International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 4
ISSN: 2088-5334

Development of Rule-Based Feature Extraction in Multi-label Text Classification

Gugun Mediamer#, Adiwijaya#, Said Al Faraby#

#School of Computing, Telkom University, Bandung, 40257, Indonesia
E-mail: gugunmediamer@student.telkomuniversity.ac.id; adiwijaya@telkomuniversity.ac.id; saidalfaraby@telkomuniversity.ac.id

Abstract— **Hadith is the second main guidelines after the Holy Quran in the Islamic religion, which was revealed through the Messenger of Allah. Today, Hadith can classified by more than one class such as advice class, prohibited, and information to facilitate readers of Hadith in filtering the appropriate classes for each Hadith of Rasulullah SAW. In the course of research, there are many kinds of data involved in a text classification study. Therefore, special handling that fit with the characteristics of certain data is required. This study investigates the handling of multi-label data—Hadith Bukhari in Indonesian translation—focusing on feature extraction, feature weighted, and preprocessing methods. This study uses a rule-based feature extraction combined with several types of preprocessing along with three types of feature-weighted methods: TF-IDF, Word2vec, and Word2vec weighted with TF-IDF, the five preprocessing stages in this research: Case Folding, Tokenization, Remove Punctuation, Stopword Removal, and Stemming. From the 13 experiments conducted in this study consist of 2000 hadiths, it was found that the best performance for multi-label classification of Hadith data produced by the combination of the proposed rule-based feature extraction, Word2vec feature weighted method, and without using Stemming and Stopword Removal in the preprocessing phase. The Hamming Loss value obtained from this combination was 0.0623. The results show that our rule-based feature extraction method better than baseline method.**

Keywords— **multi-label classification; Bukhari Hadith; feature-weighted; tf-idf; word2vec; hamming loss.**

I. Introduction

Hadith and Al-Quran are the two main guidelines in the Islamic religion, which were revealed through the Messenger of Allah [1]. Hadith is the second pillar after the Holy Quran, and refers to everything that was said and done by Rasulullah SAW. Many narrators have diligently collected Hadith. One of the foremost narrators is Bukhari, who narrated thousands of Hadith of the Prophet Muhammad.

The interesting things of this research is that a Hadith can classified by more than one class such as advice class, prohibited, and information. This study aims to facilitate readers of Hadith in filtering the appropriate classes for each Hadith of Rasulullah SAW. Text classification on Hadith proposed in [2], but the task is not to classify Hadith into multi-label class. The studies of [3], [4], [5], and [6] explained how multi-label classification task was done using different data for each study.

The bag-of-word features used in text classification research produced a very large dimension depending on the number of vocabularies that could make the programming process overly complex. Therefore, this study focused on feature extraction processes combined with several preprocessing types and three different feature weighted methods. The feature extraction process was applied according to previously defined rules. For example, a Hadith can be classified as a prohibition Hadith if it contains word rules that have the tag 'NEG' accompanied by 'lah'. For example, the word 'janganlah', in English means 'do not'. The word is composed of words 'jangan' and 'lah'. From the training data, the word 'jangan' has a 'NEG' tag; hence, this word meets the previously defined rules. There are many things as a problem limitation in this study; we used three types of classes for multi-label data such as advice label, prohibited label, and information label. Then Bukhari data that we used are 2000 Hadith.

Word2vec is an approach using word vector representation proposed by Mikolov et al. [7]. Word2vec was used in this paper as a feature-weighting method along with the TF-IDF method. The weight of each word was obtained from vector representations on the Word2vec model that was built before. Furthermore, the TF-IDF weighting method was also used in this paper along with a combination of the two, as per Lilleberg, Zhu, and Zhang [8].

The next section of this paper discusses the previous researches related to multi-label classification, weighting methods, and the method. Then, the result and discussion are explained in Section III, and finally, the conclusion of this study is presented in Section IV.

