TELKOMNIKA, Vol.16, No.3, June 2018, pp. 1345~1350
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v16i3.7751

Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion

M. Ali Fauzi*, Ro'i Fahreza Nur Firmansyah, Tri Afirianto
Faculty of Computer Science, Universitas Brawijaya, Jl. Veteran, Malang, Indonesia
*Corresponding author, e-mail: moch.ali.fauzi@ub.ac.id

Abstract

Sentiment analysis in short informal texts like product reviews is more challenging. Short texts are sparse, noisy, and lack of context information. Traditional text classification methods may not be suitable for analyzing sentiment of short texts given all those difficulties. A common approach to overcome these problems is to enrich the original texts with additional semantics to make it appear like a large document of text. Then, traditional classification methods can be applied to it. In this study, we developed an automatic sentiment analysis system of short informal Indonesian texts using Naïve Bayes and Synonym Based Feature Expansion. The system consists of three main stages, preprocessing and normalization, features expansion and classification. After preprocessing and normalization, we utilize Kateglo to find some synonyms of every words in original texts and append them. Finally, the text is classified using Naïve Bayes. The experiment shows that the proposed method can improve the performance of sentiment analysis of short informal Indonesian product reviews. The best sentiment classification performance using proposed feature expansion is obtained by accuracy of 98%.The experiment also show that feature expansion will give higher improvement in small number of training data than in the large number of them.

Keywords: *product review, sentiment analysis, short text, feature expansion, classification*

1. Introduction

Sentiment analysis is one of the fundamental problems in natural language processing (NLP). Sentiment analysis involves analyzing people's opinions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes from a piece of text [1]. Sentiment Analysis has a number of applications, including ranking products and merchants [2-3], predicting election results [4], predicting box-office revenues for movies [5-7], predicting the stock market [8], characterizing social relations [9], and etc. The rise of social media make us now dealing with much more short informal texts every day. Examples are tweets, status updates, comments, and reviews from various social platforms. Working with these short informal text genre is more challenging compared to traditional text genres because there are many limitations in this genre. Thus, there is growing interest in sentiment analysis of this kind of texts.

Sentiment classification is a special kind of text classification problem with two class, positive and negative. Since it is a text classification problem, any existing supervised learning method can be applied. In most cases, the use of statistical or machine learning techniques such as Naive Bayes, Maximum Entropy, and Support Vector Machines has proven to be successful in this field [10-16]. Some previous works also use another supervised learning such as method decision tree and K-Nearest Neighbor (KNN) to analyze sentiment within texts [17-19]. Those researches showed that standard machine learning methods using unigram (bag of words) as features perform very well in this field.

Short texts are sparse, noisy, and lack of context information. Traditional text classification methods like machine learning may not be suitable for analyzing sentiment of short texts given all those difficulties. A common approach to overcome these problems for analyzing sentiment of short texts is to enrich the original texts with additional semantics to make it appear like a large document of text. Then, traditional classification methods can be applied to it. Some of the previous works employ search engines to extract more information

Received November 7, 2017; Revised March 25, 2018; Accepted April 3, 2018

about the short text [20-22]. The other works utilize external sources such as Wikipedia and WordNet as background knowledge [23-26].

In this study, we conducted automatic sentiment analysis of short informal Indonesian product reviews. This is very useful because it allows review to be aggregated without manual intervention. Consumers can utilize this information to research products before buying. Marketers can utilize this to research public opinion of their products. Organizations can also utilize this to get critical feedback about problems in their newly released products. The reviews are gathered from a social platforms that provides reviews from users about certain product. Every review in this platform is a short informal Indonesian text that express positive or negative opinion about the product. Most of the reviews are short texts with informal language, creative spelling and punctuation, misspellings, and slang word. This paper aims to improve short text bag of words representation for sentiment analysis. We developed automatic sentiment analysis system of short informal Indonesian texts using Naïve Bayes and Synonym Based Feature Expansion. In the first step, we counduct preprocessing normalizing misspellings and slang words. In the next we use Kateglo API (kateglo.co.id) to find synonym of each word in texts to enrich the original texts. Finally, we do classification using Naïve Bayes Classifier and bag of words as the features.

2. Research Method

In general, as seen in Figure 1, the sentiment analysis system in this study consists of three main stages, preprocessing and normalization, features expansion and classification. The first stage involves several steps including tokenization, stopwords removal, stemming and misspellings words normalization. In this stage, we also counduted negation convert. In the feature expansion stage, we use Kateglo API to find synonym of each word in the review texts. Then, the synonym will be added to the original texts. Finally, in the sentiment classification stage, Naïve Bayes is trained using some training data and the expanded review texts will be classified using Naïve Bayes and bag of words as its features.

