Colloquial Indonesian Lexicon

Nikmatun Aliyah Salsabila*‡, Yosef Ardhito Winatmoko† Ali Akbar Septiandri*, Ade Jamal*
*Faculty of Science and Technology
Universitas Al Azhar Indonesia
Jakarta, Indonesia
Email: ‡salsabila@if.uai.ac.id
†Department of Industrial Engineering & Innovation Sciences
TU Eindhoven
's-Hertogenbosch, Netherlands
Email: y.ardhito.winatmoko@student.tue.nl

***Abstract*—We provide a lexicon for text normalization of Indonesian colloquial words. We gathered 3,592 unique colloquial words–also known as "bahasa alay"–and manually annotated them with the normalized form. We built this lexicon from Instagram comments provided in [1].**

***Keywords*—colloquial, slang, Indonesian, normalization, lexicon**

I. Introduction

The rise of SMS texting and social media has brought new challenges in the world of natural language processing [2], [3]. Several studies have been done in English [4], [5], French [6], and Arabic [7] languages. In this study, we focus on handling out-of-vocabulary (OOV) words, in particular due to conversational use of words and phrases in social media written in bahasa Indonesia. Text normalization is important as it will help language parser to understand lexical meaning better. The performance of language processing could be improved if we do normalization for OOV words [8].

This paper is similar to the work of Han et al. [9], but for bahasa Indonesia. Furthermore, since some of the words lack their corresponding standard form in the dictionary, our normalization method follow the guidelines provided in [10], i.e. it must be unique, euphonic, aligned to the rule of bahasa Indonesia, has no negative connotation, and is used frequently. Our motivation, as also mentioned in [9], is that dictionary-based normalization approach could outperform several previously proposed approaches [2], [5].

We divide this paper into 3 parts: similar studies in the area of text normalization; statistics of the lexicon and the colloquial words occurrences in Instagram comments; and test results. The related studies discuss text normalization in general and colloquialism in Indonesian specifically. We provide the slang and formal words analysis to understand the lexicon's characteristics. Lastly, we reproduce the Instagram spam detection and compare the result with and without using the lexicon for text normalization [1].

II. Related Studies

A. Text Normalization

To the best of our knowledge, there is no generalized text normalization corpus for bahasa Indonesia OOV although we found similar study in Arabic [7]. However, Hanafiah et al. use a dictionary of 378 slang words and have shown that text normalization can improve the accuracy of bahasa Indonesia Twitter complaint categorization [11].

OOV words can be unintentional or intentional, due to mistyping or by using colloquial language. Differentiating between unintentional and intentional OOV is beyond the scope of this lexicon at the moment.

B. Colloquialism in Indonesian

Indonesian or bahasa Indonesia have some form of colloquialism as also found in other languages. Some cases of colloquial words share similar etymology to those in English, e.g. because of the phoneme or sound changes, morphological cases like affixation [12], or even cases like "gay language" [13]. Some examples of these categories and the samples can be seen in Table I.

From linguistics point-of-view, we could also use the lexicon to observe recent trend of slang language in social media. In particular, we could compare occurrence and category of slang words in the lexicon to previous research to analyze how colloquialism in Indonesian vary from time-to-time.

TABLE I
Category for slang words with sample

Case	Formal	Slang
assimilation	kok	koq
vocal modification	sampai	sampe
naturalization	happy	hepi
clipping	lihat	liat
metathesis	bisa	sabi
abbreviation	percaya diri	pede
reversal	ucul	lucu

III. Lexicon Statistics

We built the lexicon by manually translating OOV words from 24,602 Instagram comments from public figure accounts provided by Septiandri and Wibisono [1]. Three annotators whose background is in social media research in bahasa Indonesia annotate the slang words by seeing the full comment first. Majority votes are used to break the ties. The resulting colloquial Indonesian lexicon consists of 3,628 records and

978-1-7281-1175-9/18/$31.00

are mostly slang words in bahasa Indonesia. Each record has 4 columns:

- **slang**: the slang words;
- **formal**: the corresponding formal word;
- **in-dictionary**: information whether the corresponding formal words are in Indonesian Dictionary (KBBI)[1]; and
- **context**: a sample sentence as a context of the slang word occurrence.

