Open Access Journal of Information Systems (OAJIS)

Available at http://is.its.ac.id/pubs/oajis/index.php

The Third Information Systems International Conference

Sentiment Analysis to determine Accommodation, Shopping and Culinary Location on Foursquare in Kupang City

Paulina Aliandu

Widya Mandira Catholic University, Kupang 85225, Indonesia

Abstract

The goal of this research is to determine the sentiment of accommodation, shopping and culinary location using Naive Bayes method. Research's object are the tips of accommodation, shopping and culinary location located in Kupang City, East Nusa Tenggara, Indonesia. Tips is a term on Foursquare, that written by user and it is a short messages about a location which let other users know what is good (or bad) there. Tips are taken from Foursquare. We used Naive Bayes in order to generate probability classifier model. This model used to determine sentiment of the Foursquare's tips about accommodation, shopping and culinary location. Test set accuracy is 66,22%. The result showed that Swiss Belinn Hotel, Aston Hotel are the best choices based on the tips following with On The Rock Hotel also T-More Hotel and Lounge. KFC Flobamora Mall, Palapa Resto and Café, Dapur Nekamese are the best choices for culinary. The result showed that Flobamora Mall is the best shopping place.

Keywords: sentiment analysis; naïve bayes; foursquare;

1. Introduction

The total number of foreign tourist arrivals to Indonesia in 2014 was 9.44 million, up 7.19 percent from the preceding year, meaning that the government target of welcoming 9.3 million foreign tourists last year was achieved. As usual most foreign tourist entered Indonesia through Ngurah Rai, Soekarno-Hatta or Batam airport and after that they entered each airport of their destination. Kupang City, the capital of Nusa Tenggara Timur (NTT) is the gate for the coming of tourist from outside province. Kupang has been

* Paulina Aliandu. Tel.: +6281237455405; fax: +62-380-831194
E-mail address: paulinaaliandu@gmail.com.

growing with the presence of various accommodation, culinary, sport, and art facilities. Those facilities provided make the tourist and civil feel good impression and decide to stay longer and back again to Kupang.

Badan Pusat Statistik (BPS) of NTT on its release stated that in 2010 there were 51 accommodations listed in Kupang and has been growing 3 accommodations in next year. That number is quite significant for Kupang that still in growing. BPS in 2011 stated that there were 235 restaurants in Kupang. It is amount number culinary places for Kupang. It's not easy for tourists that spent several days in Kupang to find best culinary location. According to BPS NTT, in 2011, 332.676 visitors came to NTT, which 50.170 visitors are foreign tourists and 282.506 are domestic's visitors. The number of foreign tourists entered Kupang were 8910 tourists and 137.629 domestic visitors (VHTS 2011 of BPS NTT – monthly hotels survey). If the government can provide the best service for those visitors, so it will leave good impression that can bring back them to Kupang. In order to help those visitors to choose the best accommodation, restaurant, and shopping places from various places and facilities need efforts, survey and it will take time. Market research needed to find public opinion on those choices.

Text media in communication on various review online sites, private blog, and social media sites such as Twitter, Foursquare, and Facebook has changed the way to do market research. One of the contemporary methods on market research is sentiment analysis. Opinion mining or sentiment analysis is a computational research about opinion, sentiment and emotion expressed in text (Liu, 2010). Foursquare, it is a local search and discovery service mobile application which provides search results for its users. Foursquare enabled a user to share their location with friends via the check-in- a user manually tell the application when they were at a particular location using a mobile website, text messaging, or a device specific. Check-in on Foursquare happened when user tell Foursquare where they are.

Pang et al. (2002) used machine learning in order to classify movie reviews. The research classified sentiment on movie review and determined whether the review has positive or negative sentiment. Different feature of movie review extracted and used Naïve Bayes algorithm and Support Vector Machine (SVM) to generate classification model. Aliandu (2013) also determined sentiment of Indonesian President Timeline on twitter by using 3 classes (positive, negative and neutral). This research tried to determine the sentiment of the tips on Foursquare about accommodation, culinary and shopping location in Kupang. This research used Naïve Bayes because it has high speed and accuracy when implement on large size and variation data (Han and Kamber, 2006). Larose (2012) also said the same that Naive Bayes has a high rate and accuracy when applied in big and various database. Other reasons why used Naive Bayes in this research are it fast to train and fast to classify, it is not sensitive to irrelevant features, also handles real and discrete data. The disadvantage of Naive Bayes is it assumes independence of features.

