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

Tuan Anh Le, David Moeljadi
Division of Linguistics and Multilingual Studies
Nanyang Technological University
Singapore
{H130030,D001}@ntu.edu.sg

Yasuhide Miura, Tomoko Ohkuma
Fuji Xerox Co., Ltd.
6-1, Minatomirai, Nishi-ku,
Yokohama-shi, Kanagawa
{Yasuhide.Miura,
ohkuma.tomoko}@fujixerox.co.jp

Abstract

This paper describes our attempt to build a sentiment analysis system for Indonesian tweets. With this system, we can study and identify sentiments and opinions in a text or document computationally. We used four thousand manually labeled tweets collected in February and March 2016 to build the model. Because of the variety of content in tweets, we analyze tweets into eight groups in total, including pos(itive), neg(ative), and neu(tral). Finally, we obtained 73.2% accuracy with Long Short Term Memory (LSTM) without normalizer.

1 Introduction

Millions of internet users send, post, and share text messages every day via various platforms, such as online social networking service (Twitter, Facebook, etc.), instant messaging service (WhatsApp, Telegram, etc.), online blogs, and forums. One of them, Twitter, has become popular among computational linguistics and social science researchers since it supports a policy of open data access and provides a means of getting vast amounts of linguistic, network, and other forms of data from actual human behavior with relatively little effort. Researchers can use Twitter's Application Programming Interface (API) to get the tweets in text data using particular search terms for their research, such as sentiment analysis.

At Twitter,[1] users can sign up for free, choose a 'handle' or user name, get a profile, post 'tweets' or short messages of 140 characters maximum with '@' symbols followed by other user-handles if they want their tweets to appear on the other users' profiles, and add '#' symbols for topic indicators. Unlike Facebook, Twitter users broadcast their messages to anyone who chooses to "follow" them as a broadcast medium and the 140-character limit forces users to be brief and makes it easy for anyone reading and reviewing tweets.

Carley et al. (2015) notes that since its launch in 2006, Twitter has grown to 284 million monthly active users who send about 500 million tweets per day, 80% of which are from mobile devices (as of 2014). Indonesia was ranked as the fifth most tweeting country in 2014. The number of users is increasing and it is predicted that there will be 22.8 million users in 2019.[2] The capital city of Indonesia, Jakarta, is the worlds most active Twitter city,[3] ahead of Tokyo and London.[4] Therefore, tweets data can be a good source for research on Indonesian sentiment analysis.

One of the basic tasks for sentiment analysis is polarity classification, i.e. determining whether a given text expresses positive, negative, or neutral sentiment. Much research has been done to address the problem of sentiment analysis on Indonesian tweets. Aliandu (2013) conducted research on Indonesian tweet classification into three labels: positive, negative, and neutral, using emoticons for collecting sentiment-bearing tweets as proposed by Pak and Paroubek (2010). The conclusion is that Support Vector Machine

[1] twitter.com

[2] http://www.emarketer.com/Article/Southeast-Asia-Has-Among-Highest-Social-Network-Usage-World/1013275

[3] http://blogs.wsj.com/digits/2015/09/02/twitter-looks-to-indonesia-to-boost-growth/

[4] https://www.techinasia.com/indonesia-social-jakarta-infographic

Proceedings of the 12th Workshop on Asian Language Resources,
pages 123–131, Osaka, Japan, December 12 2016.