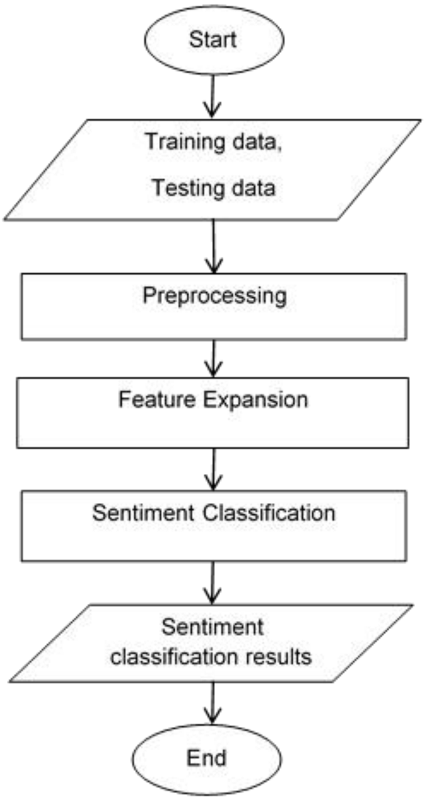


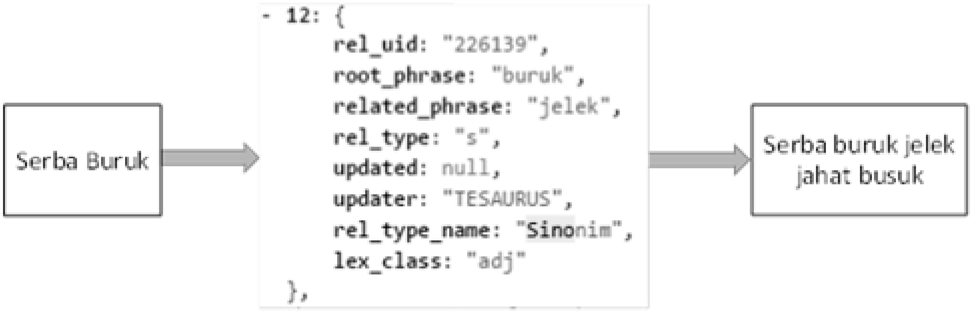
Figure 1. System Main Flowchart

2.1. Preprocessing

Preprocessing involves tokenization, stopwords removal, stemming and misspellings words normalization, and negation convert. Tokenization is an early process done to remove punctuation, numbers, and characters other than the alphabet [27-30]. Also in this stage will be coundected case folding, which is changing all capital letters into lowercase. Stopwords removal or filtering is removing uninformative words referring to the existing stopword dictionary. In this case, we use stoplist by Tala that have been used in [31]. Stemming is a process to convert every words to its root. This process is done by removing affixes such as prefix, infix and suffix. In this case, we use Nazief-Adriani Stemmer [32]. Misspellings words normalization is done by changing the words into its formal form. Examples are the word “ga” become “tidak” and the word “bisaaaa” become “bisa”. Negation convert is a process of converting negation words contained in a sentence. The negation words has influence in changing the value of sentiment in a sentence. The most used negation words in Indonesian Language are "tidak", "bukan", "tak", "tanpa", "kurang" and "jangan". The negation convert is done by finding the antonym of the word that been negated. For example, negation convert of “tidak bagus” is “jelek”.

2.2. Feature expansion

Feature expansion is process of enriching the original texts with additional semantics to make it appear like a large document of text [33]. In this study, we utilize Kateglo to find some synonyms of every words in original texts and append them. Kateglo is a dictionary website that provides API for fetching word attributes such as lexical class, root form, synonym, antonym, etc. Our system will find the synonyms of a certain word by sending some parameter to URL http://kateglo.com/api.php?format=json&phrase=[word]. Then, our system will parse the Json data received from the kateglo server and use them for feature expansion as seen in Figure 2.



```
- 12: {
      rel_uid: "226139",
      root_phrase: "buruk",
      related_phrase: "jelek",
      rel_type: "s",
      updated: null,
      updater: "TESAURUS",
      rel_type_name: "Sinonim",
      lex_class: "adj"
  },
```

Figure 2. Feature expansion steps

2.3. Classification using naïve bayes

Naïve Bayes is one of the most effective and efficient inductive learning algorithms for machine learning and data mining. The performance of Naive Bayes is competitive in the classification process although it uses the assumption of attribute independence. The assumption of the independence of these attributes on the data is rare, but although the assumption of attribute independence is violated, the performance of Naive Bayes classification is quite high, as proven by various empirical studies [34-35].

