Analyzing Tourism Information on Twitter for a Local City

Kazutaka Shimada, Shunsuke Inoue, Hiroshi Maeda and Tsutomu Endo
Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502 Japan
{*shimada, s_inoue. h_maeda, endo*}*@pluto.ai.kyutech.ac.jp*

***Abstract*—Tourism for a local city is one of the most important key industries. The Web contains much information for the tourism, such as impressions and sentiments about sightseeing areas. Analyzing the information is a significant task for tourism informatics. In this paper, we propose a tourism information analysis system for a local city. The target resource for the analysis is information on Twitter. First, we discuss a method to extract tweets (posted sentences) relating to the target locations and tourism events. Then, we analyze the polarity of the extracted tweets; positive or negative opinions. It is well-known as a P/N classification task in natural language processing. For the process, we employ an unsupervised machine learning approach that uses seed words. We evaluate and consider the extraction and P/N classification tasks. The experimental result about P/N classification shows the effectiveness of our method.**

***Keywords*-Sentiment Analysis, Tourism information on the Web, Twitter**

I. Introduction

Tourism for many local cities is one of the most important key industries. The activation of tourism leads to the activation of the local industries and communities. In this situation, the World Wide Web plays a large role [8]. Although a huge number of online documents are easily accessible on the Web, the quality of the information is a mixture of the good and bad. Finding important information relevant to the target needs has become increasingly significant. We develop a tourism information analysis system which extracts information about tourism from the Web, analyzes the extracted information in various perspectives, and visualizes the output of the analysis. Figure 1 shows the outline of the system. By using this system, people involved in the tourism can easily understand and organize significant information of the target city. The target city is Iizuka city in Fukuoka, Japan, where our university located. It is a medium-size city with population of approximately 130,000.

In this paper, we propose fundamental technologies for the tourism information analysis system. They are (1) information extraction of tourism information from the Web and (2) sentiment analysis of the extracted information. First, we explain a basic idea for the extraction process, and then consider the output of the process. Next, we discuss a sentiment analysis task. The sentiment analysis is one of the hottest topics in natural language processing [5]. Here we handle a P/N classification task. P/N classification is to classify an input sentiment into positive (P) or negative (N) opinions. In this paper, we apply an unsupervised machine learning approach based on a naive bayes classifier to the classification. The target data is extracted from Twitter[1]. It is one of the most famous microblogging services and text-based posts of up to 140 characters. The posted sentences are described as "tweets". In microblogging services such as Twitter, users tend to post tweets in real time. It denotes that tweets often contain significant information of events for tourism as lifelog data.

II. Tourism information extraction

In this section, we describe a method to extract tourism information.

A. Method

The extraction process is basically as follows:

Step1: Acquisition of basic queries,
Step2: Selection of related words,
Step3: Query generation and retrieval,
Step4: Filtering.

1) Basic query and related word:

The basic information for this process is extracted from portal sites for tourism which the city and tourist association established. Here "basic query" denotes tourist facilities, restaurants and events, such as festival, which are written in the portal sites. Figure 2 shows an example of the tourism portal site about Iizuka city[2]. It consists of (1) facility or event names, (2) a link to detailed information of each entry and (3) basic information of each entry. We define the facility or event names as the basic queries. For example, "Kaho performing theater" and "Ito Den-emon residence" are basic queries. The number of basic queries is approximately 200 words.

We need to consider a problem of basic queries. Sightseers do not always mention the basic queries, i.e., facility or event names, in tweets. Moreover, they might mention information which is related to the location or event names and does not appear in the portal site. Therefore, we need to acquire related words of the basic query, i.e., query expansion. First,

