Sentiment Analysis on Travel Destination in Indonesia

Ike Pertiwi Windasari, Dania Eridani
Department of Computer Engineering
Faculty of Engineering – Diponegoro University
Semarang, Indonesia
{ike, dania}@ce.undip.ac.id

Abstract— **Tourists usually do an online search of tourist destinations before determining their destinations. The Internet provides media to share information and experiences on tourism through electronic communication channels, such as social media, travel blogs and tourist review sites. The tourism sector has a strong influence from social media in the decision-making process and has big potential for promotions. However, it is difficult to filter the vast amount of information on the Internet. It takes a lot of time to be able to read all the reviews that are on the Internet, but reading only few reviews lead to biased evaluation. Therefore, we proposed an application that can perform sentiment analysis on travel destination reviews, especially in Indonesia.**

Keywords— Support Vector Machine (SVM), Travel Destination, Indonesia

I. Introduction

Based on a survey conducted by the Association of Indonesian Internet Network Providers (APJII) in 2016, as many as 132 million Indonesians from a total of 256 million Indonesians are now connected to the Internet. This number increased almost 51% of the survey conducted in 2014 which amounted to 88 million Internet users in Indonesia [1]. The Internet provides media to share information and experiences on travelers via electronic communication channels, such as social media and tourist review sites.

Sharing information is one among many things that can be done through the Internet. Before making a purchase, someone is likely wanting to know whether the product to buy is good or not. This can be known through reviews of people who have purchased the product before. In promotion, this is called word of mouth, or word of mouth promotion. Through the Internet, this word of mouth becomes more common. Many websites review a product, as well as an online shopping site that provides product review features to guide buyers.

In addition to product purchases, reviews also apply to specify a person's travel destination. Tourists usually do an online search of tourist destinations before determining tourist destinations. The widespread use of the Internet today, has led to a growing number of travel blogs providing travel-related information [2]. Social media also plays an important role in the selection of tourist destinations. The tourism sector has a strong influence from social media in the decision-making process and the potential for large promotions. Dina & Sabou [3] observed that young people now prefer a source from the Internet (travel blogs, travel forums & tourist review sites) to determine tourist destinations as compared to suggestions from friends. In addition to websites, social media such as Instagram also plays a role in the process of determining new tourist destinations.

However, it is difficult to filter the vast amount of information on the Internet. It takes a lot of time to be able to read all the reviews that are on the Internet, but if the reviews are read only a little then the evaluation will be biased. Therefore, we proposed an application that can perform sentiment analysis on travel destination reviews, especially in Indonesia.Sentiment is used to automatically group user reviews into positive or negative opinions.

II. Related Work

Sentiment analysis played a great role in the area of researches done by many; there are many methods to carry out sentiment analysis. Still many researches are going on to find out better alternatives due to its importance in this scenario. There have been some previous studies to analyze sentiments such as Ye's [4] research which compares three classification methods of Naïve Bayes, SVM, N-gram. This research result is that SVM and N-grams are superior than Naïve Bayes in classify the sentiment. Later research by Tan [5] on five methods of learning that is centroid classifier, K-nearest neighbor, winnow classifier, Naive Bayes and SVM show that SVM has the most superior performance.

Another research by Chandani, et al. [6] compare three classification methods of Naïve Bayes, Support Vector Machine, and Artificial Neural Network found out that SVM was the best performing method among the three. Lidya, et al [7] perform sentiment analysis in Bahasa using SVM and K-NN and found that SVM is more accurate but processed in longer time than KNN method.

Other research in sentiment analysis by Bhadane [8] focuses on the various methods used for classifying a text given and put the opinions expressed in it into negative or positive sentiment. They implemented a set of techniques for aspect classification

978-1-5386-3947-4/17/$31.00 copyright 2017 IEEE

and polarity identification of product review using machine learning (SVM) combined with domain specific lexicons. Their results indicate that the proposed techniques have achieved about 78% accuracy and are very promising in performing the tasks.

Balahur [9] in her research present a method to classify the tweets sentiment by extracting tweet features. She utilize a pre-processing stage to normalize the language and generalize the vocabulary used to express sentiment. With that approach, the system can obtain good result even tough using minimal linguistic processing. The use of such generalized features significantly improves the results of the sentiment classification, when compared to the best performing approaches that do not use affect dictionaries.

