International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 6, December 2018, pp. 5311~5317
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i6.pp5311-5317

Complaint Analysis in Indonesian Language Using WPKE and RAKE Algorithm

Rini Wongso, Novita Hanafiah, Jaka Hartanto, Alexander Kevin, Charles Sutanto, Fiona Kesuma
Computer Science Department, School of Computer Science, Bina Nusantara University, Indonesia

Article Info

Article history:

Received Dec 25, 2018
Revised Jul 3, 2018
Accepted Jul 22, 2018

Keyword:

Complaint analysis
RAKE Algorithm
Twitter
WPKE Algorithm

ABSTRACT

Social media provides convenience in communicating and can present two-way communication that allows companies to interact with their customer. Companies can use information obtained from social media to analyze how the communities respond to their services or products. The biggest challenge in processing information in social media like Twitter, is the unstructured sentences which could lead to incorrect text processing. However, this information is very important for companies' survival. In this research, we proposed a method to extract keywords from tweets in Indonesian language, WPKE. We compared it with RAKE, an algorithm that is language independent and usually used for keyword extraction. Finally, we develop a method to do clustering to groups the topics of complaints with data set obtained from Twitter using the "komplain" hashtag. Our method can obtain the accuracy of 72.92% while RAKE can only obtain 35.42%.

Corresponding Author:

Rini Wongso,
Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jl K.H. Syahdan No. 9, Palmerah, Jakarta, 11480, Indonesia.
Email: rwongso@binus.edu

1. INTRODUCTION

Social networks have become an influential means of communicating thesedays, as it allows interaction between acquaintances in different society [1]. Social media is currently a lifestyle for most people all around the world [2]. Twitter, one of social media apps, has approximately 500 million tweets and 307 million active users as stated in Live Stats on 2017 [3]. It has been known that Twitter is used for many purposes such as for protest, political campaigns, marketing, and for commenting services or products [4].

According to G. Ghedin, the high diffusion of Twitter has clearly reflected what happens in Indonesia's marketing world [5]. Dozens of companies use Twitter as the perfect media to interact with their clients. However, it is not easy to evaluate the popularity or acceptance rate of products or services as all the information is scattered and there is no way to manage it well. The amount of information that goes through Twitter takes time for managers to analyze the core of the complaints and sometimes there are tweets that are not meaningful. The process will be efficient by using a machine to extract the core of a sentence (keyword).

Extracting a keyword of a short sentence is one of the challenges in natural language processing area. It is stated by N. Hanafiah that people tends to use unstructured sentences such as incorrect grammar, contains many abbreviation, typographical errors, and emoticons in expressing their thoughts in social media [6]. The unstructured sentences need to be normalized so the machine can understand the words. Afterwards, the extracting keyword algorithm can be applied to get the core of complaints in Tweets. Despite the difficulties stated above, these data are certainly very useful for the company to know the communities' responses towards their products or services. By having these data, companies can make an appropriate decision making for their sustainability [4].

Research of keyword extraction has combined natural language processing approaches to identify part-of-speech (POS) tags which are combined with supervised learning, machine learning algorithm or statistical methods. In R. Mihalcea and P. Tarau, a system that applies syntactic filters to identify POS tags are described [7]. POS tags are used to select words to be evaluated as keywords. The co-occurences of selected words are accumulated within a word co-occurrence graph and TextRank (a graph-based ranking algorithm) are used to rank the words based on their associations in the graph. Then, keywords are selected by the top-ranking words. The research reported that TextRank performed the best when only noun and adjectives are selected as candidate keywords.

In text processing, certain algorithm is needed to obtain keywords, and one of the algorithm that is often used is RAKE (Rapid Automatic Keyword Extraction) algorithm. Recent research of [8] compares the performance of RAKE and TextRank using the same data set as in [7]. S. Rose *et al.* they described RAKE as an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents [8]. It is based on the observation that keywords frequently contain multiple words but rarely contain standard punctuation or stop words, such as the function words and, the, of, or other words with minimum lexical meaning.