[1]http://twitter.com
[2]http://portal.kankou-iizuka.jp

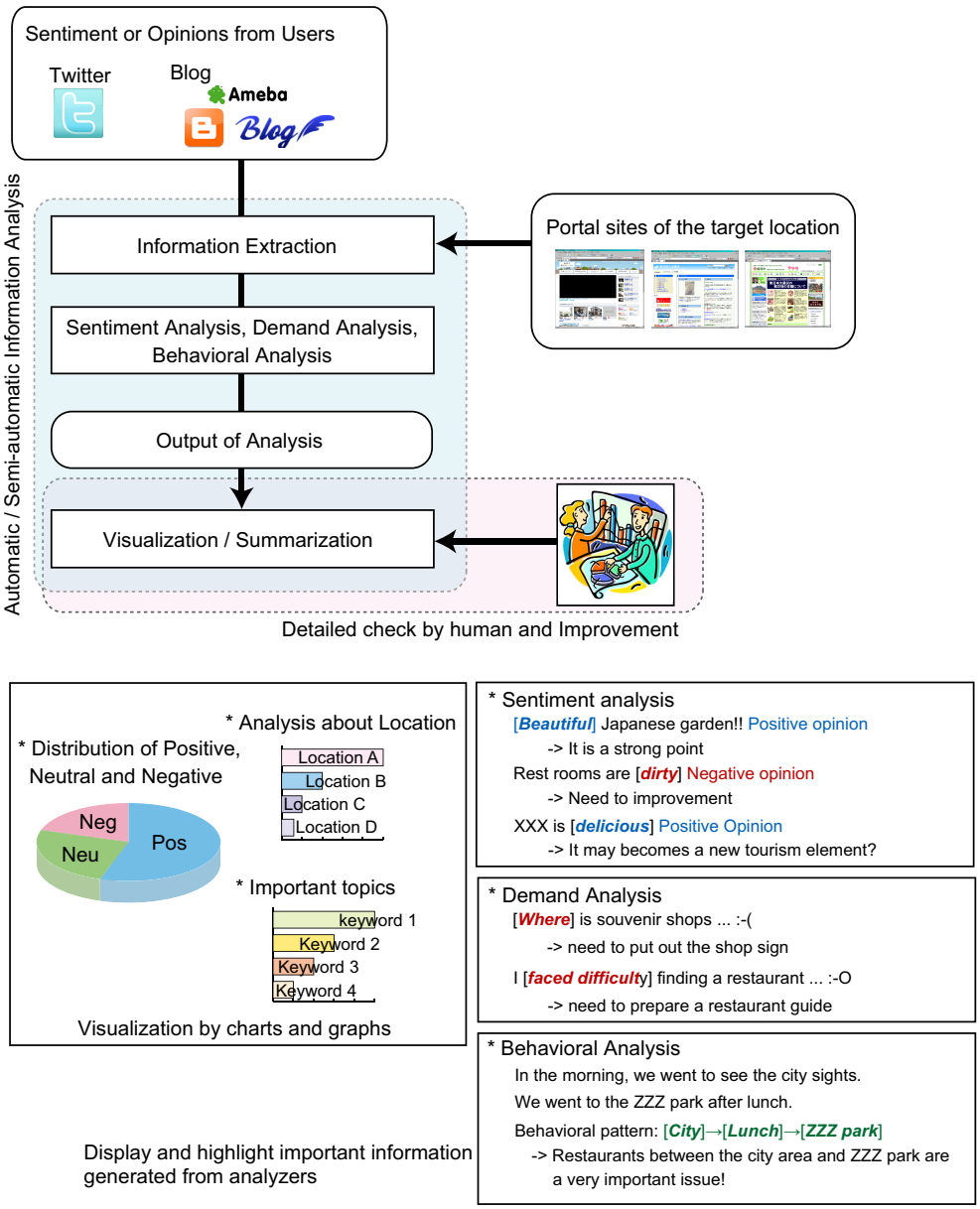

Figure 1. The outline of our prototype system.

we need to divide each sentence into words. For the process we use MeCab[3], which is one of the most famous Japanese language morphological analyzers. For the selection of related words, we introduce a weighting approach, which is well-known as Okapi-BM25 [7]. The importance of a word is computed by

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_i\cdot(1-b+b\cdot\frac{|D|}{avgdl})} \quad (1)$$

[3]http://mecab.sourceforge.net/

where

$$IDF(q_i) = log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \quad (2)$$

$f(q_i,D)$ is the frequency of a word q_i in a document D. $n(q_i)$ and $|D|$ are the number of documents containing q_i and the length of D, respectively. $avgdl$ and n are the average length of documents and the number of documents. b and k are constant factors for weighting.

