Cross-lingual Investigation of User Evaluations for Global Restaurants

Jiawen LE[†] **Hayato YAMANA**[‡]

[†]Graduate School of Fundamental Science and Engineering Waseda University, 3-4-1 Okubo, Shinjuku-ku Tokyo, Japan

[‡]Faculty of Science and Engineering Waseda University, 3-4-1 Okubo, Shinjuku-ku Tokyo, Japan

[‡]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

E-mail: [†]lejiawen818@yama.info.waseda.ac.jp [‡]yamana@yama.info.waseda.ac.jp

ABSTRACT

Twitter, as one of the most popular social network services, is now widely used to query public opinions. In this research, Twitter data, along with the reviews collected from review websites is used to carry out some basic, sentimental, and culture-based analysis, so as to figure out the cultural effects on user evaluations for global restaurants.

This research is based on the authors' previous work, which only considers posts and reviews written in English. In this research, a language expansion is carried out that more than 30 languages are taken into account. By using a range of new and standard features, a series of classifiers are trained and applied in the later steps of sentiment analysis, through which some informative results are obtained considering the relationship between user evaluations and cultural backgrounds.

Keywords

Sentiment analysis, Twitter, reviews, multi-cultural backgrounds

1. INTRODUCTION

In recent years, social network service (SNS), a newcomer in the field of social media, has drawn much attention all around the world. Twitter[1], one of the most popular social network services, owns a range of special characteristics that contribute to its huge success. It allows users to express their opinions, chat with friends, and share the updated information with each other. Because of the limited length of 140 words per tweet, users feel more free and lighthearted to send a new message, which consequently leads to the tremendous amount of information and the rapid speed of distribution, compared to the traditional media such as newspapers and televisions. Besides the characteristic of big amount, this flood of messages also enjoys a great variety in their contents. For example, some people may discuss on the currently 'hot' topics, some people may express their views towards big events, while some other people just talk about their feelings about some trivial things in their daily life, such as a joyful trip they have made, a delicious meal, or a satisfactory service they have been offered. Actually, the huge volume of tweets can be used to survey public opinions. If many users post tweets that contain complimentary words of a restaurant, it is likely that this restaurant enjoys popularity among customers.

On the other hand, the recent decades also witnessed the remarkable progress of globalization. With the increase of the number of transnational enterprises, people from all over the world can use the same product, get the same service, and savor the same meal. However, it is quite common that people from different countries may have totally different feelings about these products, service or meals, probably mostly due to their diverse cultural backgrounds. Here, take the catering services for example. In order to figure out the cultural effect on the reviews of restaurants customers from different countries, tweets, as well as some gourmet reviews, can serve as a good dataset to carry out the analysis.

This research, based on the authors' previous work, has made three main contributions.

- A language expansion has been carried out that tweets written in more than 30 languages are treated as the research subject.
- Some modifications have been made in the sequential three-step classification process.
- Considering the multi-cultural background, more than 30 countries are taken into consideration in sentiment analysis.

The rest of this paper is organized as follows. Section 2 listed prior works related with this research. Section 3 discusses the concrete methods and algorithms adopted in this research with details. Section 4 describes the process and steps of the experiment. The obtained results are also discussed and analyzed in this section, which lead to the final conclusions and summarization in Section 5.

2. RELATED WORKS

The sentiment analysis of Twitter data has been focused by many researchers in these years, and there have been a range of significant works that make

[1] http://twitter.com/

contributions to this field.

In the aspect of opinion mining, a noted work is presented Pang and Lee (2008) [1], which gives a broad view of some existing approaches for sentiment analysis and opinion retrieval. Some early research that tries to put forward new methods or improve existing approaches considering the particular study subject of tweets can be listed as followed. Go et al. (2009) [2] take usage of the emoticons to query Twitter, and then refer to these data as a training set that are divided into negative ones and positive ones according to the sentiment of the query emoticons. As for the applied models, they have built Naive Bayes, MaxEnt and Support Vector Machines (SVM), and report that SVM model outperform other models. They also have obtained the result that unigram feature model has the best performance, which cannot be gained by using bigrams and parts-of-speech feature models. The paper of Liu (2010) [3] reviews the methods and works in the field of sentiment analysis of recent years. The work of Pak and Paroubek (2010) [4] characterizes in the collecting means of objective training data. The source of this kind of data includes several popular newspapers, whose sentences are usually considered as without special sentiment polarity. In contrast with the conclusion of Go et al., this paper reports that n-gram and POS strategies both make contributions to the performance. On the other side, the research of Barbosa and Feng (2010) [5] mainly focuses on the syntax features such as hashtags, URL links, and exclamations, and then make a combination with the POS model. All the above-mentioned works only take the common English tweets into consideration, and have not touched upon the cross cultural backgrounds.

