OPEN ACCESS
J DATA SCI APPL, VOL. 3, NO. 1, PP. 001-007, JANUARY 2020

E-ISSN 2614-7408
DOI: 10.34818/JDSA.2020.3.22

JOURNAL OF DATA SCIENCE AND ITS APPLICATIONS

Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor

Novelty Octaviani Faomasi Daeli, Adiwijaya*
School of Computing, Telkom University
Bandung, Indonesia

*adiwijaya@telkomuniversity.ac.id

Received on 24-05-2019, revised on 24-10-2019, accepted on 11-05-2020

Abstract

Huge resources need effectiveness and efficiency, it can be processed by machine learning. There have been many studies conducted using machine learning method and produced quite good performance in sentiment analysis. This is because machine learning helps determine the probability of sentiment analysis of data quite well, for instance Naive Bayes (NB), K-nearest neighbor (KNN), Support vector machine (SVM), and Random forest methods. Mostly, KNN did not achieve better performance than other machine learning methods in sentiment analysis. In this study, the dataset Polarity v2.0 from Cornell movie review dataset will be used to test KNN with Information gain features selection in order to achieve good performance. The purpose of this research are to find the optimum K for KNN and compare KNN with other methods. KNN with the help of Information gain feature selection becomes the best performance method with 96.8% accuracy compared to the NB, SVM, and Random forest while the optimum K is 3.

Keywords: sentiment analysis, information gain, k-nearest neighbors

I. INTRODUCTION

Movie is a visual art that continues to grow and multiply from year to year. Through movie review, viewers can find out which films have a good quality. The higher number of films produced will make many reviews being produced. It will need much effort for viewers to read a lot of movie reviews, so they can get an information about the movie. Based on this condition, sentiment analysis in movie reviews is an interesting topic to be solved by machine learning. Machine learning can help in terms of effectiveness and efficiency, because it will automatically classify and shorten the processing time [1].

Many research about sentiment have been done with machine learning, for instance sentiment analysis of cyberbullying on instagram user comments by Naive Bayes [23], analysis sentiment for product review by Naive Bayes [22] and many more. In this research, machine learning method that will be used is K-nearest neighbor (KNN). KNN is a simple method in machine learning, but this method have a bad performance with noise features [2]. Even though KNN have a bad performance, some research showed that KNN can intensify the ability of speech recognition to help people learn reciting Al-Quran in the right way despite KNN still need to be observe to determine attributes and that atributes will be used to determine the distance [21]. The performance of KNN can be better by using a good feature selection. Information gain is one of the best feature selection [3] because it can reduce noise features better than other feature selections [5]. Also another research have showed that information gain can work well in various kinds of data including multi-label data [20]. So, information gain can helps KNN to avoid bad performance with good features. Therefore, the combination of KNN and information gain can help the viewers to get the information about movie. In this paper, KNN is used to classify movie reviews into positive or negative review.

This research use polarity v.2.0 from Cornell review dataset [12] with total 1000 documents of negative reviews and 1000 documents of positive reviews in English language. The structure of this paper is as follows in section 2 is related work, section 3 is methodology, section 4 is system design, and section 5 is evaluation.

II. RELATED WORK

Machine learning helps in organizing information better [8]. Therefore, machine learning methods are widely used to solved sentiment analysis problems. Sentiment analysis research using machine learning by Pang et al. proved that machine learning achieves better performance than humans [4]. Because of this, more research is done by using machine learning in sentiment analysis.

Support vector machine (SVM) got the best performance compared to Naive bayes (NB), K-nearest neighbor (KNN), and Decision tree at 81.75% in 1000 positive datasets and 1000 negative datasets for detecting fake reviews [16]. But in this study, SVM required the longest computing time compared to other methods. On the other hand, the sentiment analysis of Nepali language using Naive bayes (NB) achieved a better performance at 60.6% with TF-IDF feature extraction compared to SVM and Logistic regression methods with the same feature extraction in the dataset of 384 reviews [17]. However, in this study NB performance decreases when the size of corpus also decreases. So, in order for NB to achieve good performance, a large number of training data are needed. Random forest (RF) achieves an average performance of 83.16% above SVM, NB, and KNN [18]. However, the weakness of RF method is easy to over fits during training [19].

