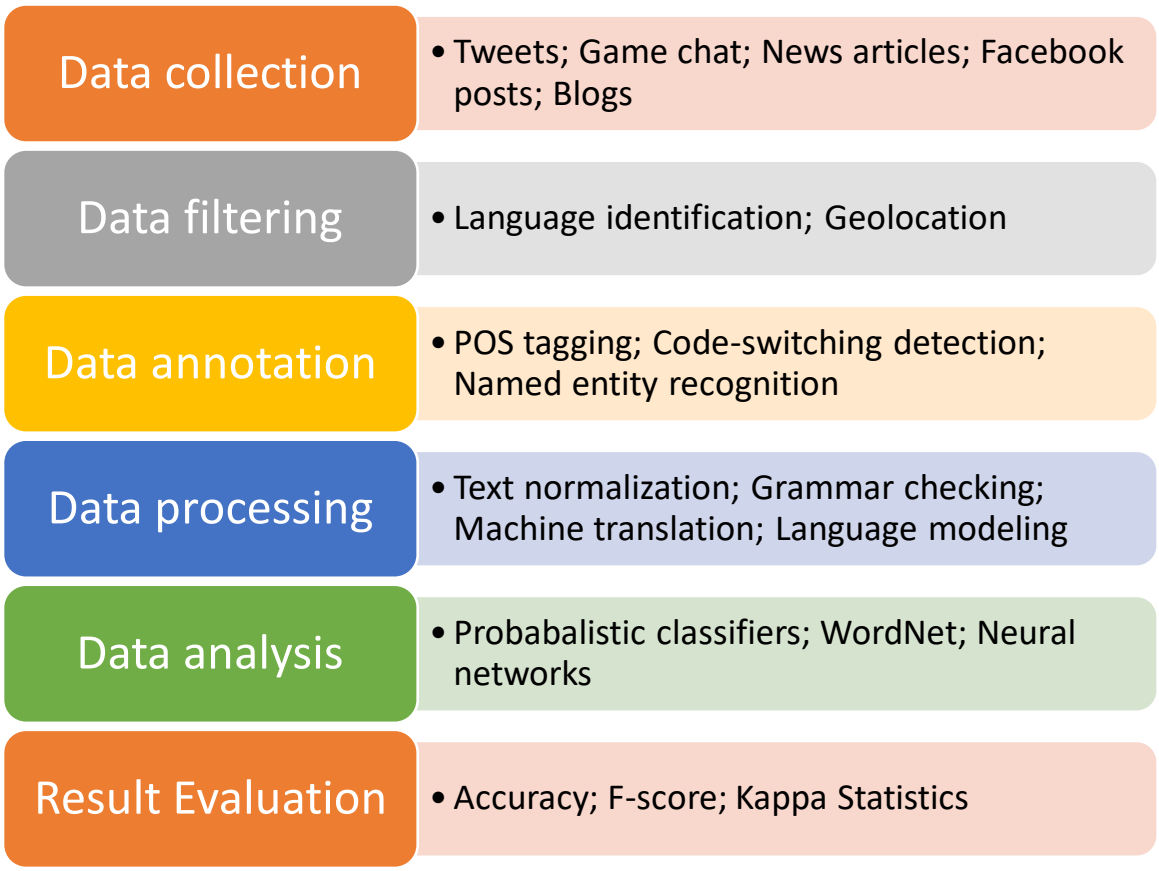
Authors	Go, Matthew Phillip; Nocon, Nicco
Title	Using Stanford Part-of-Speech Tagger for the Morphologically-rich Filipino language
Publication	Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation
Year	2017
Publisher	National University

Authors	Alcantara, Daniel; Borra, Allan
Title	Constituent Structure for Filipino; Induction through Probabilistic Approaches
Publication	Proceedings of the 22nd Pacific Asia Conference on Language, Information, and Computation
Pages	113-122