This research refined pre-processing of Aliandu (2012) by adding more stopwords, vocabulary of slangs including local language (such as *sonde* (means: *tidak*) and filtering symbol. This research used training data of Aliandu (2012). It followed what Aliandu (2012), Pak and Paurobek (2010) did in their previous researches by using 3 target classes: positive sentiment, negative sentiment and neutral class.

2. Model, Design, Analysis and Implementation

This research designed and engineered system that has ability to determine tips' sentiment on Foursquare.

2.1. Corpus and Feature Extraction

This research used corpus that has been built by Aliandu (2012). Feature extraction followed Aliandu (2012) and refined its pre-processing by adding stopwords, vocabulary of slangs including local language (such as *sonde* (means: *tidak*) and filtering symbol. The steps of feature extraction started from casefolding, filtering, tokenization, slang replacement and finished with stopword removal. This research

used TF-IDF in order to represent indexing document. Indexing is a process to generate document representation by giving a flag on text items (Salton and McGill, 1989). TF-IDF itself is a combination of term frequency and inverse document frequency. Manning (2009) define TF-IDF as a weighted scheme that give term t a weight in document d; the formula of TF-IDF see equation (1).

$$tf - idf_{t,d} = tf_{t,d}\ x\ idf_t \quad (1)$$

2.2. Naïve Bayes

Naïve Bayes Method or Naïve Bayes Classifier use probability theory. Naive Bayes method used conditional probability as a platform. This method is an induction inference statistical for classification problems. Let A and B events in a sample space. Larose (2006) said that conditional probability as shown in equation (2).

$$P(A\ |\mathrm{B}) = \frac{\mathrm{P\ (A \cap B)}}{\mathrm{P\ (B)}} \quad (2)$$

Where $\mathrm{P(A \cap B)}$ intersection probability A and B, P (B) is is probability B. We have $(B\ |\mathrm{A}) = \frac{\mathrm{P\ (A \cap B)}}{\mathrm{P\ (A)}}$, then we obtain $\mathrm{P(A \cap B)} = P(B\ |\mathrm{A})$ P(A). Value of $\mathrm{P(A \cap B)}$ then substituted to equation (1), we obtain equation (3).

$$P(A|\mathrm{B}) = \frac{P(B\ |\mathrm{A})\ \mathrm{P\ (A)}}{P(B)} \quad (3)$$

Let θ represent the parameters of unknown distributions. Larose (2006) stated posterior distribution in equation (4).

$$P(\theta|\mathrm{X}) = \frac{P(X\ |\theta)\ \mathrm{P}(\theta)}{P(X)} \quad (4)$$

Where $P(X\ |\theta)$ is likelihood function. $\mathrm{P}\ (\theta)$ is prior distribution and P(X) is normalization factor called marginal distribution of data. HMAP terminology (Hypothesis Maximum a Posteriori Probability) stated that hypotheses taken from probability value based on known prior condition. HMAP is a simplified of Bayes method called Naive Bayes. Larose (2006) stated Bayes equation for HMAP in equation (5) and equation (6).

$$\theta_{map} = argmax_\theta\, P(\theta|\mathrm{X}) = argmax_\theta\, \frac{P(X\ |\theta)\ \mathrm{P}(\theta)}{P(X)} \quad (5)$$

So then equation (5) can be written as equation (6) since it is the argument (value) that maximizes $P(\theta|\mathrm{X})$ over all θ.

$$\theta_{map} = argmax_\theta\, P(X\ |\theta)\ \mathrm{P}(\theta) \quad (6)$$

Bayes theorem when implemented on document classification d by Manning et al. (2009) written in equation (7).

$$P(c|d)\ \alpha\ P(c) \prod_{1 \le k \le n_d} P(t_k\ |\mathrm{c}) \quad (7)$$

Where $P(c|d)$ is the probability of a document d being in class c. $P(t_k|\mathrm{c})$ is the conditional probability of term t_k occurring in a document of class c. P (c) is the prior probability of a document occurring in class c.