Figure 3 shows an example of explanation about "Old Ito Den-emon's residence (the 2nd facility in Fig 2)". We obtain a weighted word list from this explanation. In this

Figure 2. The portal site on the Web.

example, appropriate words, such as “Byakuren[4]” and ”Coal mining [5]”, are located in top position in the ranking. Finally we select suitable words from the weighted word list, as related words by hand work. This selection process depends on subjective heuristics. In this process, we tend to select proper names (e.g., Byakuren) , the head of a noun (e.g., emon) and attributes of the target (e.g., Coal mining).

2) Extraction and filtering:

We retrieve tweets with Twitter API by using the manually-produced query list, namely the union of basic queries and related words. Final queries for the retrieval are combination of words in the list. However, there is a problem of the final queries. Tweets do not always contain the official name of facilities or events. For example, “Old Ito Den-emon residence”, which is one of the most famous facilities in Iizuka, is exceedingly-long words. Therefore, it’s unlikely that users input the official name itself. To solve this problem, we manually generate abbreviations of queries. For “Old Ito Den-emon residence”, we add some abbreviations such as “Ito residence” and “emon residence” to the query list.

Next, we apply post-processing to the extraction process; i.e., filtering. Extracted tweets do not always relate to the target city even if they contain queries. For example, the word “Iizuka” is used for not only location, but also person’s name. Therefore, we need to delete noise data such as tweets with person’s name “Iizuka”. For the process, we use the output of MeCab and a rule-based approach. The output

[4]She is a wife of Mr. Ito and famous poet.

[5]Mr. Ito was a rich coal mine owner and is called Coal Mine King

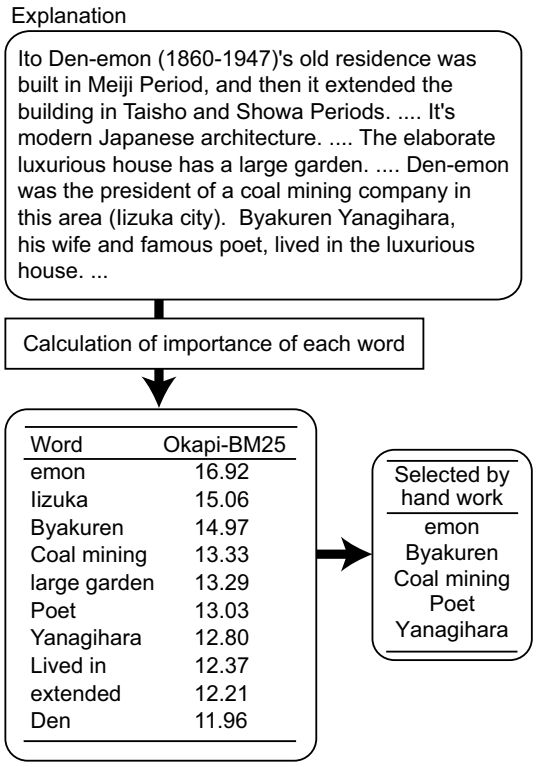

Figure 3. An example of explanation and the related words.

of MeCab includes some indicators as the part-of-speech tag; e.g., [ProperName-person’s name] and [ProperName-location’s name]. By using these tags, we judge whether each tweet is suitable or not. In addition, we prepare some suffix rules[6] manually and utilize them in this process.

B. Discussion

We implemented the approach, and extracted tweets from Twitter. The number of extracted tweets was small in amount and not sufficient. It was caused by the size of the target city. We focused on the medium-size city with some tourism properties. To obtain more information about the target, we need to consider (1) additional queries and (2) use of other resources such as weblogs.

For final query generation, we created abbreviations of queries manually; e.g., “Ito residence” from “Old Ito Den-emon residence”. However, manual generation is costly. Murayama and Okumura [4] have proposed an approach to detect abbreviations using a statistical model. Automatic generation of abbreviations is future work.

