Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets

Tuan Anh **Le**, David **Moeljadi**[1] Yasuhide **Miura**, Tomoko **Ohkuma**[2]

[1]Division of Linguistics and Multilingual Studies
Nanyang Technological University, Singapore
`{H130030,D001}@ntu.edu.sg`

[2]Fuji Xerox Co., Ltd.
`{Yasuhide.Miura,ohkuma.tomoko}@fujixerox.co.jp`

The 12th Workshop on Asian Language Resources

12 December 2016

Indonesian language

- Western Malayo-Polynesian language of the Austronesian language family
- belongs to the Malayic branch with Standard Malay in Malaysia and other Malay varieties
- spoken mainly in the Republic of Indonesia, by around 43 million people as their first language and by more than 156 million people as their second language (2010 census data)
- written in Latin script
- mildly agglutinative language, has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, and reduplication
- The lexical similarity is over 80% with Standard Malay [4]

Malay dialects in Southeast Asia

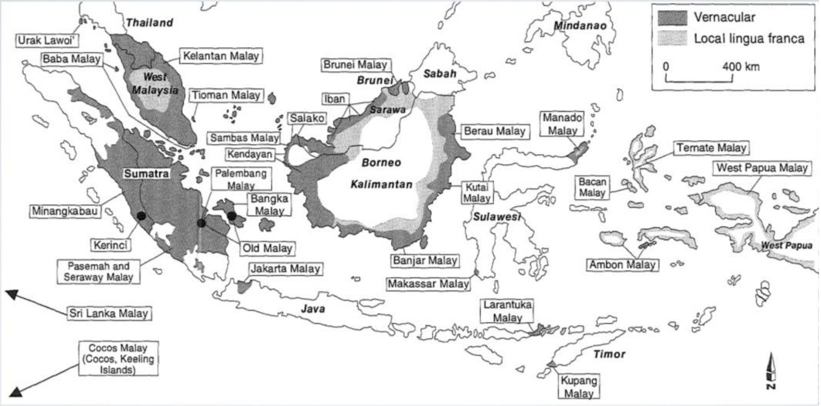

Figure: Malay dialects [1]

Diglossic nature of Indonesian

- Indonesian language is diglossic:
  - "High" variety: in education, religion, mass media, gov. activities
  - "Low" variety "Colloquial Jakartan Indonesian" [9]: for everyday communication between Indonesians
  - more than 500 regional languages spoken in various places in Indonesia: for communication at home with family and friends in the community

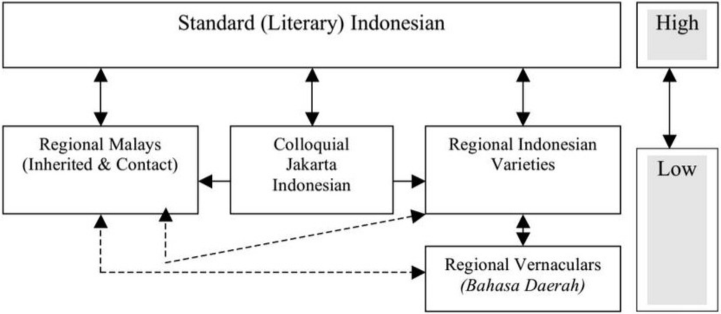
Standard (Literary) Indonesian
High
Regional Malays (Inherited & Contact)
Colloquial Jakarta Indonesian
Regional Indonesian Varieties
Low
Regional Vernaculars *(Bahasa Daerah)*

Figure: Diglossic situation in Indonesia [7]

Linguistic analysis of Indonesian tweets I

Feature	**Example**
Abbreviation	*bsk* (*besok* “tomorrow”), *bw* (*bawa* “bring”), ...
Interjection	*bhahaha* (*haha* “ha-ha”), *yahhh* (*ya* “well”), ...
Foreign word	ht (hot topic), Korean *nuna* “sister”, Japanese *ggrks* (*gugurekasu* “google it, you trash”), ...
Blending	*gamon* (***ga****gal* **mo**ve o**n** “fail to move on”), *ganchar* (***gan****ti* **char**acter “change the character”), *wotalay* (***wota****ku* *a****lay*** “exaggerative fan”), ...
Emoji	☺, ☹, ☻, ...
Emoticon	:) , :(, ;v , ...

Table: Features in Indonesian tweets

Linguistic analysis of Indonesian tweets II

Type	Example	Note
Abbreviation	*semangka* “watermelon”	abbreviated from ***semang**at, **ka**wan!* “do your best, my friend!”
Reversed word	*kuda* “horse”	reversed syllabically from *daku* “I”
	kuy	reversed letter by letter from *yuk* “let’s”
Others	*udang* “shrimp”	made from informal word *udah* “already”

Table: Word play in Indonesian tweets

Linguistic analysis of Indonesian tweets III

An example of a tweet in Indonesian:

@username @username makasih kk tfb yg paling hitz buat doa2nya :) amin yaallah aminnnn . Sukses juga buat band nya yahhh!

