Indonesian Journal of Electrical Engineering and Computer Science
Vol. 12, No. 1, October 2018, pp. 46~50
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v12.i1.pp46-50

Random Forest Approach fo Sentiment Analysis in Indonesian Language

M. Ali Fauzi
Faculty of Computer Science, Brawijaya University, Malang, Indonesia

Article Info

Article history:

Received May 5, 2018
Revised Jul 6, 2018
Accepted Jul 10, 2018

Keywords:

Random forest
Sentiment analysis
Term weighting
Text classification
TF.IDF

ABSTRACT

Sentiment analysis becomes very useful since the rise of social media and online review website and, thus, the requirement of analyzing their sentiment in an effective and efficient way. We can consider sentiment analysis as text classification problem with sentiment as its categories. In this study, we explore the use of Random Forest for sentiment classification in Indonesian language. We also explore the use of bag of words (BOW) features with some term weighting methods variation such as Binary TF, Raw TF, Logarithmic TF and TF.IDF. The experiment result showed that sentiment analysis system using random forest give good performance with average OOB score 0.829. The result also depicted that all of the four term weighting method has competitive result. Since the score difference is not very significant, we can say that the term weighting method variation in study has no remarkable effect for sentiment analysis using Random Forest.

Corresponding Author:

M. Ali Fauzi,
Faculty of Computer Science,
Brawijaya University, Malang, Indonesia.
Email: moch.ali.fauzi@ub.ac.id

1. INTRODUCTION

Nowadays, people tend to write their experience, feeling, opinions, and views about events, products or services in online platforms such as social media, blog, forum, shopping sites, or review sites. It makes online platforms become a source of highly valuable information for both consumers and producers. Customers get second opinions before purchasing some products or services. On the other hand, producers get information about what people think about their products or services and predict the public acceptance rate level. This information can be very useful for improvement and marketing strategies [1].

Sentiment analysis is a task of analyzing people's opinions from a piece of text in order to specify whether the sentiments are positive, negative or neutral. Sentiment Analysis have been obtaining popularity over the past years as a result of the rise of social media and online review website and, thus, the requirement of analyzing their sentiment in an effective and efficient way. Sentiment analysis is currently a major research field with many applications in a large number of domains such as election results prediction [2]-[4], stock market prediction [5], [6], products and merchants ranking [7], movie revenues prediction [8]-[10], learning evaluation [11], [12], and etc.

We can consider sentiment analysis as text classification problem with sentiment as its categories. Therefore, we can use supervised machine learning approaches to tackle this problem. This approach is very popular in sentiment analysis and proven to be very good in this filed. Some machine learning approach that have been used in this field for example Naive Bayes [13]-[17], Support Vector Machines [18]-[19], Maximum Entropy [20], Neural Network [21], [22] decision tree and K-Nearest Neighbor (KNN) [23]-[26].

In this study, we explore the use of Random Forest for sentiment classification in Indonesian language. Random Forest is an ensemble learning technique based on decision tree algorithm [27]. Random Forests have been incredible in recent years since the performance of this type of algorithms have surpass SVMs, Naïve Bayes and other machine learning algorithms for classification task in some domain like bioinformatics and computational biology [28]. We will try whether this type of ensemble methods still outstanding on sentiment analysis tasks. In this study, we will also explore the use of bag of words (BOW) features with some term weighting methods variation such as Binary TF, Raw TF, Logarithmic TF and TF.IDF.

2. RESEARCH METHOD

As depicted in Figure 1, sentiment analysis system in this study consists of three main stages, preprocessing, features extraction and classification using Random Forest. The ouptut of classification result is two category, positive and negative.

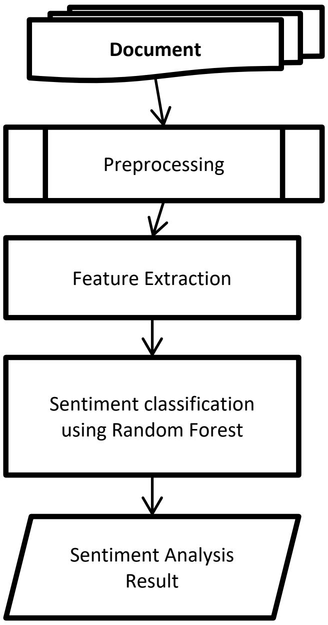
Document
Preprocessing
Feature Extraction
Sentiment classification using Random Forest
Sentiment Analysis Result

Figure 1. System main flowchart

2.1. Preprocessing

The first stage of this system is preprocessing. This stage involves several processes including tokenization, case folding and cleaning. Tokenization is a task of splitting review text into smaller units called tokens or terms [29], [30]. Case folding is a task of making all of characters in review text become lowercase [31], [32]. Meanwhile, cleaning is a task of removing punctuation, numbers, html tag and characters outside of the alphabet. In this study, we don’t employ stemming and filtering since in some previous works about sentiment analysis, stemming and filtering cannot improve classification performance.