The input parameters for RAKE are a list of stop words, and a set of phrase delimiters to parse the document text into candidate keywords. Co-occurrences of words within candidate keywords are meaningful to score the candidate keywords. RAKE begins keyword extraction on a document by parsing the text into a set of candidate keywords. A score is calculated for each candidate keyword by calculating the sum of its member word scores after every candidate keyword is identified and the graph of co-occurrences is completed. Next, the top T scoring candidates are selected as keywords of the document. In short, firstly, RAKE removed the stop-words from document and define the candidate keyword according to the domain by calculating word score based on the degree and frequency of word vertices in the graph: (1) word frequency, word degree, and ratio degree to frequency [9]. In the experiment of [8], RAKE achieves higher precision and similar recall in comparison to TextRank, as RAKE can score keywords in a single pass, while TextRank requires repeated iterations to achieve the convergence on word ranks.

A research done by Jungiewicz and Łopuszyński uses RAKE for doing keyword extraction of Polish documents in Procurement field [10]. RAKE is quite independent in terms of language as it is not developed only for a certain language. RAKE depends on the stop word list with the general idea of separating a text to group of words according to a separator or word from the stop word list. Each word will be considered as a candidate keyword and a score is calculated based on the co-occurrence graph.

According to the survey done by S. Siddiqi and A. Sharan, there are various techniques that can be used in text mining for extracting keyword and key phrase [11]. Both keyword and key phrase are needed to analyze huge number of material in form of text. Keyword and key phrase are word representation in a document which give high-level specification of the content and usually used for generating index, query refinement, and text summarization. In this method, significant words in a document are chosen without depending on any vocabulary or extracted words from the document.

Some researchers J. Greenberg *et al*. compare four open source algorithms for keyword extraction, RAKE, Tagger, Kea, and Maui [12]. According to their experiments, RAKE produce 98.57% unique words (69 of 70 unique words of 70 extracted words). Meanwhile the best result is obtained by using Tagger (100%, 50 unique words of 50 extracted words). RAKE is language independent system. However, the development of stop word list in Indonesian is not as complete as English. Hence, we proposed a WPKE (Weight Priority Keyword Extraction) algorithm which has a higher accuracy in Indonesian Tweets. The ranking process is done by giving an initial weight for each word which we have analyzed from complaint tweets. Next, the weight is being adjusted by considering the relationship between words of Indonesian grammar. The keyword is processed to grouping phase for calculating the keyword that appears in the tweets to produce the chart. Our WPKE algorithm works well in Indonesian tweets comparing with RAKE algorithm.

2. RESEARCH METHOD

Figure 1 illustrates the proposed method in the research. It begins by collecting the input of text from Twitter. Our data set consists of tweets mentioning the account of companies' customer relation center in Indonesian language. The text is pre-processed, to make the unstructured sentences can be more understandable by a machine. The normalization technique used is based on the previous research done by [6]. It developed the technique to normalize text in Indonesian language for complaint category by using data from Twitter and achieved the accuracy around 90%. The steps are divided into cleaning process, OOV detection, and word replacement. Keywords are then extracted using WPKE algorithm and grouped the word that have the similar meaning to result the complaint category. The output is visualized in a chart to provide simplicity for further analysis.

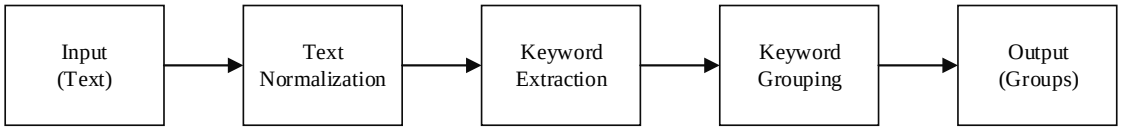


Figure 1. The proposed method

a. Keyword Extraction

Keyword extraction begins after the input text are normalized. We experimented using two algorithms. First, we apply RAKE algorithm [8] to extract the keywords from the Tweets by splitting the tweets into sentences and removing the less meaningful words using stop word list. This step generates list of candidate keywords. Next, a score is calculated for each candidate keyword according to the frequency and known as degree. An example of RAKE algorithm result is shown in Table 1.