We applied a filtering process based on some rules to the extraction. However, the noise in the output remains a huge problem. We encountered another problem although

[6]In Japanese, the suffix “-San” and “-Cho” denote person’s name adn location’s name, respectively.

we dealt with the distinction of person's names and location's names. It was different locations with an identical name. For example, the target city Iizuka has the location name "Yakiyama". This location name exists in another city Sendai. We can not distinguish between them by using only suffix rules. To distinguish between them, we need to handle contextual information. One of context information is the place of residence of each twitter's user. Another context is information in tweets before and after the target tweet. Integrating the contextual information into the filtering process is also important future work.

Although the outputs of the filtering are tweets with information about the target city, they are not always tweets with "tourism" information about the target city. The outputs often contained tweets about a daily occurrence. For example, the tweet "I'm on the way to work now. near Den-emon residence" is not useful information for tourism analysis systems. Therefore we need to judge whether a tweet contains tourism information or not; tourism likelihood estimation. For this problem, we focus on time stamp of tweets. In Twitter, users tend to post tweets in real time[7]. The time stamp of tweets is one of the most important features for the tourism likelihood estimation. For example, posts in early morning and late-evening indicate low probability of tourism information even if a tweet contains a tourism facility.

There is a problem of different events with an identical name. For example, the event "Yamakasa" is a festival that takes place in several locations in Fukuoka prefecture, which Iizuka city belongs to. In this situation, contextual information for location mentioned above is not suitable to identify the location of the event because their locations are marginal. Time information is also suitable to solve this problem because days of each event are usually different.

Tweets in Twitter include characteristic expressions, such as '"nau (now in English)" and "dan (done in English)". These expressions are also effective for the tourism likelihood estimation of each tweet. We will integrate these time features into the filtering process.

III. Sentiment analysis

We implement a sentiment analyzer for the tourism information analysis system. In this paper, we focus on P/N classification, which is to classify an input into positive or negative opinions.

A. Method

Many researchers have studied many approaches to identify P or N of an input. Pang et al. [6] have reported a P/N classification task for movie review documents. They compared several machine learning techniques, and showed the effectiveness on the review data. In general, machine learning techniques generate a high accuracy classifier by using a large amount of training data. Constructing rich training data is, however, costly.

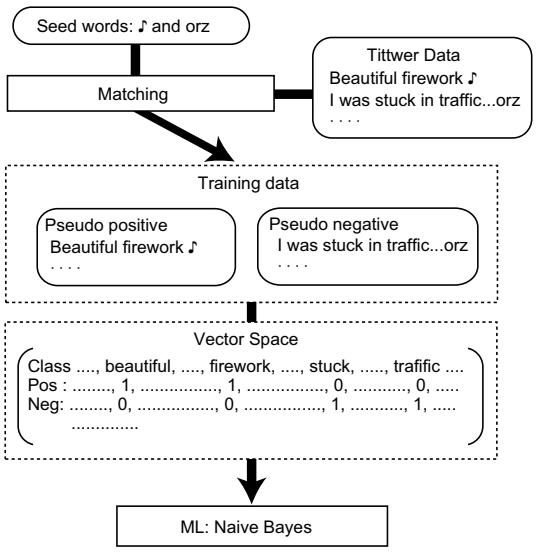

Figure 4. The outline of PN classification.

In this paper, we focus on an unsupervised machine learning approach, which does not need manually annotated training data. It is based on seed words and pseudo training data extracted from a non-tagged corpus. Turney [12] has proposed a method for classifying reviews as recommended or not recommended by using some seed words. He used the words "excellent" and "poor" as the seed words, and computed the semantic orientation of a phrase by using the Pointwise Mutual Information (PMI) between the seed word and the phrase. On the other hand, we use "♪" as the positive seed and "orz[8]" as the negative seed, and extract the pseudo training data by a simple exact match. The reason why we do not use linguistic expressions, such as "good" and "bad", is that tweets on Twitter are informal and often contain many emoticons and symbols.