Table II provides basic information about the total, unique, and number of in-dictionary words for slang and formal field. There are 3,592 unique slang words and 37 of them have more than one possible formal form: 36 have two, and 1 has three. Table III shows ten samples of slang words with more than one formal form. Among the 1,742 unique formal words, 1,159 (67%) words appear only once. Furthermore, Figure 1 demonstrates that the distribution of formal words occurrence follows Zipf's law [14].

TABLE II
NUMBER OF SLANG AND FORMAL WORDS

	Total	Unique	In-dictionary
slang	3,628	3,592	-
formal	3,628	1,742	1,284

TABLE III
SAMPLES OF SLANG WORDS WITH MULTIPLE FORMAL FORM

Slang word	Formal-1	Formal-2
K	oke	kak
bg	bang	banget
d	di	ada
da	ada	sudah
dk	dek	di
dri	dari	diri
k	ke	kak
kt	kita	kata
km	kamu	kami
anget	banget	hangat

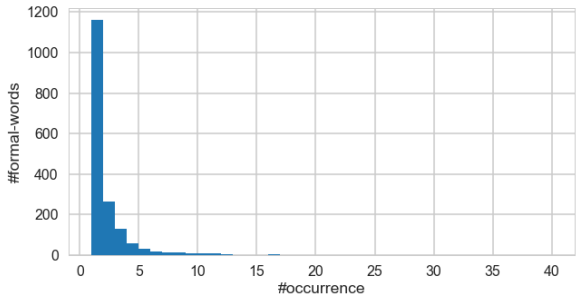

Figure 1. Frequency distribution of unique formal words

From the in-dictionary field, we observed that 459 (26%) of the unique formal words are not registered in KBBI. To understand what kind of words are on the list, we provide 10 samples in Table IV. Most of the non-registered are a registered words with suffix; the word *anak* (*child*) is registered, but *anaknya* (*their child*) is not. We also found words such as *ibunya* (*their mother*), *sayangku* (*my love*), or *begitulah* (*that's it*). Bahasa Indonesia has many types of suffixes such as '-nya' to indicate possessive pronouns and '-lah' to accentuate the meaning of the original word. Other cases include word repetition like *mirip-mirip* (*similar*) or more recent popular slang words that has no standard forms like *unyu* (*cute or endearing*).

TABLE IV
EXAMPLES OF FORMAL WORDS THAT ARE NOT IN KBBI

Slang	Formal	Lemma	Suffix
sbelahnya	sebelahnya	sebelah	nya
anak'y	anaknya	anak	nya
ibu'y	ibunya	ibu	nya
mbakx	mbaknya	mbak	nya
sygku	sayangku	sayang	ku
pipi'y	pipinya	bapak	nya
begitulahh	begitulah	begitu	lah
indah2nya	indah-indahnya	indah	nya
unyuu	unyu	unyu	-
mirip2	mirip-mirip	mirip	-

We can see from the distribution in Figure 2, formal words tend to be longer than the slang words, with the median number of characters per word is 7 for the former and 6 for the latter.

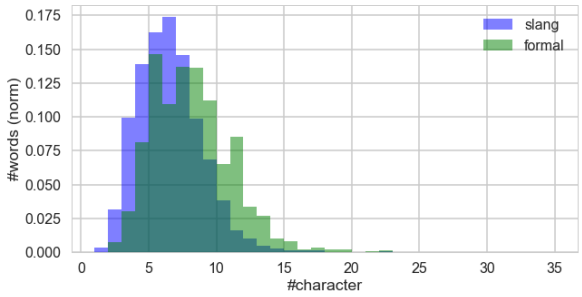

Figure 2. The normalized character length for slang and formal words

We measured the difference between slang and formal words by calculating the Levenshtein distance (edit distance) [15]. The distribution of the edit distance is shown in Figure 3. We can observe that the distribution is positively skewed with the median of 2. A small portion of the words has the edit distance value of more than 10 from character repetitions such as *gedeeeeeeeeeee* (*huge*).