As for the field of cross-lingual sentiment analysis, thenoted Oasys opinion analysis system presented by Cesarano et al. (2007) [6] allows the user to observe the change of intensity of opinion over countries and news sources. Guo et al. (2010) [7] have constructed a text mining system to detect the the different sentiment in the web texts written in different languages. The work of Cui et al. (2011) [8] uses emotion tokens to sovle the problem of cross-lingua sentiment analysis. Gao et al. (2012) [9] have researched on Twitter and the Chinese version of Twitter---Sina Weibo, and make some simple statistical comparisons in several different aspects, such as the characteristics of user behaviors and the content of messages.

Compared to all these works, this paper focuses on the analysis of cross-lingual user evaluations, which is based on the sentiment classification using the dataset of Twitter and reviews. More than 30 languages and more than 30 countries are taken into account, in order to obtain more comprehensive analysis results. All the approaches and experiment settings can be further expanded to the reviews of other fields and of other cultural backgrounds.

3. METHODOLOGY

3.1 Data Collection

The data used in this research mainly comes from two sources, Twitter and some restaurant related websites, which will be given more details in the experiment section.

First, as for the Twitter data collection, 9,523,211 restaurant-related tweets are gathered in 4 months (from Sep. 2013 to Dec. 2013), by using the Streaming API and Search API of Twitter. All this data has been restricted by the names of target restaurants, which has been translated into multi-languages.

Then, as an auxiliary dataset, the review dataset is constructed by collecting the English-written reviews from some popular review websites[234], including the text comments and their corresponding scores, from some popular gourmet sites. Totally 55,031 reviews are collected in this step.

3.2 Translation and Pre-filtering

In this research, 34 languages (i.e. en, es, id, ja, fr, pt, tl, ru, tr, zh, ar, th, et, nl, it, de, ko, bg, sv, pl, vi, sk, da, ht, lt, lv, sl, fi, is, no, fa, hu, el, uk[5]) are taken as the target languages. The selection of the target languages is based on tweet amounts, language populations, and whether can be translated by machine translation tools. Due to the complexity of tweet texts, and also the translation performances of the machine translation tools for certain languages, part of the Twitter dataset cannot be correctly translated into English, and are then discarded.

The remaining data is then filtered by the pre-defined condition of having some relationship with restaurants, which is restricted by a list of restaurant related words. These words are obtained by calculating the occurrence frequency of each word in the set of text reviews from the review dataset, and selecting out the ones that have the highest frequencies. Here, 45 most frequent words[6] are selected as the restaurant related words to filter the original Twitter data.

3.3 Location Definition

As for the tweets collected by the Twitter API, there are several items concerned with location information. For example, the ‘coordinates’ item and ’geo’ item, the ‘location’ item and ‘time zone’ item in the user profile, are all indicative for the location definition.

First, a manually constructed location name

[2] http://www.tripadvisor.com/

[3] http://www.yelp.com/

[4] http://www.zagat.com/

[5] The language code can be found in http://en.wikipedia.org/wiki/ISO_639-1/.

[6] These words are: restaurant, restaurants, food, foods, drink, drinks, dinner, dinners, lunch, lunches, breakfast, breakfasts, club, clubs, bar ,bars, pizza, pizzas, burger, burgers, coffee, cafe, cheese, grill, sushi , yelp, taco, steak, fry, fried, bbq, bakery, baked, yummy, yum, tasty, taste, tastes, delicious, eat, ate, eaten, eating, meal, meals.

dictionary[7] is used to query the names of counties or cities appeared in the above-mentioned location-related items of tweets. Then Yahoo yql API[8] is used to parse these items of the remained undefined tweets again to obtain more definitions. After these two steps, the ratio of the tweets that have been labeled with location names is 72.8%. This part of tweet data is further used in the later steps.