We utilize the naive bayes method as a means to classify tweets into positive or negative. The naive bayes method is a probabilistic model based on Bayes' theorem.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{3}$$

Here we can use only the numerator of the fraction since the denominator $P(d)$ does not depend on c.

$$\hat{c} = argmax_c P(c) \prod_{i=1}^{n} P(x_i|c) \tag{4}$$

where c is the class (P or N) and x_i is a word in a sentence.

[7]On the other hand, other web services such as weblog, in general, is not real time posting. Users tend to post an entry by a PC after returning home from the travel.

[8]This is a emoticon expressing a person who drops the knee. It usually expresses negative emotion.

Table I
THE ACCURACY RATES OF OUR METHOD AND THE SCORING METHOD.

Method	Accuracy
Our method	0.89
Scoring	0.76

Figure 4 shows the pseudo training data acquisition and the P/N classification method using it. First, our method extracts tweets containing the seed words by an exact match approach. We regard the extracted tweets as the training data for a classifier, and construct the vector space from the data. The naive bayes classifier generates the model from the vector space.

B. Experiment

In this section, we evaluated our method with a test data set. The test data consisted of 116 tweets; 64 tweets as positive and 52 tweets as negative. Our method needs a non-tagged corpus for the pseudo training data acquisition. The corpus consisted of over 10 million tweets. We extracted pseudo positive and negative sentences from them by using two seeds, namely "♪" and "orz". The numbers of pseudo positive and negative sentences were both one hundred thousand.

First, we compared our method with a simple scoring method based on polarity dictionaries. We used a noun polarity dictionary constructed by Higashiyama et al. [1], sentiment lexicon by Kaji and Kitsuregawa [2] and evaluative expressions by Kobayashi et al. [3]. Each entry in each dictionary possesses a real number score[9] The scoring method was based on the summation of the scores[10].

The experimental result is shown in Table I. The accuracy rate is computed by

$$Accuracy = \frac{\text{\# of correct outputs}}{\text{\# of tweets in the test data}} \quad (5)$$

Our method outperformed the scoring method using dictionaries. The reason was that the dictionary-based method could not handle tweets with informal expressions in the test data appropriately. This result shows the effectiveness of our method with the naive bayes classifier using the pseudo training data.

Selecting seed words is one of the most important parts in our method. We also evaluated our method with several seed sets. Table II shows the result. In the table, "bummer" and "not" are linguistic expressions. "+" denotes the combination of seed words. The seed set "P: ♪, N: orz + ;)" produced the best performance. Combining seed words is one approach to improve the accuracy. Linguistic expressions were not effective in this experiment. We need to investigate many seed combinations for high accuracy.

[9]It is positive value if the polarity of the entry is positive. Likewise, it is negative value if the polarity of the entry is negative.

[10]Actually, the ranges of scores in each dictionary were different. Therefore we normalized the scores in this experiment.

Table II
THE ACCURACY RATES IN SEVERAL SEEDS.

Seeds		Accuracy
Positive	Negative	
♪	orz	0.89
^▽^)	orz	0.81
♪	ToT)	0.81
♪	;)	0.87
♪	bummer	0.78
♪	not	0.78
♪	orz + bummer	0.89
♪	orz + ;)	**0.92**
♪ + ^▽^)	orz + ;)	0.89

IV. CONCLUSIONS

In this paper, we proposed fundamental technologies to develop the tourism information analysis system that extracts information about tourism from the Web, analyzes the extracted information in various perspectives, and visualizes the output of the analysis. The target tasks were tourism information extraction and P/N classification of the extracted information. For the extraction, we introduced basic queries and their related words based on an importance measure computed by Okapi-BM25. The number of extracted sentences was, however, small in amount and not sufficient. It was caused that we focused on the medium-size city with some tourism properties. To obtain more information about the target, we need to consider (1) additional queries and (2) use of other resources such as weblogs. Tokuhisa et al. [11] have proposed a method of extracting useful sentiment information in order to obtain useful tourism development hints. They focused on major tourism locations and facilities similar to the target. This idea is suitable for our system.