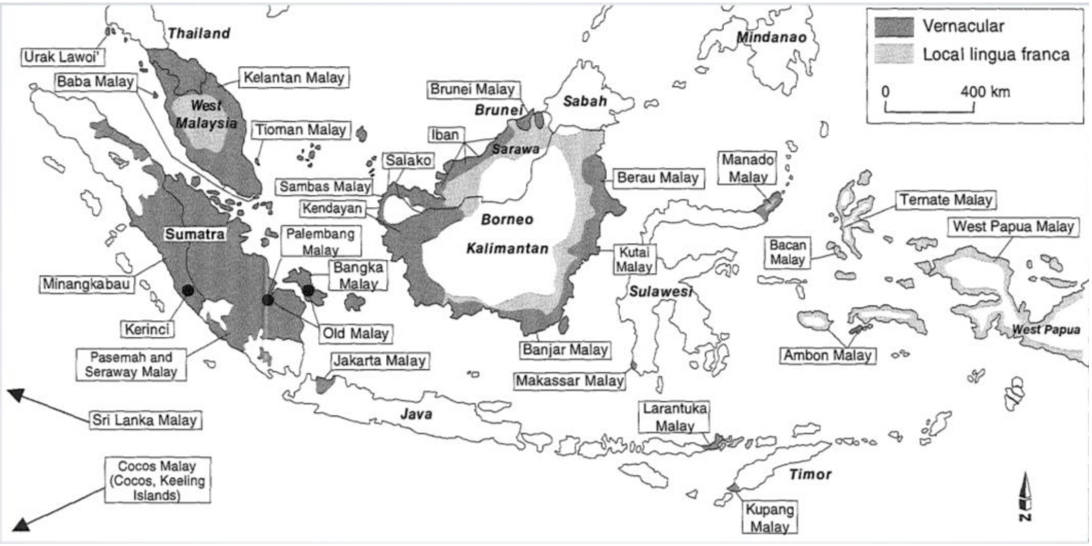

Figure 1: Malay dialects (Adelaar, 2010, p. 203)

(SVM) method (77.57% accuracy for TF-IDF and 77.79% for term frequency) was slightly better than Naive Bayes method (75.86% accuracy for TF-IDF and 77.45% for term frequency). Wicaksono et al. (2014) performed opinion tweet extraction and tweet polarity classification by automatically building a set of labeled seed corpus using opinion lexicon based technique and clustering based technique and obtaining more training instances from a huge set of unlabeled tweets by employing a classifier model. Their experiment shows that their method outperforms the baseline system which merely uses emoticons as the features for automatically building the sentiment corpus (81.13% accuracy with Naive Bayes and 86.82% accuracy with Maximum Entropy).

2 Indonesian language

Indonesian (ISO 639-3: ind), called *bahasa Indonesia* (lit. "the language of Indonesia") by its speakers, is a Western Malayo-Polynesian language of the Austronesian language family. Within this subgroup, it belongs to the Malayic branch with Standard Malay in Malaysia and other Malay varieties (Lewis, 2009) (see Figure 1). It is spoken mainly in the Republic of Indonesia as the sole official and national language and as the common language for hundreds of ethnic groups living there (Alwi et al., 2014, pp. 1-2). In Indonesia it is spoken by around 43 million people as their first language and by more than 156 million people as their second language (2010 census data). The lexical similarity is over 80% with Standard Malay (Lewis, 2009). It is written in Latin script.

Morphologically, Indonesian is a mildly agglutinative language, compared to Finnish or Turkish where the morpheme-per-word ratio is higher (Larasati et al., 2011). It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, and reduplication. Most of the affixes are derivational.

The diglossic nature of the Indonesian language exists from the very beginning of the historical record when it is called Old Malay around the seventh century A.D. to the present day (Paauw, 2009, p. 3). While much attention has been paid to the development and cultivation of the standard "High" variety of Indonesian, little attention has been particularly paid to describing and standardizing the "Low" variety of Indonesian. Sneddon (2006, pp. 4-6) calls this variety "Colloquial Jakartan Indonesian" and states that it is the prestige variety of colloquial Indonesian in Jakarta, the capital city of Indonesia, and is becoming the standard informal style. Paauw (2009, p. 40) mentions that Colloquial Jakartan Indonesian is a variety which has only been recognized as a separate variety recently. Historically, it developed from the Low Malay varieties spoken in Java by Chinese immigrant communities, which have been termed "Java Malay". It has also been influenced by the Betawi language of Jakarta, a Low Malay variety which is thought to have been spoken in the Jakarta region for over one thousand years.

In addition to this "Low" variety, the more than 500 regional languages spoken in various places in

