Twitter User Classification with Posting Locations

Naoto Takeda[1(✉)] and Yohei Seki[2(✉)]

[1] Graduate School of Library, Information and Media Studies, University of Tsukuba, Tsukuba-shi, Ibaraki 305-8550, Japan
s1621623@u.tsukuba.ac.jp
[2] Faculty of Library, Information and Media Science, University of Tsukuba, Tsukuba-shi, Ibaraki 305-8550, Japan
yohei@slis.tsukuba.ac.jp

Abstract. Twitter contains a large number of postings related to the reputation of products and services. Analyzing these data can provide useful marketing information. Inferring the user class would make it possible to extract opinions related to each class. In this paper, we propose a method that treats each user's posting location for a tweet as a feature in the analysis of user classes. The proposed method creates clusters of geotags (obtained from Twitter tags) to identify the locations most often visited by the target user, which are then used as features. As an example, we conducted experiments to classify targets based on three classes: "student," "working member of society," and "housewife." We obtained an average F-measure of 0.779, which represents an improvement on baseline results.

Keywords: Inferring occupation · Geolocation · Twitter

1 Introduction

Twitter[1] is a popular microblogging service that enables its users to read and write short messages of up to 140 characters. It is now a mainstream social networking service and is becoming a significant element of social infrastructure. Because of its ability to reflect the user's thoughts and actions in real time, Twitter has attracted much research interest in recent years. For this reason, identifying user attributes such as gender, age, and occupation could enable analysts to answer questions such as "What brands are popular among young female users?" or "Which cars do working members of society prefer?" using attribute-based trend analysis. Extracting opinions through Twitter would be less expensive than carrying out traditional questionnaire-based surveys and would permit real-time assessment.

However, attributes such as gender, age, and occupation are not usually disclosed on Twitter. From previous research, only 24.4 % of users disclose their

[1] https://twitter.com/.

A. Morishima et al. (Eds.): ICADL 2016, LNCS 10075, pp. 297–310, 2016.
DOI: 10.1007/978-3-319-49304-6_35

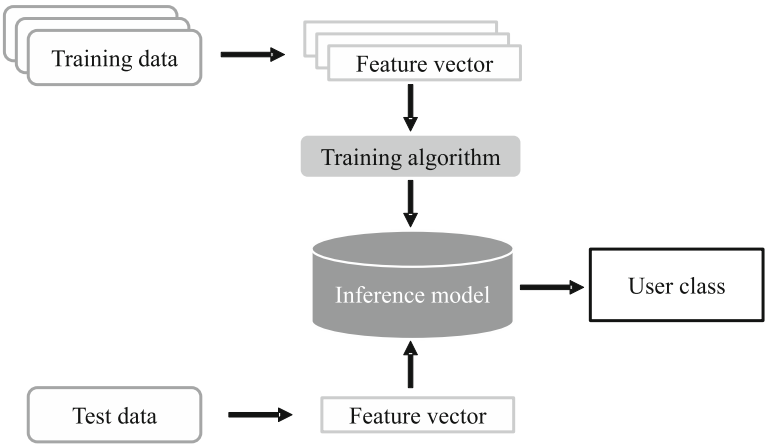

Fig. 1. Process of inferring the user class

occupation in their profiles [7]. For this reason, it is difficult to extract opinions or relate them to the user class as traditional surveys can do. Therefore, some research projects are focusing on approaches to this problem of inferring user classes from user profiles [2,12,13].

In recent years, new services such as Google Now[2] and WebPlaces[3] have offered information based on location data. Research projects and approaches are also beginning to use location information (explained in more detail in Sect. 2.2). In particular, it has been shown that combining location information with the content and time of a posting can reveal new insights, enabling the creation of inference models more accurate than existing methods [4]. In this paper, using location information from users, we infer the user class by a method that focuses on differences between posting locations.