Year	2008

Authors	Guillaume Genthial
Title	Sequence Tagging with Tensorflow
Link	https://guillaumegenthial.github.io/sequence-tagging-with-tensorflow.html
Year	2017

Sentiment Analysis



Possible Resource Person(s) from NU:

- Mr. Joseph Marvin R. Imperial
- Mr. Manolito Octaviano Jr.
- Ms. Angelica De La Cruz
- Prof. Rachel Edita Roxas

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):

Authors	Lam, Alron Jan;
Title	Improving Twitter Community Detection through Contextual Sentiment Analysis of Tweets

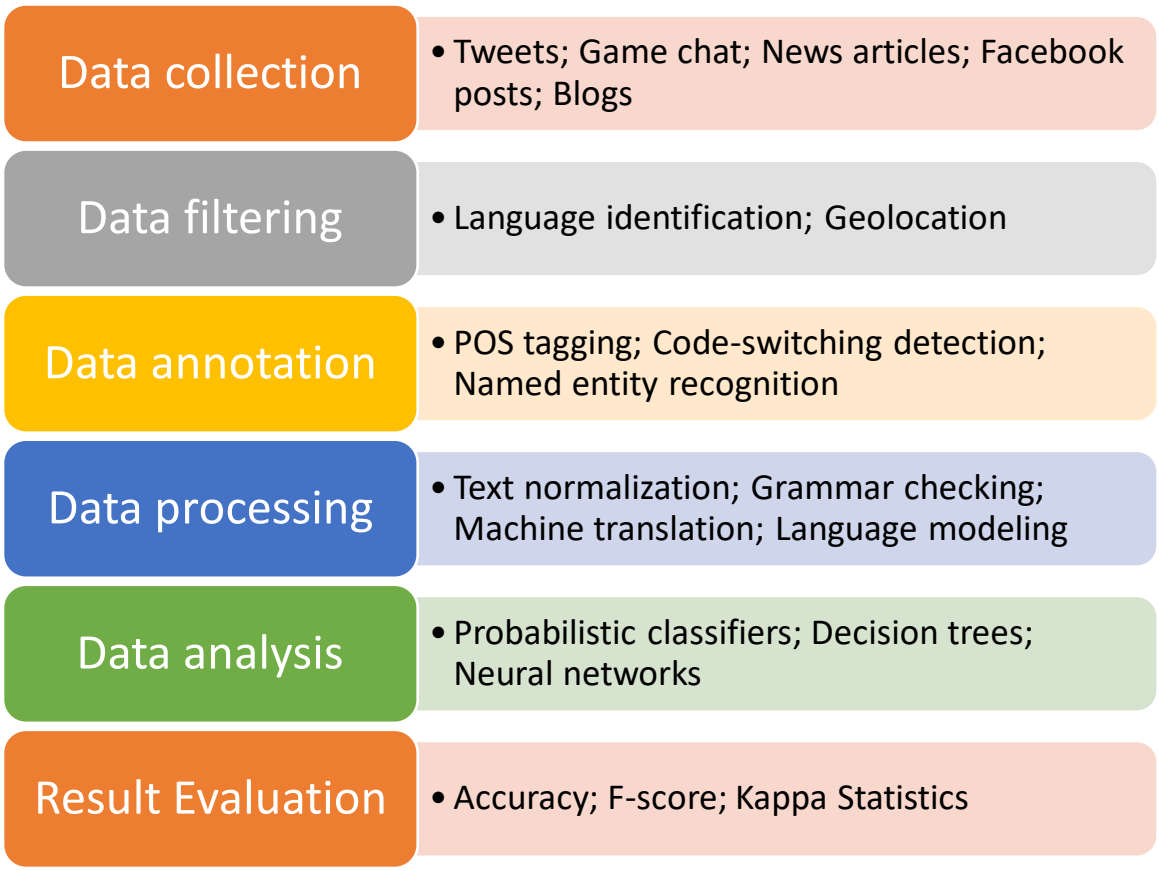
Publication	54th Annual Meeting of the Association for Computational Linguistics
Pages	30-36
Year	2016
Publisher	ACL

