Multi-Language Sentiment Analysis of SNS across Different Cities

Muga Yoshikawa Hiroaki Saito
Keio Universiry
{muga, hxs}@nak.ics.keio.ac.jp

1 Introduction

A sentiment analysis system is a classifier that determines sentiment in a sentence. Given a sentence as input, it distinguishes if the input means something positive, negative or neutral for instance. These days this system is applied to many areas. For example, movie or product reviews [9] and opinion surveys for elections [6]. Since there are billions of active users, sentiment analysis of Social Network Service (SNS) especially has potential to see what the users are thinking of, what images they are having on some certain topics. Billions of people in the world are now using the Internet and smartphones. Thus sentiment analysis of SNS can be helpful to see opinions for many kinds of topic (e.g. TV show, concerts etc.).

The EU referendum in the U.K in June 2016 showed that the result varied clearly by area in the country [5]. People in London and Scotland supported for remain but the other area did the opposite. Not only this vote in the U.K, this phenomenon was observed also in the U.S. presidential election in November 2016 [4]. The results of these votes both turned out that people even in the same country have different opinion depending on where they live. Not only a large country such as the U.S, but even in the U.K, which is way smaller than the U.S, the gap is widening. Hence, sentimental analysis of different locations is a meaningful task.

In addition, we focus on four languages because even people in the same country or city speak different languages, which happens usually in many places nowadays. A psychology research [11] shows language effects on action cognition depending on the language the participant speaks or hears. According to another psychology research [8], if people speak in foreign languages they react differently when they are asked to bet on gamble. Therefore a perspective of only one language is not enough for observing trends on SNS especially in big cities where many people from different countries around the world gather.

This paper proposes a system of multi-language sentiment analysis of SNS, specifically Twitter, and analyses tweets across different locations around the world. The reason why we choose Twitter is that thousands of tweets can be searched easily via Twitter API. With public available twitter data that are annotated by humans, we build classifiers to label the tweets obtained by location. We pick 12 cities where English, French, German or Spanish is spoken as the official language. We explore the difference of distribution of sentiment value between cities and languages.

2 Related Work

There is a research of multi-language sentiment analysis of social media [13] and the research focuses on three languages, English, Spanish and Russian. In the research, Volkova et al. try to explore the gender bias in the use of subjective language and incorporate this bias into models to improve sentiment analysis for three languages. They build a model for the analysis with corpus-based language independent approach proposed by [14].

Narr et al. propose a classification method for sentiment analysis that can be applied for multiple languages [10]. In the research the authors analyse four languages, English, French, German and Portuguese. They use emoticons as noisy labels to generate training data and train Naïve Bayes classifier on the collected data. The tweets classified by the classifier are evaluated on the human annotated data that we use in this paper.

Bautin et al. apply another way for multi-language sentiment analysis [7]. They explore an approach utilizing machine translation method and perform sentiment analysis on the English translation of a foreign language text. The target of their sentiment analysis is News and blogs, not social media like this paper.

Although all of these three researches belong to multi-language sentiment analysis, they do not consider location information. We, however, collect tweets in four languages by city around the world and build classifiers so that we perform multi-language sentiment analysis between different cities.

3 Twitter Place ID

In this paper, we use the place ID in Twitter API [2] in order to obtain tweets by each location. Figure 1 shows a part of the screenshot that is displayed

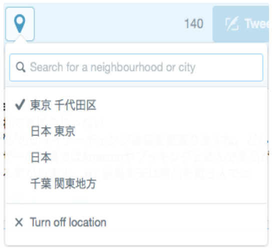

Figure 1: Suggested places in Tokyo based on Twitter place ID

when a user tweets somewhere in Tokyo. The suggested places are shown based on place ID where the user tweets. If the user turns on location service, the user can set one of the suggested areas as location information. Even when users don't turn on GPS service on their phone, these places are shown because they are displayed based on place ID in Twitter API. Named locations or districts in a city are given each place ID by Twitter. For example, the place ID of Twitter head quarter is "07d9cd6afd884001". There are 3 types of place ID. Like the previous example, one of them represents a named location. Another represents a whole city (e.g. Tokyo Japan) and the other is district (e.g. Shibuya Tokyo, Japan) And each place ID contains an attribute of its own coordinates, latitude and gratitude.

With place ID of an area, tweets from the area can be searched via Twitter API. For searching as many tweets as possible, we search tweets via the search API by each district or borough of a city. We list up basically all of the districts of a city and then search the coordinate of each district because place ID is found by specifying the coordinate of the place or nearby. Place ID that represents a whole city are also searched. But to avoid to be too complicated, we do not use place ID of named locations. Otherwise it is necessary to set the criteria of choosing searched locations and this can be different depending on cities. By collecting the coordinates of the areas in each city and searching the place ID of them, we set up place ID lists of the cities.

4 Sentiment Analysis by City

Figure 2 describes the overview of our system. The system takes collected tweets by city as input. They are searched by the Twitter API, specifically with the place ID explained in Section 3. Tweets are searched by each language by specifying language parameter of the Twitter search API. Before the tweets are labelled by the classifiers, they are pre-processed to replace emoji unicode in them. The classifiers are built by machine learning and they require training data of annotated tweets, which are labelled positive, negative or neutral. In the end, classifiers label input tweets as one of sentimental values, which are positive, negative or neutral.