Table 1. Keywords extracted by using RAKE

Input	Keywords
"kecewa order Goxxx tapi tidak ada tnggapan dari supir supir saya telepon tidak menjawab si supir sndri juga tidak menelpon saya"	[('kecewa order goxxx, 10.0), ('si supir sndri', 7.026666666666667), ('tnggapan', 2.0), ('menelpon', 2.0), ('telepon', 2.0), ('supir supir', 0.7200000000000001)]

There are 6 candidate keywords extracted from the Tweets each with its score. Accordingly, with the score, the final keyword of this sentence is keyword with the highest value which is "kecewa order goxxx". We then evaluate by requesting some people who understand Indonesian languange to review the keywords obtained by this algorithm, and the result is not satisfactory. After some experiments and analysis, we proposed a method called WPKE (Weight Priority Keyword Extraction) which works based on certain weighting schemes. Steps of the method is illustrated by Figure 2.

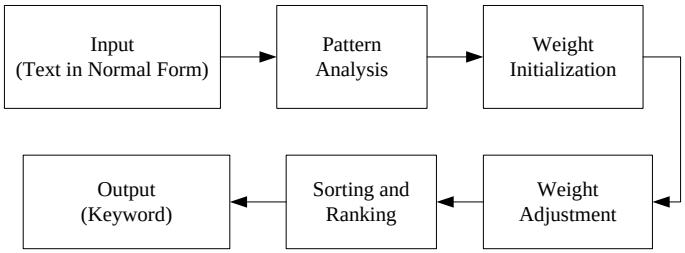


Figure 2. Our proposed WPKE method

WPKE method begins by defining the pattern of the tweet. Based on our analysis, the typically variation of a keyword in a *tweet* is constructed from the following patterns: (1) Noun + Noun, (2) Noun, (3) Verb + Noun, (4) Pronoun. At first, we give an initial weight for each word type according to pattern with the following rules described in Table 2. We give a value of 1 for Noun, because usually the thing to discuss in a sentence is about an object. Meanwhile, the initial weight for an adjective is 0, since the adjective is not a typical keyword which can show the essence of a tweet, but it is usually used in a sentence related to a noun. A value of 0.5 and 0.1 are given to Verb and Pronoun respectively, according to the possibility a keyword in a tweet is a Verb or Pronoun. Other type of words that are not described in Table 2 is ignored.

Table 2. Initial Weight Rules in WPKE method

Type	Initial Weight
Noun	1
Verb	0.5
Pronoun	0.1
Adjective	0

Furthermore, the current weight is adjusted based on the order in which the word is formed as shown in table 3. The candidate score is calculated using the formula: *Candidate Score = Initial Weight – Word Distance + Additional Weight.* For example, a normalized tweets "Saya kecewa nasinya bau" has a type of word "Pronoun Adjective Noun Adjective". Each word in the tweet is processed started from the first word "saya" produces phrase combinations of "saya kecewa", "saya nasinya", and "saya bau". The second word "kecewa" produces phrase combinations of "kecewa nasinya" and "kecewa bau", while the third word only produces one phrase combination of "nasinya bau". The score for candidate "saya kecewa" is -0.9 where the initial weight of the word "saya" (Pronoun) is 0.1, the word distance between "saya" and "kecewa" is 1 and the additional weight is 0 (because there is no pattern formed as shown in Table 3 below). For the remaining candidate: "saya nasinya", "saya bau", "kecewa nasinya", "kecewa bau", and "nasinya bau" have candidate score -1.9, -2.9, -1, -2, and 0.5 respectively. The final keyword we obtained is the one with the highest candidate score.

Table 3. Weight Adjustment Rules in WPKE method

Patterns	Additional Weight
Noun+"tidak" + Adjective	1
Noun+Noun/Verb+Adjective	0.5
Noun + Noun	0.5
Others	0

b. Keyword Grouping

Typically, several tweets have the same main topic, therefore we want to group those similar keywords into a group and rank them. The grouping method is based on the given input of Twitter account or topic. We prepared the data in prior to keyword grouping by querying from database, the data of: "kepada", "perihal", "tweet_hasil", and "inti". These data are obtained from the normalization process, except for "inti", which we get from keyword extraction process as mentioned in Section 2.1. The query result is filtered according to certain criterias of: (1) finding tweets containing the input in "kepada" or "perihal" without having perfect match (%input%), (2) the input must not be preceded by any other characters to avoid irrelevant tweets being processed. The details of the second criteria can be seen in Table 4 below. The last row shows "rejected" status since the "ab" appears in a word "akr*ab*" in the tweet.