In text classification, the goal is to find the best class for the document or sentence. Manning et al. (2009) mentioned that the best class in Naive Bayes classification is the most likely or maximum posteriori (MAP) class c_{map}, written in equation (8). Equation (8) used to determine the best class for TF-IDF feature.

$$class(t_i) = argmax_{c \in C}\ [log\ \hat{P}(c) + \sum_n f_{ni} log \hat{P}(t_k|\mathrm{c})] \quad (8)$$

$\hat{P}(c)$ and $\hat{P}(t_k|\mathrm{c})$ are the values that we got from training data. By using TF-IDF term weighting we estimate that two values. Kibriya et al. (2005) wrote that maximum a posteriori (MAP) of class c_{map} by using TF-IDF feature in equation (9).

$$class(t_i) = argmax_{c \in C}\ [log\ \hat{P}(c) + \sum_n f_{ni} log \hat{P}(t_k|\mathrm{c})] \quad (9)$$

Prior probability for TF-IDF feature defined in equation (10).

$$\hat{P}(c) = log\ \frac{N_c}{N'} \quad (10)$$

3. Results and Analysis

Classification model built by using corpus from Aliandu (2012), with additional on refined preprocessing. There are 3 classes target : positive, negative and neutral sentiment. This new preprocessing has additional on slangs and local language, vocabulary and filtering symbols. The corpus trained to gain new model classifier. Performance of classifier used accuracy.

3.1. Feature Extraction of Tips

Foursquare eschews the traditional concept of letting users leave long form reviews, and instead encourages the writing of 'tips' – short messages about a location which let other know about what is good (or bad) there. Tips are limited to 200 characters in length, but can include a URL to link an external site with more information. The process of obtaining clean data (pre-processing) as follows:

a. Case folding. This stage changes all the capital letter in text becomes lower case.
b. Removed URL link
c. Filtering, by eliminating illegal character in such as %, /,* and so on.
d. Removed any replacing slang words with formal words listed in local dictionary. This research refined local dictionary by adding more local language such as *sonde* became *tidak*, *besong* menjadi *kalian* and soon.
e. Stoplist removal, by removing character and words listed as stopwords or words with high frequency availability (such as "dan" and "yang").

Testing scenario followed what Go et al. (2009) did. Test set data collected manually, data were took from Foursquare tips about accommodation, culinary and shopping places in Kupang from beginning tips on its location until 5 June 2014. 408 positive sentiment tips, 123 negative sentiment tips and 152 neutral sentiment tips have been manually annotated. List of Foursquare's location that used to build test set, see Table 1. Total column is a sum of positive, negative and neutral sentiment that already manually annotated.

Table 1. List of some Fourssquare's location

Query	Manually annotated			Classification result			Total	Category
	Pos.	Neg.	Neu.	Pos.	Neg.	Neu.		
Swiss-Belinn Kristal	22	10	12	18	16	9	43	Hotel
Aston Kupang Hotel	3	2	1	4	1	1	6	Hotel
Hotel on The Rock	9	1	1	7	2	2	11	Hotel
T-More Hotel and Lounge	9	0	3	7	1	4	12	Hotel
Hotel Sasando	4	2	2	5	3	0	8	Hotel
Hotel Astiti	8	4	5	7	3	7	17	Hotel
Hotel Sylvia	7	5	0	5	4	3	12	Hotel
KFC Flobamora Mall	39	9	10	27	19	10	58	Culinary
Rotterdam Steak House	35	1	4	20	7	9	40	Culinary
Borneo Bakery and Café	24	6	4	17	11	5	34	Culinary
Kampung Solor	24	1	7	14	11	7	32	Culinary

Restauran Nelayan	19	1	1	13	2	5	21	Culinary
Jagung Bakar El Tari	12	2	7	10	8	2	21	Culinary
Bambu Kuning	8	2	2	5	4	3	12	Culinary
Solaria Flobamora Mall	7	3	1	6	3	2	11	Culinary
Mokko Donut And Coffee	10	5	2	10	3	4	17	Culinary
Palapa Resto and Café	4	1	1	4	2	0	6	Culinary
Sari Pitaka	5	1	3	3	4	2	9	Culinary
Tanjung Restaurant	9	0	1	6	0	4	10	Culinary
Kit's Restaurant Rica-rica	6	3	0	5	3	1	9	Culinary
Dapur Nakamese	6	0	0	4	0	2	6	Culinary
Flobamora Mall	10	9	18	17	8	11	37	Shopping
Hypermart Kupang	2	13	9	10	8	6	24	Shopping
Gramedia Kupang	4	7	6	8	5	4	17	Shopping
Matahari Supermarket	5	1	2	3	3	2	8	Shopping
Rukun Jaya Swalayan	5	6	3	4	5	5	14	Shopping

3.2. Analysis and Accuracy

Performance of testing result used accuracy with holdout method on training data showed that the result not much different from what Aliandu (2012) got. Accuracy of Naïve Bayes is 77,48% while Aliandu (2012) was 71,21%. The addition of vocabulary, list of slangs, and filtering symbols in pre-processing did not give significant impact.