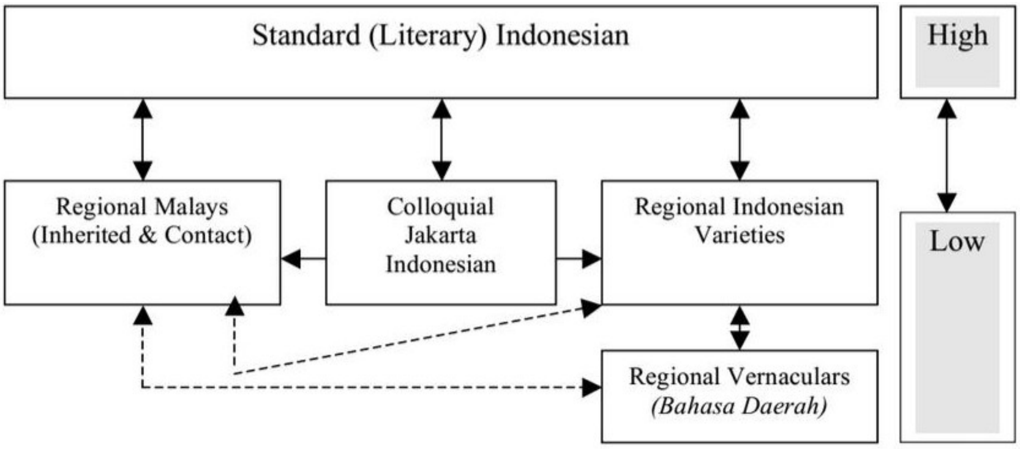
Standard (Literary) Indonesian
High
Regional Malays (Inherited & Contact)
Colloquial Jakarta Indonesian
Regional Indonesian Varieties
Low
Regional Vernaculars (Bahasa Daerah)

Figure 2: Diglossic situation in Indonesia (Paauw, 2009, p. 16)

Feature	Example
Abbreviation	*yg* (*yang* “REL”), *bsk* (*besok* “tomorrow”), *bw* (*bawa* “bring”), . . .
Interjection	*bhahaha* (*haha* “ha-ha”), *wkwk* (*haha* “ha-ha”), *yahhh* (*ya* “well”), . . .
Foreign word	ht (hot topic), Korean *nuna* “sister”, Japanese *ggrks* (*gugurekasu* “google it, you trash”), . . .
Blending	*gamon* (*gagal* move on “fail to move on”), *ganchar* (*ganti* character “change the character”), *wotalay* (*wotaku alay* “exaggerative fan”), . . .
Emoji	☺, ☹, ☻, . . .
Emoticon	:) , :(, ;v , . . .

Table 1: Features in Indonesian tweets

Indonesia add to the complexity of the sociolinguistic situation in Indonesia. The “High” variety of Indonesian is used in the context of education, religion, mass media, and government activities. The “Low” variety of Indonesian is used for everyday communication between Indonesians. The regional vernaculars (*bahasa daerah*) are used for communication at home with family and friends in the community. In some areas, Indonesian coexists with yet another regional lingua franca, which is often a Malay variety. For example, in the city of Kalabahi of Alor Island, locals speak Indonesian, Kupang Malay, and one of the local languages such as Abui in different contexts. This complex situation is well described in Paauw (2009) and shown in Figure 2.

3 Linguistic analysis of Indonesian tweets

Tweets in Indonesian reflect the diglossic nature of the Indonesian language, as mentioned in Section 2. In addition to the “High” and “Low” registers in spoken Indonesian, many informal features in contemporary written text appear, such as abbreviations, interjections, foreign words (sometimes abbreviated), blending of Indonesian and foreign words, emoji, and emoticons, as shown in Table 1. Interesting phenomena, such as word play, also appear, as shown in Table 2.

Type	Example	Note
Abbreviation	*semangka* “watermelon”	abbreviated from *semangat, kawan!* “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 2: Word play in Indonesian tweets

Below is an example of a tweet in Indonesian with some features mentioned above.

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

makasih "thanks" and *buat* "for" are low register words. *kk*, *tfb*, and *yg* are abbreviations of *kakak* "brother", True Friend Band, and *yang* "REL" respectively. *hitz* and *band* are English words. Reduplication is represented as number two "2" in *doa2nya* "the prayers". Repetition of letters appears in a word *aminnnn* "amen" and a discourse particle *yahhh*. A space is inserted between the word *band* and enclitic *=nya* "DEF" but a particle *ya* and the word *allah* "God" are written without a space in between. Also, there is one emoticon :).

If the tweet above is translated into standard, high register Indonesian, it would be as follows.

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

4 Sentiment Analysis Approach

The problem with sentimental information is that it is often vague and mixed. There may be more than one opinion or sentiment in a tweet. For example "I like this product but I do not like the price". To simplify the problem, we assume there is only one major sentiment in any given tweet. This sentiment must be either negative (NEG), positive (POS) or neutral (NEU). With this assumption, we transformed the sentiment analysis task into a single-label text classification problem.

We want to automate the sentiment analysis task as much as possible. To do that we use supervised machine learning approach. First, we prepare labeled tweet data set. Each data entry in the data set is a pair of tweet (textual data) and corresponding label. Next, we transform this data set into a suitable format to train the classifier model. After the model is trained, it can assign label to new tweets automatically.

