Twitter Social Media Sentiment Analysis in Tourist Destinations Using Algorithms Naive Bayes Classifier

D T Hermanto¹, M Ziaurrahman², M A Bianto³ and A Setyanto⁴

¹,²,³,⁴ Magister Teknik Informatika, Universitas AMIKOM Sleman, Yogyakarta, Indonesia.

E-mail: deditrihermanto@outlook.com

Abstract. The current tourism destination in Indonesia has developed rapidly which is affected by the role of technology that has a major influence on freedom of access in seeking information. A tweet also has the possibility of containing information or condition about the tourist destinations that they will or have visited, such as the experience of visitors in the tour, the visitor's opinion on a tourist spot, and other tourist attractions. The Naive Bayes method employs training data in order to create the probability of each criterion for different classes; nevertheless, the probability values of the criteria can be elevated to produce sentiment analysis towards opinions on twitter. The research has been done by classifying an opinion in the form of comments into two classes, which is positive and negative with the level of accuracy that is influenced by the training process. Based on that, it can be concluded clearly that the public sentiment information to the tourist attractions included in the positive sentiment.

1. Introduction

Indonesia's tourism world is continuing to increase. According to the data from the Central Bureau of Statistics (CBS), the number of foreign visitors during January - April 2017 reached up to 4.20 million people. The number increases by 19.34 percent compared to the same period in the previous year, which was 3.52 million visits. Based on data obtained from the World Economic Forum (WEF) on 6 April 2017, Indonesia's tourism competitiveness ranked 42 in 2017, wherein the previous year Indonesia was ranked 50th in 2015 [1]. It is also worth to know that according to the Central Bureau of Statistics (CBS), the tourism sector is the second largest foreign exchange contributor in 2017, while the sector is still ranked fourth in 2015 [2].

Technology plays an essential role in the tourism industry rises in Indonesia. Nowadays is the digital era, accessing information is extremely easy. Almost all of the Indonesian citizen is connected to the internet using various devices, either mobile, pc or other devices when they want to go or travel [3]. Among those devices, smartphones are the ultimate choice that provides many needs and conveniences when traveling, for instance like a review of the sights, planning, reservations, and many other things like finding a place to eat and a hotel using GPS navigation [3]. Internet network is also a source of information when traveling which is the most commonly used.
Per the development of the Web 2.0 era (social media) has changed the market conditions for tourism [3]. In recent years, social media has grown tremendously fast. Growth from users to facilities offered by the social media. Plenty of social media used by the community, one of them is Twitter which becomes a favorite for Indonesian people [4]. In microblogging services such as Twitter, users tend to post tweets in real time. Tweets often contain significant information from an event, which is very valuable as a crucial tool for finding out the influence of opinions on tourist destinations. Besides, we can know the events or any discussion topics in a city which is related to tourism through a popular hashtag. Moreover, it also can be understood the sentiment or public opinion about tourism in the city.

2. Methodology

This study is an observational study where the data comes from social media. This study process large amount of data frame processes which consisting of several steps. Data collection, data cleaning, feature extraction, modeling, machine learning, and evaluation. The output of this research is the accuracy of food trend prediction compared to restaurant sales data. The results of this study are the trend of web-based food based on social media.

2.1 Preprocessing Data

Preprocessing data is a process of transforming unstructured data forms into structured data based on need, for further mining processes (sentiment analysis, summary, clustering of documents, etc.). In this preprocessing stage, there are four steps has to be done as follows:

1. Selecting comments. In this pre-processing step, the tweet comment without certain hashtag is filtered out, and the comment containing the word or hashtag (#) and the name of the destination location in Yogyakarta city is selected.
2. Data Cleansing. The result of the sentence selection found that there are affixes or conjunction, or random error and variant on the measured variable [5], consequently need to remove characters such as HTML, keywords, emotion icons, hashtag (#), username (@username), URL (http://website.com), and email (name@website.com).
3. Parsing This process was done to break the document into pieces of words for later analysis of the word collection by separating the word and determining the syntactic structure of each word [6].
4. Normalizing Sentences. Normalizing sentences that are inconsistent with KBBI (Indonesian dictionary) Nevertheless that less commonly understood sentences become normal and can be understood easily [7].

Here are some things needed to normalize the sentence:

1. Spaces in punctuation and symbols other than the alphabet.
2. Spaces the punctuation of words after or before, subsequently they can be neatly arranged during the tokenization process.
3. Change all words by lowercase
4. Normalization of words
2.2 Naïve Bayes Classifier

Naïve Bayes Classifier is a classification method rooted in Bayes's theorem. The method of classification employs the probability and statistical methods proposed by the British scientist Thomas Bayes, predicting future opportunities based on experience, is known as Bayes's Theorem. There are several processes performed in this method, such as feature formation, probability calculations for each sentiment class, calculating the maximum Vmap. The formula used in determining the value of the equation naïve bayes classifier as follows:

\[
P(C|F_1, ..., F_n) = \frac{p(C)p(F_1, ..., F_n|C)}{p(F_1, ..., F_n)}
\]