IV. COLLOQUIAL WORDS STATISTICS

In this section, we analyze the frequency of colloquial words in social media texts. The frequency is important to determine whether colloquialism is prominent in processing instagram comments and tweets. For instagram comments, we utilize the same dataset that we use for building lexicon [1]. For

[1]See https://kbbi.kemdikbud.go.id/

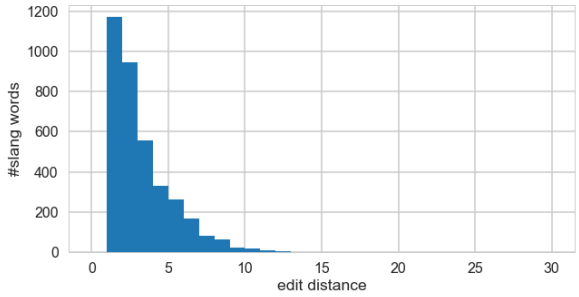

Figure 3. Distribution of edit distance between slang and formal words

each comment in the corpus, we tokenize and count how many token exist in the lexicon to find the proportion of slang words. Table V presents examples for 2 comments:

1) *Ca nnti kl anak ny sekolah nulis nama ny lama tuh, pnjang bnr* (*Ca, later, when the kid is going to school, it will take some time to write her name. Such a long name.*).
2) *wisudanya sederhana bgt make up sm style nya.. bedanyaaa sm wisuda di Indonesia...* (*Such a simple make up and style for the graduation, so different with in Indonesia*)

TABLE V
SAMPLE TOKENS WITH THE CORRESPONDING FORMAL WORDS

Token	Formal	Token	Formal
Ca	-	wisudanya	-
nnti	nanti	sederhana	-
kl	kalau	bgt	banget
anak	-	make	memakai
ny	nya	up	-
sekolah	-	sm	sama
nulis	menulis	style	-
nama	-	nya	-
ny	nya	bedanyaaa	bedanya
lama	-	sm	sama
tuh	-	wisuda	-
pnjang	panjang	di	-
bnr	benar	Indonesia	-

Based on Table V, both comments have 13 tokens, but the number of tokens that exist in the lexicon is different: 7 and 5. Therefore, the proportion for the first and second comments are 0.54 and 0.38 respectively. Note that the phrase *make-up* in the second comment is mistakenly considered as a slang word, because it contains *make* which is also commonly use a slang word for *memakai* (*use*).

The slang words distribution for all comments is shown in Figure 4. Most of the comments are 20% slang words, while we can also observe there is a separate cluster of comments with 0% slang words. In addition, the length of comments is negatively correlated with the slang words proportion, but the Pearson coefficient is weak. This means Instagram comments tend to contain around 20% slang words regardless of the length.

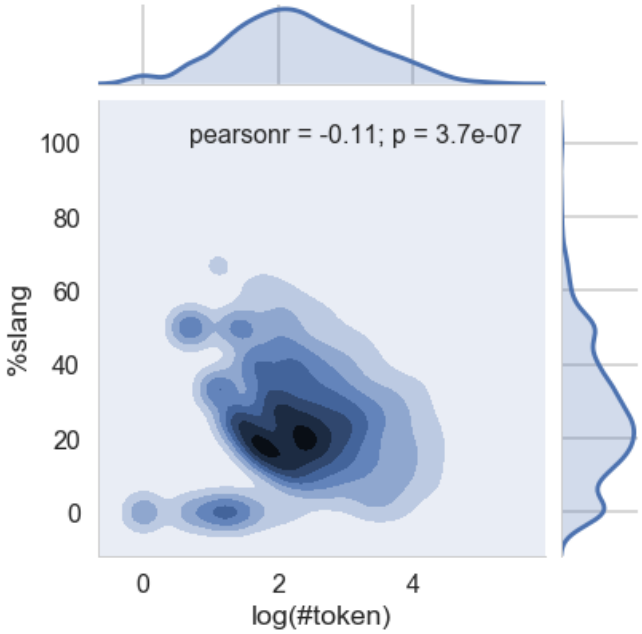

Figure 4. Proportion of slang words in comments

V. LEXICON EVALUATION

To evaluate the lexicon, we reproduced the experiment done in [1] to detect spam comments on Instagram. We built normalized comments by translating slang tokens with the lexicon presented in this paper. If the slang token has more than one possible formal form, we choose the more common one. Table VI shows the best F1-scores from all algorithms for each feature set. To recap, the features and the methods are:

1) *basic*: the number of tokens, number of upper case words, number of numerical characters, percentage of emoji, and the length of the text.
2) *keywords*: hand-engineered keywords provided in the paper
3) *bag-of-words*: binary bag-of-words with latent semantic analysis (LSA)
4) *TF-IDF*: term frequency inverse document frequency with LSA
5) *FastText*: word2vec via skip-gram model using the implementation provided in [16]

Overall, the F1-scores are similar in the raw and normalized versions. We did not see significant improvement by introducing text normalization in this task. The best scores are still FastText+Basic+Keywords, both raw and with normalization, with F1-scores of around 0.96. Normalization only yields better score in TFIDF and FastText+Basic.