Based on the approaches previous researches on sentiment analysis using machine learning, the machine learning have a good performances but it take much time to process the model in order to fits the data train [9]. To shorten the time they used, a combination between machine learning and feature selection methods is needed. Besides that, feature selection can improve machine learning performance [10].

Abdul Samad Hasan Basari used machine learning and swarm intelligence method for sentiment analysis [6]. In that research, Particle swarm optimization (PSO) is needed to improve the parameters used in SVM and this research achieves an accuracy of 77%. Another research conducted by Tim O'Kefee and Irena Koprinska shows that SVM achieve an accuracy of 87.15% with the Proportional difference feature selection method [7]. Moreover, Songbo Tan and Jin Zhang research shows that machine learning achieves better performance with feature selection [5]. In that research, Information gain (IG) is a better feature selection than Mutual information (MI), CHI, and Document frequency (DF). Pascal Soucy and Guy W. Mineau showed that KNN had a good performance when the number of features were reduced, even with these conditions KNN has a better performance than NB [11]. From the previous work, the author concludes that machine learning requires the help of other methods to maximize processing features to achieve better performance. Based on previous research, SVM, NB, and RF methods have good performance [16, 17, 18]. In other hand, KNN achieves lower performance than SVM, NB and RF. Hence, the author would like to use KNN with the help of IG to improve its performance and compare the result to SVM, NB and RF.

III. RESEARCH METHOD

1) Information gain

Information gain in machine learning is used to select features that have good relevance. Features with Information gain values below the threshold will be deleted [13]. Information gain will measure the level of relevance of features in a class. A Good feature is a feature that has high relevance to certain classes [3].

In (1) it can be seen that Y is a class, X is an attribute, values(X) shows the amount of data that contains atribute X, and $|Y_V|$ denotes the amount of data with class v, in this case v consists of positive and negative class. $|Y|$ is the amount of data in the positive class and the negative class. *Entropy* (Y) is the entropy of class Y. Entropy is the level of importance of attributes to a class.

$$GAIN\ (Y,X) = Entropy(Y) - \sum_{v \in Values(X)} \frac{|Y_v|}{|Y|} \times Entropy\ (Y_v) \tag{2}$$

$$Entropy\ (Y) = \sum_{v}^{c} - P_v\ Log_2\ P_v$$

c is the number of classes. P_v is the ratio between the amount of data with class v to the overall amount of data.

The dataset used in this study has a balanced number of classes between positive and negative classes. So, the simplification of the formula can be done and it can be seen in (3)

$$GAIN\ (Y,X) = 1 - \sum_{v \in Values(X)} \frac{|Y_v|}{|Y|} \times Entropy\ (Y_v) \quad (3)$$

2) K-nearest neighbor

K-nearest neighbor or commonly abbreviated as KNN is a distance-based classifier [11]. Generally, KNN calculates the distance of one test data with all existing data train using Euclidean distance. The formula for Euclidean distance equation can be seen in (4)

$$dist = \sqrt{\sum_{i=1}^{d} |Te_i - Tr_i|^2} \quad (4)$$

d is the number of dimensions or features, Te_i is feature i in *data test* and Tr_i is feature i in *data train* [14].

In this research the author will use Euclidean distance. The process after calculate the distance for each data train is voting. Voting aims to determine the class or label of a data test. Voting is done by taking as much as K-nearest distance and count how many of each class is contain. If the results of voting are more positive classes, then the data tested is a positive class and vice versa. The disadvantage of KNN is one need to determine the right K for all data test and data train that are not overfitted (only good for training data). While the advantages of the method are the concept is easy to understand and effective on large training data [11].