After getting the result of the formation on the feature, afterwards, the next step is to do the calculation of each class. The formula used in determining the value of the calculation equation of each class as follows:

\[
P(C_i) = \frac{fd(c_i)}{|D|}
\]

Information:
fd (ci) = number of documents present in ci

\[|D| = \text{amount of training data}\]

After the formation of the feature is completed, the next step that has to be done is to calculate every possibility that exists in each feature on the sentiment class by doing the equation. The formula used as follows:

\[
p(w_k|c_i) = \frac{f(w_k|c_i)+1}{f(c_i)+|W|}
\]

Information:
f (wki, ci) = value of occurrence wk on ci
f (ci) = total occurrence of keyword ci
\[|W| = \text{number of wk elongation}\]

3. Result and Discussion

In this study, the determination of favorite tourist destinations in Yogyakarta is obtained from Yogyakarta Tourism Statistic Book of 2016 [8]. The destination used is the top 10 ranking of the statistics of the number of visitors. The list of destinations in this study can be seen in Table 1.

| No | Destination Name        |
|----|-------------------------|
| 1  | Pantai Parangtritis     |

Table 1. Tourism Destination
Based on tourist destinations that exist in table 1, then will be conducted opinion search of each tourist destinations on Twitter by determining the dataset in the form of positive words and negative words as a target sentiment of tweet obtained, such as table 2 below.

**Table 2. Result Of sentiment analysis**

| Data | Tweet | Positive | Negative | Value |
|------|-------|----------|----------|-------|
| Pantai parangtritis mempunyai keunikan pemandangan yang tidak terdapat pada objek wisata lainnya | T1 | 1 | 0 | 1 |
| Korban shooting ditengah hari bolong di pantai parangtritis | T2 | 0 | 1 | -1 |
| ini gimana ini bayarnya kok salah pahamwith immfe at pintu masuk kawasan wisata pantai parangtritis | T3 | 0 | 1 | -1 |
| Indahnyoo terimakasih sayang pantai parangtritis yogyakarta | T4 | 1 | 0 | 1 |
| Menikmati sore dipantai parangtritis | T5 | 1 | 0 | 1 |
| Perjuangan siang cetar ini dibayar kelapa | T6 | 2 | 0 | 2 |
muda segar.
Bisa buat cerita
w Bayu at Pantai
Parangtritis

![Pie Chart Sentiment Analysis Result](image)

**Table 3.** Sentiment analysis result

| No | Destination                | Average  |
|----|----------------------------|----------|
|    |                            | Negative | Neutral | Positive |
| 1  | Pantai Parangtritis        | 0.18     | 0.54    | 0.28     |
| 2  | Kawasan Pantai Baron       | 0.18     | 0.60    | 0.23     |
| 3  | Candi Prambanan            | 0.10     | 0.69    | 0.21     |
| 4  | Gembira Loka               | 0.00     | 0.25    | 0.75     |
| 5  | Taman Pintar               | 0.08     | 0.32    | 0.60     |
| 6  | Kaliurang                  | 0.28     | 0.59    | 0.12     |
| 7  | De Mata Art Museum         | 1.00     | 0.00    | 0.00     |
| 8  | Hutan Pinus                | 0.13     | 0.76    | 0.11     |
| 9  | Kraton Yogyakarta          | 0.23     | 0.61    | 0.16     |
| 10 | Musium Benteng Vredeburg   | 0.19     | 0.53    | 0.27     |

3.1 Train Data Formation

The formation of train data with keywords that will be the parameter of the unit is the opinion of twitter to tourist destinations that have been determined by its value. The result value is neutral (0), positive (1) and rated negatively (Value -1). Table 4 shows the data training was done by the pre-processing process beforehand.

**Table 4.** Train data Formation

| Data | Emergence | Sentiment Class |
|------|-----------|-----------------|
| T1   | Unik [1]  | Positive        |
| T2   | Korban [-1]| Negative       |
| T3   | Salah [-1]| Negative       |
| T4   | Indah [1]| Positive        |
### 3.2 Calculating Probability

After the founding of the feature, the next stage is to calculate probability from each destination. The results of the probability of the naïve Bayes training class can be seen in table 5.

**Table 5. Probability Class Training**

| Sentiment Class | Data (j) | fd(cj) | p(cj) |
|-----------------|----------|--------|-------|
| Positive        | T1 0 T2 0 T3 1 T4 1 T5 1 T6 4 | 4/6   |
| Negative        | T1 0 T2 1 T3 0 T4 0 T5 1 | 2/6   |

### 3.3 Determining Train Data Probability

After calculating the probability of each class, the next step is to determine the probability of each training data in each class. The result of probability can be seen in table 6.

