International Journal of Engineering & Technology, 7 (4.38) (2018) 1499-1501

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Sentiment Analysis of Indonesian Movie Review using K-Nearest Neighbors and Information Gain

Ria Ine Pristiyanti, M. Ali Fauzi, Lailil Muflikhah

Faculty of Computer Science, Brawijaya University
**Corresponding author: moch.ali.fauzi@ub.ac.id*

Abstract

Movie review is a necessity for movie lover to get information about people opinion on the movie to watch. However, movie lover cannot read all of the movie review manually. It will be costly and time consuming. Therefore, automatic way to analyze them is needed. In this study, we use bag of word (BOW) model and utilize IG to select the best features before KNN is employed to classify the review into positive or negative. The result for using all of term for classification is better than the feature selection due to the elimination of term having low information gain value with 92% accuracy.

Keywords: *Movie Review, Sentiment Analysis, Information Gain, K-Nearest Neighbors*

1. Introduction

Movie review is a necessity for movie lover to get information about people opinion on the movie to watch. This information can be used as a consideration in determining the quality of a movie so that the movie lovers can decide whether the movie is worth watching or not. However, movie lover cannot read all of the movie review manually. It will be costly and time consuming. Therefore, automatic way to analyze them is needed. Analyzing people opinion in the movie review is called sentiment analysis. Sentiment analysis is the process of applying natural language processing (NLP) and text analysis to identify and extract subjective information from a text (Hussein, 2016; Rofiqoh et al., 2017; Antinasari et al., 2017; Gunawan et al., 2017; Fauzi et al., 2018). Sentiment analysis can be applied using a classification method to facilitate in grouping data in the form of positive sentiment or negative sentiment. One of the popular method in classification is k-nearest neighbor (KNN) (Nurjanah et al., 2017; Mentari et al., 2018; Claudy et al., 2018). KNN is used in this study because it's simplicity in which the process is based on a simple weighting approach. Nevertheless, it still has high accuracy value.

Sentiment analysis, just like text classification in general, also has a high feature dimension. High feature dimension in KNN can lead into high computational complexity as well. Hence, we need to reduce the feature so as to avoid the high dimensionality problem during the classification process (Khan, et al. 2016). The process of reducing features is usually called as feature selection. One of the best performing method in feature selection is Information Gain (IG).

IG is a method that based on entropy to select the most important features (Fauzi., Et al, 2017). In a previous study conducted on the reuters dataset, feature selection using IG succeeded in generating a f-measure value of 0.86 (Yang and Pederson, 1997). In this study, we used information gain as feature selection method and KNN as the classification method for Indonesian movie review sentiment analysis. We use bag of word (BOW) model and utilize IG to select the best features before KNN is employed to classify the review into positive or negative.

2. Problem Statement

In this study, we want to analyze the effect of feature selection method on movie review sentiment analysis accuracy. We will also wanto to know which the amount of feature to be used for sentiment analysis to get the best performance.

3. The Aim of Research

This paper conducted research to analyze the effect of feature selection method on movie review sentiment analysis accuracy. This paper is also conducted to get the information of the amount of feature to be used for sentiment analysis to get the best performance.

4. Method of Research

There are three process in this study, as seen in Figure 1, i.e. preprocessing, feature selection using IG and classification using k-nearest neighbor method. The fist process, preprocessing, we conducted some steps including tokenization, stoword removal, stemming and term weighting. In the tokenization step, all of words were converted into lowercase and some characters like punctuation and numbers were removed (Fauzi, et al., 2013; Fannisa et al., 2018). The next step is stopwords removal or removing uninformative words based on the existing dictionary. In this case, we use stoplist by Tala (2003). The last step step in preprocessing is stemming or a process of reducing every words to its root by removing affixes such as prefix, infix and suffix.

The second process is feature selection. Feature selection is the process of selecting the most relevant features. One of the most popular feature selection in text classification is information gain. IG will calculate the importance of terms w to each existed document d and defined as follow (Uguz, 2011):

$$IG(w_i;d) = P(w_i)\sum_k P(d_k|w_i)\log P(d_k|w_i) + P(\overline{w_i})\sum_k P(d_k|\overline{w_i})\log P(d|\overline{w_i}) \quad (1$$

where $p(w_i)$ is the probability that word w_i appear, $\overline{w_i}$ means that word w_i does not occur, $p(d_k)$ is the probability of the kth document value, $p(d_k|w_{i,j})$ is the conditional probability of the document k value given that w_i appear, $p(w_{i,j})$ is the probability that w_i and w_j appear together, and $\overline{w}_{i,j}$ means that w_i and w_j do not appear together but w_i or w_j can appear.