4.1 Data Collection

From February to March 2016, we collected 900 thousands Indonesian tweets from Twitter Public Streams[5] using Python script and Tweepy package.[6] The script listens to Twitter's public stream and download any tweet with language code equals to 'id'. We also downloaded and processed 1,694 Emoji definitions for normalization as well as 61,374,640 Indonesian tokens from Wikipedia for building word2vec model (Mikolov et al., 2013).

4.2 Data labeling

We decided to add five more labels to categorize the tweets better. In total, we made eight labels (POS for positive, NEG for negative, NEU for neutral, FOR for foreign, RET for retweet, ADV for advertisement, INF for information, and XXX for others) for classifying Indonesian tweets, as shown in Table 3. Because of resource limitations, we chose 4,000 tweets as data and labeled them manually using the eight labels. Tweets written in languages such as English, Standard Malay, regional Malays, regional Indonesian varieties, and regional vernaculars such as Javanese and Sundanese are given FOR label. We only used tweets written in Standard Indonesian and Colloquial Jakarta Indonesian for POS, NEG, and NEU labels. Tweets containing news, tips, and quotations are given INF label. We found difficulties in labeling because of the absence of context, ambiguity, and new slangs.

Out of 4,000 tweets, we got about 25% or 1,005 tweets having sentiments (positive, negative, or neutral). More than half of them (569 tweets) are neutral, the rest of them have positive or negative sentiments with roughly the same number, as shown in Figure 3.

[5] `https://dev.twitter.com/streaming/public`

[6] `http://www.tweepy.org/`

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 ! Jadi jangan sok ngatur !!* “You know nothing about me! So don’t control 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* https://xxx “Tips for Caring for Xiaomi Mobile Phone Screen”
	Date and time	*@username Selasa, 01 Maret 2016 Pukul 12:33 [Indonesia]* “Tuesday, 1 March 2016 12:33”
	Quote	*waktu memang bukan dokter yang baik ..., tapi dia adalah guru terbaik ...* “time is indeed not a good doctor ..., but it is the best teacher ...”
	Story	*(cont.) duduk di kursi taman ... sambil memegang ponselnya ... (cont.)* “sitting on a bench ... holding his phone ...”
FOR	Foreign language	Polisi Yaua Majambazi Watatu....Baada ya Kupekuliwa Walikutwa...
ADV	Advertisement	DELL Desktop C2D 2.66GHz-CPU 3Gb-RAM... https://xxx...
XXX	Others	EEEEEEHEHEHEHEHE TIRURITUTURURURURUTURURUTUT

Table 3: Eight labels used in labeling tweets and examples of tweets

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

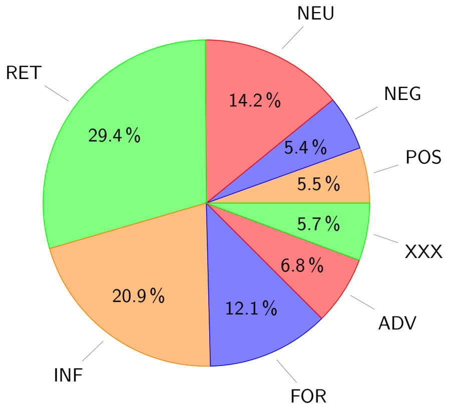

Figure 3: Manual tweets labeling with eight labels, their numbers, and percentage

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

Table 4: Adjustments before tokenization

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	*mannaaaa*	*mana*
Make several groups of same two letters to two groups	ABABABA → ABAB	*hahahahah*	*haha*

Table 5: Normalizing tweets

4.3 Feature Design

Machine learning algorithms do not work directly on textual data. In order to use machine learning algorithm, we have to convert textual data into numerical format. First, we split tweets into tokens and normalize them (replacing informal words, etc.). We then use the word2vec representation to represent tokens. If the token can be found in word2vec model (i.e. we have vector representation for a word), we will use word2vec vector to represent the token. However, if we cannot find the word, we will use a zero vector instead. The input then will be a vector of $n * m$ dimensions where n is the maximum number of words in a tweet and m is the dimension of a word vector. In this system, we assumed that the longest tweet may has up to 72 words and we used a 200 dimensions word2vec model. Therefore the input will be 72 x 200 = 14400 dimensions.

