Towards Understanding Cross-Cultural Crowd Sentiment Using Social Media

Yuanyuan Wang[1](✉), Panote Siriaraya[2], Muhammad Syafiq Mohd Pozi[3], Yukiko Kawai[2], and Adam Jatowt[4]

[1] Yamaguchi University, 2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan
y.wang@yamaguchi-u.ac.jp

[2] Kyoto Sangyo University, Motoyama, Kamigamo, Kita-ku, Kyoto 603-8555, Japan
spanote@gmail.com, kawai@cc.kyoto-su.ac.jp

[3] Universiti Tenaga Nasional, Jalan Ikram-Uniten, 43000 Kajang, Selangor, Malaysia
syafiq.pozi@uniten.edu.my

[4] Kyoto University, Yoshida-homachi, Sakyo-ku, Kyoto 606-8501, Japan
adam@dl.kuis.kyoto-u.ac.jp

Abstract. Social media such as Twitter has been frequently used for expressing personal opinions and sentiments at different places. In this paper, we propose a novel crowd sentiment analysis for fostering cross-cultural studies. In particular, we aim to find similar meanings but different sentiments between tweets collected over geographical areas. For this, we detect sentiments and topics of each tweet by applying neural network based approaches, and we assign sentiments to each topic based on the sentiments of the corresponding tweets. This permits finding cross-cultural patterns by computing topic and sentiment correspondence. The proposed methods enable to analyze tweets from diverse geographical areas sentimentally in order to explore cross-cultural differences.

Keywords: Crowd sentiment analysis
Similar but sentimentally different · Cross-cultural studies

1 Introduction

Social media offers many possibilities for analyzing cross-cultural differences. For example, Silva et al. [8] compared cultural boundaries and similarities across populations in food and drink consumption based on Foursquare data. Park et al. [6] attempted to demonstrate cultural differences in the use of emoticons on Twitter. Other researches focused on cultural differences related to user multilingualism in Twitter [4,5]. In this context, sentiment analysis has become a popular tool for data analysts, especially those who deal with social media data. It has been recently quite common to analyze public opinions and reviews of events, products and so on social media using computational approaches. However, most of the existing sentiment analysis methods were designed based on a single language, like English, without the focus on particular geographic

G. Chowdhury et al. (Eds.): iConference 2018, LNCS 10766, pp. 67–73, 2018.
https://doi.org/10.1007/978-3-319-78105-1_8

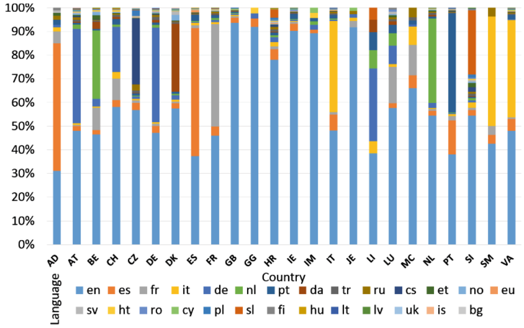

Fig. 1. European language distribution across different European countries in Twitter.

areas and on inter-regional comparisons. It is however necessary to develop new technology to be able to adapt sentiment analysis to a wide number of other cultures and areas [7] and to be able to compare the results. Most current methods cannot explore sentiment differences between diverse geographical areas to provide customized location-based approaches.

To foster cross-cultural studies between different spatial areas, we propose a novel crowd sentiment analysis to find similar semantics which are characterized by different sentiments based on social media data. We use data derived from different geographic places such as different prefectures, municipalities, or countries. In particular as an underlying dataset in our study, we utilize Twitter data gathered using Twitter Streaming API over Western and Central part of Europe issued during approximately 8 months in 2016. The data consists of 16.5 million tweets accumulating to 5 GB memory size. Fig. 1 shows the distribution of languages in our dataset (we show only European languages) accumulated from all users from each analyzed country. We can observe that English is a commonly used language across European countries in Twitter. Therefore, in this paper, for simplicity, we focus on English tweets. We then explore cross-cultural differences based on similar semantics but different sentiments in different geographical areas. Our method delivers two kinds of output based on the proposed crowd sentiment analysis: similar-but-sentimentally-different topics and terms.

For start, users need to select two locations. The method then returns the ranked list of similar-but-sentimentally-different topics (terms) in the form of term clouds, as well as the list of representative tweets for the extracted topics in both the locations. User can also select a time period (e.g., one of seasons) and, by this, the ranked topic (term) list, the term clouds, and the tweet list can be updated. When a user clicks a given term, the method presents the list of its most related tweets. We believe that such data could provide complementary knowledge to many social media studies interested in location-based sentiment analysis of user activities or in sentiment-based recommendation. The ranked term list could also help to improve methods that rely on sentiment analysis by adjusting and correcting sentiment lexicons. Note that although we focus on Twitter, our cross-cultural sentiment analysis can accept any datasets, e.g., services, products, or facilities, for discovering sentiments of topics over tweets. This should be useful for better recommending particular activities, products, services, events, or places to visit for a given segment of users.