II. Material and Method

A problem that is often found in the study of text classification is high-dimensional data, so a method to deal with this problem is needed. Many researcher have been doing research in the text classification field, such as [9], [5], [2], [6], and [10]. The feature extraction and feature selection is the most often used methods in the study of text classification for reducing the number of high-dimensional data [11]. There are many types of feature selections that can be used to handle high dimensions, such as rule-based feature selection to improve efficiency in built systems without reducing the value of accuracy that should be obtained [12].

Component analysis based on the feature selection method was implemented in prior research to eliminate duplicate features and irrelevant ones [13]. The study selected a subset of feature using a Genetic algorithm, and used the results for the classification process. Previous research also utilized several types of feature selection methods, including the Symmetrical Uncertainty (SU) method, Information Gain (IG), and Correlation Coefficient (CC) [3]. SU and CC are the methods resulting from the normalization of the IG method. The combinations of tf-idf, SU, Calibrate Label Ranking, and SVM produce the best result of F-measure.

In another work [14], the Part of Speech Tagger (PoSTag) method was combined in a sentiment analysis, and used as a feature selection method to select the best set of features from the dataset. Similar to the above work, the PoSTag method was found to be the best method for selecting features that could stand alone, because other features, namely phrase features, would only be able to increase accuracy when combined with the unigram method [15].

Problem transformation is one approach for solving multi-label classification. One of the most popular problem transformation approaches is Binary Relevance (BR), which aims to transform multi-label classification into a single-label classification, after which the data is generally classified using single-label algorithms. Label Powerset (LP) and Ranking by Pairwise Comparison (RPC) are also some examples of problem transformation approaches [11].

The vector representation method using Word2vec has been shown to yield good results, as per previous researches [8], [4]. Both studies used vector representation as the feature weighting method. Unlike another study [16], Word2vec was actually used for the feature selection process by clustering features based on similarities. Besides Word2vec, other researches [3], [8], [4] also used the TF-IDF method as the feature weighting process. The study of Lilleberg et al. [8] and Rahmawati and Khodra [4] also used both feature weighting methods by combining Word2vec vector representation and the TF-IDF vector. Lilleberg et al. [8] produced the best results using this feature weighting combination but Rahmawati and Khodra [4] did not.

In the research [17], expalained a study related to rule-based feature recognition using if-then rules which is part of the Logic Rules. For each rule produce a uniqueness of form feature definition. So, there are no two rules define the same form feature. The disadvantage is when the form feature has been extracted, it does not match the expected pattern. There are many Logic Rules that can be used to pattern recognition, the difference between them is a model representation for identifying form features: (1) syntactic pattern recognition; (2) state transition diagrams and automata; (3) logic (if–then) rules and expert systems; (4) graph-based approach; (5) convex hull volumetric decomposition; (6) cell based volumetric decomposition; (7) hint-based approach; and (8) hybrid approach.

SVM proven to yield good results for the multi-label classification process, where the CLR approach was used to as the problem transformation approach [3] [4]. Rahmawati and Khodra [4] used SVM and CLR because, as shown in their previous study [3], the combination of these classifications were able to produce the best value. Lilleberg et al. [8] also used the SVM Linear method for their classification process. Hamming Loss is a method often used to evaluate classification results that have been obtained before. In Fu et al. [18], Hamming Loss was used as the method to measure the extent of the misclassification of all pairs of data and labels. In this case, a smaller Hamming Loss value indicated a better model built from the data used.