3.4 Spam Filtering

Because of the fact that the Twitter dataset contains so many undesired spam tweets, a filtering step to discard the obvious spam tweets is necessary. Strictly, whether a tweet is spam or not in this research should depend on whether the content of the tweet text contains some useful information to indicate the subjective opinions towards the restaurants. In this research, however, a simple spam filtering technique is applied. Firstly, advertisements and pure 'check-in' tweets (e.g. *'I'm at Burger King (Stationsplein 4, Groningen) http://t.co/gWSyUMLD'*) are regarded as 'spam'. In addition, tweets posted in a certain short time period which have exactly the same contents are also considered as 'spam'.

A Bayesian classifier is used here, because Bayesian classification is usually robust to the noisy information. The training features include the number of the followers and friends of the user, the ratio of the number of followers and friends, the date of the registration, average number of new friends and followers per day, the latest 20 posted tweets, and also some syntax characteristics. 1000 and 200 manual labeled tweets are taken as the training set and test set respectively, and the performance of the 'spam' classifier turns out to be of an accuracy of 97.8%. This trained classifier is then applied to the whole dataset to filter out the 'spam' tweets.

In this experiment, the result indicates that 9.8% samples are classified as 'spam', which is a higher proportion compared to some previous study[9]. This may be explained by the specialized definition of a 'spam' tweet, and also the special focus on the restaurant field may contribute to the large amount of advertisements. Finally, these 9.8% 'spam' tweets are filtered out of the dataset, and the remained samples are to be processed in the later analysis.

3.5 Features for Sentiment Classification

3.5.1 Dictionaries Construction

Before the features selection step, two dictionaries are constructed beforehand.

First, a total word dictionary is needed. This dictionary records all the words appeared in the total Twitter dataset, with their occurrence frequencies. Here, in order to remove some noisy items, a restriction of frequency no less than 3 is applied. The size of this dictionary (*tw_total_dict*) turns out to be 58,615 entries.

Then, for the later sentiment analysis, an initiative polarity dictionary is required. As research resources, there already exist several popular authoritative polarity dictionaries on the Internet, which contain certain amount of words with their polarities. In this step, some of these online dictionaries are combined to construct the initiative polarity dictionary (*pol_dict_ini*) so as to have a better coverage of vocabulary, and the structure of the dictionary is listed in Table 1.

Table 1: The structure of the initial polarity dictionary

Label in the dictionary	Source
Positive	Positive Score > 0.75, or Positive Score – Negative Score > 0.5 (SentiWordNet[10]), Strong Positive (MPQA[11]), Positiv category (the General Inquirer[12])
Negative	Negative Score > 0.75, or Negative Score – Positive Score > 0.5 (SentiWordNet), Strong Negative (MPQA), Negativ category (the General Inquirer)
Neutral	Positive Score = 0 and Negative Score = 0 (SentiWordNet)

These entries in *pol_dict_ini* totally amount to 125,277.

3.5.2 Syntax Features

A big difference between tweets and common texts lies in the syntax characteristics of them. Tweets own many unique syntax characteristics that other sentences do not have, including the at mark, the retweet mark, the URL link, and the hashtag. These characteristics bring about some inconvenience while preprocessing the tweet texts, but on the other hand, they are quite informative in the task of sentiment analysis.

In this research, totally 10 syntax characteristics are taken into consideration. They are exclamation marks (!), question marks (?), upper-case words, capitalized words, hashtags (#), at marks (@), retweet marks (RT), URL links, emoticons, and slang words. All these characteristics are counted by their occurrences in one tweet, and this 10-dimension vector is regarded as the '*syn*' feature. Here, a manually built emoticon dictionary

[7] The country names and country name codes in the dictionary are collected from http://en.wikipedia.org/wiki/ISO_3166-1/.

[8] http://developer.yahoo.com/yql/.

[9] This study is retrieved from Pear Analytics, Twitter Study. http://www.pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf/.

[10] http://sentiwordnet.isti.cnr.it/

[11] http://mpqa.cs.pitt.edu/

[12] http://www.wjh.harvard.edu/~inquirer/

and slang dictionary are referred to during the counting process, and they are both built from some online resources[1].