2 Crowd Sentiment Analysis

The processing flow of our crowd sentiment analysis is shown in Fig. 2 on Twitter datasets for two geographical areas (e.g., France and Italy). Our approach consists of 3 stages: (1) *Sentiment Modeling* for categorizing tweets into positive and negative by applying neural networks, (2) *Topic Modeling (1, 2)* for detecting tweet topics through utilizing LDA model, and (3) *Topic-Topic Similarity Estimation* for finding similar topics based on output from *Topic Modeling 2*.

In order to identify each tweet's sentiment, we developed a sentiment classification model based on existing labeled tweet dataset used in [2]. The dataset consists of 1,600k tweets used as the training set and 498 tweets for the testing set. Re-tweets and tweets that contain URL have been removed from the dataset. We then use the deep learning approach to implement the classification model. There are three necessary steps in this stage: *preprocessing*, *transformation*, and *learning*. In the preprocessing step, every tweet is cleaned from non-word symbols and converted into a list of terms. Then, these lists are transformed into a vector representation before being fed into the learning algorithm.

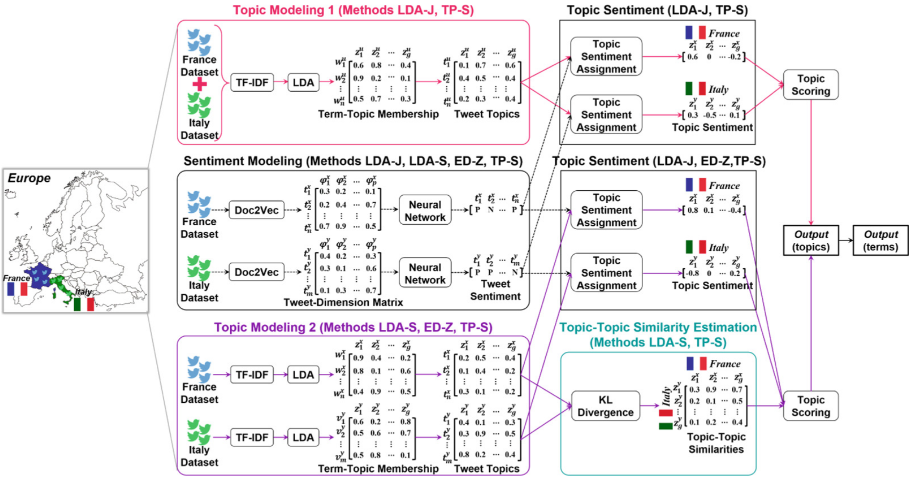
Europe
France
Italy
Topic Modeling 1 (Methods LDA-J, TP-S)
France Dataset
Italy Dataset
TF-IDF
LDA
Term-Topic Membership
Tweet Topics
Topic Sentiment (LDA-J, TP-S)
Topic Sentiment Assignment
France
Italy
Topic Sentiment
Topic Scoring
Sentiment Modeling (Methods LDA-J, LDA-S, ED-Z, TP-S)
Doc2Vec
Neural Network
Tweet-Dimension Matrix
Tweet Sentiment
Topic Sentiment (LDA-J, ED-Z,TP-S)
Output (topics)
Output (terms)
Topic Modeling 2 (Methods LDA-S, ED-Z, TP-S)
Topic-Topic Similarity Estimation (Methods LDA-S, TP-S)
KL Divergence
Topic-Topic Similarities

Fig. 2. Cross-cultural crowd sentiment analysis (e.g., France vs. Italy). For topic output, we propose two methods as listed in Sect. 3.2: LDA-J which is based on *Topic Modeling 1*, *Sentiment Modeling*, and *Topic Sentiment*; and LDA-S based on *Sentiment Modeling*, *Topic Modeling 2*, *Topic Sentiment*, and *Topic-Topic Similarity Estimation*. For term output, we propose ED-Z based on *Sentiment Modeling*, *Topic Modeling 2*, and *Topic Sentiment*; and TP-S based on LDA-S.

Next, every tweet is transformed into a feature vector using Doc2Vec algorithm. It can identify tweets that have similar meaning, which could not be well represented by other feature representation such as bag of words (BoW). Unlike Doc2Vec, BoW, or *TF-IDF* have tendency to produce sparse data. However, the set of human vocabulary consists of almost unlimited number of elements. Hence, representing a single instance over a set of universal vocabulary will always result in sparse vector. Doc2Vec allows large number of features (typically thousands of terms) to be represented in a lower dimensional space. We limit the feature number to 300 features. Each tweet will then have its own vector representation. These representations will be fed into a fully connected neural network for supervised learning.