Twitter postings reveal user features such as specific gender-related terms and various posting locations related to the user's occupation. In this study, we analyzed the classes of Twitter users by selecting features and designing an inference model based on machine learning. In one example, inferring the "housewife" class would offer an opportunity to collect housewives' opinions about childcare support. Figure 1 provides an outline of the proposed method. In this research, we used the following three features to infer user class:

– proportion of the postings for each hour of the day at each location,
– characteristic terms for each class in tweets, and
– proportion of the postings for each hour of the day.

These features are described in detail in Sect. 3.

[2] https://www.google.com/intl/ja/landing/now/.
[3] http://www.webplaces.com/.

2 Related Work

2.1 Research on Inferring the Classes of Twitter Users

In this research, a feature vector that includes characteristic terms for each class has been used as an indicator of identity. Cheng et al. [2] extracted terms characteristic of each region and inferred the user's place of residence with 51.0 % accuracy. Rao et al. [13] inferred the user's gender, age, political outlook, and regional origin using the number of followers, the tweet's contents, and the retweet frequency. Preoţiuc-Pietro et al. [12] proposed a method for occupation inference that used general user information, statistics about the tweets (for example, number of followers or average number of tweets/day), and the topics of tweets. The nine target occupations were decided using the Standard Occupational Classification[4]. An accuracy evaluation was carried out, resulting in an average accuracy of 52.7 %.

2.2 Research on Tweet Posting Times and Locations

In our research, we focus on how locations visited by the user each day of the week and each hour of the day lead to differences in target classes. For instance, users belonging to the "housewife" class are likely to tweet during the daytime on weekdays near their place of residence, whereas "students" and "working members of society" rarely do this. By using the proportion of postings per hour and per day of the week as features, it is possible to make highly accurate inferences about the user's class. Gao et al. [4] used time patterns to show that higher accuracy could be achieved by combining the posting time and day of the week with the location, when compared with a Markov-process method that used only location information. Ye et al. [15] analyzed posting times and tweet content, showing that different topics (for example, "university" or "parties") were more likely to be mentioned in particular time slots and on particular days of the week.

3 Method

3.1 The Number of Postings per Hour for Each Location

First, by clustering geotags attached to tweets, we can extract the locations frequently visited by a user. The clustering method used in this research is a density-based spatial clustering algorithm with noise (DBSCAN), as proposed by Ester et al. [3]. Because DBSCAN is a clustering method based on the density of data sets, it enables researchers to obtain high-density clusters. DBSCAN has two parameters, namely $MinPts$ (a threshold for the quantity of data belonging to a cluster) and Eps (a threshold for distances between data points). In the clustering procedure, a cluster is considered to be a set of points containing at

[4] http://www.bls.gov/soc/.

Table 1. Example of a location-information API response

Name	Category	...	Score
Tokyo midtown	Shopping center, mall, commercial complexes	...	87.910
Galleria	Shopping center, mall, commercial complexes	...	87.366
Presse premium Tokyo midtown	Other supermarkets	...	87.015

least $MinPts$ data points within a radius of Eps. Data not belonging to any cluster are considered as noise.

We used the following algorithm, which enables the application of DBSCAN to Twitter geotags and extracts the locations often visited by users.

1. Collect tweets containing geotags posted by a target user.
2. Compute the geographical distances between data points within geotag sets and perform clustering using DBSCAN.
3. From the extracted clusters, detect those with posting dates spread over 7 days or more and consider them "often visited places".

The parameter values used for DBSCAN were $MinPts = 5$ and $Eps = 100\,\mathrm{m}$.

Next, we attached labels to the clusters. We used the Yahoo! Open Local Platform location-information application-programmer-interface (API)[5] provided by Yahoo! Japan. The location-information API takes latitude and longitude as required parameters and returns the names of main landmarks and locations in the area. Several major spots may appear in the response field with scores that take into consideration the level of importance and scope of influence defined for each type of location. For instance, an input latitude = 35.66521320007564 and longitude = 139.7300114513391 (Akasaka 9, Minato-ku, Tokyo) could return the responses shown in Table 1.