3.5.3 Modified Unigram

Compared to the standard unigram model, an additional dimension-reduction is applied while processing the modified unigram features. First, for each word in *tw_total_dict*, set the polarity score as 2, -2, and 0 if it is labeled as Positive, Negative, and Neutral in *pol_dict_ini* respectively. Then parse all the tweets to calculate out the PMI (Pointwise Mutual Information) values of all the pairs of words in *tw_total_dict.* The PMI value of word w_1 and w_2 is given by

$$PMI(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1) \cdot p(w_2)}$$

where, $p(w_1, w_2)$ is the co-occurrence probability of word $w_1$and w_2 in one tweet, and $p(w_1)$ and $p(w_2)$ are the occurrence probabilities of word $w_1$and w_2 in one tweet respectively.

Then, for each word NOT in *pol_dict_ini*, sort its PMI values with the words in *pol_dict_ini*, and carry out majority voting among the top 10 sorted items. The 'positive inclined' word is then score as 1, the 'negative inclined' word is then scored as -1, and other words (i.e. the top 10 corresponding words are all from the Neutral category in *pol_dict_ini*) is then scored as 0. The output of this step is a new polarity dictionary *(pol_dict*) with the vocabulary of total_dict, and each word in it is mapped to a score of 5 scales (i.e. 2, 1, 0, -1, and 2).The comparison of the word counts for each scale before and after this step is showed in Table 2.

Table 2: The comparison of polarity word counts between *pol_dict_in*i and *pol_dict*

	2	1	0	-1	-2	Total
pol_dict_ini	17,494	0	19,581	0	4,734	58,615
pol_dict	17,494	7,108	24,332	4,957	4,734	58,615

Based on *pol_dict*, each tweet can be projected to a 5-dimension vector, and each dimension records the count of the unigram words in this category. This vector is named as the *'5s'* feature.

3.5.4 The Review Dataset-based Average Score

While users always express their opinions in their tweets, they also give out the clear evaluation towards the products and services on some special review websites. These reviews are much more detailed in the user experience, and are always with a corresponding concrete score, such as the most common 5 scale score. This kind of information can be quite useful if taken full advantage of.

In this research, as mentioned in section 3.1, a corpus of reviews is previously constructed, and each entry in this dataset has a tuple structure, which can be described as $(text, score)$. In this step, all the text parts are first processed as a BoW model, and the total vocabulary of the review dataset is described as W_{rv}. For each word w_i in W_{rv}, the review dataset-based polarity score is calculated by

$$pol_{w_i} = \frac{\sum_{text_j \in TX_{w_i}} score_j}{|TX_{w_i}|}$$

where, TX_{w_i} is the set of review texts, in which the word w_i occurs, $text_j$ is a review text in TX_{w_i}, and $score_j$ is the corresponding score of $text_j$.

Then for each tweet tw i in the Twitter dataset, the review dataset-based average score is given by

$$avg_{tw_i} = \frac{\sum_{w_j \in W_{tw_i}} pol_{w_j}}{|W_{tw_i}|}$$

where, W_{tw_i} is the word set of tw_i, and pol_{w_j} is the polarity score of w_j given by the last step. This float average score for each tweet is named as the '*rv*' feature.

This feature selection method seems quite simple and plain, but both in this research and some previous research, it turns out to have quite a good performance.

3.5.5 The Review Dataset-based CCA Score

Canonical correlation analysis (CCA) is a classical statistics method to figure out the latent relations among multiple variables. It can also be used to analyze the potential relation between the text format evaluation and the score format evaluation when they are referring to the very same issue.

In this case, the review dataset constructed from gourmet websites has the characteristic that each entry in it consist a comment text and a 5-scale score, which is described by the format $(text, score)$. For the reason that there must be some consistency in the comment text and score from the same person, it can be safely concluded that there is some latent relationship between them. Thus, the CCA method can be used here to get the latent relationship between the users' sentiment and some polarity words. Here, the adopted measure criterion is the first correlated variable. The review dataset is taken as the condition set, and the first correlated variable parameters are decided by the CCA process. Then, for each tweet in the Twitter dataset, the first correlated variable is applied and calculated. Finally, a float number is given to each tweet as the '*cca*' feature.