Table 4. Tweet Processing Criteria

Input	Tweet	Status
ab	saya kecewa dengan ab	Accepted
ab	saya kecewa dengan abcare	Accepted
ab	ab mengecewakan	Accepted
ab	abcare mengecewakan	Accepted
ab	Saya kecewa tidak akrab	Rejected

Keyword grouping begins by adding a flag to give a status whether a tweet has been processed. The keyword phrase obtained from "inti" is divided into word to search for a keyword without having to have a perfect phrase match. These words are used to search for the same group of words in other tweets. The algorithm calculates the frequency of occurrences of a keyword against keywords from other tweets. This process produces the order of words that most often appear, whereas the same result is ordered by the length of the word in ascending order. The example results of keyword grouping is shown in Figure 3. The keyword in tweet-5 does not give a contribution value to the word 1 and word 2, therefore this tweet goes to keyword grouping again (loop). The original sentence of tweet-5 is "internetnya lambat payah buat kecewa berat aja" where the word "internet" makes this tweet displayed from the query results. When there is no result of querying word 1 and 2 of tweet-5, this keyword goes to the Extended Grouping phase.

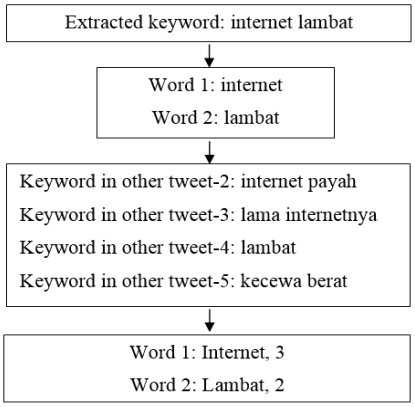


Figure 3. Keyword Grouping Example

We enhanced the grouping process to get similar tweets that has not been processed to be grouped together. This is needed to reduce the number of complaint keywords that is shown in the chart. Extended grouping basically uses the list of keywords that has the most similar set of words to search for the right keyword in the tweet. This process multiplies keyword that has a high score and let the keyword that has a low score to be filtered. Sorting in the previous step is important to determine which keyword will be given the first opportunity to find the keyword in the tweet that has not been processed. Extended grouping process will get keyword with the highest score and search for a match towards the sentences that has the same keyword, and just like previous keyword grouping process, the score will be added, and the flag status will be updated.

3. RESULTS AND ANALYSIS

Our experiment is done on a small set of data (50 tweets) retrieved from tweets mentioning two telecommunication service companies in Indonesia, where in this paper the names are dissembled as: '@com1' and '@com2' using the hashtag "komplain" (#komplain). Other company we discussed as example in this section is commercial company in Indonesia, dissembled as '@comm3'. All the tweets are in Indonesian language. Based on the experiment of keyword extraction RAKE and WPKE, we obtain the following result as shown in Table 5.

Table 5. Comparison of Keyword Extraction using RAKE and WPKE

	Number of Correct Keywords	Number of Incorrect Keywords	Percentage
RAKE	6	44	12%
WPKE	18	32	36%

Based on the result described in Table 5, WPKE method successfully obtained 18 correct keywords (36%) from 50 tweets, meanwhile RAKE algorithm can only obtain 6 correct keywords (12%). Checking is done manually by selected reviewers by giving them the data tweets and ask them for the keywords. Correct keywords from 26 tweets (52%) cannot be found either by RAKE or WPKE. For example, in a tweet of: "(@com1) jaringan comm1 kenapa nih leletnya super (#komplain)", using RAKE we obtained keyword of "leletnya super" while using WPKE we obtained "jaringan comm1". In this example, WPKE is considered correct, and RAKE is incorrect. The error in RAKE algorithm can be seen from the selection of words that failed to be normalized because of the English word "super", while WPKE successfully extract the complaint keyword as the pattern of "Noun + Noun" is found.