Here, we use "score" and "category" within the responses as location labels. "Score" indicates the probability of it being the right location. "Category" represents a category associated with the location, such as "university or graduate school," or "shopping center or mall, commercial complexes." We assign location labels using the location-information API as follows:

1. Input the center of gravity of the extracted cluster via the API.
2. Consider responses for which the response field "score" is at least 70, with the "category" of the spot having the highest score becoming the location label for the cluster.
3. If the maximum value of the response field "score" is less than 70, the location label for that cluster is set to "none."
4. For the clusters tagged as "none," select the cluster with the largest number of points, and tag it as "around the place of residence."

[5] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/placeinfo.html.

However, if there are a large number of labels involved, some may best be treated as noise. Therefore, in Sect. 4.3, we consider in more detail the location labels to be used. In addition, because often-visited places may vary according to the time and day of the week, we compute the proportion of postings for each hour of each day of the week to use as a feature.

3.2 Terms Characterizing Specific Classes that Appear in Tweets

We also consider as features the characteristic terms that appear in tweets from user groups that pertain to the target class. Characteristic terms used as features are selected based on mutual information [8], which expresses the mutual dependency between two random variables. The mutual information $I_{(N)}$ between a class and a term is computed using Eq. (1). This method is often used by researchers for text classification [9,11]. Eq. (1) is applied to all the target classes and terms appearing in the training data.

$$I_{(N)} = \frac{N_{11}}{N} \log 2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log 2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log 2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log 2 \frac{NN_{00}}{N_{0.}N_{.0}} \quad (1)$$

Here, N_{11} represents the number of Twitter data items that pertain to the class and contain the term among the totality of items in the training data. N_{10} is the number of Twitter data items that do not pertain to the class but do contain the term in the training data. N_{01} represents the number of Twitter data items that pertain to the class but do not contain the term in the training data. Finally, N_{00} is the number of Twitter data items that do not pertain to the class and do not contain the term in the training data. Therefore, $N_{.1}$ equals $N_{01} + N_{11}$. In Eq. (1), a large value indicates an output of terms biased toward the class. Terms are considered characteristic features of classes when they are ranked in the top 2,000 in terms of computed mutual information. We use a term's relative frequency within the target user's entire set of tweets as the feature's value. If the same term appears for more than one class, the terms chosen as features are those pertaining to the class with the higher score.

3.3 The Proportion of Tweets per Hour

Finally, we use the time of posting as a feature. We extract the posting time of all of the target user's tweets and compute the number of tweets posted for every hour and every day of the week. However, even in the case of users from the same class, the number of postings can vary considerably from user to user. For this reason, we consider as a feature the proportion of postings rather than the actual number of postings per hour.

4 Classes for Inference and Feature Selection

4.1 Selecting Classes for Inference

In selecting classes for inference, we manually labeled the occupations for 600 accounts selected randomly from the data obtained by Twitter crawling. In cases

where it was difficult to determine a user's occupation from the user profile, we referred to past tweets before choosing labels. The "student" class was identified as representing users who go to school in the daytime on weekdays. The "working member of society" class involved full-time employees who travel to companies in the daytime on weekdays. The "housewife" class comprised married and female users. The "company/group" class represented groups who used their accounts for advertising. For accounts that only retweet postings or post about specific hobbies, we considered the occupation "unknown." In addition, because relatively few accounts were labeled "no occupation" or "part-time worker," we classified them as "others," together with those considered as "unknown." In cases where the user might belong to multiple classes, we prioritized "working member of society", "housewife", "student", and other classes in that order (to best suit marketing applications). In this experiment, no users belonged to multiple classes.

These results are shown in Table 2. Based on this table, as an example of inferring user classes from posting locations, we analyzed the following three classes, "student," "working member of society," and "housewife" (close to 80 % of accounts). Because accounts owned by companies or groups rarely contained tweets with geotags and did not represent individual users, these were eliminated from the target classes.

Table 2. Proportions of manually labeled "occupations"

Class	Proportion
Student	**0.502**
Working member of society	**0.290**
Company/group	0.108
Housewife	**0.032**
Twitter bot	0.017
Others	0.051

4.2 Selecting Location Labels

To obtain the number of postings per location, we identified location labels. For the 300 accounts that posted at least 200 geotags, we clustered the geotags using the method described in Sect. 3.1 and input the center of gravity of each cluster into the location-information API. Table 3 shows the top 15 location categories obtained.