3.5.6 The Window Co-occurrence-based Average Score

In the modified unigram model in section 3.3.3, the polarity score of each word is calculated without the consideration of the neighboring relationship among words. Since that this relationship may contain some indicative information for sentiment analysis, the score based on the co-occurrence in a three-word window is calculated out in this section.

Inspired by the research of Brody and Elhadad [10], in which a propagation algorithm is applied to analyze the sentiment of online reviews, a modified graph-based propagation algorithm is also adopted here to obtain the polarity score of each word in *tw_total_dict* based on the three-word window neighboring relationship.

First, a co-occurrence dictionary is constructed by parsing all the tweets in the Twitter dataset. The key of the item in this dictionary is the word pair $w_i\_w_j$, and the value of the item in this dictionary is the times $t(w_i, w_j)$ these two words appeared in the three-word window.

Then, as an initial propagation graph, all the words in *tw_total_dict* are taken as the nodes of the graph. The value of each node is initiated as 1, and -1 for the words in the Positive category and Negative category of *pol_dict_ini* respectively. For other words, the initiated node value is set as 0. Then, for each iteration, the value of each node is updated by

$$v'_{n_i} = (1-\alpha) \cdot \frac{\sum_{n_j \in NEI_{n_i}} v_{n_j} \cdot \left(1 + \log\left(t(n_i, n_j)\right)\right)}{\sum_{n_j \in NEI_{n_i}} \left(1 + \log\left(t(n_i, n_j)\right)\right)} + \alpha \cdot v_{n_i}$$

where, NEI_{n_i} is the set of the nodes neighbored with node n_i, and $t(n_i, n_j)$ is the co-occurrence times of the words of node n_i and n_j, according to the previous built co-occurrence dictionary. α is a tuning parameter, which is set as 0.6 in this step. In the final graph after running the iterations to convergence, each node has a float value indicating the polarity of the word of this node. A polarity dictionary can be obtained by this final graph, and the formula to calculate the average score of each tweet in section 3.5.4 can also be applied here based on the newly constructed polarity dictionary. This float score for each tweet is named as the '*win3*' feature.

3.5.7 The POS-based Feature

Except for the above mentioned models, the POS (part-of-speech) information is also usually used in the NLP analysis. It has been reported that some part-of-speech pairs are especially sentiment expressive in some previous research. Here, all the tweets are first processed by the Stanford Parser [13] to get the dependencies trees. Then some typical pairs of POS are extracted from these dependency trees. In this research, 10 most common and sentiment expressive POS pairs are chosen manually, and the sentiment expressed in these pairs are decided according to some manually constructed rules. The applied POS pairs include '*acomp'*, *'advmod', 'amod', 'conj', 'dobj', 'neg', 'nsubj', 'purpcl', 'rcmod',* and *'xcomp'*. For each tweet in the Twitter dataset, each above-mentioned POS pair that appears in this tweet is given with a polarity label. Then, to decide the polarity of the tweet, a simple majority voting method is applied, which means that the polarity label that has the biggest POS pair count passes its polarity to the tweet. This feature is called *'pos'* in the later analysis steps.

4. EXPERIMENT AND RESULTS

4.1 Overview

The main steps of the whole experiment are described in the flow chart below (Figure 1).

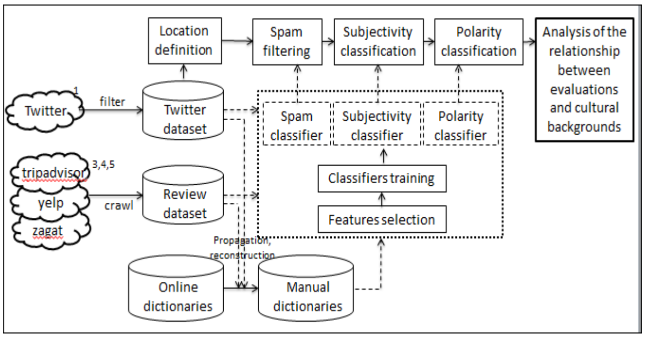

Figure 1: The main flow of the experiment

After collecting the Twitter and review data, the location definition step is carried out, and two dictionaries are manually constructed based on the datasets and some online dictionaries. Then three main classifiers are trained and used to classify and the tweets. Finally, based on the classification results and the data analysis results, the cultural effect on evaluations is clarified. The implementation details of the main process will be further introduced in the following sub-sections.