Authors	Regalado, Ralph Vincent J; Chua, Jenina L; Co, Justin L; Tiam-Lee, Thomas James Z;
Title	Subjectivity Classification of Filipino Text with Features Based on Term Frequency--Inverse Document Frequency
Publication	2013 International Conference on Asian Language Processing (IALP)
Pages	113-116
Year	2013
Publisher	IEEE

Authors	Regalado, Ralph Vincent J; Cheng, Charibeth K;
Title	Feature-Based Subjectivity Classification of Filipino Text
Publication	2012 International Conference on Asian Language Processing (IALP)
Pages	57-60
Year	2012
Publisher	IEEE

Authors	Imperial, Joseph Marvin; Orosco, Jeyrome; Mazo, Shiela Mae; Maceda, Lany
Title	Sentiment Analysis of Typhoon Related Tweets using Standard and Bidirectional Recurrent Neural Networks
Publication	Cornell University - ArXiv
Year	2019
Publisher	Arxiv

Classification



Possible Topic: Classification of Typhoon-related Tweets

Twitter has been found to be a potentially useful source of information in times of disaster. As a microblogging platform, users tend to use it for near-real-time updates. Specifically, in the context of disasters, some use it to report damage, request for assistance, find missing persons, etc. These could be useful for concerned entities like government agencies that conduct disaster response. However, with the large multitude of tweets, it is hard for people to manually scour through them; the task is sometimes likened to finding a needle in a haystack. Thus, automatic classification of relevant tweets will be useful for situations like these. Students interested in this area will be involved in experimenting with different features (like word embeddings) and classification algorithms to achieve this end goal.

Possible Resource Person(s) from NU:

- Mr. Joseph Marvin R. Imperial
- Mr. Manolito Octaviano Jr.
- Ms. Angelica De La Cruz
- Prof. Rachel Edita Roxas

Target Venue(s):