In this study authors will use SVM classification method to perform sentiment analysis with the topic of tourist destinations in Indonesia.

III. Working Procedures

In this paper, the sentiment analysis of travel destination in Indonesia is obtained through several processes. The analysis sentiment process is shown in Fig. 1.

First, we collect the dataset from ten Indonesia travel destination review in Bahasa from TripAdvisor. Next, we do manual labeling to classify text into positive and negative sentiment. This label will be used as train data in SVM.

The next step is text preprocessing. Text preprocessing is the process of preparing the initial text data and change it into basic text form and eliminating the noise. This stage aims to obtain more optimal calculation results. The steps taken in preprocessing text in this research are the convert emoticon, cleansing and case folding.

Following is feature extraction which is stemming, convert negation, stop word removal, and tokenizing. We then do term weighting. The term weighting method used in this study is using Term Frequency - Inverse Document Frequency (TF-IDF) method. Term weighting is assigning value to each term derived from the training data. The scoring aims to find out how important a term in representing a sentence. The determination of term score in the TF-IDF method is based on the frequency of terms occurrence in a document.

The TF-IDF value of the term can be found using the following equation:

$$Wi = TF(\omega i, d) \times IDF(\omega i) \quad (1)$$

where Wi is the weight of word ωi in document $d \in D$, $TF(\omega i, d)$ is the term frequency, or the number of ωi in d, and

$$IDF(\omega i) = \log(|D|/DF(\omega i)) \quad (2)$$

where IDF is the inverse document frequency and $DF(\omega i)$ represents the appearance of ωi in D. The largest value of $IDF(\omega i)$ occurs when ωi appears only in one document and its effect is particularly substantial.

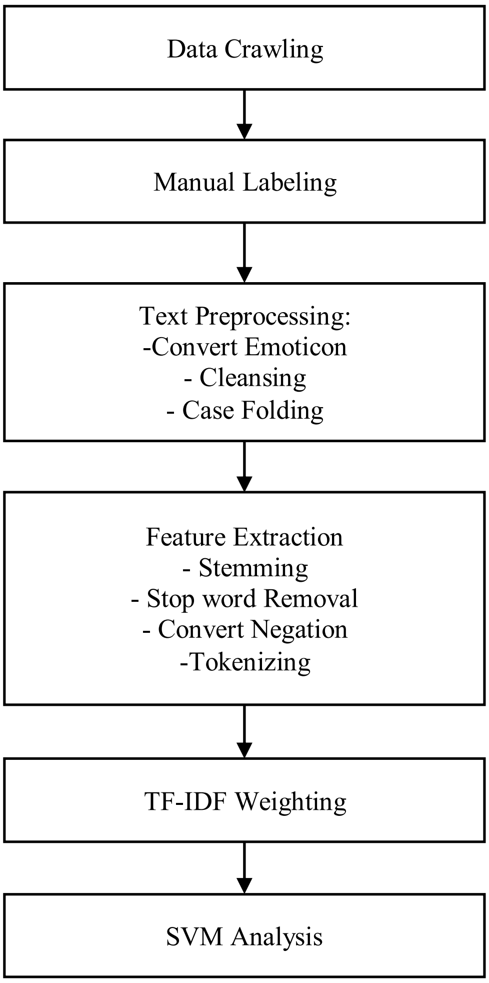


Fig. 1. Sentiment Analysis Process

We use Support Vector Machine (SVM) in this research to analyze the dataset provided. In this research, we divide the data into two classes: positive sentiment class and negative sentiment class. Every data included in the positive sentiment class have the label "1" while the negative sentiment class has the label "-1".

IV. Result and Discussion

The system has gathered review in Indonesian language from ten travel destination in Indonesia as a training data. This program use PHP and utilized LibSVM library by Ian Barber. LibSVM is a library that supports SVM functions such as training and classification. The input data format for SVM processing is an array.

The index page where user can select the destination is shown in Fig 2.