According to our analysis, in several cases RAKE tends to take most of the tweet as the keyword, as in tweet of "(@comm3) kalau emang tidak bisa nanogram barang pas hari sabtu tidak usah email pesan kirim barang buang-buang waktu aja (#SAMPAH) (#Komplain)", RAKE gives a long output as "email pesan kirim barang buang-buang" that affects the keyword search result. Based on the result, WPKE provide more precise output because of the ability to recognize patterns that have been adjusted to the keyword patterns of complaint in Indonesian language, while the RAKE algorithm has a disadvantage due to the use of a stop word list that is not supported with a complete list of stop words in Indonesian language. The lack of stop

words makes the RAKE algorithm tends to produce longer, less precise keyword, and leads to larger computational loads as the number of word checks increases with the longer keyword obtained. The RAKE algorithm also has a more suitable method used for keyword extraction of a document, not for text like microblogs that only has a maximum length of 140 characters. Moreover, RAKE tokenized word according to a list of stop words and continued with frequency calculation, degrees, and the appearance of word in the document.

We obtained the following result as can be seen in Table 6 for the grouping process using both *keyword grouping* and *extended grouping.*

Table 6. Comparison of Keyword Grouping and Extended Grouping

	Keyword Grouping	Extended Grouping
Total	21	28
Percentage	52.25%	70%

The result in Table 6 above is obtained from a total of 40 data, taken randomly from Twitter with the "komplain" hashtag in Indonesian language, a different data set from the one used for keyword extraction above. Based on Table 6 above, it can be seen that by doing the iteration twice, keyword grouping that is continued by extended grouping, there is an increase of 17.25%. Total of data described above is the number of data successfully grouped together, and percentage is the number of data that is successfully grouped in proportion to the total number of data. There are 12 data that cannot be found by extended grouping because no data passes to create a new set. For example, there is a complaint about com2 name', which shows complaints of disappointment of why, using the name of com2, but because only a few (in this case only 1) complain, then the data is not feasible, and a new group is not created as it will bring up an unprocessed tweet.

Based on the result of Table 5 and 6, we do another experiment by comparing the result of RAKE combined with Extended Grouping with result of WPKE with Keyword Grouping and Extended Grouping, and we obtained the following result as can be seen on Table 7 below, using another data set of 48 data from Twitter account of @com1 and @com2, both are telecommunication service companies in Indonesia.

Table 7. Experiment Result

	WPKE + Keyword Grouping	WPKE + Keyword Grouping + Extended Grouping	RAKE + Extended Grouping
Total	27	35	17
Percentage	56.25%	72.92%	35.42%

An example of a successful tweet processed by extended grouping is the tweet of "(@com2) saya mau tanya ini kenapa jaringan com2 di telepon saya di daerah kabupaten Kendal kok sinyalnya tidak ada (#komplain) plggn", in this keyword grouping, we obtained a keyword of "daerah kabupaten", meanwhile after extended grouping process, we obtained "jaringan". This is due to keyword that is grouped in first tweet can change the keyword to "jaringan" as in contains the word "jaringan" and the position of "jaringan" that has the highest score, so it is prioritized in doing searching.

We visualized the results of complaint analysis by using a chart where the horizontal bar describes the number of tweets that contains complaints of certain topics, as can be seen in the following Figure 4.

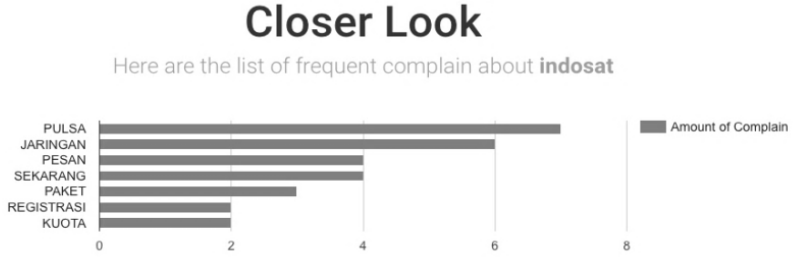


Figure 4. Bar Chart for Complaint Analysis towards @com2

4. CONCLUSION

Based on the experiments conducted, we conclude the following: (1) WPKE method as proposed in the research shows a significant increase in accuracy compared to RAKE Algorithm due to the fact that stop word list in Indonesian language is not well developed yet. (2) The result obtained using WPKE + Keyword Grouping + Extended Grouping has the accuracy of 72.92% which exceeds RAKE + Extended Grouping with only 35.42%. In the future, we plan to develop stop word list in Indonesian language as it can lead to a significant improvement for all-natural language processing in Indonesian language.