2.2. Feature Extraction

Bag-of-word (BOW) features will be used in this study. Each document would be represented as a vector in a space terms with the unique terms from preprocessing stage become its features. The feature vector value is determined using some term weighting method. The most popular term weighting methods are Term Frequency (TF), Inverse Document Frequency (IDF) and the combination of the two, Term Frequency Inverse Document Frequency (TF.IDF) [33].

Term Frequency is assigning weights by assuming that each term have a contribution that is proportional to the number of its occurrences in the document [34], [35]. There are some popular variation of TF such as Binary TF, Raw TF, and Logarithmic TF. Using Binary TF, each document is represented as a binary vector. A term that occurs in a document will get value 1 in the document vector, otherwise a term that never occurs in a document will get value 0. This kind of term weighting does not consider the number of term occurrences, only 0/1 values. In contrast to Binary TF, Raw TF method does consider the number of term occurrences. A term will get value based on how many times it appears in the document. Meanwhile Logarithmic TF also consider the number of term occurrences. The difference is Logarithmic TF assume that the importance of a term in a document does not increase proportionally with term how many times it occurs. The weights of term t in document d using Logarithmic TF can be counted as follows:

$$TF(t,d) = 1 + \log\left(f_{t,d}\right) \tag{1}$$

where $f_{t,d}$ is the number of the how many times term t appears in the document d.

Meanwhile, Inverse Document Frequency is a global term weighting that been counted by regarding the distribution of the term in the dataset. This term weighting will give higher value for a rare term, a term that only appears in certain documents. The weights of term t using IDF formulated as follows:

$$IDF(t) = 1 + \log\left(\frac{N_d}{df_t}\right) \tag{2}$$

where N_d is the number of documents in dataset and df_t is the number of documents in dataset that where term t appears.

The most popular term weighting is TF.IDF. TF.IDF is a multiplication of TF and IDF. The weight combination of term t in document d can be counted as follows [36]:

$$TF \bullet IDF(t,d) = TF(t,d) \bullet IDF(t) \tag{3}$$

where $TF(t,d)$ is the TF value of term t in document d and $IDF(t)$ is the IDF value of term t.

2.3. Sentiment Classification using Random Forest

The last stage is sentiment classification. Each review will be classified into positive or negative category. In this study, we employ random forest for the classification task. Random forest algorithm is a supervised classification algorithm. It is an ensemble learning technique based on decision tree algorithm [27]. This Ensemble technique combines the predictions of some base estimators constructed with decision tree algorithm to enhance robustness over an individual estimator. Random Forest grows a lot of classification trees, which is called forest. If we want to classify a new data, each tree gives its category prediction as one vote. The forest chooses the category that has majority voting. In general, the more trees in the random forest the higher accuracy results given.

Random Forests have been gaining popularity in recent years since the performance of this type of algorithms have outstanding for classification task in some domain like bioinformatics and computational biology. There also some works in text classification using Random forest such as for hatespeech detection [37] and authorship profiling [38].

3. RESULTS AND ANALYSIS

Experiment conducted by using 386 reviews taken from FemaleDaily. All of the reviews is in Indonesian language. Instead of using cross validation, Random Forest use out-of-bag (OOB) error estimate to get an unbiased estimate of the classification performance. OOB score range form 0 to 1. The higher OOB score the better classification performance, otherwise the lower OOB score indicates worse classification performance. In the experiment, Random Forest will be tested using several term weighting method including Binary TF, Raw TF, Logarithmic TF, and TF.IDF. The experiment is conducted using Scikit-learn library [39]. Theresult can be seen in Figure 2.

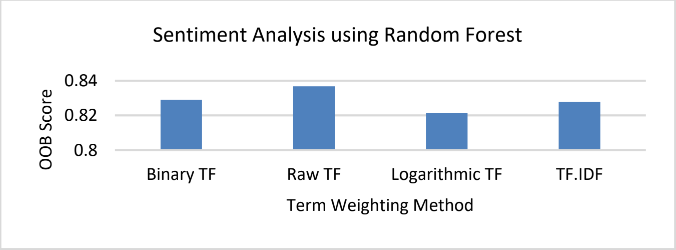

Figure 2. Sentiment analysis experiment reuslt using random forest

Figure 2 show that sentiment analysis using random forest give good performance with average OOB score 0.829. We can also see form Figure 2 that all of the four term weighting method has competitive result. The OOB score between is just slightly different. The best OOB score is gained by Raw TF by 0.837. The lowest OOB score is gained by Logarithmic TF by 0.821. In the second place is Binary TF with OOB score 0.829 and the third place is TF.IDF with OOB score 0.828. This result is actually surprising because usually TF.IDF can outperform any other term weighting method. However, since the score difference is not very significant, we can say that the term weighting method variation in study has no remarkable effect for sentiment analysis using Random Forest.