Many feature extraction methods used in the case of text classification. In this paper, we propose a pre-defined rule-based method in the hopes that the rules made will provide better accuracy. In addition, our method also attempts to weigh the features using TF-IDF and vector representations using the Word2vec method, as per previous research [8], [4]. Figure 1 explains the method used to build the system in this study. We implemented two different types of corpus for this research. First corpus was used to build a vector representation model using the Word2vec method and the second corpus was used to build the PoSTag model. The built model was then used in the feature weighting process and feature extraction process. After that, we performed a rule-based feature extraction process, as described in Tables I, II, and III. The second process produced a collection of features for use in the next process, namely the weighting of features. Then, the following process involved the k-fold cross validation, used to divide the data into training data and testing data, where the number of k used was 10. After that, the classification process was performed using the SVM method. The last stage is the evaluation process, where the evaluation method used for this multi-label case was Hamming Loss. The rules for feature extraction were used to define the advice labels presented in Table I: The rules for feature extraction were used to define the prohibited labels outlined in Table II: The rules for feature extraction were used to define the information labels outlined in Table III: The inside () is a tag of words defined in the PoSTag corpus that the previous learning process had applied.

The rules in Table 1, 2, and 3 are obtained by observing the sentence pattern from a word and word sequences of several many Hadiths to represent all Hadith. Then the identification of the word whether the word as a characteristic of Advice Hadith, Prohibited Hadith, or Information Hadith, refer to the label of the Hadith. To identify word that matches the sentence pattern, we use word class obtained from the PoSTag model. The disadvantage of this approach, if the identification of the word class is incorrect, the output word will not match the label characteristics. Figure 1 is presented below.

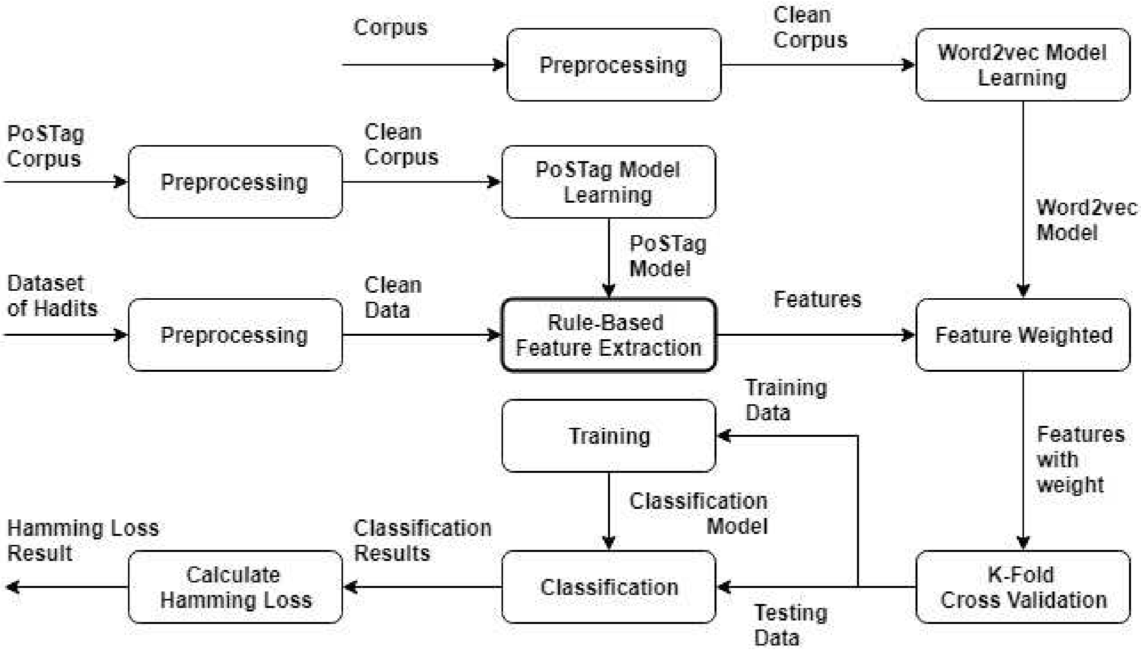

Fig. 1 System Overview

TABLE I
CHARACTERISTICS OF ADVICE LABELS

Rules	Word Examples
Words with the word tag (VB) + lah	Makanlah
Words with the word tag (RB) + lah	Segeralah
Words with the word tag (JJ) + lah	Beramallah
Words that contain ‘anjur’ and has the word tag (VB)	Menganjurkan
Words that contain ‘wajib’ and has the word tag (VB)	Mewajibkan
Word ‘kewajiban’ and has the word tag (NN)	Kewajiban
Word ‘Hendaknya’ or ‘Hendak’	Hendaknya or Hendak
Word ‘Mentaati’ or ‘Mentaatiku’	Mentaati or Mentaatiku
Word ‘Sebaik-baiknya’	Sebaik-baiknya