The last processs is classification using k-nearest neighbors (KNN). Classification is the process of dividing data into groups which have dependent and independent characteristics in which each group acts as a class. Classifying documents is the process of automatically grouping a document into classes that have been previously known based on the contents of the document. In this sentiment analysis study, there are two classes, i.e. positive and negative.

K-nearest neighbor algorithm is one of the methods to classify an object based on training data that has the closest distance to it (Suharsono et al., 2017). Usually the distance used in KNN is Euclidean distance (Nilson, 1996). However, we can use another similarity or dissimilarity measure like jaccard, cosine similarity, manhattan distance, and etc. In this study, we use TF.IDF as term weighting and cosine similarity as similarity measure.

TF.IDF is the most popular term weighting method in text classification (Fauzi et al. 2017; Fauzi et al. 2018). TF.IDF is a multiplication of term frequency (TF) and inverse document frequency (IDF). TF is a simple local term weighting method that give each term a weight based on to the number of its occurrences in the document. Meanwhile, IDF is a global term weighting method that gives high weight for rare terms, terms that occurs in only a few document, and small weight for frequent term that only in many documents. The TF.IDF weight of term t in document d is counted as follows:

$$TF \cdot IDF(t,d) = (1+\log(f_{t,d})) \cdot (1+\log\left(\frac{N_d}{df_t}\right))$$

where $f_{t,d}$ is the number term t that occur in document d and N_d is the number of documents in corpus and df_t is the number of documents in corpus that contains term t.

For the similarity measure in KNN, we use cosine similarity instead of Euclidian distance because of it has better performance in text mining. Cosine similarity calculates the cosine value of the angel between two documents. Since it is based on the cosine of the angle between two vectors, the value ranges from 0 to 1. The greater the cosine value, the more the similarity between the two documents (Pramukantoro, 2016).

The Cosine similarity of document q and document dj can be formulated as follows:

$$\cos(q,d_j) = \frac{\sum_{t_k}[Weight(t_k,q)] \bullet [Weight(t_k,d_j)]}{\sqrt{\sum |Weight(q)|^2} \bullet \sqrt{\sum |Weight(d_j)|^2}},$$

where $\cos(q,d_j)$ is the cosine value between docuemnt q and document d_j, $Weight(t_k,q)$, $Weight(t_k,d_j)$ are weighted words t_k tk on query q and document d_j respectively. Meanwhile $\sqrt{\sum |Weight(q)|^2}$ and $\sqrt{\sum |Weight(d_j)|^2}$ is the length of the document vector q and document vector d_j respectively. In this study, the weight used is TF.IDF like the explanation before.

Finally, the classification results will be evaluated by using accuracy, precision, recall and f-measure. We will evaluate the use of information gain as feature selection in movie review sentiment analysis.

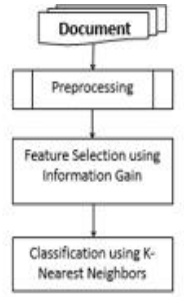

figure 1. System Main Flowchart

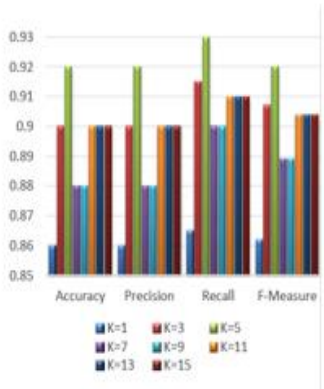

Figure 2. Experiment result for k value variatio

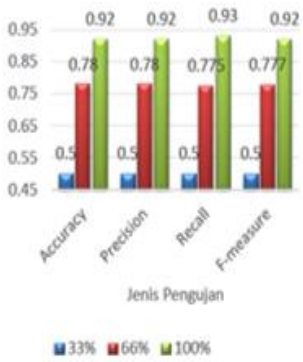

Figure 3. Experiment result for number of features used variation

5. Result and Analysis

In this study, we conducted two experiments. The first one is an experiment to see the effect of k value on the accuracy of sentiment analysis using KNN and to find the most optimal k value. The second experiment is to see the effect of the number of features used in sentiments analysis accuracy using KNN and information gain based feature selection feature and to find the number of features used that provide the best accuracy.

The first experiment is done by using different k values for KNN including k=1, k=3, k=5, k=7, k=9, k=11, k=13, and k=15. The experiment result can be seen in Figure 2.

As seen in Figure 2, the worst performance is obtained at k = 1 with accuracy, precision, recall, and f-measure only 0.86, 0.86, 0.86, and 0.86 respectively. The greater the value of k, as in k = 3 and k = 5, the classification performance increases. However, after the value of k grows larger like in k=7 until k=15, the classification performance decreases. The most optimal k is when k = 5 because it has the best performance by accuracy, precision,

recall and f-measure value is 0.92, 0.92 , 0.93 and 0.92 respectively.