4. CONCLUSION

In this study, we explore Random Forest with several term weighting method for sentiment analysis in Indonesian Language. This system in this study consists of three main stages, preprocessing, features extraction and classification using random forest. The ouptut of classification result is two category, positive and negative. The experiment result showed that sentiment analysis using random forest give good performance with average OOB score 0.829. The result also depicted that all of the four term weighting method has competitive result. Since the score difference is not very significant, we can say that the term weighting method variation in study has no remarkable effect for sentiment analysis using Random Forest.

REFERENCES

[1] Jansen BJ, Zhang M, Sobel K, Chowdury A. Twitter power: Tweets as electronic word of mouth. Journal of the Association for Information Science and Technology. 2009 Nov 1;60(11):2169-88.

[2] Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. Predicting elections with twitter: What 140 characters reveal about political sentiment. Icwsm. 2010 May 23;10(1):178-85.

[3] Bermingham A, Smeaton A. On using Twitter to monitor political sentiment and predict election results. InProceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011) 2011 (pp. 2-10).

[4] Sang ET, Bos J. Predicting the 2011 dutch senate election results with twitter. InProceedings of the workshop on semantic analysis in social media 2012 Apr 23 (pp. 53-60). Association for Computational Linguistics.

[5] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of computational science. 2011 Mar 31;2(1):1-8.

[6] Zhang X, Fuehres H, Gloor PA. Predicting stock market indicators through twitter “I hope it is not as bad as I fear”. Procedia-Social and Behavioral Sciences. 2011 Jan 1;26:55-62.

[7] McGlohon M, Glance NS, Reiter Z. Star Quality: Aggregating Reviews to Rank Products and Merchants. InICWSM 2010 May 16.

[8] Mishne G, Glance NS. Predicting Movie Sales from Blogger Sentiment. InAAAI Spring Symposium: Computational Approaches to Analyzing Weblogs 2006 Mar 27 (pp. 155-158).

[9] Joshi M, Das D, Gimpel K, Smith NA. Movie reviews and revenues: An experiment in text regression. InHuman Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics 2010 Jun 2 (pp. 293-296). Association for Computational Linguistics.

[10] Sadikov E, Parameswaran AG, Venetis P. Blogs as Predictors of Movie Success. InICWSM 2009 Mar 20.

[11] Ortigosa A, Martín JM, Carro RM. Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior. 2014 Feb 28;31:527-41.

[12] Munezero M, Montero CS, Mozgovoy M, Sutinen E. Exploiting sentiment analysis to track emotions in students' learning diaries. InProceedings of the 13th Koli Calling International Conference on Computing Education Research 2013 Nov 14 (pp. 145-152). ACM.

[13] Kang H, Yoo SJ, Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. 2012 Apr 30;39(5):6000-10.

[14] 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.

[15] 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.

[16] 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).

[17] 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.

[18] Mullen T, Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. InEMNLP 2004 Jul (Vol. 4, pp. 412-418).

[19] 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.

[20] Batista F, Ribeiro R. Sentiment analysis and topic classification based on binary maximum entropy classifiers.

[21] Munir MM, Fauzi MA, Perdana RS. Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN. 2017;2548:964X.

[22] Lam SL, Lee DL. Feature reduction for neural network based text categorization. InDatabase Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on 1999 (pp. 195-202). IEEE.

[23] Bilal M, Israr H, Shahid M, Khan A. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University-Computer and Information Sciences. 2016 Jul 31;28(3):330-44.

[24] 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.

[25] 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.

[26] 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.

[27] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

[28] Statnikov A, Aliferis CF. Are random forests better than support vector machines for microarray-based cancer classification?. InAMIA annual symposium proceedings 2007 (Vol. 2007, p. 686). American Medical Informatics Association.

[29] 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).

[30] Rosi F, Fauzi MA, Perdana RS. Prediksi Rating Pada Review Produk Kecantikan Menggunakan Metode Naïve Bayes dan Categorical Proportional Difference (CPD). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2018; 2(5):1991-97.

[31] Lestari AR, Perdana RS, Fauzi MA. Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2017; 1(12):1718-24.

[32] M. Ali Fauzi, Djoko Cahyo Utomo, Budi Darma Setiawan, and Eko Sakti Pramukantoro. 2017. Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning. In Proceedings of the International Conference on Advances in Image Processing (ICAIP 2017). ACM, New York, NY, USA, 151-155. DOI: https://doi.org/10.1145/3133264.3133303

[33] 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.

[34] 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).

[35] 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.

[36] 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.

[37] Alfina I, Mulia R, Fanany MI, Ekanata Y. Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study.

[38] Palomino-Garibay A, Camacho-González AT, Fierro-Villaneda RA, Hernández-Farias I, Buscaldi D, Meza-Ruiz IV. A random forest approach for authorship profiling. Cappellato et al.[8]. 2015.

[39] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825-30.