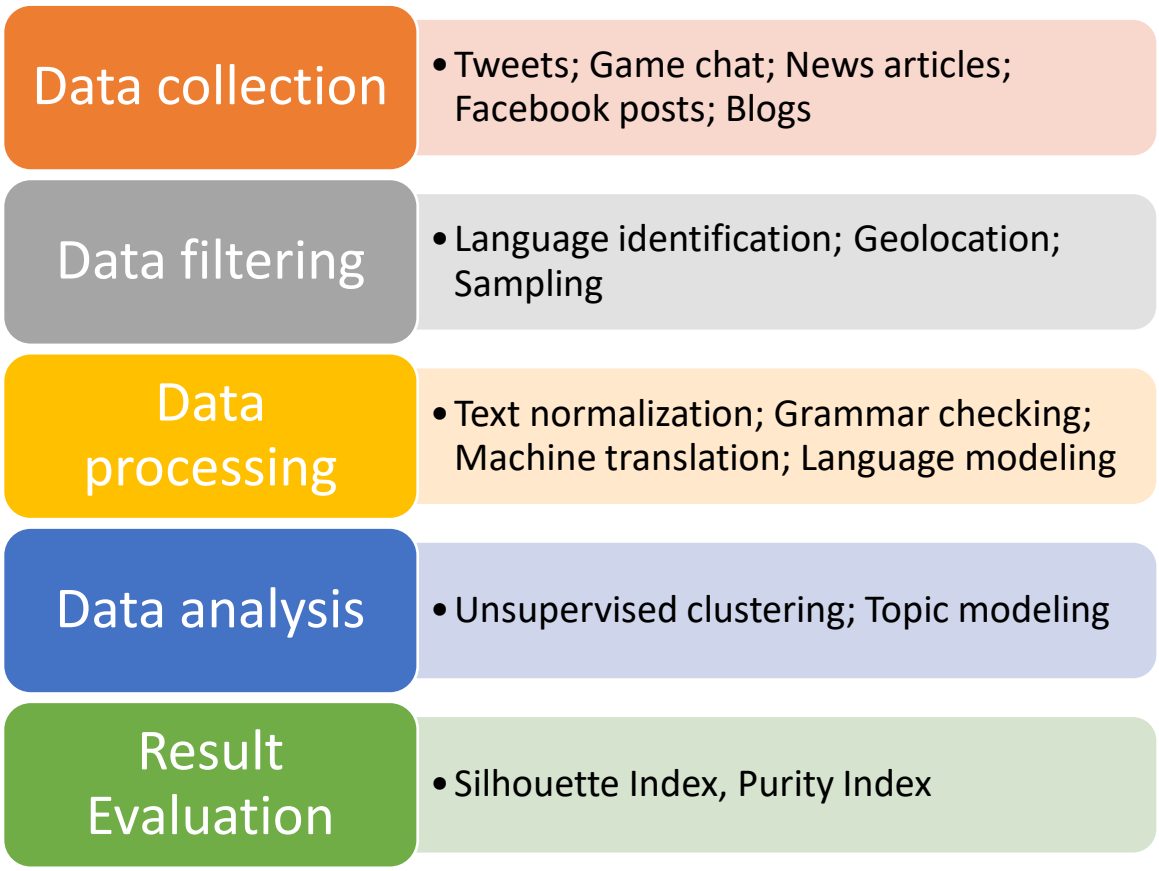
- Local: PCSC, NNLPRS

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

Starting Reference(s):

10NNLPRS Proceedings and 11NNLPRS Proceedings (https://sites.google.com/site/11nnlprs/past-symposia)

Topic Discovery



Possible Resource Person(s) from NU:

- Mr. Joseph Marvin R. Imperial
- Mr. Manolito Octaviano Jr.
- Ms. Angelica De La Cruz
- Prof. Rachel Edita Roxas

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):

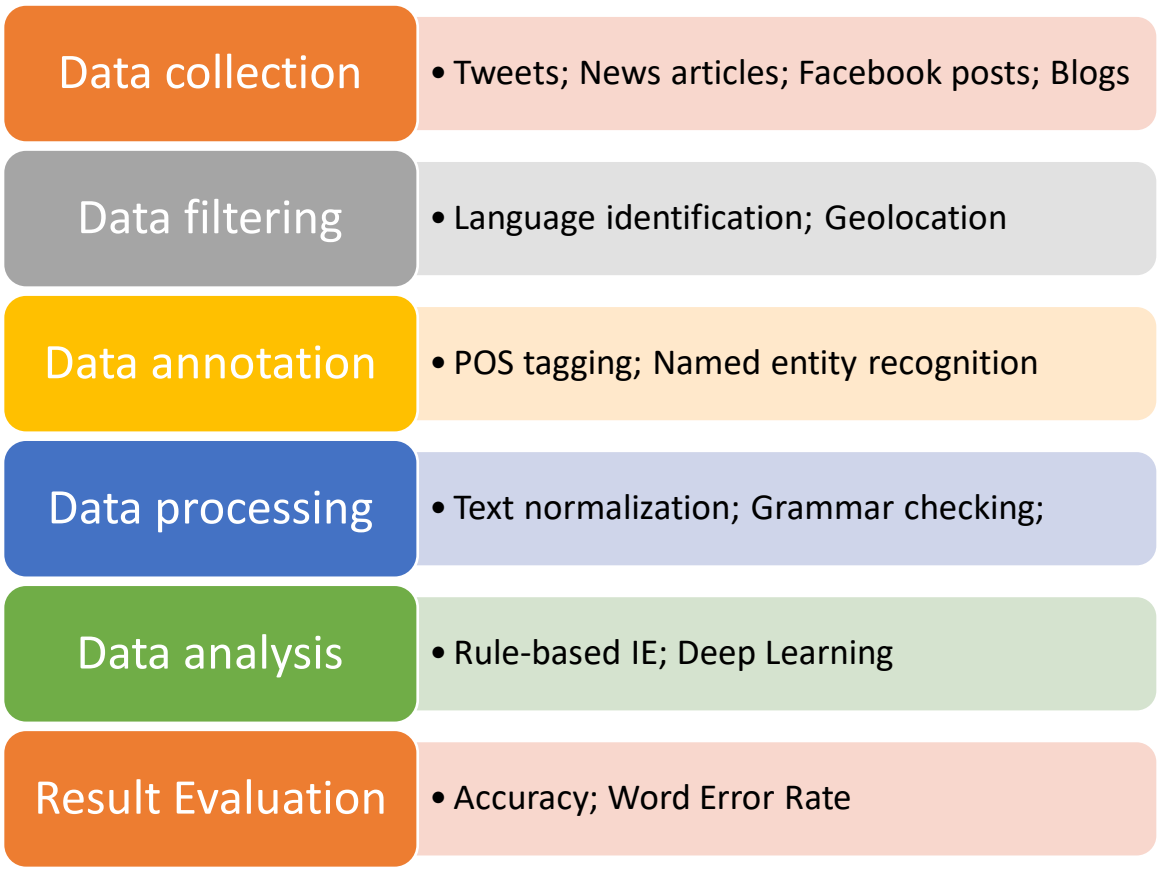
Authors	Ligutom III, Cerino; Orio, Jay Vincent; Ramacho, Dyannah Alexa Marie; Montenegro, Chuchi; Roxas, Rachel Edita; Oco, Nathaniel;
Title	Using Topic Modelling to make sense of typhoon-related tweets