The Naïve Bayes classification is incorporated into the Bayes learning algorithm constructed by training data to estimate the probability of each category contained in the characteristics of the testing document. In general, the classification process using the Naïve Bayes method can be seen in equation 1.

$$P(c_j \mid w_i) = \frac{P(c_j) \cdot P(w_i \mid c_j)}{P(w_i)} \tag{1}$$

where $P(c_j \mid w_i)$ is posterior or the probability of class c_j given word w_i. $P(c_j)$ is prior or the probability of class c_j. $P(w_i \mid c_j)$ is likelihood or the probability of word w_i given class c_j. Then, $P(w_i)$ is evidence or the probability of word w_i.

The probability of word w_i actually can be eliminated in the classification process because the probability of word w_i will not affect the comparison of classification results between categories. Thus, the process of classification can be simplified to be equation 2.

$$P(c_j \mid w_i) = P(c_j) \cdot P(w_i \mid c_j) \tag{2}$$

3. Results and Analysis

Dataset that was used in the experiment is mobile banking app reviews. There are 100 testing data consisting of 50 positive reviews and 50 negative reviews. Meanwhile, the training data used varies from 50, 100, 400 to 1000 training data. This experiment was conducted to determine the effect of feature expansion and the number of training data on the sentiment classification performance. The evaluation method used in this experiment is accuracy. Experiment results shown in Figure 3.

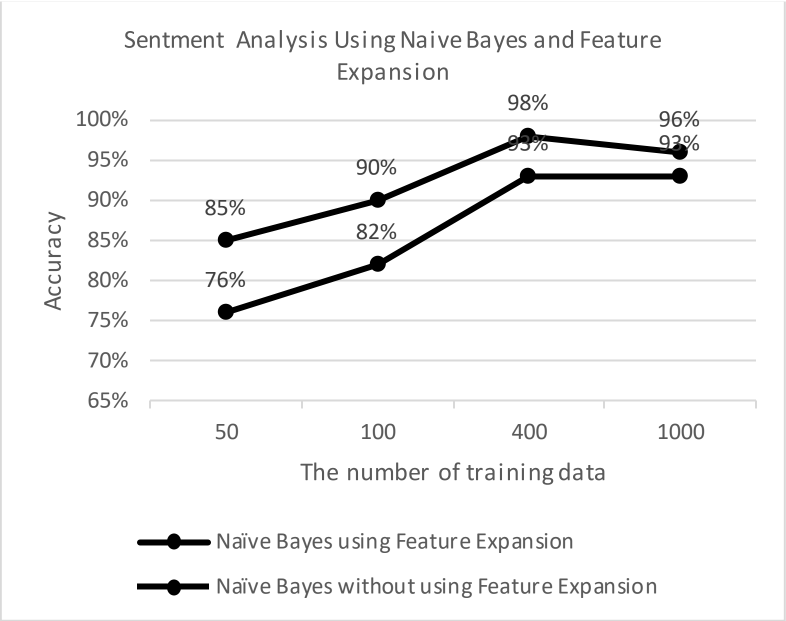


Figure 3. The experiment results

As displayed in Figure 3, the best sentiment classification accuracy is obtained when the training data is 400 using feature expansion by 98%. The experiment results using feature expansion on every number of training data always have better classification accuracy compared to the ones that not using feature expansion. This results show that feature expansion increase the sentiment classification performance. In short-text classification, many words in the testing data never appear in the training data. It can damage the sentiment classification performance. Using feature expansion, the system append some new words, in this case the synonyms of each word in testing data, to the testing data before the classification stage. Therefore, the vocabulary on the testing data will be richer and the probability of training data and testing data share the same words will increase. Thus, sentiment classification process will produce better performance compared without using feature expansion.

Also from Figure 3, can be seen that the number of training data does have effect on the sentiment classification performance either with feature expansion, or without feature expansion. The more the training data, the higher accuracy obtained. The highest accuracy difference between sentiment classifications using feature expansion and not using feature expansion occurs when training data used is minimal. This difference will be closer along with the increasing number of training data. This result show that feature expansion will have bigger influence in small training data. In the large number of training data, the word in testing data that will be expanded most likely has already appeared in the train data. Hence, in this case, using feature expansion does not increase the sentiment classification performance significantly.