**Table 6. Probability Of Exercise Data**

| Data f (Wki, Ci) | Sentiment Class | Positive | Negative |
|------------------|-----------------|----------|----------|
|                  | Unik            | 1+1/5+7=2 | 0+1/5+7=1 |
|                  | Indah           | 1+1/5+7=2 | 0+1/5+7=1 |
|                  | Korban          | 0+1/5+7=1 | 1+1/5+7=2 |
|                  | Salah           | 0+1/5+7=1 | 1+1/5+7=2 |
|                  | Nikmat          | 1+1/5+7=2 | 0+1/5+7=1 |
|                  | Cetar           | 1+1/5+7=2 | 0+1/5+7=1 |
|                  | Muda            | 1+1/5+7=2 | 0+1/5+7=1 |
Table 7. Classification Test Data

| Comment                                                                 | Before preprocessing | After preprocessing |
|------------------------------------------------------------------------|----------------------|---------------------|
| pantaiparangtritismempunyaikeunikanpemandangan                         | Parangtritis         | Unik                |
| yang tidakderapatpadaobjekwisatalainnya                                |                      | Wisata              |

From the flow of the classification process, it has several important stages such as follows:

Calculate the Vmap to be used naive bayes classification in the determination of probabilities based on the learning process on the test data of each class, to obtain the greatest probability value. The calculation to be used is as follows:

\[ V_{\text{map}} = \arg \max (\text{positive}) \times p(w_k | c) \times p(c) \]

Based on the results of the training following the results of the calculation:

\[ V_{\text{map}} = \arg \max (\text{positive}) \times p(w_k | c) \times p(c) \]

\[
p(\text{“parangtritis”}|c) \times p(\text{“unik”}|c) \times p(\text{“wisata”}|c)
\]

1. Vmap for positive sentiment
   \[ V_{\text{map}}(\text{“positive”}) = p(\text{“positive”}) \times p(\text{“parangtritis”}|\text{positive}) \times p(\text{“unik”}|\text{positive}) \times p(\text{“wisata”}|\text{positive}). \]
   \[ = \frac{4}{6} \times \frac{1}{12} \times \frac{2}{12} \times \frac{1}{12} = 0.0007716 \]

2. Vmap for positive sentiment
   \[ V_{\text{map}}(\text{“negative”}) = p(\text{“negative”}) \times p(\text{“parangtritis”}|\text{negative}) \times p(\text{“unik”}|\text{negative}) \times p(\text{“tourism”}|\text{negative}). \]
   \[ = \frac{2}{6} \times \frac{1}{9} \times \frac{1}{9} \times \frac{1}{9} = 0.0004572 \]

Determining the maximum Vmap, from the results of the calculations that have been done before, it will produce a positive Vmap value more significant than the negative Vmap. Nevertheless, it can be concluded that public opinion about tourist destinations in Yogyakarta are classified into positive sentiments.

4. Conclusion

Based on the research, it can be concluded that:

1. The Naive Bayes method employs training data to produce the probability of each criterion for different classes, subsequently the probability values of the criteria can be optimized to generate sentiment analysis on Twitter opinions on tourist destinations in Yogyakarta based on
the classification process that has been done consequently as to classify an opinion in the form of comments into two classes that is positive and negative with the level of accuracy that is influenced by the training process. Based on the findings, this can be concluded clearly that the public sentiment information on the tourist attractions included in the positive sentiment.

2. For further research, it is expected to be able to apply other algorithms to get more accurate results in decision making as recommendations for tourist attractions.

Acknowledgments
First of all, We would like to give our gratitude to Allah SWT Who has guided us until today. His blessing and guidance have led us to live this life, especially to accomplish this papers. Then, We would like to thanks to our University of Amikom Yogyakarta that give us a chance. We also would like to thank our beloved classmate Informatics Engineering Postgraduate program, Big data course in the University of Amikom Yogyakarta.

5. References

[1] Anonym 2018 Begini Perkembangan Pariwisata Indonesia 2018 Era Jokowi (tagarnusantara: Indonesia)
[2] E Prima 2017 Luhut: Pariwisata Jadi Penyumbang Devisa Nomor Wahid 2019 (Bisnis Tempo: Indonesia)
[3] R Minazzi 2015 Social Media Marketing in Tourism and Hospitality (Springer International Publishing: Switzerland) pp 127-135
[4] S D Prihadi 2015 Berapa Jumlah Pengguna Facebook dan Twitter di Indonesia? (CNN Indonesia)
[5] I A Putu 2014 (Formaldehid Journal: Bali) Studi Analis Kesehatan ,Sekolah Tinggi Kesehatan Wira Medika PPNI.
[6] M Y N S Putro 2011 Analisis Sentimen pada Dokumen berbahasa Indonesia dengan Pendekatan Support Vector Machine (Masters: Binus)
[7] N Adiyasa 2011 Analisis Sentimen Pada Opini Berbahasa Indonesia Menggunakan Pendekatan Lexicon-Based (Catatan Kecil : Adiyasan)
[8] Dinas Pariwisata Daerah Istimewa Yogyakarta 2016 Statisti Keparwisataan 2016 (Visit Jogja: Indonesia)