IV. DESIGN SYSTEM

The dataset will go through preprocessing. The results from preprocessing will be selected and extracted the dataset to get a best features. Then, the best features will be used to classify data. The results of the classification will be evaluated to determine the performance of the method used.

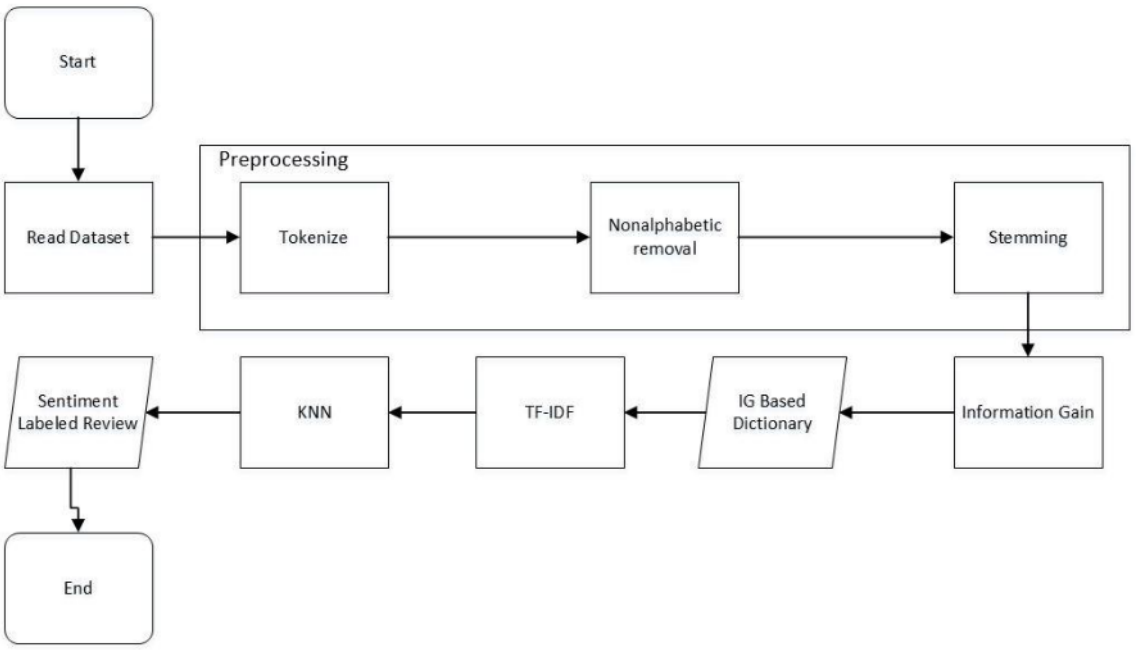

Fig. 1. Classification Flowchart

Figure 1 shows the processes of proposed sentiment analysis system. The first process is to prepare the dataset into Information Gain Based Dictionary is by preprocessing the dataset to get a structured dataset and continued with Information Gain to reduce unimportant features. Feature in Information Gain Based Dictionary will be constructed by TF-IDF for feature extraction process. K-nearest neighbor will be used after all the preprocessing, feature selection and feature extraction are done.

1) Preprocessing

Preprocessing will transform previously unstructured dataset into structured form. The purpose of converting data into structured fom so that dataset can be processed. The dataset will go through the tokenization process. In the tokenization process, each review test will be divided into sequence of words. Next, each word will go through a lowercase process to equalize the structure of each word. Then, a non alphabetic removal process is carried out. Non alphabetic removal process contains stopword removal process and punctuation removal process. The last process is stemming for simplify the word with affixes. Stemming may results the words that are not in dictionary, but the purpose of stemming is to bring variant forms of a word together and not to map a word onto its paradigm form. So, we need to add additional rule that might be included after stemming is done to calculate the relevance of the features to be used as sentiment analysis features such like feature selection and feature extraction. The result of this preprocessing process is a unique word dictionary and will be used as features.