4.2 Preprocessing

As for the Twitter dataset, the preprocessing basically contains 12 steps. While listed in a sequential order, they are (1) 'RT' and URL links deletion; (2) Emoticons conversion; (3) Lower-casing; (4) HTML transcoding; (5) Hashtags conversion; (6) Punctuation deletion; (7) Word segmentation; (8) Non-alphabet words and single alphabet words deletion; (9) Stop words discard; (10) Repeated alphabets reduction; (11) Chat words conversion; (12) Lemmatization. However, the processing is also task-specific in some steps. For example, as input for the Stanford Parser, only (1) ~ (5) are carried out.

As for the review dataset, the preprocessing is much simpler. Basically 6 steps are needed, including (1) Lower-casing; (2) Word segmentation; (3) Non-alphabet words and single alphabet words deletion; (4) Stop words discard; (5) Chat words conversion; (6) Lemmatization.

4.3 Sentiment Classification

In this research, sentiment classification is divided by two steps. The first step, subjectivity classification, is to classify the spam filtered dataset into the subjective dataset and the objective dataset. The second step, polarity classification, is to further classify the subjectivity dataset into the positive dataset and the negative dataset. In each of these two steps, a pre-trained classifier is applied to carry out the classification. The training process of these two classifiers is described in details as follows.

Features selection

In section 3.3, 6 groups of features are introduced.

[13] http://nlp.stanford.edu/software/lex-parser.shtml/.

They are the '*syn*', '*5s*', '*rv*', '*cca*', '*win3*', and '*pos*' features. All the combinations of these 6 groups of features are implemented in this experiment.

Training method

The SVM (Linear, RBF, and Polynomial) methods and the Naïve Bayes (Gaussian, Multinomial, and Bernoulli) methods are used in this experiment.

Training implementation

The total number of implementation variations turns out to be

$$(2^6 - 1) \cdot 6 = 378$$

Validation method

The standard 10-fold cross-validation is applied here.

Training set

For the subjectivity classifier: 1000 manually labeled tweets (500 subjective, 500 objective).

For the polarity classifier: 1000 manually labeled subjective tweets (500 positive, 500 negative).

Test results

Part of the test results of the subjectivity classifiers and the polarity classifiers are showed in Table 3 and Table 4.

Table 3: Subjectivity classifiers performance

syn	*5s*	*rv*	*win3*	*cca*	*pos*	**accurracy**
•		•		•		74.7%
	•		•	•		74.9%
	•	•		•	•	75.8%
	•	•	•			76.4%
•	•	•	•		•	76.5%
•		•	•	•	•	77.5%
•		•	•		•	**78.4%**

Table 4: Polarity classifiers performance

syn	*5s*	*rv*	*win3*	*cca*	*pos*	**accurracy**
•	•					82.2%
			•	•	•	85.3%
•			•	•	•	87.2%
•	•	•		•	•	89.6%
		•	•		•	89.9%
		•	•	•		90.6%
		•	•	•	•	**91.1%**

The first column is the number of the implementation, and the last column records the highest accuracy of the 6 training methods in this implementation. A field with the circle mark indicates that this group of features is applied. According to these tables, the best-performed subjectivity classifier is obtained by the features combination of *'syn', '5s', 'rv',* and '*cca*', with SVM polynomial training method, while the best-performed polarity classifier is obtained by the features combination of '*rv*', *'win3', 'cca',* and '*pos*', with the SVM linear training method. These two classifiers are used in the classification step for the whole spam filtered Twitter set.