Not only on Twitter but also on SNS as a whole

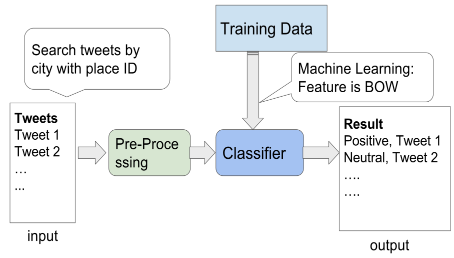

Figure 2: Overview of sentiment analysis by city

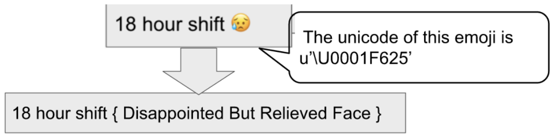

Figure 3: How an emoji is replaced in a tweet

many people are using emoji nowadays [3]. Sometimes it takes an important role of a sentence. In Figure 3, an emoji at the end of the sentence is replaced with the explanation of the emoji. This way, the tweet can have more meaningful characters for classifiers than the unicode of the emoji, which is "u\U0001F625" This replacement is applied to all of the languages.

We do not prepare training datasets ourselves but instead take them from other works. The numbers of dataset of each language are listed in Table 1. Spanish dataset is much larger than the others because this is the only dataset from [12] and the others are from [10]. All of the datasets are annotated by people and importantly the tweets in the datasets are labelled sentimental values. Therefore they can be considered to be trustful.

Table 1: Number of dataset of each language

Language	# of dataset
English	7,200
French	1,797
German	1,800
Spanish	68,000

We employ three kinds of classifiers in this paper. Originally SVM is a binary classification method but it can be used for multi-class classification as well. There are two ways for multi-class classification SVM, one versus one and one versus all. One versus one compares two classes of N classes and constructs a hyperplane while one versus all compares one of N classes and the others. Random forest is an ensemble learning method that samples random data and constructs decision trees. It classifies based on the results of the multiple decision trees. We apply the API of scikit-learn [1] to build these classifiers in Python 2.7.12.

5 Experiment

Before the experiment of sentiment analysis across cities, we measure F1 scores of each classifier in the four languages to confirm whether the classifiers built with the given dataset work well enough. The dataset is divided to 80 % training data and 20 % test data. Table 2 shows F1 scores of the classifiers. As these scores show, we believe these classifiers work well enough to do experiment of sentiment analysis by city.

Table 2: F1 scores of each classifier in the four languages

Language	One vs. One SVM	One vs. All SVM	Random Forest
English	0.83	0.82	0.83
French	0.78	0.82	0.72
German	0.81	0.82	0.82
Spanish	0.82	0.81	0.63

We choose 12 cities that have a large population, mostly capitals in fact, because it is likely to retrieve many tweets from such cities. Another reason is that the more data we get, the more precise result we can reach. One of the four languages has to be spoken as official language in the chosen cities. If possible we choose two cities from the same country. (e.g. New York and San Francisco)

We managed to retrieve about a million tweets in total from the beginning of November to 26 December 2016. During this period, we searched and retrieved tweets from the 12 cities in the four languages by specifying the language parameter in the Twitter search API. In Table 3 the cities are listed with the numbers of the retrieved tweets during the period.

After preparing retrieved tweet data for classification as explained in Section 4, the three classifiers trained with each language dataset put sentiment values to the retrieved tweets. A tweet has to be labelled as one of the sentiment values, positive, negative or neutral.

Table 3: Number of retrieved tweets of each city

City	# of tweets	Language
Barcelona	17,963	English
Berlin	20,141	English
Hamburg	3,992	English
London	356,227	English
Madrid	17,348	English
New York	80,182	English
Paris	27,955	English
Quebec	43,171	English
San Francisco	198,359	English
Lille	26,705	French
Paris	139,260	French
Quebec	18,995	French
Berlin	30,804	German
Hamburg	10,720	German
Vienna	4,997	German
Barcelona	27,064	Spanish
Buenos Aires	683,182	Spanish
Madrid	29,720	Spanish

6 Evaluation & Result

For evaluation we employ chi square test of independence in each language to see if the distributions of sentiment values of two cities are independent or not. Apparently different things happen in each city. Even though people in the two cities speak the same language, it is assumed that the distributions of sentiment values ought not to be independent of location. If it turns out that they are not independent in many cases, this validates the sentiment analysis proposed in this paper. The reason we decide to use this test is that the output belongs to categorical data in statistics. The output consists of a tweet and its sentiment value put by the classifier. Hence it is only allowed to count the observed frequency. We consider a cross-tabulation table of the observed frequency of sentiment values of two cities and an example of such table is Table 4, which shows the classification result of New York (N.Y) and London. As it shows, the results of SVM (One vs. One and One vs. All) are similar but they are both different from the one of Random Forest. This happens to almost all of the results whatever language it is. Because the results between the classifiers are different, we decide to do chi square test on all of the results.