REFERENCES

[1] K. A. Al-Enezi, I. F. T. Al Shaikhli, and S. S. M. AlDabbagh, "The Influence of Internet and Social Media on Purchasing Decisions in Kuwait," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 10, no. 2, pp. 792–797, 2018.

[2] C.-L. Hsu, C.-C. Yu, and C.-C. Wu, "Exploring the continuance intention of social networking websites: an empirical research," *Inf. Syst. E-bus. Manag.*, vol. 12, no. 2, pp. 139–163, 2014.

[3] R. A. Setiawan and D. B. Setyohadi, "Analisis Komunikasi Sosial Media Twitter sebagai Saluran Layanan Pelanggan Provider Internet dan Seluler di Indonesia," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 3, no. 1, pp. 16–25, 2017.

[4] P. K. Kumar and S. Nandagopalan, "Insights to Problems, Research Trend and Progress in Techniques of Sentiment Analysis," *International Journal Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2818–2822, 2017.

[5] G. Ghedin, "From customer care to religion, the Twitter explosion in Indonesia | Digital in the round." [Online]. Available: http://www.digitalintheround.com/indonesia-twitter/. [Accessed: 03-Jul-2018].

[6] N. Hanafiah, A. Kevin, C. Sutanto, Y. Arifin, and J. Hartanto, "Text Normalization Algorithm on Twitter in Complaint Category," *Procedia Comput. Sci.*, vol. 116, pp. 20–26, 2017.

[7] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[8] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," *Text Min. Appl. Theory*, pp. 1–20, 2010.

[9] N. Naw and E. E. Hlaing, "Relevant words extraction method for recommendation system," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 2, no. 3, pp. 169–176, 2013.

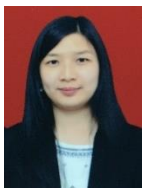
[10] M. Jungiewicz and M. Łopuszyński, "Unsupervised keyword extraction from Polish legal texts," in *International Conference on Natural Language Processing*, 2014, pp. 65–70.

[11] S. Siddiqi and A. Sharan, "Keyword and keyphrase extraction techniques: a literature review," *Int. J. Comput. Appl.*, vol. 109, no. 2, 2015.

[12] J. Greenberg, Y. Zhang, A. Ogletree, G. J. Tucker, and D. Foley, "Threshold Determination and Engaging Materials Scientists in Ontology Design," in *Research Conference on Metadata and Semantics Research*, 2015, pp. 39–50.

BIOGRAPHIES OF AUTHORS

Rini Wongso has completed her bachelor and master degree majoring Computer Science in Bina Nusantara University, Jakarta, Indonesia in 2014. She is a lecturer and researcher in Artificial Intelligence field in Bina Nusantara University, Jakarta, Indonesia. She previously worked as a Java Developer, developing Banking and HR System. She is interested in the field of Machine Learning, Computer Vision, Natural Language Processing, Artificial Intelligence Applications, and Software System.



Novita Hanafiah received the M.Sc degree in software system engineering from KMNUTNB, Thailand, in 2013. The research about entity recognition was conducted in RWTH Aachen in 2012. She is currently a lecturer and subject content coordinator in Bina Nusantara University. The main areas of research interest are artificial intelligence, natural language processing and software system.

Jaka Hartanto has completed his bachelor degree majoring in Computer Science, and his master degree majoring General Management in Bina Nusantara University, Jakarta, Indonesia in 2007. He is a a lecturer and researcher in Software Engineering field in Bina Nusantara University. He is also a founder of PT BIG, and System Analyst at JJ know it.