2) Feature selection and feature extraction

The selection feature method used is Information gain. Information gain is a method that shows the importance of a feature to a class. In this paper, threshold for IG value divided into 5 which is greater than 0.1, 0.2, 0.3, 0.4 and 0.5.

After going through the feature selection process, the features will enter the feature extraction process. There are 3 types of feature extraction which is probabilistic, geometric, and logic. In this study, the selected extraction feature is a geometric type because the classification method used is K-nearest neighbor (KNN) which is a type of geometric classification. The selected geometric type is Term frequency Inverse document frequency (TF-IDF) so the result will be fewer than TF or IDF method. TF-IDF formula can be seen in 5

$$TF\ IDF_a = TF_a \times \log\left(\frac{|\mathrm{N}|}{|DF_a|}\right) \tag{5}$$

Term frequency (TF) is the number occurrences of words in a document. N is the total number of dataset. DF is number of documents containing related features [15], and a is attribute being weighted. The words that have gone through the preprocessing will be weighted using the TFIDF formula. The end result of this feature extraction process is unique words that have been weighted.

3) K-nearest neighbor

At the classification stage, the K-nearest neighbor classification method will be selected. Before the classification process, the best features have been selected from preprocessing, extraction features and selection features. The distance calculation based on TF-IDF is done using Euclidean distance. The best K on KNN will be searched by 10-fold Cross validation. K-closest distance will be voted. Most results will be the class of the data being tested.

4) Testing

On this research there are 2 tests. In the first test, KNN without Information gain feature selection performance will be compared to NB, SVM and RF which also without Information gain feature selection. In the second test, KNN with Information gain feature selection performance will be compared to NB, SVM and RF which also with Information gain feature selection. In second test, the threshold for IG divided into 5 which is greater than 0.1, 0.2, 0.3, 0.4 and 0.5. In the SVM method, the Polynomial kernel will be tested. Whereas in KNN method, testing will be carried out with the values of K based on the highest average accuracy from IG threshold. The RF method that will be used in this study will generate 500 trees. RF method always uses the feature selection, in this case when RF does not use the IG feature selection, RF will use Gini impurity feature selection. 10 fold Cross validation will be used to know the performance of each method.

V. RESULT AND DISCUSSION

In this study, the dataset used is polarity dataset v2.0. This dataset consists of 2000 movie reviews with 1000 positive reviews and 1000 negative reviews.

The purpose of this paper are to find the optimal K for KNN based on IG threshold, and to find the best IG threshold. Threshold has a function to get the feature that have higher relevance to a class. In Table 1 shows that the highest average accuracy is K = 3 with average accuracy equal to 83.45%. The higher K for KNN, the lower average accuracy that KNN will get. It's because only the 3 closest data train will define the class for the data being tested.

TABLE I
Comparison of K in KNN based on IG threshold.

IG threshold	K		
	3	5	7
0.1	61.55	57.10	56.45
0.2	65.95	57.70	54.05
0.3	96.15	96.20	96.25
0.4	96.80	96.80	96.80
0.5	96.80	96.80	96.80
Average	83.45	80.92	80.07

After finding the optimal K for KNN. KNN will be compared to SVM, NB and RF. The cases are divided into 2. First is KNN with K = 3 will be compared to SVM, NB and RF without IG while the other case is with IG. In Figure 2 shows that KNN with threshold of 0.4 and 0.5 have the highest accuracy with 96.8%. SVM without IG got the lowest accuracy with 50.05% and SVM has the lowest accuracy than other method. On the other hand, NB has the highest average accuracy than the other methods. RF has a stable performance around 80%.

