ICIC Express Letters
Volume **16**, Number **11**, November **2022**

ICIC International ©2022 ISSN 1881-803X
pp. **1169–1177**

SENTIMENT ANALYSIS WITH SLANG DICTIONARY IN INDONESIAN SOCIAL MEDIA USING MACHINE LEARNING APPROACH

Jason Andreas Widjaja* and Antoni Wibowo

Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
anwibowo@binus.edu
*Corresponding author: jason.widjaja@binus.ac.id

Received November 2021; accepted February 2022

Abstract. *Indonesia has one of the biggest social media users in this world. In social media, most Indonesians use slang words in their interaction. The usage of slang makes it difficult to understand correctly what it means. The most common way to analyze these texts is sentiment analysis. However, the problem in using sentiment analysis on text in social media is also strongly related to slang expressions. These slang expressions can have various meanings, making it hard to interpret directly. In this paper, we propose a sentiment analysis with Indonesian slang dictionary with machine learning approach to address this issue. This slang dictionary will replace most of slang words with official words. Our experiment shows that this proposed slang dictionary has successfully increased the performance of machine learning classifier in doing sentiment analysis. The proposed slang dictionary has increased the accuracy and F1 score by 8% and 9%, respectively.*

Keywords: Sentiment analysis, Slang, Dictionary, Social media, Indonesian, Machine learning

1. **Introduction.** In recent years, Internet technology in Indonesia has developed very rapidly. Based on data from Hootsuite, Internet users in Indonesia reach 175 million people, which is 64% of the total population of Indonesia. In addition, social media users are also the majority of Internet users with a total of 160 million [1]. Social media itself is an online media that is used by each other where users can easily participate, interact, share, and create content for blogs, social networks, wikis, forums and virtual worlds without regard for space or time. People all over the world use blogs, social networks, and wikis as their primary form of social media.

The problem in this scope is that in the use of social media, users generally do not pay attention to grammar, standard words, and spelling. Moreover, there is an abundance of slang words in social media [2]. This causes many people not to understand what the slang words mean, which can lead to a lot of miscommunications or misunderstanding. When people use slang words, it becomes a complex thing to classify text sentiment. This is because words cannot be interpreted directly because there are some expressions that are expressed implicitly. Furthermore, these slang words are also changing with time. One way to solve this problem is to use sentiment analysis.

Sentiment analysis is the use of text mining and computational linguistics to study people's opinions and attitudes [3,4]. Sentiment analysis is widely applied to voice customer materials such as survey reviews and responses, online and social media, for applications

DOI: 10.24507/icicel.16.11.1169

that range from marketing to customer service to clinical medicine [5]. Nowadays sentiment analysis can also be applied in tourism, owing to big data [6]. Sentiment analysis can also be applied to finance [7].

However, the problem in using sentiment analysis on social media is strongly related to the use of slang expressions. Because this slang expression is difficult to interpret and analyze directly, it takes an extra process to analyze the slang expression [8]. Therefore, in this study, before the classification process is carried out, a preprocessing will be carried out where a dictionary of slang words will be made and if the sentences on social media find these slang words, they will be changed according to the intended meaning with the standard words used accordingly [9].

In Indonesian, these slang expressions are very common. These slang expressions can be in the form of informal words that have special meanings. In addition, in Indonesian texts that are used daily, people often make abbreviations which make these words unable to be properly stemmed. To sum it up, these colloquial texts are ubiquitous and have unstructured syntax [10]. Therefore, this proposed dictionary not only replaces the slang words, but also changes the abbreviated words.

There have been several studies on the topic of sentiment analysis conducted using a machine learning approach with a combination of feature extraction and feature selection. The combination of feature extraction and feature selection methods is used to improve the classification performance of machine learning methods. This can be seen in [11], which compared two feature extraction techniques, namely: TF-IDF and N-gram before the classification technique was carried out. The best results are obtained using the TF-IDF technique and the logistic regression classification and it has better accuracy results than the N-gram. As for feature selection, it can be seen in [12], which compared the feature selection methods: Chi-Square, Correlation, Regularized Locality Preserving Indexing (RLPI), and Information Gain and the results found that Correlation has good results.