4. Conclusion and Future Works

The proposed method, Synonym based feature expansion, had been proven can improve the performance of sentiment analysis of short informal Indonesian product reviews. Based on the experiment, Naïve Bayes classifier that use feature expansion always have better classification accuracy compared to the ones that not using feature expansion. The best sentiment classification performance is obtained when the training data is 400 using feature expansion by accuraty of 98%. The number of training data also affect on the sentiment classification performance either with feature expansion, or without feature expansion. The more the training data, the higher the accuracy obtained. The highest accuracy difference between sentiment classifications using feature expansion and not using feature expansion occurs when training data used is minimal. This difference is decreasing along with the increasing number of training data. This result show that feature expansion will give bigger improvement in small training data than in the large number of training data.

References

[1] Liu B. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*. 2012; 5(1):1-67.

[2] McGlohon M, Glance NS, Reiter Z. Star Quality: Aggregating Reviews to Rank Products and Merchants. *In ICWSM*. 2010.

[3] 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): 1991-1997.

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

[5] Asur S, Huberman BA. *Predicting the future with social media.* In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2010; 01:492-499.

[6] 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; :293-296.

[7] Sadikov E, Parameswaran AG, Venetis P. Blogs as Predictors of Movie Success. I*n ICWSM.* 2009.

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

[9] Groh G, Hauffa J. Characterizing Social Relations Via NLP-Based Sentiment Analysis. *In ICWSM.* 2011.

[10] Aue, A. and Gamon, M. *Customizing sentiment classifiers to new domains: A case study.* In Proceedings of recent advances in natural language processing (RANLP). 2005; 1(3.1):2-1.

[11] Tan, S., Wu, G., Tang, H. and Cheng, X. *A novel scheme for domain-transfer problem in the context of sentiment analysis.* In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. 2007; :979-982.

[12] 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):1733-1741.

[13] 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;3(2):1-6.

[14] 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 - 2770.

[15] 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): 1725-1732.

[16] 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): 1718-1724.

[17] 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):1750-1757.

[18] 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 – 2743.

[19] Bilal, M., Israr, H., Shahid, M. and 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; *28*(3):330-344.

[20] Bollegala, D., Matsuo, Y. and Ishizuka, M. Measuring semantic similarity between words using web search engines. *www*. 2007; *7*(757-766).

[21] Sahami, M. and Heilman, T.D. *A web-based kernel function for measuring the similarity of short text snippets*. In Proceedings of the 15th international conference on World Wide Web. 2006; :377-386.

[22] Yih, W.T. and Meek, C. Improving similarity measures for short segments of text. In *AAAI*. 2007; 7(7):1489-1494.

[23] Chen, M., Jin, X. and Shen, D. Short text classification improved by learning multi-granularity topics. In *IJCAI*. 2011; :1776-1781.

[24] Hu, X., Sun, N., Zhang, C. and Chua, T.S. *Exploiting internal and external semantics for the clustering of short texts using world knowledge*. In Proceedings of the 18th ACM conference on Information and knowledge management. 2009; :919-928.

[25] Phan, X.H., Nguyen, L.M. and Horiguchi, S. *Learning to classify short and sparse text & web with hidden topics from large-scale data collections*. In Proceedings of the 17th international conference on World Wide Web. 2008; :91-100.

[26] Jin, O., Liu, N.N., Zhao, K., Yu, Y. and Yang, Q. *Transferring topical knowledge from auxiliary long texts for short text clustering.* In Proceedings of the 20th ACM international conference on Information and knowledge management. 2011; :775-784.

[27] Fauzi, M.A., Arifin, A.Z. and Yuniarti, A. Arabic Book Retrieval using Class and Book Index Based Term Weighting. *International Journal of Electrical and Computer Engineering (IJECE).* 2017; *7*(6).

[28] Pramukantoro, E.S. and Fauzi, M.A. *Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification.* In Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on. 2016; :149-155.

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

[30] Fauzi, M.A., Utomo, D.C., Setiawan, B.D. and Pramukantoro, E.S. 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*. 2017; :151-155.

[31] 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.

[32] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M. and Williams, H.E. Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*. 2007; *6*(4):1-33.

[33] Yusuf S, Fauzi MA, Brata KC. Sistem Temu Kembali Informasi Pasal-Pasal KUHP (Kitab Undang-Undang Hukum Pidana) Berbasis Android Menggunakan Metode Synonym Recognition dan Cosine Similarity. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.* 2018;2(2):838-847.

[34] Qiang, G. *An effective algorithm for improving the performance of Naïve Bayes for text classification.* In 2010 Second International Conference on Computer Research and Development. 2010.

[35] 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):610-5.