2.1 Topic Modeling

We perform a topic modeling by using LDA model with *TF-IDF* scored terms of either the joint dataset of different geographical areas (*Topic Modeling 1*) or on separate datasets, each for a given geographical area (*Topic Modeling 2*).

LDA is a generative model in which the topic distribution is assumed to have a Dirichlet prior. After learning is completed, the probability of a term w to belong to a topic z_g ($g \in [1, G]$), $P(w|z_g)$, is known, where G denotes the topic number (G is set to 300 in the experiments). Then, the probability of z_g given a term w can be easily inferred by applying Bayes' rule, $P(z_g|w) \propto P(w|z_g)P(z_g)$, where $P(z_g)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution [1]. Therefore, through the LDA model, we can obtain the probabilistic distribution of topics given the joint dataset of two different geographical areas in *Topic Modeling 1*, or given the datasets of each geographical area treated separately as in *Topic Modeling 2*.

2.2 Topic-Topic Similarity Estimation

Since we have two separate tweet datasets in two different geographical areas for *Topic Modeling 2*, we need to synchronize topics from these datasets. In the next stage, we measure the similarities between topics in two datasets by computing the topic distributions of each dataset using the LDA model, and then computing Kullback-Leibler (KL) divergence [3] between the topic distributions of a pair of topics in two datasets by $D_{KL}(P||Q) = \sum_w P(w) \cdot \log \frac{P(w)}{Q(w)}$.

We consider a topic z_i^x in area x (e.g., France) to be similar to z_j^y in area y (e.g., Italy) if $D_{KL}(P||Q) \leq 0.0002$ for this topic pair. Hence, tweets that belong to such topics are assumed to be semantically similar. Note that for computing KL divergence we always use joint vocabulary from the two datasets.

Finally, we assign sentiment to each topic based on the number of positive and negative tweets covered by the topic by computing the weighted average sentiment score over topics. Based on the computed sentiment scores of topics and the similarities of topics, we can then find semantically similar topics that have different sentiments. The topic pairs in two datasets of two geographical areas x and y are ranked by the Euclidean distance as follows:

$$dist(z_i^x, z_j^y) = \sqrt{(\#pos(z_i^x) - \#pos(z_j^y))^2 + (\#neg(z_i^x) - \#neg(z_j^y))^2} \quad (1)$$

Here, $\#pos(z_i^x)$ ($\#pos(z_j^y)$) returns the number of positive tweets about a topic z_i^x (z_j^y) in the dataset of geographical area x (y), and $\#neg(z_i^x)$ ($\#neg(z_j^y)$) returns the number of negative tweets about z_i^x (z_j^y).

3 Experiments

3.1 Dataset

We collected 8.81×10^6 English tweets produced by 7.41×10^5 unique Twitter's users in South-West Europe during 2016/4/30–12/21. Currently, we test the datasets of two countries: France and Italy. Table 1 shows the dataset statistics.

Table 1. Dataset statistics.

	France	Italy	Total
#Tweets	484,450	470,916	955,366
#Total unique terms	44,970	39,762	84,732
#Ave. unique terms per tweet	9.78	9.58	–
#Positive tweets: #Negative tweets	54k:27k	29k:12k	–

3.2 Metrics and Tested Methods

We use normalized Discounted Cumulated Gain (nDCG) at the following ranks: @5, @10, @20 and @30. Each result is judged using the 1-to-5 Likert scale, where 5 means the highest quality result and 1 indicates the lowest quality. We also compare all the methods using Mean Reciprocal Rank (MRR). The reciprocal rank of scored topics or terms is the multiplicative inverse of the rank of the first correct answer being the highest ranked result whose score is equal or above 4.

Topic Output Evaluation. For cultural studies of different geographical areas to show semantically similar but sentimentally different topics in those areas, we test two methods based on *Topic Modeling (1, 2)*:

1. **LDA without topic-topic similarity (LDA-J).** This method ranks topics on the joint dataset of different geographical areas by *Topic Modeling 1* using LDA based on their sentiment scores.
2. **LDA with topic-topic similarity (LDA-S).** This method ranks topic pairs on two datasets of different geographical areas by *Topic Modeling 2* using LDA based on their sentiment scores and topic-topic similarity.