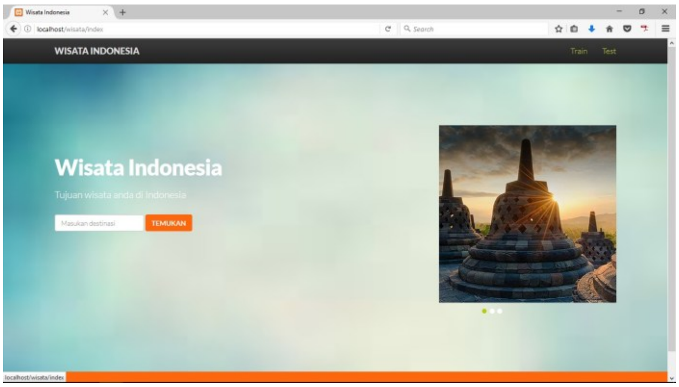
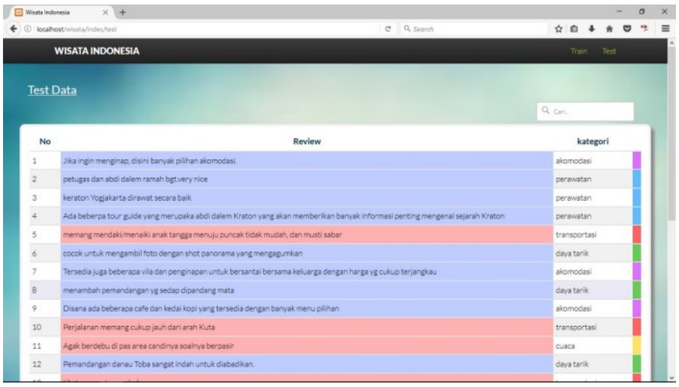


Fig 2. Index Page

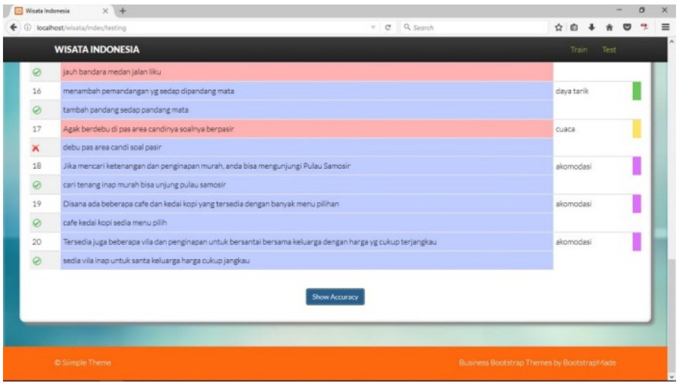
After user select the destination area in Indonesia and click enter, user will be shown selected review from TripAdvisor that have been collected before about the destination chosen. Fig. 3. indicate the training data with manual labeling and Fig. 4 show the classification result using SVM.



Test Data

No	Review	kategori
1	Jika ingin menginap, disini banyak pilihan akomodasi.	akomodasi
2	petugas dan abdi dalem ramah bgt.very nice	perawatan
3	keraton Yogjakarta dirawat secara baik	perawatan
4	Ada beberpa tour guide yang merupaka abdi dalem Kraton yang akan memberikan banyak informasi penting mengenai sejarah Kraton	perawatan
5	memang mendaki/menaiki anak tangga menuju puncak tidak mudah, dan musti sabar	transportasi
6	cocok untuk mengambil foto dengan shot panorama yang mengagumkan	daya tarik
7	Tersedia juga beberapa vila dan penginapan untuk bersantai bersama keluarga dengan harga yg cukup terjangkau	akomodasi
8	menambah pemandangan yg sedap dipandang mata	daya tarik
9	Disana ada beberapa cafe dan kedai kopi yang tersedia dengan banyak menu pilihan	akomodasi
10	Perjalanan memang cukup jauh dari arah Kuta	transportasi
11	Agak berdebu di pas area candinya soalnya berpasir	cuaca
12	Pemandangan danau Toba sangat indah untuk diabadikan.	daya tarik