In Table 3, because "None" indicates that no location has been identified, this cannot be used as a feature. Instead, the cluster with the largest number of postings classified as "none" was reassigned the location label "around the

Table 3. Location categories obtained from the location-information API

Category	Proportion
None	0.331
Other supermarkets	0.046
Shopping centers, malls, commercial complexes	0.044
Hotels	0.028
Bookstores	0.025
McDonald's	0.023
Drugstores	0.022
Other casual restaurants	0.020
Stations (JR local lines)	0.019
Stations (other lines)	0.019
Universities and graduate schools	0.015
Elementary schools	0.015
Family Mart	0.017
Lawson	0.014
Seven-Eleven	0.013

place of residence," and this was used as a feature. The location label "elementary schools" and those involving convenience stores, such as "Family Mart," "Lawson," and "Seven-Eleven," were eliminated because there are branches all over the country, which could lead to an incorrect analysis of locations near the clusters adopted. Another observation is that "station" may refer to "subway station," "Japan Railway (JR) station," or "other railway-company station." All of these categories were grouped together under the "stations" label. In the same way, "McDonald's" and "other casual restaurants" were merged. Finally, the categories "high school" and "junior high school" were merged with "universities and graduate schools" under a general "school" label. To summarize the above procedure, we adopted nine location labels, namely "around the place of residence," "other supermarkets," "shopping centers, malls, commercial complexes," "hotels," "bookstores," "drugstores," "other casual restaurants," "stations," and "schools."

4.3 The Selection of Characteristic Terms

We carried out an experiment to select the features of terms to characterize a class (as explained in Sect. 3.2). To select the characteristic terms, we collected 100 Twitter users for each class. To determine the user's class, we checked the user profile, where "20-year-old student", for example, would identify the "student" class. This group comprised 300 users whose tweets were then used as training data. We performed a morphological analysis of the training data

tweets, using MeCab [6]. Japanese is an agglutinative language that should be analyzed using a word segmentation tool. MeCab is an open-source morphological analysis tool for segmenting words. We also computed the mutual information per class for all of the terms that appeared in the training data. The top 2,000 terms for each class (6,000 terms in total) were used to make a feature vector. Table 4 presents examples of the characteristic terms for each class.

Table 4. Examples of the characteristic terms associated with classes

Student	Working member of society	Housewife
Post	News	Good morning
NAVER[a]	"Konkatsu"[b]	Son
Curation	Net	Happiness
Pokemon	Activity	Husband
Japanese "sake"	Working member of society	My family

[a] http://matome.naver.jp/
[b] Search for a marriage partner

5 User Class Inference Using the Posting Location

5.1 Objective

To investigate the validity of inferring three classes ("student," "working member of society," and "housewife") from the posting location, we compared them against a baseline method. The proposed method used a combined vector that contained the proportion of postings for each hour and the location as features, together with a baseline feature vector. The number of dimensions of the vector in the proposed method was 24 (per hour) $\times$ 7 (per day) $\times$ 9 (the number of the location categories) = 1,512 dimensions. The baseline comprised a combined vector containing terms that characterize each class, and a vector containing the proportion of postings for each hour. The number of dimensions of the vector in the baseline was 24 (per hour) $\times$ 7 (per day) = 168 dimensions. The difference between the proposed method and the baseline is therefore only the addition of location information. In previous research [10,14], characteristic terms that infer gender or age have been used as features of the baseline. In contrast, we conducted additional research to infer the user occupation. The day of the week and time of the posting are important data for inferring user occupation because posting time and occupation are closely related. Therefore, the features of the day of the week and time of the posting were added to the baseline.

Based on the results of these experiments, we investigated the locations and times biased toward specific classes and considered them as new features to improve accuracy.

