Text Mining and Support Vector Machine for Sentiment Analysis of Tourist Reviews in Bangkalan Regency

Imamah*, Husni, Eka Malasari Rachman, Ika Oktavia Suzanti, and Fifin Ayu Mufarroha

1Engineering Faculty, Universitas Trunojoyo Madura, Bangkalan, Indonesia

*i2m@trunojoyo.ac.id

Abstract. TripAdvisor is a travel site that offers reviews of hotels, flights, restaurants and tourist attractions. Reviews from tourists are indispensable for developing tourism, but the number of comments will complicate the owner to analyze the important aspects of the review so that the reviews should be beneficial to develop spot, overlooked or unreadable. This research aims to facilitate the owner of tourist places in the Bangkalan regency to classify negative opinion, positive opinion, and to know the target opinion using techniques sentiment analysis. The initial stage of the sentiment process analysis on this research is Web scrapping on the TripAdvisor site, the purpose of this stage to collect user review data. The review Data obtained will then be classified using the Support Vector Machine (SVM) method. Further review data classification results will be processed using the Text Mining method, to find the target opinion that is considered important in the review. Based on the research that has been done, obtained the accuracy of the classification process with SVM method of 70.22% for Indonesian-language reviews.

1. Introduction

Tripadvisor is a travel site that offers reviews of hotels, flights, restaurants and tourist attractions. Reviews from tourists are indispensable for developing tourism, but the number of comments will complicate the owner to analyzing the important aspects of the review so that the reviews should be beneficial to develop spot, overlooked or unreadable. This research aims to facilitate the owner of tourist places in the Bangkalan regency to classify negative opinion, positive opinion, and to know the target opinion using techniques sentiment analysis. A sentiment analysis system generally consists of two parts: an aspect and opinion word extraction model and an aspect-level sentiment classification model. The former aims to identify the aspect and opinion words mentioned in reviews[1][2], while the latter aims to infer the sentiment polarity of specific aspects[3][4].

2. Related Work

This section provides a review of articles mainly studying sentiment analysis. In 2018, Nanda Cahyo conducted an analysis of the sentiment review of mobile Banking application users using the Fuzzy K-Nearest Neighbor. In his research, Nanda Cahyo uses TF-IDF and cosine Similarity to calculate the distance between the data, and the highest F-Measure of 0.9604and the lowest 0.8349[5]. In 2017, Wilianto analyzed the tourism sentiment in West Java using the Naïve Bayes classifier method and was charged with Vmap value for positive sentiment of 0.0084 and 0.0064 for negativevmap sentiment[6].

In 2018, Fanissa conducted research using the naïve Bayes method to analyse tourism sentiment in
Malang City using training data from TripAdvisor user reviews[7]. Haryanto also conducted a study using SVM and Query Expansion for analysis sentiment review of Indonesian goods obtained the accuracy value of 94.75%, whereas if optimized by adding Expansion Query obtained accuracy 96.25%[8]

3. Methods
In this study, we use Support Vector Machine (SVM) as a method for classifying positive sentiment and negative sentiment. From several literature as reference, here is a description of the method used in this research.

3.1. Support Vector Machine (SVM)
The Support Vector Machine (SVM) is a relatively new technique for predictor, both in case of classification and regression. The Support Vector Machine (SVM) is a set of guided learning methods that analyzes data and recognizes patterns, used for regression classification and analysis. The original SVM algorithm was created by Vladimir Vapnik and the current standard derivative (soft margins) proposed by Corinna Cortes and Vapnik Vladimir[9].