Testing set is the object of the research that needs to determine its sentiment. Testing set contains of 683 tips. After pre-processing step, the test set left 671 tips. Table 2. showed confusion matrices of test set Accuracy of test set is 66,22%.

Table 2. Confusion matrices of test set

Positive vs. All Matrices			
TF-IDF		Response	
		Positive	Other
Ref.	Positive	223	175
	Other	103	170

Neutral vs. All Matrices			
TF-IDF		Response	
		Neutral	Other
Ref.	Neutral	52	99
	Other	112	408

Negative vs. All Matrices			
TF-IDF		Response	
		Negative	Other
Ref.	Negative	56	66
	Other	125	424

Sum of One vs. All Matrices			
TF-IDF		Response	
		True Pos	True Neg
Ref.	True Pos	331	340
	True Neg	340	1002

Sentiment determination on culinary location, accommodation and shopping at Table 1 shown that Swiss-Belinn Hotel is the best choice for accommodation followed by On The Rock Hotel, T-More Hotel and Lounge with the highest number of positive sentiment. But based on sentiment proportion, it showed that the best choice belongs to Aston Hotel followed by On The Rock Hotel, on the other hand Hotel Ima had the highest negative sentiment. Class proportion counted by dividing count of sentiment in each class with all data about that object.Class proportion consider the high number of negative sentiment.KFC Flobamora Mall has the highest number of positive sentiment but based on sentiment proportion, so the

highest positive sentiment belong to Palapa Resto and Café and then Dapur Nekamese. On the other hand Sari Pitaka had the highest negative sentiment for culinary location. Flobamora Mall has the highest number of positive sentiment but based on sentiment proportion, the highest positive sentiment belong to Gramedia. Matahari Supermarket had the highest negative sentiment for shopping place.

The conclusion and suggestion based on implementation and analysis phase listed below.

3.3. Conclusion

The accuracy of training data after refined pre-processing when using Naïve Bayes with TF-IDF feature is 77,48%. Accuracy on test set is 66,22%. Except tips on Foursquare has unique characteristic compared to other but machine-learning algorithms shown to classify tweet sentiment with good performance like other text data when using Naïve Bayes. Social media generate big data daily so Naive Bayes is the best method applied on big data like that because of its speed faster on training and accuracy. Sentiment classification result showed that Swiss-Belinn Hotel, Aston Hotel are the best accommodation followed by On The Rock Hotel also T-More Hotel and Lounge. Hotel Ima has the highest negative sentiment. KFC Flobamora Mall, Palapa Resto and Café, Dapur Nekamese are the best culinary location. Sari Pitaka has the highest negative sentiment for culinary location. On the other hand Flobamora Mall and Gramedia are the best shopping places but Matahari Supermarket has the highest negative sentiment.

3.4. Further Research

Manual data training labeling can be used in order to see how if it result can be better than using the presence of emoticon in a sentence to express sentiment. Sentiment determination can use Part Of Speech Tagger (POS Tagger).

References

[1] Liu B, *Handbook of Natural Language Processing, chapter Sentiment Analysis and Analysis, 2nd Edition*, 2010.

[2] Pang B, Lee L, Opinion mining and sentiment analysis, *Foundation and Trends in Information Retrieval*, 2(1-2): 1-135, 2008.

[3] Han J, Kamber M, *Data Mining : Concepts and Techniques (2nd edition)*, Morgan Kaufmann Publishers, 2006.

[4] Larose D, 2006, *Data Mining Methods and Models*, John Wiley & Sons, Inc. New Jersey, USA, 2006.

[5] Aliandu, P, Twitter Used by Indonesian President: A Sentiment Analysis of Timeline, *Proceedings of The 2nd Information System International Conference (ISICO)*, 2013.

[6] Aliandu P, Sentiment Analysis on Indonesian Tweet, *Analisis Sentimen Tweet Berbahasa Indonesia di Twitter*, Theses, Universitas Gadjah Mada, 2012.

[7] Pak A, Paurobek P, Twitter as Corpus for Sentiment Analysis and Opinion Mining, *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[8] Salton G, McGill, MJ, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1989.

[9] Manning C, Raghavan P, Schutze H, *Introduction to Information Retrieval*, Cambridge University Press, 2009.

[10] Kibriya AM, Frank E, Phahringer B, Holmes G, Multinomial Naive Bayes for Text Categorization Revisited, *Proceedings of 17th Australian Joint Conference on Artificial Intelligence*, 2004.

[11] Go A, Bhayani R, Huang L, *Twitter Sentiment Classification using Distant Supervision*, CS224N Project Report, Stanford, 2009.