Based on research that has been done, it can be seen that feature extraction and feature selection can improve machine learning classification performance. Therefore, in this study, a comparison will be made between feature extraction, feature selection, and machine learning classifier methods. For feature extraction, TF-IDF and N-gram will be used. For feature selection, Information Gain will be used. For machine learning classifier, we will use Naïve Bayes, Decision Tree, Random Forest, SVM, k-NN, and Max Entropy as comparisons.

This paper is structured as follows. The related works are listed in Section 2 of this paper. The methodology is explained in Section 3. The result and discussion are explained in Section 4. The conclusion is explained in Section 5.

2. **Related Works.**

There are a lot of studies about sentiment analysis using various methods and using various datasets. Social media is one of the most used sources of datasets. The reason is because social media provides newest and wide array of information [13]. Most popular social media as dataset tools includes Twitter [11,14,15], Facebook [16,17] and other social media [13,18]. Another popular source of datasets is reviews, in which movie reviews are the most sought-after [12,19,20].

One of methods to do sentiment analysis is by using machine learning. Employing machine learning method has produced various results. The results are in a range of F1 value 75%-80% [14,18]. Some machine learning methods have better performance in processing sentiment analysis such as OneR [20]. Moreover, machine learning methods can also be combined with feature extraction methods to further boost their performance as conducted by [11,12]. Machine learning methods can also be employed on cloud services [17].

Another method to do sentiment analysis is lexical method. [16] compared the performance of lexicon-based and machine learning method and found no significant differences. In [21], a comparison between lexicon-based approaches for sentiment analysis was studied, with 77% accuracy as the best result. In [22], a sentiment analysis using combination of lexicon and neural networks was studied and achieved a final accuracy of 91%. Another study [23] compared machine learning and lexicon-based approach and found that lexicon-based approach has superior result.

In recent years, the rapid development of deep learning has caught interest of many researchers to develop this method on sentiment analysis. In [24], a deep learning model to detect radicalism via sentiment analysis has achieved an 81% accuracy. In [25], a study on sentiment analysis on comment text using BiLSTM method was proposed and proved that it has higher accuracy than RNN, CNN, LSTM, and NB. Another research [26] combined sentiment lexicon and deep learning CNN and GRU. This proposed model achieved a high accuracy of 93%.

In this study, we have decided to apply our proposed slang dictionary in machine learning method as it is the most standard method of doing sentiment analysis.

The summary of related works can be seen in Table 1.

Table 1. Summary of related works

No	Journal	Dataset	Methods	Accuracy
1	[11]	Twitter	TF-IDF, N-gram	TF-IDF has a higher accuracy than N-gram with a difference of 3%-4%
2	[12]	IMDb	Hybrid feature extraction	Hybrid feature extraction method improves performance
3	[14]	Twitter	Machine learning	79% accuracy
4	[15]	Twitter	Hybrid	F1 value 84%
5	[16]	Facebook	Machine learning & lexicon	No significant difference between lexicon and machine learning
6	[17]	Facebook	Cloud machine learning	Improve the performance of the existing system
7	[18]	Social media	Machine learning	F1 value 75%
8	[19]	Amazon & IMDb	Contextual analysis for supervised machine learning	Estimated error 2.75-3.94
9	[20]	-Amazon -IMDb	Naïve Bayes, J48, BFTree, and OneR	OneR has the highest level of accuracy
10	[21]	Movie reviews	Lexicon	Best accuracy 77%
11	[22]	IMDb	Lexicon and neural networks	91% accuracy
12	[23]	Blogs	Machine learning and lexicon-based	Lexicon-based has better performance
13	[24]	Twitter	Deep learning	81.63% accuracy
14	[25]	Hotel review	BiLSTM	F1 score 92.18%
15	[26]	Book review	Lexicon and deep learning	Accuracy 93.3%

3. **Methodology.** The general overview of the method consists of these several steps: 1) Collecting data; 2) Preprocessing; 3) Feature extraction; 4) Feature selection; 5) Classification; 6) Evaluation.