4.4 Normalizer

We used the default word tokenizer from NLTK 3 (Python) and normalized the tweets. In order to get only the tweet, we removed page links which begin with `https` and user names which begin with '@' symbols. We did not remove the topic indicators which begin with '#' symbols because they can be a feature for the supervised machine learning algorithm. Since emojis are important for sentiment analysis, we added spaces between emojis to make them easier to tokenize. Table 4 summarizes these adjustments. Afterwards, we used NLTK (Bird et al., 2009) word tokenizer to tokenize the tweets.

After tokenizing the tweets, we removed the enclitic *nya* "DEF" and its orthographic variant *ny* and made reduplicated words, letters, and syllables into its non-reduplicated counterparts as shown in Table 5. Since there are many informal words and orthographic variants in tweets as mentioned in Section 3, we compiled a list of 376 frequent informal words in tweets and their corresponding formal, standard Indonesian words. Since most informal words are written in their short forms, we also listed down the full forms, in addition to the corresponding formal words, as shown in Table 6. Some informal words, such as *peje* or *pajak jadian*, do not have a corresponding word or compound in formal Indonesian and thus we translated them into many words. In addition, since tweets use various emojis which are essential for sentiment analysis, we made a file which contains a list of emojis and their English equivalents. One emoji may have two or more equivalents, for example ↘ has two equivalents: "arrow lower right" and "south east arrow".

For each tokenized word, we checked whether it is listed in the informal word list. If yes, it is changed to its formal counterpart and tokenized. If it is in emoji list, each word in each English definition of the emoji is translated into Indonesian word(s) using WordNet in NLTK (Bird et al., 2009). If the English word is in Princeton WordNet and has Indonesian translation(s), it is translated into Indonesian. Thus,

Informal word	Full form	Standard Indonesian word	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 for food after someone is officially in a relationship"

Table 6: Some examples of informal Indonesian words and the corresponding formal words

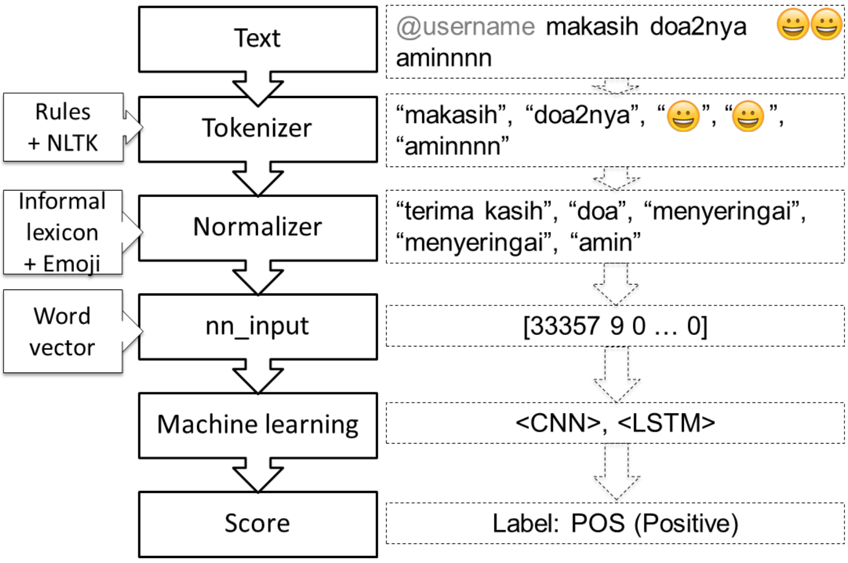

Figure 4: Summary of our system architecture with examples

we get a list of formal Indonesian words from each tweet which is used for the next step.

4.5 Text and word2vec

We downloaded Indonesian Wikipedia data and use Python to convert it into text format. We then use word2vec tool[7] to train the word2vec model. Each word in this model is represented by a vector of 200 dimensions.

4.6 Machine learning

We used Python and Theano package[8] to build the classification model. The input is 72 dimensions × 200 dimensions per word. The output is 8 dimensions (labels). We experimented with two algorithms, i.e. Convolutional Neural Network (CNN) and LSTM.

As the data set that we could prepare was small, we used k-fold cross-validation method with k=10. We split the data set into 10 groups and test the model 10 times. Each time we use one group for testing and the other 9 groups for training. We take the average of the accuracy as the final accuracy for each method.