5.2 Data and Methods

The experimental data involved all tweets in Japanese posted during a 1-year period (from July 22, 2014 through July 21, 2015) from users with a Japanese geotag who had posted at least 200 tweets (as well as their own tweets). Twitter streaming APIs[6] were used for data collection. Users with the terms "student," "working member of society," and "housewife" in their profiles were extracted, and 200 users for each class (600 users in total) were selected as sources of experimental data. We checked that the accounts used for this data did not overlap with those used for the analysis in Sect. 4.3. Their classes were manually confirmed. Users without clusters of frequently visited places and those involving 50 clusters or more were removed as representing noise.

We used the accuracy, precision, recall, and F-measure of each class in the evaluation. The assessment was based on a tenfold cross-validation. Support-vector-machine (SVM) and random-forest models were used for classification because they had been used in previous research [1,16] as high-performance classifiers. In this experiment, we used a multilabel classification function for both classifiers. LIBSVM[7] and scikit-learn[8] were used in our implementation. LIBSVM and scikit-learn are open source machine learning libraries. Because the results of the random-forest model can vary because of the random sampling of initial values, a tenfold cross-validation was carried out, and the average value was used in the evaluation.

5.3 Results

Experimental results for the classifiers are shown in Tables 5 and 6. An improvement was found in the precision, recall, F-measure, and accuracy of each class using the SVM model. In comparing the classifiers, the SVM model obtained higher inference accuracy than the random-forest model.

5.4 Discussion

Based on the results obtained, and analyzing the cases of incorrect class inference, examples were found where the location used for inference was incorrect. Particularly in cases where the registered location was not near the center of gravity of the adopted cluster, the location inference could fail. For instance, if there is a school nearby, the location "school" may be incorrectly assigned to tweets that happened to be posted outside a school. Short phrases such as "good morning" and "good night" were often found in postings made by incorrectly classified users. There were many tweets that used the Swarm[9] check-in function. Such tweets and comments contain only information related to the current location or place, such as "I'm in Shinjuku Station in Shinjuku-ku, Tokyo." Users

[6] https://dev.twitter.com/streaming/overview.
[7] https://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[8] http://scikit-learn.org/stable/.
[9] https://www.swarmapp.com/.

Table 5. Inference accuracy based on the SVM model

Method	Class	Precision	Recall	F-measure	F-measure average	Accuracy
Proposed method	Student	**0.800**	**0.776**	**0.784**	**0.768**	**0.772**
	Working member of society	**0.751**	**0.700**	**0.718**		
	Housewife	**0.767**	**0.856**	**0.801**		
Baseline	Student	0.778	0.745	0.759	0.750	0.750
	Working member of society	0.720	0.688	0.701		
	Housewife	0.762	0.832	0.791		

Table 6. Inference accuracy based on the random-forest model

Method	Class	Precision	Recall	F-measure	F-measure average	Accuracy
Proposed Method	Student	0.721	0.810	0.758	**0.728**	**0.732**
	Working member of society	**0.681**	0.667	**0.665**		
	Housewife	0.811	**0.727**	**0.760**		
Baseline	Student	0.727	0.815	0.763	0.726	0.730
	Working member of society	0.669	0.675	0.664		
	Housewife	0.814	0.708	0.751		

with a large number of such tweets are not well suited to inference methods involving vectors containing characteristic terms as features, and this may have reduced the accuracy.

Considering separately the results for the three classes, in the case of "student," improvement was caused by postings within clusters with the location label "school," which did not appear in other classes. Conversely, for the class "housewife," the proposed method showed a high F-measure. This can be explained by the strong relation between the posting location, posting time, and the terms posted by those with the "housewife" class in comparison with the

Table 7. Correct classes and classes inferred using the proposed method

Correct classes / Class inferred	**Student**	**Working member of society**	**Housewife**
Student	**160**	35	11
Working member of society	30	**149**	35
Housewife	10	16	**154**

Table 8. Posting times and locations exhibiting a tendency toward certain classes, according to mutual information