The result of second experiment, which is the experiment of the number of features used using information gain based feature selection, is displayed in Figure 2. In this experiment, we use 33%, 66%, and 100% of total number of feature for classification.

By using 33% of total term, the experiment result show a low accuracy value by only 0.5. When the term used is 66%, the performance got better with 0.78 accuracy. The best accuracy is gained when using 100% or all of the terms with accuracy value 0.92. This is influenced by several factors such as the term that should be used for the classification has a low information gain value so that the term is removed and not used during the classification process.

In this study 33% term with the value of the highest information gain value is a term that appears once in the training data while the lowest term is the term that appears in almost all training data. In addition, the limit on the number of term takings influences the term to be taken, where the term having the same value of the gain information is deleted when the limit has been determined.

6. Conclusion

In this study, we use bag of word (BOW) model and utilize IG to select the best features before KNN is employed to classify the review into positive or negative. The result for using all of term for classification is better than the feature selection due to the elimination of term having low information gain value with 92% accuracy.

In the k value experiment, k = 5 is the most optimal for the classification process using k-nearest neighbor method which yields an accuracy value of 92%. Meanwhile, in the second experiment, the result show that the number of features used for the classification process is directly proportional to the result of accuracy, where the fewer the number of terms used the smaller the result of accuracy otherwise the greater the number of terms used then the greater the value of accuracy. While on testing many number of terms used based on threshold value of information gain use 66% of term number has lower accuracy value when compared with 66% usage without regard to value of same information gain. Testing the use of term results of information gains affect the results of classification using k-nearest neighbor. The result of the combination between the use of information gain method with k-nearest neighbor method produces a low accuracy compared with k-nearest neighbor method.

References

[1] Antinasari P, Perdana RS, Fauzi MA. Analisis Sentimen Tentang Opini Film Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naive Bayes Dengan Perbaikan Kata Tidak Baku. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2017; 1(12):1733-41.

[2] Claudy YI, Perdana RS, Fauzi MA. Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2018; 2(8):2761-65.

[3] Fanissa S, Fauzi MA, Adinugroho S. Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.2018; 2(8):2766-70.

[4] Fauzi MA, Arifin AZ, Gosaria SC. Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model. Indonesian Journal of Electrical Engineering and Computer Science. 2017 Dec 1;8(3).

[5] Fauzi MA, Arifin A, Yuniarti A. Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi. 2013;5(2).

[6] Fauzi MA, Arifin AZ, Yuniarti A. Arabic Book Retrieval using Class and Book Index Based Term Weighting. International Journal of Electrical and Computer Engineering (IJECE). 2017 Dec 1;7(6):3705-10.

[7] Fauzi MA, Afirianto T. Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2018 Jun 1;16(3).

[8] Fauzi MA, Yuniarti A. Ensemble Method for Indonesian Twitter Hate Speech Detection. Indonesian Journal of Electrical Engineering and Computer Science. 2018 Jul 1;11(1).

[9] Gunawan F, Fauzi MA, Adikara PP. Analisis Sentimen Pada Ulasan Aplikasi Mobile Menggunakan Naive Bayes Dan Normalisasi Kata Berbasis Levenshtein Distance (Studi Kasus Aplikasi BCA Mobile). Systemic: Information System and Informatics Journal. 2017 Des 31; 3(2):1-6.

[10] Hussein DM. A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences. 2016 Apr 26.

[11] Khan, M.T., Durrani, M., Ali, A., Inayat, I., Khalid, S., Khan, H.,. Sentiment analysis and the complex natural language. 2016. Pakista: Complex adaptive system modeling.

[12] Mentari ND, Fauzi MA, Muflikhah L. Analisis Sentimen Kurikulum 2013 Pada Sosial Media Twitter Menggunakan Metode K-Nearest Neighbor dan Feature Selection Query Expansion Ranking. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2018; 2 (8):2739-43.

[13] Nilsson, NJ. Introduction To Machine Learning. 1996. Stanford Univercity.

[14] Nurjanah WE, Perdana RS, Fauzi MA. Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2017; 1 (12), 1750-57.

[15] Pramukantoro ES, Fauzi MA. Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification. InAdvanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on 2016 Oct 15 (pp. 149-155). IEEE.

[16] Rofiqoh U, Perdana RS, Fauzi MA. Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2017; 1(12):1725-32.

[17] Suharno CF, Fauzi MA, Perdana RS. Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors dan Chi-Square. Systemic: Information System and Informatics Journal. 2017 Dec 7;3(1):25-32.

[18] Tala FZ. A study of stemming effects on information retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands. 2003 Jul.

[19] Uğuz H. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems. 2011 Oct 31;24(7):1024-32.

[20] Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. InIcml 1997 Jul 8 (Vol. 97, pp. 412-420*)*.