TABLE II
CHARACTERISTICS OF PROHIBITED LABELS

Rules	Word Examples
Words with the word tag (NEG) + lah	Janganlah
Words that contain ‘larang’	Melarang
Words that contain ‘haram’	Diharamkan
Word squence ‘tidak pernah’ and after that words with the word tag (VB)	tidak pernah dihalalkan
Word squence ‘tidak menyekutukannya’	Tidak menyekutukannya
Word ‘Jangan’	Jangan
Word ‘Membangkang’	Membangkang
Word ‘Mendurhakaiku’	Mendurhakaiku

TABLE III
CHARACTERISTICS OF INFORMATION LABELS

Rules	Word Examples
Words that begin with ‘me-’ and end with ‘-kan’ and that word has the word tag (VB)	Mempekerjakan
Word ‘Niscaya’	Niscaya
Word ‘Sesungguhnya	Sesungguhnya
Word ‘Bahwa’	Bahwa
Word ‘Apabila’	Apabila
Word ‘Maka’	Maka
Word ‘Barangsiapa’	Barangsiapa

This research classifies different feature weighting methods three scenarios. The first feature-weighting method was TF-IDF. The Term Frequency (TF) value was obtained by the number of occurrences of the term (*t*) in the i-th document (d_i) [19]. Meanwhile, the Inverse Document Frequency (IDF) value was obtained using the equation outlined in Schütze et al. [19].

$$idf_t = log(\frac{N}{df_t}) \quad (1)$$

Where: *N* is the number of documents (*d*) and df_t is the number of documents (*d*), which contains the term (*t*).

The next classification feature used was the extraction results using the Average of Word2vec method. This method has been used in [8] and [4] research, and shown in [4] that Average of Word2vec was increased the accuracy result. The feature was obtained by summing all vector representation of the token contained in the document. Then, the number of the vector divided the sum. Therefore, the vector produced by the process was average vector used as a feature vector. In this case, we used 100 and 300 vector lengths. Suppose a vector length is 100, the feature vector will also have 100 dimensions. Fig 2 explains the detail of feature vector process [8].

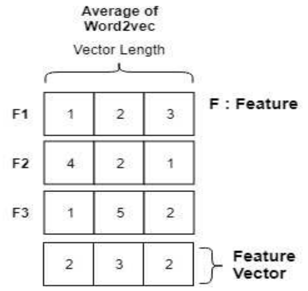

Fig. 2 Detail of Feature Vector Process

The last feature weighting method was obtained by combining the results vector of TF-IDF and Word2vec, as per prior studies [8] [4]. Equations (2) and (3) referred from Lilleberg et al. [8] were used.

$$w\_R(d_i) = \Sigma_t w_t w2v(t), \quad where\ w_t = \text{tf-idf weight of } t \tag{2}$$

$$C(d_i) = \text{concatenate}(\text{tf-idf}(d_i),\ w\_R(d_i)) \tag{3}$$

Where: d is the document obtained from our data, and t is the feature contained in d_i; w2v(t) shows the vector representation of t, and w_t is the weight of the feature obtained from TF-IDF for each document. The first step, for each w_t was multiplied by w2v(t). In the third step, the result from the second step was combined with the vector obtained from TF-IDF. The process of combining vectors, for example, the length of the vector TF-IDF (d_i) 2000 (number of features) was combined with the length of the vector in Word2vec, which was 200, which gave a vector length d_i of 2200 [8]. The concatenation between word2vec and tf-idf illustrated in the Figure 3.