4.4 Basic Data Analysis

Based on the 'list of restaurant chains' on Wikipedia, 6 restaurants (i.e. *Burger King, Mcdonald's, KFC, Pizza Hut*, *Subway*, and *Starbucks*) who have most locations worldwide are chosen as the research subject. Filter the original Twitter dataset with these keywords, and carry out the location definition process described in section 3.3. Then 34 countries (i.e. United states (US), United Kingdom (GB), Australia (AU), Indonesia (ID), Malaysia (MY), Canada (CA), Philippines (PH), Singapore (SG), Brazil (BR), India (IN), South Africa (ZA), Japan (JP), Mexico (MX), France (FR), and Netherlands (NL), Greece (GR), Thailand (TH), China (CN), Russia (RU), Spain (ES), Argentina (AR), Chile (CL), South Korea (KR), Germany (DE), Italy (IT), Ireland (IE), Venezuela (VE), Colombia (CO), Poland (PL), Egypt (EG), Ukraine (UA), New Zealand (NZ), Viet Nam (VN))[14] are selected out, and only tweets from these 34 countries and areas are remained to be processed by the spam filter introduced in section 3.4. While the original Twitter dataset amounts to 10 million, the size of this pre-filtered and spam-filtered Twitter dataset has been reduced to approximately 3 million. This dataset is the input of the later steps of sentiment classification.

In this section, some basic statistical analysis is taken out to obtain a general overview of these restaurants in the 33 countries. Figure 2 shows the distribution of tweets over the 6 restaurants in each country.

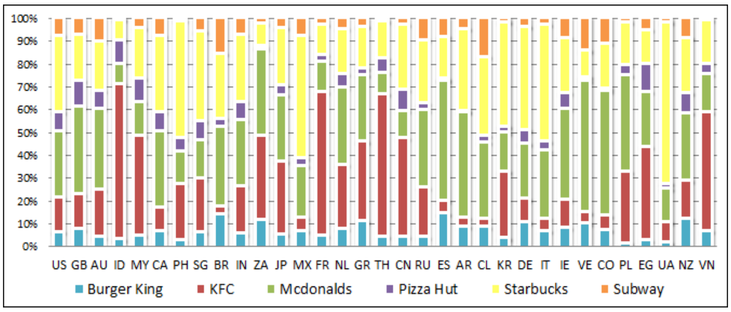

Figure 2: General distribution of tweets

From Figure 2, the following conclusions can be obtained: (1) In different countries, the distribution of tweets over the 6 restaurants is quite different; (2) For each restaurant, the tweets distribution over the 15 countries is quite different and biased; (3) These distributions may give information for the popularities of each restaurant in each country.

4.5 Sentiment Analysis

In this section, the one-quarter spam filtered dataset is further processed by the optimal subjectivity classifier and polarity classifier introduced in section 4.3, according to certain steps in Figure 1. These two classification steps divide the spam filtered dataset into 3 polarity groups, i.e. positive, negative, and objective. While the positive tweet, negative tweet, and objective

[14] The alpha-2 code in the brackets is a simplification of the full country name. The standard adopted here is ISO 3166-1.

tweet is given a polarity score of 1, -1, and 0 respectively, the average sentiment score lines for each restaurant and for each country are described in Figure 3.

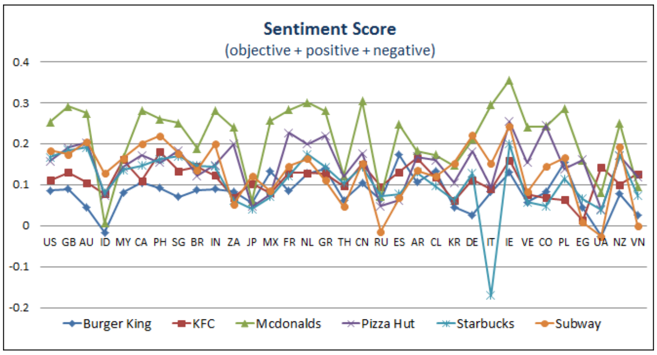

Figure 3: Sentiment score lines for each restaurant

As for the sentiment scores for the 34 countries, the k-means method is applied to cluster these countries into several groups. Here, k is empirically set as 10, and the10 clusters turns out to be:

- US, CA, PH, SG, DE, AU, IN, NZ;
- JP, ID, KR;
- ES, MX;
- IT;
- RU, UA;
- GB, NL, FR, GR, CN, IE, PL;
- MY, BR, AR, CL;
- TH, VN;
- EG, ZA;
- CO, VE.