Translated into standard, high register Indonesian:

@username @username terima kasih, kakak TFB yang paling hit, untuk doa-doanya :) amin, ya Allah, amin . Sukses juga untuk band-nya, ya!

Translated into English:

@username @username thank you, the most popular TFB brothers, for the prayers :) amen, o God, amen. Success for the band, too!

Sentiment Analysis Approach

- Assumption: there is only one major sentiment in any given tweet
- must be either negative (NEG), positive (POS) or neutral (NEU)
- sentiment analysis task = a single-label text classification problem
- automate the sentiment analysis task
- supervised machine learning approach
  1. prepare labeled tweet data set: a pair of tweet (textual data) and corresponding label
  2. transform this data set into a suitable format
  3. train the classifier model
  4. assign label to new tweets automatically

Data collection

- 900,000 Indonesian tweets
- from February to March 2016
- Twitter Public Streams (`https://dev.twitter.com/streaming/public`) using Python script and Tweepy package (`http://www.tweepy.org/`)
- 1,694 Emoji definitions for normalization
- 61,374,640 Indonesian tokens from Wikipedia for building word2vec model [5]

Data labeling I

- eight labels (**POS**itive, **NEG**ative, **NEU**tral, **FOR**eign, **RET**weet, **ADV**ertisement, **INF**ormation, and **XXX** for others) for classifying Indonesian tweets
- 4,000 tweets as data and labeled manually using the eight labels
- we only used tweets written in **Standard Indonesian** and **Colloquial Jakarta Indonesian** for POS, NEG, and NEU
- difficulties in labeling because of the absence of context, ambiguity, and new slangs
- 25% or 1,005 tweets having sentiments (POS, NEG, or NEU)

Data labeling II

Label	Type	Example
POS	Positive	*Seger banget ini buat mata...* "This is very fresh for eyes..."
NEG	Negative	*Lo gak tau apa-apa tntang gue !* "You know nothing about me! "
NEU	Neutral	*cara daftar teman ahok gimana ya* "how to register for teman ahok?"
RET	Retweet	*RT @username: Menarik nih!* "This is interesting!"
INF	Article title	*Tips Merawat Layar Ponsel Xiaomi* "Tips for Caring for Xiaomi Screen"
FOR	Foreign language	Polisi Yaua Majambazi Watatu....
ADV	Advertisement	DELL Desktop C2D 2.66GHz-CPU
XXX	Others	EEEEEEHEHEHEHEHE TIRURITUTURU

Table: Eight labels used in labeling tweets and examples of tweets

Data labeling III

Label	Type	Number
POS	Positive	221
NEG	Negative	215
NEU	Neutral	569
RET	Retweet	1176
INF	Information	837
FOR	Foreign language	483
ADV	Advertisement	272
XXX	Others	227
	Total	4000

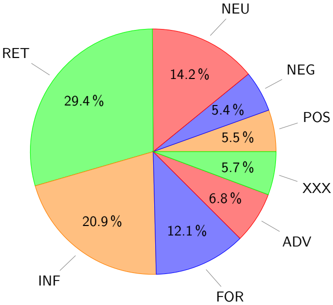

Figure: Manual tweets labeling with eight labels, their numbers, and percentage

Feature design

- convert textual data into numerical format for machine learning algorithm
  1. split tweets into tokens and normalize
  2. use the word2vec representation to represent tokens
  3. If the token can be found in word2vec model
     - ★ use word2vec vector to represent the token
     - ★ else use a zero vector
  4. input = a vector of $n * m$ dimensions (n = the maximum number of words in a tweet, m = the dimension of a word vector)
- Assumption: the longest tweet has up to 72 words, we used a 200 dimensions word2vec model
- Input = 72 x 200 = 14400 dimensions

Normalization I

Action	**Example**	
	Before	**After**
Remove page links	*mantep*-https://xxx...	*mantep*-
Remove user names	@username *asek dah :**	*asek dah :**
Add spaces between emoji	*terlalu semangat*☺☺	*terlalu semangat* ☺ ☺

Table: Adjustments before tokenization

- we used NLTK [2] word tokenizer to tokenize the tweets

Normalization II

Action	**Pattern**	**Example**	
		Before	**After**
Remove *nya* or *ny*	ABC*nya* → ABC	*doa2nya*	*doa2*
	ABC*ny* → ABC	*ujanny*	*ujan*
Remove reduplication with hyphen (-) or 2	ABC-ABC → ABC	*ular-ular*	*ular*
	ABC2 → ABC	*doa2*	*doa*
Remove reduplicated letters	AABBBCC → ABC	*mannaaa*	*mana*
Make several groups of same two letters to two groups	ABABABA → ABAB	*hahahah*	*haha*

Table: Normalizing tweets

Normalization III

- we compiled a list of 376 frequent informal words in tweets, their full forms, and their corresponding formal, standard Indonesian words
- we made a file which contains a list of emojis and their English equivalents. One emoji may have two or more equivalents, e.g. ↘ has two equivalents: "arrow lower right" and "south east arrow"