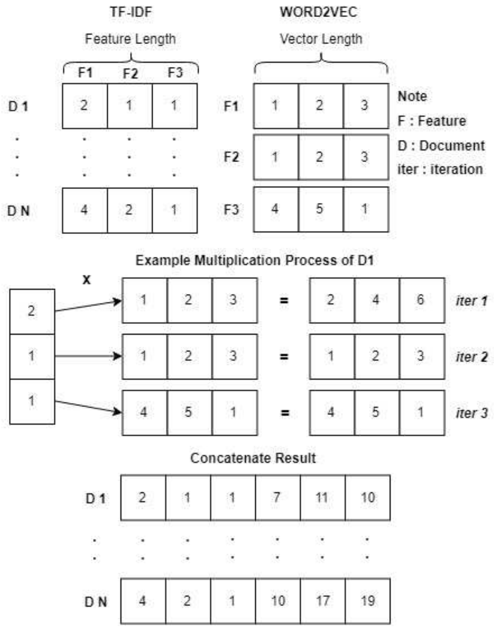

Fig. 3 Concatenation of Word2vec on TF-IDF Illustration

As explained before, we used hamming loss metrics to know how good the model is built. This metrics has been used in [5] and [6]. The equation shows as follow.

$$Hamming\ Loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}\left[\hat{Y}_j^{(i)} \neq Y_j^{(i)}\right] \tag{4}$$

The equation (4) how to calculate the hamming loss. Where N is the number of document that will be classified, L is the number of label type in dataset; the j-th output label belongs to i-th notated by $\hat{Y}_j^{(i)}$ and $Y_j^{(i)}$ is the j-th actual label belongs to i-th that obtained from the dataset. This reasearh only use the hamming loss metrics because it metrics was popular used in multi-label classification reseach, among them are [5] and [6].

III. Result and Discussion

The data used in this study is Bukhari Hadith data in Indonesian translation, which consists of 2000 Hadiths. The corpus to build the Word2vec model consisted of data from the Indonesian Wikipedia that combined 7008 Bukhari Indonesian Hadith data without including labels. On the other hand, the corpus to build the PoSTag model is open data taken from research [20]. This study focused on three factors: preprocessing, feature extraction, and feature weighting. This study combined each choice contained in these three factors. These choices are explained in Table IV below.

TABLE IV
Research Focus

Factor	Alternatives
Feature Weighted	TF-IDF, Word2vec, Word2vec weighted by TF-IDF
Feature Extraction	Proposed Rule-based Feature Extraction, All Vocabulary
Preprocessing	Stemming, Stopword Removal, No Stemming and Stopword Removal

Five preprocessing stages were done in this study. The first three stages include case folding, tokenization, and removal of punctuation, while the other two stages involve the factors focused on in this study, as mentioned in Table IV. The five preprocessing stages were completed using the help of a NLTK library for the first three stages, and the Sastrawi library for the other two. The following process was the classification method, where a multi-label binary relevance approach was used to change the multi-label classification into a binary classification, because the simple SVM method could not classify multi-label data directly. We performed binary classification using SVM with the help of the Scikit-learn library. Then, to build the Word2vec model, we used the help of the Gensim library. Finally, we used the python-crfsuite library to build the PoSTag model, which we used for extracting features. Other libraries that we used include Pandas and Numpy.

Table V describes the results of the experiments that were conducted in this study. It can be seen in the first six rows of Table V, bag-of-word was used as features. The first three rows used the Stemming process for preprocessing and produced a Hamming Loss value higher than the three rows below them, which only used Stopword removal. This is probably because the Stemming process can change the meaning of a sentence because it cuts off any affixes contained in a word. For example, the sentence "Hendaklah kalian mulai yang sebelah kanan anggota wudhunya" means, "You must start ablution with the right side of your body", in English. If the sentence went through a Stemming process, the words that have affixes such as the word "hendaklah" and "sebelah" will be removed, so the sentence will become "hendak kalian mulai yang belah kanan anggota wudhunya", which, in English, means "You want to start with the right side of your body for ablution". From this example, the words "hendaklah" and "hendak" and the words "sebelah" and "belah" clearly have very different meanings. From the experiment, it is known that the impact of preprocessing on hamming loss is very big, meaning that we have to choose the type of preprocessing carefully. Therefore, in the next experiment, the Stemming process was not used in the preprocessing stages.