3.2. Text Mining
Text mining is an application of data mining concepts and techniques to find patterns in the text, which is the process of analyzing text to obtain useful information for a particular purpose. Based on the irregularities of the text data structure, the process of text mining requires some initial stages in which the point is to prepare the text to be more structured[10]. In this research, there are several phases to do text mining in the process of Sentiment analysis, namely:
1. Tokenizing
   The tokenizing stage is the cutting phase of the input string based on each word that follows it.
2. Filtering
   Stage filtering is a step taking the important words of the token result.
3. Stemming
   The stemming stage is the search for the root word of each word filtering result.
4. Tagging
   Tagging is a stage for searching the original/root form of each word stemming results.
5. Analyzing
   The analyzing stage is the defining stage of how far between the words between the documents exist.

4. Result and Discussion
The data used in this study is a review of the tourism data in the Bangkalan on the TripAdvisor site taken on the october 2019 range. The process of web scraping using tools data miner. Bangkalan Tourism Data is displayed on table 1. Many of the tourist data have not been reviewed, which means the tourist spots are not much known or no one is writing their reviews on TripAdvisor.

| No | Tourist Attractions               | Total |
|----|----------------------------------|-------|
| 1  | Mercusuar Sembilangan            | 31    |
| 2  | Pantai sirep kemuning            | 16    |
| 3  | Bukit kapur jaddih               | 138   |
| 4  | Bukit kapur pelalangan arosbaya  | 42    |
| 5  | Panorama alam bukit geger        | 0     |
| 6  | Wisata hutan mangrove sepulu     | 0     |
| 7  | Bukit lambion beramah            | 0     |
| 8  | Wisata pantai rongkang           | 0     |
Table 2. Preprocessing data

| No | Tourist Attractions                      | Total |
|----|-----------------------------------------|-------|
| 9  | Pasarean makam sultan abdul kadirun     | 0     |
| 10 | Pasarean syaichona choli                | 0     |
| 11 | Pasarean aer mata ebu arosbaya          | 16    |
| 12 | Pasarean sunan cendana kwanyar         | 0     |
| 13 | Makam agung arosbaya                   | 0     |
| 14 | Taman rekreasi kota bangkalan          | 0     |
| 15 | Taman paseban alun-alun bangkalan      | 0     |
| 17 | Pusat kerajinan batik tanjung bumi     | 0     |
| 18 | Jembatan suramadu                      | 1151  |
| 19 | Museum bangkalan                       | 0     |
|    | Total                                   | 1394  |

4.1. Preprocessing data

Data reviews that obtained with Data Miner is not normalized, unstructured and there are many noise. So we need to normalized, cleansing, eliminating numbers, emoticons, punctuation marks, case folding, filtering and tokenizing. We used python for data preprocessing which runs on Google Colabs. Example of preprocessing are shown in table 2.

| Reviews                                            | Data cleansing                          |
|----------------------------------------------------|-----------------------------------------|
| Incredible thrill and heat, there are traders on the edge of the hill, good selfie place, suitable for the holiday place | incredible thrill heat traders edge good selfie place suitable holiday place |

4.2. Data labeling

The next process is opinion class labelling which aims to give positive or negative labels on reviews. In this study, we used Lexicon to calculate score of sentiment. Reviews will be positive labelled if a score greater or equal to 0. Otherwise if a score is less than or equal to 0, it will be negative labelled. The result of the labelling based on lexicon in the review is shown in Fig. 1.

Once we obtained opinion class labelling, than we classified data based on sentiment. The percentages of data classifying is shown in fig. 3. Based on pictures 2 and 3 it is known that of a total of 1394 reviews, there are 1235 positive reviews and 159 negative sentiments.

Figure 1. Data Labeling Results

Figure 2. Class of sentiment
Figure 3. Percentages of class sentiment

4.3. Wordcloud sentiment
Wordcloud is important to find out the word that often appears in the sentiment reviews. In this study, we used wordcloud from positive class as shown in fig.4.

Figure 4. Positive Wordcloud

The next step after labelling process is classification stage using the SVM method. However before the classification process, we must calculate the weighted of words using TF-IDF. TF-IDF is the rate frequency of a term in reviews, while IDF is the relationship between a term in reviews. In this study, the kernel used to classify is a linear kernel. The Dataset is divided into 80% of training data and 20% for data testing. So the total amount for the training data is 1115 reviews, and 279 for data testing.