VI. CONCLUSIONS

In this paper, we present a lexicon of normalized colloquial words. We believe that this lexicon will be useful for natural language processing tasks in bahasa Indonesia. We have

TABLE VI
F1-SCORE FOR EACH FEATURE SET WITH XGBOOST (*) OR SVM (**) AS THE BEST PERFORMING ALGORITHM

Feature	Unnormalized	Normalized
Basic	**0.7775***	0.7580*
Keywords	**0.8726****	0.8711**
Basic+Keywords	**0.9093****	0.9065*
Bag-of-words (BoW)	**0.9121***	0.9043*
TFIDF	0.9089*	**0.9166****
FastText	**0.9398****	0.9316**
BoW+Basic	**0.9399****	0.9328**
BoW+Basic+Keywords	**0.9381***	0.9370*
TFIDF+Basic	**0.9377****	0.9356**
TFIDF+Basic+Keywords	**0.9436***	0.9423*
FastText+Basic	0.9523**	**0.9547****
FastText+Basic+Keywords	**0.9601****	0.9599**

provided the basic statistics and the lexicon will be freely available on GitHub[2] under the MIT License. There are at least two possibilities to use the lexicon: (1) as a dictionary for a text normalization step, and (2) as a dataset to build a text normalization model.

From our simple evaluation, we found no significant improvement when we introduced normalization with our lexicon to detect spam in Instagram comments. However, it is still inconclusive whether normalization affects Indonesian language processing in general. We need more extensive research to utilize the lexicon for other cases such as sentiment analysis, topic modelling, or question answering and explore whether slang normalization improves performance in Indonesian social media analysis.

ACKNOWLEDGMENTS

This work was supported by Research and Public Services Institution (LP2M) publication grant funded by Universitas Al Azhar Indonesia.

REFERENCES

[1] A. A. Septiandri and O. Wibisono, "Detecting spam comments on indonesias instagram posts," in *Journal of Physics: Conference Series*, vol. 801, no. 1. IOP Publishing, 2017, p. 012069.

[2] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a# twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 368–378.

[3] B. Han, P. Cook, and T. Baldwin, "Lexical normalization for social media text," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 4, no. 1, p. 5, 2013.

[4] M. Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu, "Investigation and modeling of the structure of texting language," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 10, no. 3-4, pp. 157–174, 2007.

[5] P. Cook and S. Stevenson, "An unsupervised model for text message normalization," in *Proceedings of the workshop on computational approaches to linguistic creativity*. Association for Computational Linguistics, 2009, pp. 71–78.

[6] C. Kobus, F. Yvon, and G. Damnati, "Normalizing sms: are two metaphors better than one?" in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 441–448.

[2]https://github.com/nasalsabila/kamus-alay

[7] H. A. Bakr, K. Shaalan, and I. Ziedan, "A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic," in *The 6th international conference on informatics and systems, infos2008. cairo university*, 2008.

[8] A. Aw, M. Zhang, J. Xiao, and J. Su, "A phrase-based statistical model for sms text normalization," in *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006, pp. 33–40.

[9] B. Han, P. Cook, and T. Baldwin, "Automatically constructing a normalisation dictionary for microblogs," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, 2012, pp. 421–432.

[10] Bidang Pengembangan, Pusat Pengembangan dan Pelindungan, "Bagaimana Sebuah Kata Masuk ke KBBI," 2018. [Online]. Available: http://badanbahasa.kemdikbud.go.id/lamanbahasa/artikel/2547/bagaimana-sebuah-kata-masuk-ke-kbbi

[11] N. Hanafiah, A. Kevin, C. Sutanto, Y. Arifin, J. Hartanto *et al.*, "Text normalization algorithm on twitter in complaint category," *Procedia Computer Science*, vol. 116, pp. 20–26, 2017.

[12] S. A. A. Noehilasari, "Periodisasi dan proses pembentukan kosakata bahasa gaul tahun 1990-2012," Ph.D. dissertation, Universitas Negeri Yogyakarta, 2014.

[13] T. Boellstorff, "Gay language and indonesia: Registering belonging," *Journal of Linguistic Anthropology*, vol. 14, no. 2, pp. 248–268, 2004.

[14] G. K. Zipf, "Selected studies of the principle of relative frequency in language," 1932.

[15] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[16] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.