TABLE V
EXPERIMENT RESULTS

No	Word2vec algorithm	Vector length	Preprocessing		Feature Extraction Description	Feature Weighted Description	Hamming Loss
			Stemming	Stop word Removal			
1	-	-	Yes	Yes	Bag-of-word	Tf-idf	0.0902
2	Skipgram	100	Yes	Yes	Bag-of-word	Average of Word2vec	0.0995
3	Skipgram	100	Yes	Yes	Bag-of-word	Word2vec weighted by TF-IDF	0.0997
4	-	-	No	Yes	Bag-of-word	Tf-idf	0.0762
5	Skipgram	100	No	Yes	Bag-of-word	Average of Word2vec	0.0983
6	Skipgram	100	No	Yes	Bag-of-word	Word2vec weighted by TF-IDF	0.0857
7	-	-	No	No	Rule-based feature extraction	Tf-idf	0.0647
8	Skipgram	100	No	No	Rule-based feature extraction	Average of Word2vec	0.0709
9	Skipgram	100	No	No	Rule-based feature extraction	Word2vec weighted by TF-IDF	0.0677
10	Skipgram	300	No	No	Rule-based feature extraction	Average of Word2vec	**0.0623**
11	Skipgram	300	No	No	Rule-based feature extraction	Word2vec weighted by TF-IDF	0.0710
12	CBOW	100	No	No	Rule-based feature extraction	Average of Word2vec	0.0743
13	CBOW	100	No	No	Rule-based feature extraction	Word2vec weighted by TF-IDF	0.0747

In the next experiment, we used the rule-based feature extraction proposed in this paper. The stages of preprocessing did not involve Stemming and Stopword removal. The Stopword removal process will not impact the results if used, because the rule-based feature extraction that we proposed will not extract Stopwords as features automatically. By using the feature extraction that we proposed, the results we obtained were better than the results of bag-of-word features.

The next factor that we considered was the length of the vector model in Word2vec. As we explained in Section III, we used 100 and 300 vector lengths. Then, we changed the vector length in the 10th experiment to 300 from the previous vector length of 100. From the experiment it was shown that the length 300 of the vector in word2vec is better than 100. Combination of the experiment produced the best result of all experiments.

Besides considering vector length, the learning method when building the Word2vec model was also one of the things we considered. Out next experiment used the Word2vec model with the Continuous Bag of Words (cbow) learning method. In the 12th row of Table V, it can be observed that the Hamming Loss value with Word2vec using the cbow model is not better than Skipgram.

Another factor that we observed was the weighting method. Of the three methods used, shows that the best result is obtained by using the features of Average of Word2vec with vector length of 300. This occurs because vector representation of Word2vec represents the similarity of the features. The last experiment we performed was to combine Word2vec vectors with TF-IDF. The obtained results were not change significantly, and did not show the best result.

IV. CONCLUSION

In this paper, we proposed rule-based features to obtain better research results for multi-label classification of Bukhari Hadith data in Indonesian Language translation. From the experiments conducted, we can conclude that the proposed rule-based feature extraction method was significantly better than the baseline method.

Our research also shows that not using Stemming in the preprocessing stage especially for the data we used would yield better results, because Stemming can change the meaning of a sentence. The Stopword removal process did not provide any significance in this paper, so this process is also not recommended. Afterwards, the best weighting method in this paper was found to be Average of Word2vec with vector length of 300, because it always produced the best result of all the results.