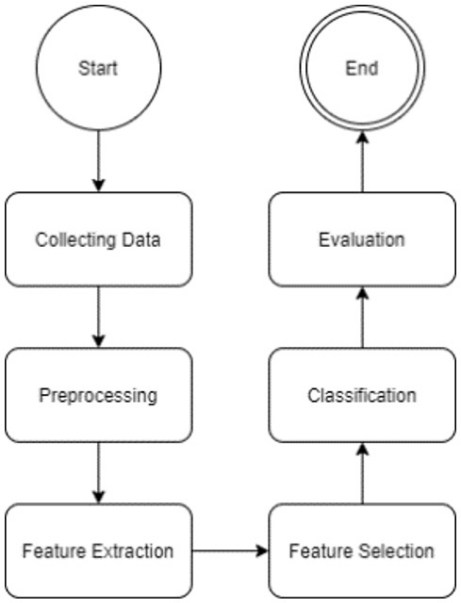

FIGURE 1. Methodology

A. Collecting data

The dataset used will be taken from Twitter. The keywords that will be used as data searches on Twitter are 15 slang words that often appear on social media. The reason why these keywords were chosen is to make sure the slang dictionary to be used consistently. These words can be seen in Table 2.

TABLE 2. Slang keywords

	Slang words
1	wkwk
2	baper
3	galau
4	jomblo
5	php
6	lebay
7	alay
8	santuy
9	kuy
10	sabi
11	kepo
12	gaje
13	mantul
14	mager
15	pw

However, the proposed slang dictionary is not limited to these fifteen words.

Tweets taken from Twitter start from August 1, 2021, to October 1, 2021. Tweets from Twitter were obtained using the Twitter API with the python programming language with tweepy library. In this research a total of 1000 Tweets were collected.

B. Preprocessing

In general, there are 7 preprocessing steps in this research. These steps are 1) Change emoji & emoticons; 2) Data labeling; 3) Remove noise; 4) Normalize Case; 5) Remove stopwords; 6) Create a slang dictionary; 7) Tokenization & Stemming.

In analyzing sentiment, we cannot forget about emoji and emoticons. These emojis and emoticons function as representations of the sentiments felt by humans. Therefore, these emojis and emoticons should not be deleted because they have important meanings. In order for processing to be carried out, these emojis and emoticons must be converted into text. For example, ":)" will be changed to "smiley".

Because the data taken from the data is raw data, it is necessary to perform data labeling. The results of this data labeling the data will be divided into 3 separate labels, namely: positive data, negative data, and neutral data.

The noise in question is RT, mentions, numbers, punctuation marks and links in tweets. This removed noise has no function in sentiment analysis.

Normalize Case is the process of converting a text into a standard format. In this study, the Normalize Case process is carried out by changing all uppercase letters to lowercase. This is useful for making the process more consistent.

Stopwords are short words that are often used in language. These stopwords have no meaning by themselves so they will be deleted. Examples of stopwords in Indonesian are "yang", "for", "pada", "ke", "para", etc.

At this stage, a slang dictionary will be created. A slang dictionary is a list of slang words with meanings and definitions. At this stage, the slang words found will be converted into the closest meaning or definition that exist in the official dictionary of Indonesian (KBBI). For example, in this slang dictionary, the word "gaje" will be changed to "tidak jelas". This result will make it possible to stem the word properly.

Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be words, characters, or subwords. The tokenization used in this experiment is word tokenization. Word tokenization is the most commonly used tokenization algorithm. It divides a piece of text into individual words based on certain delimiters. Depending on the delimiter, different word-level tokens are generated. An example of this tokenization is the sentence "I am eating", this sentence will have 3 tokens, namely: "I-am-eating". Stemming is the process of reducing a word into stems that are affixed with suffixes and prefixes or to the root of a word known as lemma. An example of stemming in Indonesian is that "pemberi" will be changed to "beri".

C. Feature extraction

After doing the preprocessing stage, the steps taken before doing the classification are feature extraction and feature selection. Feature extraction and feature selection are useful for increasing the level of accuracy at the classification stage. Feature extraction used in this research is TF-IDF and N-gram.