5 Results and evaluation

Having conducted our sentiment analysis with CNN and LSTM, we obtained the results as shown in Table 7. The best accuracy we got was 73.22% using LSTM without normalizer. We have not conducted LSTM with normalizer, but looking at the results for CNN, it seems that the normalizer we made at

[7] https://code.google.com/archive/p/word2vec/

[8] http://deeplearning.net/software/theano/

	Matched	Sentences	Accuracy	STD
CNN without normalizer	3,440	4,920	69.92%	1.87
CNN with normalizer	2,898	4,428	65.45%	2.12
LSTM without normalizer	3,440	4,428	73.22%	1.39

Table 7: Results of sentiment analysis with CNN and LSTM

the present stage does not make the accuracy higher. The reason is perhaps because it covers very few informal words.

5.1 Discussion

We developed a sentiment analysis system and it has good accuracy according to our test set. In our opinion, it has yet to be practical enough for many real-life applications.

However, this system proved to be useful in aiding us to generate labeled data much faster. We noticed that by using the output of the system as the starting point, our annotator can annotate much faster compare to manual labeling. This finding can be helpful for generating data for low resource languages such as Indonesian.

6 Conclusions and future works

We have built a system architecture which includes tokenizer, normalizer, CNN and LSTM. The result is that we obtained 73.2% accuracy with LSTM without normalizer. The model can be used as a baseline to build a more complex state-of-the arts neural networks model in Indonesian. Since the result of the current model is comparable to results in English and Japanese, some known cross-lingual extensions using a multilingual resource are possible future directions of the model. We plan to put more efforts in building a dictionary for informal words because the normalizer contains very few informal words. We believe that this can make the accuracy higher and it is maybe better to perform an error analysis of the normalization rules. We used only emojis in our system, in the future we will use emoticons, too.

In addition, we plan to use Indonesian SentiWordnet Barasa[9] which was built based on SentiWordNet[10] (Baccianella et al., 2010) and Wordnet Bahasa (Bond et al., 2014). We will focus more on the constructions or sentence structures in Indonesian. Franky et al. (2015) present a few things to note related to the words and sentence structures, such as word sense disambiguation and multi-word expressions. They also list some features for sentiment prediction, such as negation words and question words. In order to do this, we will use an Indonesian POS Tagger (Rashel et al., 2014). In the future, we plan to employ a computational grammar for Indonesian, such as Indonesian Resource Grammar (INDRA) (Moeljadi et al., 2015), to obtain higher accuracy and better results.

Acknowledgements

Thanks to Francis Bond for his support and precious advice.

References

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

Paulina Aliandu. 2013. Sentiment analysis on Indonesian tweet. In *The Proceedings of International Conferences of Information, Communication, Technology, and Systems*, pages 203–208.

Hasan Alwi, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 2014. *Tata Bahasa Baku Bahasa Indonesia*. Balai Pustaka, Jakarta, 3 edition.

[9] `https://github.com/neocl/barasa`

[10] `http://sentiwordnet.isti.cnr.it/`

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Francis Bond, Lian Tze Lim, Enya Kong Tang, and Hammam Riza. 2014. The combined wordnet bahasa. *NUSA: Linguistic studies of languages in and around Indonesia*, 57:83–100.

Kathleen M. Carley, Momin Malik, Mike Kowalchuk, Jrgen Pfeffer, and Peter Landwehr. 2015. Twitter usage in Indonesia. Technical report, Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, December.

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

Septina Dian Larasati, Vladislav Kubo, and Daniel Zeman. 2011. Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Springer CCIS proceedings of the Workshop on Systems and Frameworks for Computational Morphology*, pages 119–129, August.

M. Paul Lewis. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16 edition.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

David Moeljadi, Francis Bond, and Sanghoun Song. 2015. Building an HPSG-based Indonesian Resource Grammar (INDRA). In *Proceedings of the GEAF Workshop, ACL 2015*, pages 9–16.

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

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Fam Rashel, Andry Luthfi, Arawinda Dinakaramani, and Ruli Manurung. 2014. Building an Indonesian Rule-Based Part-of-Speech Tagger. Kuching.

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

Alfan Farizki Wicaksono, Clara Vania, Bayu Distiawan, and Mirna Adriani. 2014. Automatically building a corpus for sentiment analysis on Indonesian tweets. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation*, pages 185–194, Phuket,Thailand, December. Department of Linguistics, Chulalongkorn University.