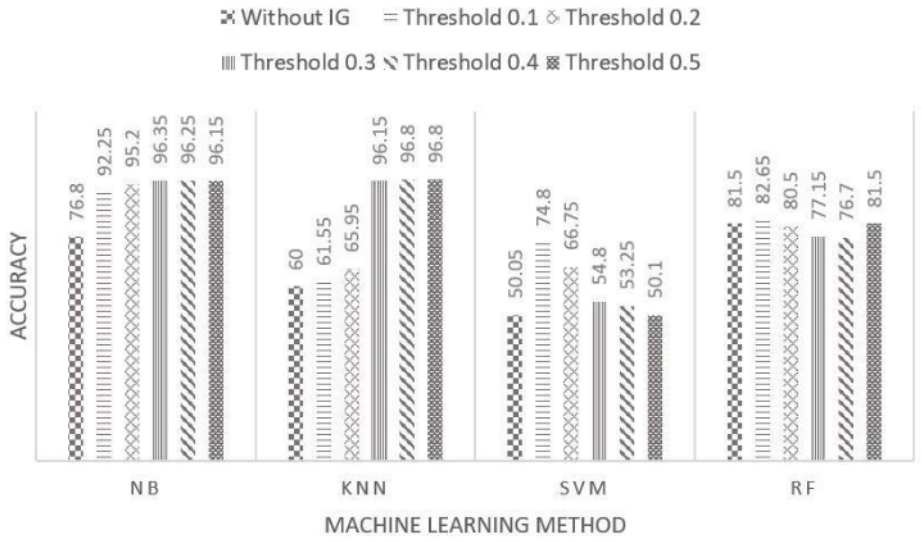

Fig. 2. Comparison accuracy of machine learning methods with and without Information gain.

All the methods use the same type of feature. IG with threshold 0.1, 0.2, 0.3, 0.4 and 0.5 can improve SVM performance. It can be seen in the Figure 2, the highest accuracy SVM with IG or whithout IG is when the threshold of IG is equal to 0.1 and the lowest accuracy SVM with IG is when the threshold of IG is equal to 0.5. It is because the higher value of threshold the less feature will be selected. However, the reason for this also the quality of term dataset like SVM need clearly to be separable and it is hard for SVM to build a model if the feature clearly not linearly separable. The highest accuracy of NB is when the threshold of IG is equal to 0.3. NB accuracy always going up until the threshold is equal to 0.3, but going down when threshold more than 0.3. It shows that the best variety of features to classify is when the threshold equal to 0.3. It is because NB method is quite stable in the 0.1 to 0.5 threshold it shows that information gain can choose a best feature for NB to calculate the probability of the

feature. K in table 1 means the amount of nearby features used to classify data test. KNN got the highest accuracy with 96.8% than the other methods. It can be seen in Figure 2, KNN accuracy always get better if the threshold going up. The higher value of threshold, the less feature will be selected. It is show that KNN has a better accuracy when using high relevance features. In this research, feature selection is a part of process in RF, so we can not cut the features selection process of RF. In this research, Gini Impurity as a feature selection is used as feature selection only for RF when the author try to remove information gain selection feature in all machine learning method that used in this research. This condition resulted not having a features selection make it worse for other classifiers except random forest. Also every classifiers have different threshold because every method have different technique in building a classification model.

VI. CONCLUSION

It can be concluded that the best K for KNN is equal to 3 for Polarity v2.0 dataset. KNN has been be compared to another machine learning method such as NB, SVM and RF. Comparison between KNN, NB, SVM and RF without IG and with IG were done by using 10 fold Cross validation to know the performance.

Without feature selection KNN only achieved the performance equal to 60% with a value of K=3. After using Information gain, KNN has a better performance which also the highest performance compared to other methods with 96.8% at K=3. It can be concluded that the reduction of irrelevant features has a greater effect on KNN method than other methods. Feature selection with IG improve all machine learning methods performance. It's because IG can reduce features that are less relevant to the class. Although KNN only compared the distance between an object to the others, but KNN result can be more better than the other because the performance of KNN is based on good features and the appropriate amount of features. Processing features such as information gain that succeed in reducing unnecessary features is needed to get features with a good amount and quality of features. However, the results of the comparison of machine learning may differ in the case of different datasets. For further research, author recommend to find the optimal threshold by a method. It's because in this paper the optimal threshold is set manually.