In this paper, the rule-based feature extraction proposed for Hadith data in Indonesian Language translation still needs further development. This is because the rule cannot cover all words or word sequences that are a characteristic of the three labels contained in the dataset, yet. We believe that these rules can still be developed to improve the accuracy of multi-label classifications of Bukhari Hadith data in Indonesian Language translation. In addition, for further research, we suggest to compare this proposed method with other special methods for Hadith.

References

[1] M. N. Al-Kabi, H. A. Wahsheh, I. M. Alsmadi, and A. Moh'd Ali Al-Akhras, "Extended Topical Classification of Hadith Arabic Text," Int. J. Islam. Appl. Comput. Sci. Technol., vol. 3, no. 3, pp. 13–23, 2015.

[2] S. Al Faraby, E. R. R. Jasin, A. Kusumaningrum, and others, "Classification of hadith into positive suggestion, negative suggestion, and information," in Journal of Physics: Conference Series, 2018, vol. 971, no. 1, p. 12046.

[3] D. Rahmawati and M. L. Khodra, "Automatic multi-label classification for Indonesian news articles," in Advanced Informatics: Concepts, Theory and Applications (ICAICTA), 2015 2nd International Conference on, 2015, pp. 1–6.

[4] D. Rahmawati and M. L. Khodra, "Word2vec semantic representation in multi-label classification for Indonesian news article," in Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On, 2016, pp. 1–6.

[5] R. A. Pane, M. S. Mubarok, N. S. Huda, and others, "A Multi-Lable Classification on Topics of Quranic Verses in English Translation Using Multinomial Naive Bayes," in 2018 6th International Conference on Information and Communication Technology (ICoICT), 2018, pp. 481–484.

[6] A. M. K. Izzaty, M. S. Mubarok, N. S. Huda, and Adiwijaya, "A Multi-label Classification on Topics of Quranic Verses in English Translation Using Tree Augmented Na□ve Bayes," in 2018 6th International Conference on Information and Communication Technology (ICoICT), 2018.

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv Prepr. arXiv1301.3781, 2013.

[8] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, 2015, pp. 136–140.

[9] A. I. Pratiwi and others, "On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis," Appl. Comput. Intell. Soft Comput., vol. 2018, 2018.

[10] M. S. Mubarok, Adiwijaya, and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Na{\"\i}ve Bayes," in AIP Conference Proceedings, 2017, vol. 1867, no. 1, p. 20060.

[11] M. S. Sorower, "A literature survey on algorithms for multi-label learning," Oregon State Univ. Corvallis, vol. 18, 2010.

[12] Z. Hao and B. Liu, "A rule based feature selection approach for target classification in wireless sensor networks with sensitive data applications," Int. J. Distrib. Sens. Networks, vol. 10, no. 4, p. 429651, 2014.

[13] M.-L. Zhang, J. M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," Inf. Sci. (Ny)., vol. 179, no. 19, pp. 3218–3229, 2009.

[14] N. D. Patel and C. Chand, "Selecting Best Features Using Combined Approach in POS Tagging for Sentiment Analysis." IJCSMC, 2014.

[15] B. M. Badr and S. S. Fatima, "Using skipgrams, bigrams, and part of speech features for sentiment classification of twitter messages," in Proceedings of the 12th International Conference on Natural Language Processing, 2015, pp. 268–275.

[16] Z. Su, H. Xu, D. Zhang, and Y. Xu, "Chinese sentiment classification using a neural network tool &amp;#x2014; Word2vec," in 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014, pp. 1–6.

[17] B. Babic, N. Nesic, and Z. Miljkovic, "A review of automated feature recognition with rule-based pattern recognition," Comput. Ind., vol. 59, pp. 321–337, 2008.

[18] D. Fu, B. Zhou, and J. Hu, "Improving SVM based multi-label classification by using label relationship," in Neural Networks (IJCNN), 2015 International Joint Conference on, 2015, pp. 1–6.

[19] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval," vol. 39, 2008.

[20] A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, "Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus," in Asian Language Processing (IALP), 2014 International Conference on, 2014, pp. 66–69.