The N-gram feature extraction method will group each successive N-character in a token. Then the results of the grouping will be the features that will go through a weighting process in the TF-IDF feature extraction method.

D. Feature selection

The next step is feature selection. This feature selection is used to optimize and improve accuracy results. Feature selection that will be used in this research is Information Gain. This feature selection method is expected to make the model have a faster and better performance.

E. Classification

At this stage, a machine learning classification method will be applied. In this study, Random Forest, k-NN, Decision Tree, Naïve Bayes, Max Entropy, and SVM will be used.

F. Evaluation

The next step is evaluation. To find out the results of the evaluation, a confusion matrix will be used to measure the values of accuracy, precision, recall, and F1 score.

4. **Results and Discussion.** In this study, python programming language was used. The library used includes tweepy, sklearn, and nltk. After doing all the steps mentioned

in Section 3. In Table 3, we compiled the result of machine learning classification without slang dictionary.

Table 3. Result without slang dictionary

N-gram	Classification method	*Accuracy*	*Precision*	*Recall*	*F1 score*
Unigram	Naïve Bayes	0.56	0.5	0.56	0.48
	Decision Tree	0.45	0.45	0.45	0.43
	Random Forest	0.5	0.47	0.5	0.46
	k-NN	0.47	0.45	0.47	0.44
	SVM	0.56	0.57	0.56	0.42
	Max Entropy	0.57	0.52	0.57	0.48
Bigram	Naïve Bayes	0.46	0.39	0.46	0.41
	Decision Tree	0.5	0.35	0.5	0.4
	Random Forest	0.51	0.35	0.51	0.4
	k-NN	0.55	0.37	0.55	0.42
	SVM	0.55	0.39	0.55	0.41
	Max Entropy	0.55	0.44	0.55	0.43
Trigram	Naïve Bayes	0.5	0.46	0.5	0.47
	Decision Tree	0.51	0.34	0.51	0.39
	Random Forest	0.52	0.38	0.52	0.4
	k-NN	0.28	0.5	0.28	0.2
	SVM	0.56	0.32	0.56	0.41
	Max Entropy	0.56	0.32	0.56	0.41

We can compare it to Table 4, which applies the slang dictionary.

Table 4. Result with slang dictionary

N-gram	Classification method	*Accuracy*	*Precision*	*Recall*	*F1 score*
Unigram	Naïve Bayes	0.54	0.48	0.54	0.49
	Decision Tree	0.55	0.5	0.55	0.51
	Random Forest	0.54	0.49	0.54	0.5
	k-NN	0.53	0.46	0.53	0.48
	SVM	0.6	0.55	0.6	0.47
	Max Entropy	0.55	0.46	0.55	0.47
Bigram	Naïve Bayes	0.54	0.53	0.54	0.5
	Decision Tree	0.58	0.49	0.58	0.49
	Random Forest	0.59	0.51	0.59	0.5
	k-NN	0.6	0.45	0.6	0.48
	SVM	0.62	0.69	0.62	0.48
	Max Entropy	0.6	0.47	0.6	0.48
Trigram	Naïve Bayes	0.55	0.46	0.55	0.5
	Decision Tree	0.63	0.47	0.63	0.52
	Random Forest	0.62	0.46	0.62	0.52
	k-NN	0.64	0.46	0.64	0.53
	SVM	0.65	0.43	0.65	0.52
	Max Entropy	0.65	0.43	0.65	0.52

To better visualize the difference between the performance of the machine learning without slang dictionary and the proposed slang dictionary, we have compiled the average accuracy and F1 score between the two in Table 5.

Table 5. Results comparison

Average	Without slang dictionary	With slang dictionary
Accuracy	0.51	0.59
F1 score	0.41	0.50

From these tables, we can see that the slang dictionary has significantly improved the performance of machine learning classifier in analyzing sentiment. The average accuracy without slang dictionary is only at 51% compared to with slang dictionary at 59%. Furthermore, the F1 score also differs significantly with 50% compared to 41% favoring the one with slang dictionary.

Moreover, from the results, we can see that the results with bigram and trigram are slightly better than the one with unigram. The reason for that is because in this slang dictionary, most of the slang words are changed to two or more words, which results in more features to be extracted.