REFERENCES

[1] Sebastiani F *Machine learning in automated text categorization* 2002 CSUR 34(1) 1 47
[2] Jiang S, Pang G, Wu M and Kuang L *An improved K-nearest-neighbor algorithm for text categorization* 2012 Expert Systems with Applications 39(1) 1503 09
[3] Pratiwi A I *On the feature selection and classification based on information gain for document sentiment analysis* 2018 Applied Computational Intelligence and Soft Computing
[4] Pang B, Lee L, and Vaithyanathan S *Thumbs up?: sentiment classification using machine learning techniques* 2002 July EMNLP ACL 10 79 86
[5] Tan S and Zhang J *An empirical study of sentiment analysis for chinese documents* 2008 Expert Systems with applications 34(4) 2622 29
[6] Basari A S H, Hussin B, Ananta I G P and Zeniarja J *Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization* 2013 Procedia Engineering 53 453 62
[7] OKeefe T and Koprinska I *Feature selection and weighting methods in sentiment analysis* 2009 December In Proceedings of the 14th ADCS 67 74
[8] Sahami M and Koller D *Using machine learning to improve information access* 1998 (Doctoral dissertation, Stanford University, Department of Computer Science).
[9] Chaovalit P and Zhou L *Movie review mining: A comparison between supervised and unsupervised classification approaches* 2005 January HICSS 112c- 112c IEEE.
[10] Lee C and Lee G G *Information gain and divergence-based feature selection for machine learning-based text categorization* 2006 Information processing and management 42(1) 155 65.
[11] Soucy P and Mineau G W *A simple KNN algorithm for text categorization* 2001 ICDM 2001 647 48)
[12] Pang B and Lee L *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts* 2004 July ACL 271
[13] Yang Y and Pedersen J O *A comparative study on feature selection in text categorization* 1997 July Icml 97 412 20
[14] Cha S H *Comprehensive survey on distance/similarity measures between probability density functions* 2007 1(2) 1
[15] Na J C, Sui H, Khoo C S and Zhou Y *Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews* 2004
[16] Elmurngi E, and Gherbi A *An empirical study on detecting fake reviews using machine learning techniques* 2017 In International Conference on Innovative Computing Technology (pp. 107-114).
[17] Thapa L B , and Bal B K *Classifying sentiments in Nepali subjective texts In Information Intelligence, Systems & Applications* 2006 IISA.
[18] Xu Guo X and Ye Y and Cheng, J *An Improved Random Forest Classifier for Text Categorization* 2012 JCP 7(12) 2913 2920.
[19] Gupte A, Joshi S, Gadgul P and Kadam A *Comparative study of classification algorithms used in sentiment analysis* JCSIT 5(5) 6261 64

[20] Bakar, M.Y.A. Adiwijaya, and AI Faraby,S.,2018. *Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language Translation) Using Information Gain and Backpropagation Neural Network* 2018 In International Conference on Asian Language Processing (IALP) (pp. 344-350) IEEE.

[21] Adiwijaya, Aulia, M. N., Mubarok, M. S., Novia, W. U., and Nhita, F. *A comparative study of MFCC- KNN and LPC-KNN for hijaiyyah letters pronounciation classification system.* 2017 In 5th International Conference on Information and Communication Technology (ICoIC7) (pp. 1-5). IEEE.

[22] Mubarok, M.S., Adiwijaya and Aldhi, M.D. *Aspect-based sentiment analysis to review products using Nave Bayes.* In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060). AIP Publishing.

[23] Naf'an, M. Z., Bimantara, A. A., Larasati, A., Risondang, E. M., and Nugraha, N. A. S. T *Sentiment Analysis of Cyberbullying on Instagram User Comments.* Journal of Data Science and Its Applications, 2(1), 88-98.