Publication	2016 International Conference on Asian Language Processing (IALP)
Pages	362 - 365
Year	2017
Publisher	IEEE

Authors	Soriano, Cheryll Ruth; Roldan, Ma Divina Gracia; Cheng, Charibeth; Oco, Nathaniel;
Title	Social media and civic engagement during calamities: the case of Twitter use during typhoon Yolanda
Publication	Philippine Political Science Journal
Volume	37
Number	1
Pages	06-25
Year	2016
Publisher	Routledge

Authors	Syliongka, Leif Romeritch; Oco, Nathaniel; Lam, Alron Jan; Soriano, Cheryll Ruth; Roldan, Ma Divina Gracia; Magno, Francisco; Cheng, Charibeth;
Title	Combining Automatic and Manual Approaches: Towards a Framework for Discovering Themes in Disaster-related Tweets
Publication	Proceedings of the 24th International Conference on World Wide Web
Pages	1239-1244
Year	2015
Publisher	ACM

Information Extraction



Possible Topic: Visualizing Disaster Information Extracted from Philippine News Articles / Tweets

News articles and tweets contain loads of information on disasters before, during, and after it happens. These information sources contain typhoon names, date range of occurrence, locations hit, casualties, financial and material needs of the victims, and others. They also contain information about donations (and of what type) provided by countries, organizations, individuals to the victims. In this research, students will create an automated way of extracting this information from these sources and displaying them in a visual way showing the series of events related to each typhoon.

Target Venue(s):

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

Starting References:

- https://www.aclweb.org/anthology/W/W14/W14-2905.pdf
- https://www.aclweb.org/anthology/W/W16/W16-3906.pdf
- https://www.aclweb.org/anthology/C/C08/C08-3001.pdf

Resources

http://bit.ly/PLOCDATA

- Tweets – Swardspeak, Yolanda, Election 2013
- 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

Projects

LanguageTool: https://languagetool.org/

ASEANMT: http://aseanmt.org/

Opinion Space: http://opinion.berkeley.edu/

Online Tools

Twitter 4J: http://twitter4j.org/en/

Moses SMT Engine: http://www.statmt.org/moses/

SentiWordNet: http://sentiwordnet.isti.cnr.it/

Weka: http://www.cs.waikato.ac.nz/ml/weka/