The null hypothesis of the chi square test is "If it is the same language, the distributions of sentiment values of the two different cities are independent." The χ^2 statistic is calculated using the expected frequency and the values in the cross-tabulation. Since there are two rows and three columns in the tabulation, the degree of freedom is 2. If the χ^2 statistic is greater than 5.991, which is the appropriate critical value at a 5 % level of significance when degree of freedom is 2, the null hypothesis is rejected. And it is concluded that the distributions are not independent.

Table 4: Classification results of New York and London

City	Positive (tweets)	Negative (tweets)	Neutral (tweets)	Total (tweets)
One vs. One SVM				
N.Y	18,504 (23.07%)	21,645 (26.99%)	40,033 (49.94%)	80,182
London	84,300 (23.66%)	80,419 (22.56%)	191,508 (53.76%)	356,227
One vs. All SVM				
N.Y	16,878 (21.05%)	21,355 (26.63%)	41,949 (52.31%)	80,182
London	76,704 (21.53%)	78,556 (22.05%)	200,967 (56.42%)	356,227
Random Forest				
N.Y	8,538 (10.65 %)	10,175 (12.69%)	61,469 (76.66%)	80,182
London	36,804 (10.33%)	28,896 (8.11%)	290,527 (81.56%)	356,227

(The numbers in parentheses represent percentage of each sentiment value)

Table 5 shows the χ^2 statistics of English, French, German and Spanish tweets of each combination of the cities respectively. We also compare English tweets for three combinations, Paris & Quebec, Berlin & Hamburg and Barcelona & Madrid to see if the result of chi square test is different from the one of the official language of each city. For instance, if it is Barcelona we compare its result both in English and Spanish. In Table 5, the numbers coloured red are smaller than 5.991. There are eight cases out of 45 where such condition is met. In these eight cases, the null hypothesis is accepted and in the other cases it is rejected. These results imply that except the eight cases location has something to do with emotional reaction among the two cities on Twitter when it is the same language. As for Paris &Quebec, Berlin & Hamburg and Barcelona & Madrid, the results are consistent between the two languages in the 16 cases out of 18. Another way of saying, the results are same in the both languages except for the case of Barcelona & Madrid.

Table 5: χ^2 statistics of each language

	One vs. One SVM	One vs. All SVM	Random Forest
English			
London/N.Y	737.407	804.5223	1,733.021
London/S.F	1,321.374	1,518.266	2,880.292
N.Y/S.F	3.172	3.316	2.278
Paris/Quebec	980.719	862.486	358.153
Berlin/Hamburg	62.883	73.823	15.674
Barcelona/Madrid	13.010	14.136	2.779
French			
Lille/Paris	19.522	28.281	85.578
Lille/Quebec	210.083	270.096	219.323
Paris/Quebec	299.681	394.692	144.962
German			
Berlin/Hamburg	299.189	298.030	8.322
Berlin/Vienna	230.744	201.563	24.614
Hamburg/Vienna	5.103	1.306	7.831
Spanish			
Barcelona/B.A	38.228	35.798	21.298
Barcelona/Madrid	13.688	4.895	6.350
B.A/Madrid	2.770	9.626	73.001

(N.Y: New York, S.F: San Francisco, B.A: Buenos Aires)

7 Conclusion

We proposed multi-language sentiment analysis on Twitter across different cities. Searching tweets by Twitter place ID, we managed to retrieve about a million tweets by city in the four languages. We employ chi square test as evaluation and for each language the results show the distributions of sentiment values are not independent between two cities in most of the cases. Since this is acceptable in terms of a cultural aspect, it approves the validity of this sentiment analysis.

One of the things to be fixed is changing the feature vector. In this paper we use BOW as feature vector but distributed representation is also an alternative solution. We just take a look at the whole of the tweets but for more precise analysis it is necessary to choose tweets by topic and compare them between cities. We believe we can accomplish more precise multi-language sentiment analysis by resolving these tasks.

References

[1] http://scikit-learn.org/stable/index.html.

[2] https://dev.twitter.com/rest/reference/get/geo/id/

[3] http://www.adweek.com/socialtimes/report-92-of-online-consumers-use-emoji-infographic/627521.

[4] http://www.bbc.com/news/election-us-2016-37889032, 2016.

[5] http://www.bbc.com/news/uk-politics-36616028, 2016.

[6] http://www.cnn.co.jp/tech/35092265.html, 2016.

[7] Bautin et al. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

[8] Boaz Keysar et al. The foreign-language effect. *Psychological Science*, 2011.

[9] Hatem Ghorbel et al. Sentiment analysis of french movie reviews. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[10] Narr et al. Language-independent twitter sentiment analysis. In *KDML/LWDA*, 2012.

[11] Panos Athanasopoulos et al. Two languages, two minds. *Psychological Science*, 2014.

[12] Romn et al. The challenge of aspect-based sentiment analysis. In *TASS*, 2014.

[13] Volkova et al. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *EMNLP*, 2013.

[14] Volvoka et al. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *ACL*, 2013.