4.6 Culture-based Analysis

As one of the main objectives of this research, the relationship between the user evaluations and cultural background is taken as the analysis subject in this section. Figure 4 shows the culture map after adding the cluster results into a blank world map. The countries in same cluster are of the same color.

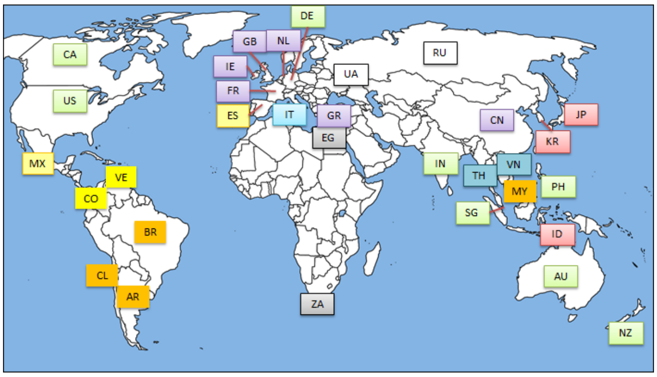

Figure 4: The culture map

Upon observing the country clusters in Figure 4, some explanations are as follows: (1) As for some countries, the location-based cultural effects are quite obvious. For example, the cluster of BR, CL, AR, the cluster of RU, UA, the cluster of ZA, EG, the cluster of TH, VN, and the cluster of most of the Western European countries, have been clustered into the same cluster according to their location and basic cultural background; (2) Some of the English-speaking Asian countries are clustered into the same group with North American countries, which suggest that the language-based cultural background may have some effects; (3) Comparing to most of the European countries, some countries, such as ES and IT, seem to have quite different opinions for these restaurants, which may suggest that they have some special attitudes considering the food culture. (4) However, some confusing results still exist. For example, CN is clustered into the Western European cultural background group, and MY is clustered into the South American cultural background group. (5) These confusing results may be explained by some other effective element except for general cultural background, such as the eating patterns, the brand reputation, marketing strategies, and some locally specialized products and services. (6) Some limitations of the experiment, such as the fact that only fast food restaurants are taken as targets, may also contribute to the unexpected results.

5. CONCLUSION

In this research, the relationship between user evaluations and cultural backgrounds is investigated. This investigation is based on more than 30 countries around the world, and tweets written in more than 30 languages are analyzed. The main steps include preprocessing, location definition, spam filtering, subjectivity classification, polarity classification, and a series of analysis. Three key classifiers (i.e. spam classifier, subjectivity classifier, and polarity classifier) are trained with a range of different implementations, and have obtained the accuracy of 97.8%, 78.4%, and 91.1% respectively. The results in later steps of basic analysis, sentiment analysis, and culture-based analysis indicate that the cultural effects on user evaluations for restaurants actually exist, and are quite obvious for some countries and cultural backgrounds.

As the next steps, first, some latent elements other than the cultural background should be further investigated, so as to figure out the underlined facts that can explain for some unexpected results of certain countries. Then, some other possible expansions, including the expansion from catering business to other fields, should be carried out.

Acknowledgements

The authors thank the students in Yamana Laboratory of Waseda University for their valuable comments and suggestions.

References

[1] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[2] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N Project Report, Stanford, 2009: 1-12.

[3] Liu B. Sentiment analysis and subjectivity[J]. Handbook of natural language processing, 2010, 2: 568.

[4] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]//LREC. 2010.

[5] Barbosa L, Feng J. Robust sentiment detection on twitter from biased and noisy data[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 36-44.

[6] Cesarano C, Picariello A, Recupero D R, et al. The OASYS 2.0 Opinion Analysis System[J]. ICWSM, 2007, 7: 313-314.

[7] Guo H, Zhu H, Guo Z, et al. OpinionIt: a text mining system for cross-lingual opinion analysis[C]//Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010: 1199-1208.

[8] Cui A, Zhang M, Liu Y, et al. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis[M]//Information Retrieval Technology. Springer Berlin Heidelberg, 2011: 238-249.

[9] Gao Q, Abel F, Houben G J, et al. A comparative study of users' microblogging behavior on Sina Weibo and Twitter[M]//User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, 2012: 88-101.

[10] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 804-812.