Informal	**Full form**	**Standard Ind.**	**Meaning**
acc	account	*akun*	"account"
blg	*bilang*	*berkata*	"say"
mager	*malas gerak*	*malas bergerak*	"lazy to move"
peje	*pajak jadian* (lit. "dating tax")	*uang traktir teman saat resmi berpacaran*	"money to treat friends after someone is officially in a relationship"

Table: Some examples of informal Indonesian words and the corresponding formal words

Normalization IV

- For each tokenized word:
  - It is listed in the informal word list → change to its formal counterpart and tokenize
  - It is in emoji list → each word in each English definition of the emoji is translated into Indonesian word(s) using WordNet in NLTK [2]
- we get a list of formal Indonesian words from each tweet

Experiment setups

- we use word2vec tool (`https://code.google.com/archive/p/word2vec/`) to train the word2vec model
- we use Python and Theano package (`http://deeplearning.net/software/theano/`) to build the classification model
  - input = 72 × 200 dimensions per word
  - output = 8 dimensions (labels)
- we experiment with two algorithms: Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM)
- we used k-fold cross-validation method with k=10

Summary of system architecture

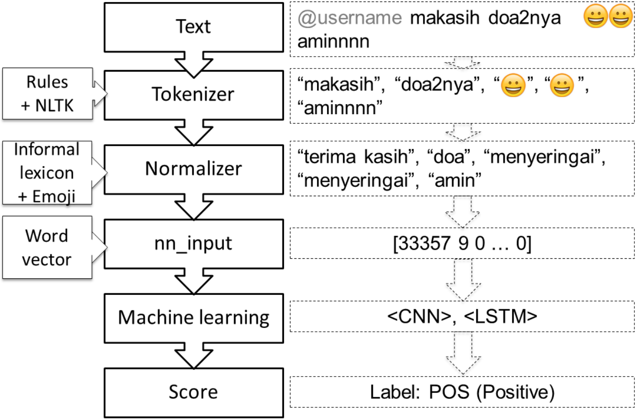

Figure: Summary of our system architecture with examples

Results and evaluation

- it seems that the normalizer we made does not make the accuracy higher, perhaps because it covers very few informal words
- useful in aiding us to generate labeled data much faster, annotate much faster compare to manual labeling
- helpful for generating data for low resource languages such as Indonesian

	Matched	**Sentences**	**Accuracy**	**STD**
CNN without normalizer	3,102	4,428	70.05%	1.93
CNN with normalizer	2,898	4,428	65.45%	2.12
LSTM without normalizer	3,440	4,428	**73.22%**	1.39

Table: Results of sentiment analysis with CNN and LSTM

Conclusions and future works

- a system architecture which includes tokenizer, normalizer, CNN and LSTM
- Result: 73.2% accuracy with LSTM without normalizer
- as a baseline to build a more complex state-of-the arts neural networks model in Indonesian
- cross-lingual extensions using a multilingual resource
- Future works:
  - a dictionary for informal words
  - emoticons
  - Indonesian SentiWordnet Barasa (`https://github.com/neocl/barasa`)
  - Indonesian constructions or sentence structures [3]: negation words and question words
  - Indonesian POS Tagger [8]
  - Indonesian Resource Grammar (INDRA) [6]

Acknowledgments

- Thanks to Francis Bond for his support and precious advice.

References I

Alexander Adelaar. “Structural Diversity in The Malayic Subgroup”. In: *The Austronesian languages of Asia and Madagascar*. London and New York: Routledge Language Family Series, 2010, pp. 202–226.

Steven Bird, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. O’Reilly Media Inc., 2009. URL: `http://www.nltk.org/book/` (visited on 11/24/2014).

Franky, Ondřej Bojar, and Kateřina Veselovská. “Resources for Indonesian Sentiment Analysis”. In: *The Prague Bulletin of Mathematical Linguistics 103*. Prague: Charles University in Prague, Faculty of Mathematics, Physics, Institute of Formal, and Applied Linguistics, 2015, pp. 21–41.

M. Paul Lewis. *Ethnologue: Languages of the World*. 16th ed. Dallas, Texas: SIL International, 2009. URL: `http://www.ethnologue.com` (visited on 12/01/2014).

References II

Tomas Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., 2013, pp. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

David Moeljadi, Francis Bond, and Sanghoun Song. “Building an HPSG-based Indonesian Resource Grammar (INDRA)”. In: *Proceedings of the GEAF Workshop, ACL 2015*. 2015, pp. 9–16. URL: http://aclweb.org/anthology/W/W15/W15-3302.pdf.

Scott H. Paauw. “The Malay contact varieties of Eastern Indonesia: A typological comparison”. PhD dissertation. State University of New York at Buffalo, 2009.

References III

Fam Rashel et al. “Building an Indonesian Rule-Based Part-of-Speech Tagger”. In: Kuching, 2014.

James Neil Sneddon. *Colloquial Jakartan Indonesian*. Canberra: Pacific Linguistics, 2006.