As for the overall best performance results in this experiment, we can see them in Table 6.

Table 6. Best performance

N-gram	Classifier	Accuracy	F1 score
Trigram	SVM	0.65	0.52
	Max Entropy	0.65	0.52

From table above, we can see the best performed classifiers are SVM and Max Entropy with trigram N-gram. For SVM, the reason why it has good performance is that in this dataset, there are a lot of features extracted; hence SVM performs well [27]. Since the principle of Max Entropy is determined by probability distribution by words [28], it can have good performance in this experiment because most of the slang words have consistent sentiment.

Overall, we believe that this sentiment analysis with machine learning combined with slang dictionary method on Indonesian social media has improved the accuracy significantly. We believe that the slang dictionary can be enhanced more by adding more slang and abbreviated words.

5. **Conclusions.** In this paper, we have applied the proposed slang dictionary on sentiment analysis on Indonesian social media. After the experiments, we have found that the slang dictionary is able to boost machine learning classifiers performance. The average accuracy with slang dictionary is 8% higher compared to without slang dictionary. As for the F1 score, the average F1 score with slang dictionary has increased 9% compared to without slang dictionary.

In this regard, we believe that the accuracy can be improved more if this slang dictionary is enhanced. Although the words in this slang dictionary are not small, there are still many Indonesian slang expressions that are not in this dictionary. Moreover, creating an official dictionary that contains the meanings of Indonesian slang expressions is advisable. With this official dictionary, the slang translation results will be more consistent.

REFERENCES

[1] *Hootsuite*, www.hootsuite.com, 2020.

[2] K. Matsumoto, F. Ren, M. Matsuoka, M. Yoshida and K. Kita, Slang feature extraction by analyzing topic change on social media, *CAAI Trans. Intelligence Technology*, vol.4, no.1, pp.64-71, DOI: 10.1049/trit.2018.1060, 2019.