Term Output Evaluation. We also return terms that have different sentiment values, while having the same semantics and syntactic forms. Such terms can be used for improving sentiment lexicons by geo-based customization. In this context, we set up one baseline and we propose two methods:

1. **Euclidean distance using tweet sentiments (ED-T).** This baseline ranks terms to find semantically similar but sentimentally different terms by the Euclidean distance scores using Eq. (1) where $\#pos$ ($\#neg$) are simply the numbers of positive/negative tweets from the two datasets of different geographical areas, respectively. Here, we remove stopwords and low frequency terms if the frequency is less than 50 times in both datasets.
2. **Euclidean distance using topic sentiments (ED-Z).** This method ranks semantically similar but sentimentally different terms by the Euclidean distance scores in Eq. (1) where $\#pos$ ($\#neg$) means the number of positive/negative topics on two datasets of different geographical areas. Here, we consider a term to belong to a given topic if $P(w|z) > 0.001$.
3. **Term probabilities with topic-topic similarity (TP-S).** We match topics in two datasets of different geographical areas by their similarity and then obtain top-ranked n ($n = 30$ by default) topic pairs (same as in **LDA-S**). Finally, this method ranks terms of the top-ranked topic pairs by computing the sum of their probabilities in the two datasets as given by LDA output within the top-ranked n topic pairs. The score of each term is the sum of its probabilities: $\sum_w P(w|z_i^x) \cdot P(w|z_j^y)$. Here, we remove stopwords and low frequency terms if the frequency is less than 50 times in both datasets.

3.3 Experimental Results

Results of Topic Output Evaluation. The main observation is that our proposed method **LDA-S** based on *Topic Modeling 2* outperforms **LDA-J** based on *Topic Modeling 1* and that **LDA-S** performs best according to nDCG@10, @20, @30, and MRR (see Table 2). Note that **LDA-J** does not perform topic-topic similarity but instead it is using the joint dataset of different geographical areas. Although **LDA-J** performs better than **LDA-S** according to nDCG@5, less important common topics in the joint dataset. Future work will improve **LDA-J** by using a new topic modeling based on Wikipedia corpus.

Results of Term Output Evaluation. The main observation is that our proposed methods **ED-Z** and **TP-S** outperform the baseline **ED-T** and that **ED-Z** performs best according to nDCG@5, @10, @20, and @30 (see Table 2). **ED-T** baseline does not perform any topic modeling. Instead it is just considering

Table 2. Results of topic (term) output evaluation in nDCG@5, 10, 20, 30, and MRR.

Output	Method	@5	@10	@20	@30	MRR
Topic	LDA-J	**0.898**	0.768	0.792	0.816	0.1
	LDA-S	0.861	**0.874**	**0.883**	**0.831**	**0.188**
Term	ED-T	0.826	0.763	0.762	0.784	0.077
	ED-Z	**0.887**	**0.893**	**0.835**	**0.836**	0.063
	TP-S	0.827	0.774	0.796	0.774	**0.1**

the difference of sentiments of the tweets containing a target term in the two datasets. This has the drawback of considering tweets where the terms do not have important role. It is necessary to detect topics and their key representative terms by using a topic modeling as our proposed methods. Comparing the results of the proposed methods **ED-Z** and **TP-S**, we found that **ED-Z** is better than **TP-S** according to nDCG@5, @10, @20, @30. Future work will combine **ED-Z** and **TP-S** to rank terms of top-ranked topic pairs based on **LDA-S** and compute the score of each term by the Euclidean distance scores of the number of positive/negative topics in the top-ranked topic pairs.

4 Conclusion

In this research, we have proposed a cross-cultural crowd sentiment analysis for finding similar topics or identical terms that are however subject to different sentiments as a part of wider cross-cultural study. In future, we will experiment using social media data in other geographical areas (e.g., Asia and America). We will also try to analyze cross-cultural crowd sentiment on each location based on the multilingual analysis of Twitter data similar to [5]. Furthermore, we plan to expand the current analysis method to recommend particular activities, products, services, events, or places to visit for a given segment of users.

Acknowledgments. This work was partially supported by MIC SCOPE (#171507010), and JSPS KAKENHI Grant Numbers 16H01722, 17K12686, 17H01822.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
2. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford 1, 12 (2009)
3. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
4. McCollister, C.: Predicting author traits through topic modeling of multilingual social media text. Ph.D. thesis, University of Kansas (2016)
5. Mohd Pozi, M.S., Kawai, Y., Jatowt, A., Akiyama, T.: Sketching linguistic borders: mobility analysis on multilingual microbloggers. In: WWW 2017, pp. 825–826 (2017)
6. Park, J., Baek, Y.M., Cha, M.: Cross-cultural comparison of nonverbal cues in emoticons on twitter: evidence from big data analysis. J. Commun. **64**(2), 333–354 (2014)
7. Rudra, K., Rijhwani, S., Begum, R., Bali, K., Choudhury, M.: Understanding language preference for expression of opinion and sentiment: what do Hindi-English speakers do on twitter? In: EMNLP 2016, pp. 1131–1141 (2016)
8. Silva, T.H., de Melo, P.O.S.V., Almeida, J., Musolesi, M., Loureiro, A.: You are what you eat (and drink): identifying cultural boundaries by analyzing food and drink habits in foursquare. In: ICWSM 2014, (2014)