4.4. Accuracy testing
We used Confusion matrix to measure our proposed models. Confusion matrix is a method to perform accuracy of data mining or decision support system. In the measurement of performance using confusion matrix, there are 4 (four) terms as a representation of classification process results. The four terms are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The True Negative (TN) value is the amount of negative data detected correctly, whereas False Positive (FP) is a negative data but is detected as positive data. Meanwhile, True Positive (TP) is a positive data that is
detected correctly. By using the confusion matrix obtained in this research, will also be calculated accuracy, precision and recall.

| Predictions | Positive | Negative |
|-------------|----------|----------|
| Positive    | 153      | 50       |
| Negative    | 81       | 156      |

Table 3. Confusion Matrix

Based on table 3, we could find out accuracy, precision, and recall by using the value of confusion matrix which calculated as following:

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% = 70.22\%
\]

\[
\text{Precision} = \frac{(TP)}{(TP + FP)} \times 100\% = 65.38\%
\]

\[
\text{Recall} = \frac{(TP)}{(TP + FN)} \times 100\% = 75.36\%
\]

5. Conclusion

Based on the research that has been done, it can be concluded that TripAdvisor users gave many positive reviews related to tourism in Bangkalan district. Positive sentiment percentage is 89% while the negative sentiment is 11%. Data Reviews that have been labeled positive and negative then classified using the Linear Support Vector Machine (SVM) method, has an accuracy 70.22% for Indonesian-language reviews.

Acknowledgments

We thanks to University of Trunojoyo Madura, which has funded this research through research programs for lecturers.

References

[1] P. Liu, S. Joty, and H. Meng, “Fine-grained opinion mining with recurrent neural networks and word embeddings,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1433–1443.

[2] W. Wang, S. J. Pan, D. Dahlmeier, and X. Xiao, “Coupled multi-layer attentions for co-extraction of aspect and opinion terms,” in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3316–3322.

[3] Y. Wang, M. Huang, L. Zhao, and E. Al, “Attention-based lstm for aspect-level sentiment classification Methods,” in Proceedings of the Conference on Empirical Processing, in Natural Language, 2016, pp. 606–615.

[4] D. Ma, S. Li, X. Zhang, and H. Wang, “Interactive attention networks for aspect-level sentiment classification,” in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 4068–4074.

[5] N. C. Wirawan and P. P. Adikara, “Analisis Sentimen Dengan Query Expansion Pada Review Aplikasi M- Banking Menggunakan Metode Fuzzy K-Nearest Neighbor ( Fuzzy k-NN ),” J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya, vol. 2, no. 1, pp. 362–368, 2018.

[6] Li. Wilianto, T. H. Pudjiantoro, and F. R. Umbara, “Analisis Sentimen terhadap tempat wisata dari komentar pengunjung dengan menggunakan metode Naive bayes Classifer (studi kasus Jawa Barat),” in SNATIF, 2017, pp. 439–448.
[7] S. Fannisa, M. Fauzi, and S. Adinugroho, “Analisis Sentimen Pariwisata di Kota Malang Menggunakan Metode Naive Bayes dan Seleksi Fitur Query Expansion Ranking.,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 8, pp. 2766–2770, 2018.

[8] D. J. Haryanto, L. Muflikhah, and M. A. Fauzi, “Analisis Sentimen Review Barang Berbahasa Indonesia Dengan Metode Support Vector Machine Dan Query Expansion,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 9, pp. 2909–2916, 2018.

[9] G. A. Buntoro, “Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter.,” *Integer J. Maret*, vol. 1, no. 1, pp. 32–41, 2017.

[10] O. Somantri and D. Dairoh, “Analisis Sentimen Penilaian Tempat Tujuan Wisata Kota Tegal Berbasis Text Mining,” *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 2, p. 191, 2019.