[3] W. Medhat, A. Hassan and H. Korashy, Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal*, vol.5, no.4, pp.1093-1113, DOI: 10.1016/j.asej.2014.04.011, 2014.
[4] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, 2020.
[5] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, Opinion mining and sentiment analysis, *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp.452-455, 2016.
[6] A. R. Alaei, S. Becken and B. Stantic, Sentiment analysis in tourism: Capitalizing on big data, *Journal of Travel Research*, vol.58, no.2, pp.175-191, DOI: 10.1177/0047287517747753, 2019.
[7] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev and D. Trajanov, Evaluation of sentiment analysis in finance: From lexicons to transformers, *IEEE Access*, vol.8, pp.131662-131682, DOI: 10.1109/ACCESS.2020.3009626, 2020.
[8] L. Wu, F. Morstatter and H. Liu, SlangSD: Building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification, *Language Resources and Evaluation*, vol.52, no.3, pp.839-852, DOI: 10.1007/s10579-018-9416-0, 2018.
[9] A. G. Prasad, S. Sanjana, S. M. Bhat and B. S. Harish, Sentiment analysis for sarcasm detection on streaming short text data, *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pp.1-5, DOI: 10.1109/ICKEA.2017.8169892, 2017.
[10] D. S. Maylawati, W. B. Zulfikar, C. Slamet, M. A. Ramdhani and Y. A. Gerhana, An improved of stemming algorithm for mining Indonesian text with slang on social media, *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, pp.1-6, DOI: 10.1109/CITSM.2018.8674054, 2018.
[11] R. Ahuja, A. Chug, S. Kohli, S. Gupta and P. Ahuja, The impact of features extraction on the sentiment analysis, *Procedia Computer Science*, vol.152, pp.341-348, DOI: 10.1016/j.procs.2019.05.008, 2019.
[12] K. Kumar, B. S. Harish and H. K. Darshan, Sentiment analysis on IMDb movie reviews using hybrid feature extraction method, *International Journal of Interactive Multimedia and Artificial Intelligence*, vol.5, no.5, p.109, DOI: 10.9781/ijimai.2018.12.005, 2019.
[13] L. Yue, W. Chen, X. Li, W. Zuo and M. Yin, A survey of sentiment analysis in social media, *Knowledge and Information Systems*, vol.60, no.2, pp.617-663, DOI: 10.1007/s10115-018-1236-4, 2019.
[14] A. Hasan, S. Moin, A. Karim and S. Shamshirband, Machine learning-based sentiment analysis for Twitter accounts, *Mathematical and Computational Applications*, vol.23, no.1, DOI: 10.3390/mca23010011, 2018.
[15] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, Sentiment analysis of tweets using machine learning approach, *2018 11th International Conference on Contemporary Computing (IC3)*, pp.1-3, DOI: 10.1109/IC3.2018.8530517, 2018.
[16] C. Dhaoui, C. M. Webster and L. P. Tan, Social media sentiment analysis: Lexicon versus machine learning, *Journal of Consumer Marketing*, vol.34, no.6, pp.480-488, DOI: 10.1108/JCM-03-2017-2141, 2017.
[17] R. Arulmurugan, K. R. Sabarmathi and H. Anandakumar, Classification of sentence level sentiment analysis using cloud machine learning techniques, *Cluster Computing*, vol.22, no.S1, pp.1199-1209, DOI: 10.1007/s10586-017-1200-1, 2019.
[18] D. Li, R. Rzepka, M. Ptaszynski and K. Araki, A novel machine learning-based sentiment analysis method for Chinese social media considering Chinese slang lexicon and emoticons, *The AAAI-19 Workshop on Affective Content Analysis AFFCON2019: Modeling Affect-In-Action*, HI, Hawaii, USA, 2019.
[19] A. A. Aziz and A. Starkey, Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches, *IEEE Access*, vol.8, pp.17722-17733, DOI: 10.1109/ACCESS.2019.2958702, 2020.
[20] J. Singh, G. Singh and R. Singh, Optimization of sentiment analysis using machine learning classifiers, *Human-Centric Computing and Information Sciences*, vol.7, no.1, p.32, DOI: 10.1186/s13673-017-0116-3, 2017.
[21] V. Bonta, N. Kumaresh and N. Janardhan, A comprehensive study on lexicon based approaches for sentiment analysis, *Asian Journal of Computer Science and Technology*, vol.8, no.S2, pp.1-6, DOI: 10.51983/ajcst-2019.8.S2.2037, 2019.
[22] Z. Shaukat, A. A. Zulfiqar, C. Xiao, M. Azeem and T. Mahmood, Sentiment analysis on IMDB using lexicon and neural networks, *SN Applied Sciences*, vol.2, no.2, p.148, DOI: 10.1007/s42452-019-1926-x, 2020.

[23] N. Mukhtar, M. A. Khan and N. Chiragh, Lexicon-based approach outperforms supervised machine learning approach for Urdu sentiment analysis in multiple domains, *Telematics and Informatics*, vol.35, no.8, pp.2173-2183, DOI: 10.1016/j.tele.2018.08.003, 2018.

[24] A. Iriani, Hendry, D. H. F. Manongga and R.-C. Chen, Mining public opinion on radicalism in social media via sentiment analysis, *International Journal of Innovative Computing, Information and Control*, vol.16, no.5, pp.1787-1800, 2020.

[25] G. Xu, Y. Meng, X. Qiu, Z. Yu and X. Wu, Sentiment analysis of comment texts based on BiLSTM, *IEEE Access*, vol.7, pp.51522-51532, DOI: 10.1109/ACCESS.2019.2909919, 2019.

[26] L. Yang, Y. Li, J. Wang and R. S. Sherratt, Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning, *IEEE Access*, vol.8, pp.23522-23530, DOI: 10.1109/ACCESS.2020.2969854, 2020.

[27] M. R. Huq, A. Ali and A. Rahman, Sentiment analysis on Twitter data using KNN and SVM, *International Journal of Advanced Computer Science and Applications*, vol.8, no.6, pp.19-25, 2017.

[28] C. Yin and J. Xi, Maximum entropy model for mobile text classification in cloud computing using improved information gain algorithm, *Multimedia Tools and Applications*, vol.76, no.16, pp.16875-16891, DOI: 10.1007/s11042-016-3545-5, 2017.