For the P/N classification, we applied an unsupervised machine learning approach based on a naive bayes classifier. It was based on seed words and pseudo training data extracted from a non-tagged corpus by using the seed words. In the experiment, we compared our method with a dictionary-based method first. The accuracy of our method was 13% higher than that of the dictionary-based method. This result shows the validness of our method based on the pseudo training data acquisition. Then, we compared several seed sets for our method. Suitably-combined seed words increased the accuracy from 0.89 to 0.92. However, the size of our test set was small; approximately 100 tweets. We need to evaluate our method with a large scale data set. Moreover, we dealt with only positive/negative classification. Naturally, there is 3rd class, i.e., neutral sentences, which do not contain sentiment information. Three-class problem, namely positive, neutral and negative, is unavoidable work in the future.

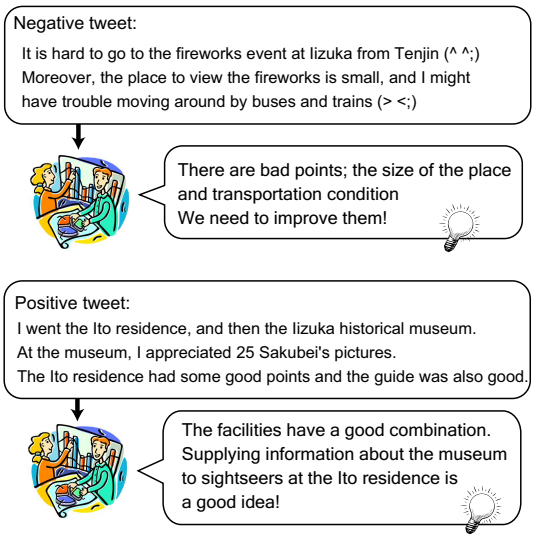

Figure 5. The outputs of our system.

The goal of our task is to develop the tourism information analysis system. Figure 5 shows two actual examples of the output of our system. Our system might help to detect good or weak points of facilities or events by checking the outputs. There are many output styles to visualize the data. We have studied an interactive summarization system of sentiment information [10] and developed a sentiment summarizer that treated objective information [9]. Construction of the tourism information analysis system with knowledge from these researches is the most important future work.

ACKNOWLEDGMENT

This work was supported by a research grant from Iizuka City.

REFERENCES

[1] Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. Acquiring noun polarity knowledge using selectional preferences. In *Proceedings of the 14nd Annual Meeting of The Association for Natural Language Processing*, pages 584–587, 2008.

[2] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-CoNLL2007)*, 2007.

[3] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. Collecting evaluative expressions for opinion extraction. In *In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 584–589, 2004.

[4] Norifumi Murayama and Manabu Okumura. Statistical model for japanese abbreviations. In *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, pages 260–272, 2008.

[5] Bo Pang and Lillian Lee. *Opinion mining and sentiment analysis*, volume 2. Foundations and Trends in Information Retrieval, 2008.

[6] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[7] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *In Proceedings of the Third Text REtrieval Conference (TREC 1994)*, 1994.

[8] Hajime Saito. Analysis of tourism informatics on web. *Journal of the Japanese Society for Artificial Intelligence*, 26(3):234–240, 2011.

[9] Kazutaka Shimada, Ryosuke Tadano, and Tsutomu Endo. Multi-aspects review summarization with objective information. In *Proceedings of the 12th Conference of the Pacific Association for Computational Linguistics (PACLING2011)*, 2011.

[10] Kazutaka Shimada, Masashi Yamaumi, Ryosuke Tadano, Masashi Hadano, and Tsutomu Endo. Interactive aspect summarization using word-aspect relations for review documents. In *Proceedings of the 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS & ISIS 2010),*, pages 189–188, 2010.

[11] Masato Tokuhisa, Hideto Okumura, and Masaki Murata. Sentiment analysis of weblog articles to support tourism development. *Journal of Society for Tourism Informatics*, 7(1):85–98, 2011.

[12] Peter D. Turney. Thumbs up? or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, 2002.