Fig 3. Training Data



✓	jauh bandara medan jalan liku	
16	menambah pemandangan yg sedap dipandang mata	daya tarik
✓	tambah pandang sedap pandang mata	
17	Agak berdebu di pas area candinya soalnya berpasir	cuaca
✗	debu pas area candi soal pasir	
18	Jika mencari ketenangan dan penginapan murah, anda bisa mengunjungi Pulau Samosir	akomodasi
✓	cari tenang inap murah bisa unjung pulau samosir	
19	Disana ada beberapa cafe dan kedai kopi yang tersedia dengan banyak menu pilihan	akomodasi
✓	cafe kedai kopi sedia menu pilih	
20	Tersedia juga beberapa vila dan penginapan untuk bersantai bersama keluarga dengan harga yg cukup terjangkau	akomodasi
✓	sedia vila inap untuk santa keluarga harga cukup jangkau	

Fig 4. Classification Result

We used sample of 30 reviews which is then tested manually for accuracy, recall, and precision.

A. Accuracy

Accuracy is the overall correctness of the model and is calculated as the sum of true positive and true negative divided by the total number of classifications.

Accuracy of system = ((12+5) / 20) x 100% = 85 %
Misprediction Rate = ((3+0)/20) x 100% = 15 %

B. Recall

Recall measures the completeness, or sensitivity, of a classifier. Higher recall means less false negatives, while lower recall means more false negatives. Improving recall can often decrease precision because it gets increasingly harder to be precise as the sample space increases.

Recall (Positive) = (12/12) x 100 % = 100 %
Recall (Negative) = (5/8) x 100 % = 62,5 %

C. Precision

Precision measures the exactness of a classifier. A higher precision means less false positives, while a lower precision means more false positives. This is often at odds with recall, as an easy way to improve precision is to decrease recall.

Precision (Positive) = (12/15) x 100 % = 80 %
Precision (Negative) = (5/5) x 100 % = 100 %

V. Conclusion

We preprocess the text obtained from TripAdvisor review in Indonesia travel destination in Bahasa, to eliminate the noise, then we use n-gram unigram and TF-IDF as feature extraction methods and use SVM algorithm for classification methods. The result is a prediction of dataset that is considered to be a positive or negative sentiment. From the test results obtained an accuracy of 85% which is considered good result. The positive sentiment recall value is 100% and negative sentiment recall value is 62.5%. The precision value for positive sentiment is 80% and precision value for negative sentiment is 100% which is good but we need to improve train data for positive sentiment to improve accuracy and precision.

In the future, we would like to improve the accuracy of our study by performing larger scale experiments by having larger data set and collect review from more website. We would also like to categorize each review into different categories suitable to travel needs such as weather, peak season, expected expenditure, hospitality, etc.

Acknowledgement

This research is supported by Strategic Research Fund of Faculty of Engineering, Diponegoro University.

References

[1] Widiartanto, Y. H. *Pengguna Internet di Indonesia Capai 132 Juta*. 2016. Accessed on 10 April 2017

[2] Ayeh, J. K., Leung D., Au N., Law R. *Perceptions and strategies of hospitality and tourism practitioners on social media: an exploratory study*, Information and communication technologies in tourism 2012, pp. 1-12. 2012

[3] Dina R., Sabou G. *Influence of social media in choice of touristic destination*. Cactus Tourism Journal. 3(2), pp. 24-30. 2012

[4] Ye, Q., Zhang Z., Law B. *Sentiment classification of online reviews to travel destinations by supervised machine learning approaches.* Expert Systems with Applications, 36, pp. 6527–6535. 2009

[5] Chandani V., Wahono R. S. Purwanto. *Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film.* Vinita Chandani Journal of Intelligent Systems, 1(1), pp. 55-59. 2015

[6] Tan, S., & Zhang, J. (2008). *An empirical study of sentiment analysis for chinese documents.* Expert Systems with Applications, 34(4), pp. 2622–2629. 2015

[7] Lidya, S.K., Sitompul, O.S., Efendi, S. *Sentimen analysis pada teks bahasa Indonesia menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN).* Seminar Nasional Teknologi Informasi dan Komunikasi. 2015.

[8] Balahur, A., “Sentiment Analysis in Social Media Texts,” pp. 120–128, June 2013.

[9] Bhadane, C., Dalal, H., & Doshi, H., “Sentiment analysis: Measuring opinions,” Procedia - Procedia Computer Science, *45*, pp. 808–814. 2015.