Student	Working member of society	Housewife
Monday 2 pm, school	Thursday 11 pm, around place of residence	Tuesday 5 pm, around place of residence
Wednesday 3 pm, school	Tuesday 10 pm, around place of residence	Wednesday 11 am, around place of residence
Tuesday 6 am, around place of residence	Monday 11 pm, around place of residence	Monday 4 pm, around place of residence
Wednesday 7 am, around place of residence	Sunday 1 am, around place of residence	Thursday 3 pm, around place of residence
Thursday 12 pm, school	Wednesday 10 pm, around place of residence	Wednesday 10 am, around place of residence

other classes. Another observation is that the results for "working member of society" were relatively inaccurate for both the baseline and proposed methods. Table 7 shows the numbers of correct classes and the numbers of classes inferred by the proposed method based on the SVM model. The results indicate that the class "working member of society" was susceptible to incorrect classification as either "student" or "housewife." We considered that, because "working member of society" covers a wide range of ages, it is difficult to find common features when inferring terms and visited places.

We carried out another experiment to investigate location labels with posting times for those cases where people in a certain class visited frequently. The data for the investigation involved 100 different users from those used in previous experiments. Table 8 shows the five highest mutual-information values relating to each class and posting time/location.

Within the class "student," it was found that the label "school," which does not appear in other classes, occupies a high rank. In the "working member of society" case, all location labels were "around place of residence." Regarding posting times, it was found that a large number of postings occurred at night. One possible explanation is that working members of society probably tweet after they have come home from work. In the case of "housewife," all of the location labels were "around place of residence," as was the case with "working members of society." However, one peculiarity was that these tweets were posted on weekday afternoons (after 12:00 noon), a pattern that was not observed for the other classes.

Based on these observations, inferences were made by considering a vector formed by the number of postings in the 100 cases where posting locations and

Table 9. Classification accuracy after adding features based on posting times and locations that exhibit a tendency toward certain classes

Method	Class	Precision	Recall	F-measure	F-measure average	Accuracy
Proposed method (adding features)	Student	**0.800**	**0.798**	**0.792**	**0.779**	**0.782**
	Working member of society	**0.760**	**0.723**	**0.733**		
	Housewife	**0.783**	**0.852**	**0.812**		
Baseline	Student	0.778	0.745	0.759	0.750	0.750
	Working member of society	0.720	0.688	0.701		
	Housewife	0.762	0.832	0.791		

times were most affected by each class. This vector was then merged with the existing vector for the proposed method. The baseline for comparison used a combined vector comprising a vector containing characteristic terms for each class as features and a vector formed by the relative number of posting in each hourly slot. This is the same as the baseline described in Sect. 5.3. Classification was carried out using LIBSVM. The results of the additional experiments are shown in Table 9. By adding the vector whose features involve posting times and locations that are biased toward each class, it was possible to improve the precision, recall, F-measure, and accuracy.

6 Conclusions

In this paper, experiments were conducted using actual data, and the validity of the proposed method (using posting locations) was verified. An improvement was found in the precision, recall, F-measure, and accuracy for each class. However, for the "working member of society" group, the analysis revealed frequent misclassification errors, resulting in a relatively low F-measure. We consider that this was caused by the large variation found within the "working member of society" class in terms of age, making it difficult to find common characteristics for term usage and places visited.

The discussion also included ways to improve the accuracy of location inference. In future work, we plan to identify the posting location from content without geotags because tweets with geotags are generally rare [2,5]. Specifically, by using location labels as correct-answer data, we can select characteristic terms related to specific places as features for inferring the posting location. It will then be possible to apply the proposed method to users who do not post tweets with geotags. Finally, we plan to improve the assessment accuracy by considering the distances between clusters.

Acknowledgment. This work was partially supported by a JSPS Grant-in-Aid for Scientific Research (B) (#16H02913).

References

1. Brdar, S., Ćulibrk, D., Crnojević, V.: Demographic attributes prediction on the real-world mobile data. In: Proceedings of the Mobile Data Challenge by Nokia Workshop in conjunction with International Conference on Pervasive Computing, Newcastle, UK, June 2012
2. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM2010), Toronto, ON, Canada, pp. 759–768, October 2010
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD 1996), Portland, OR, USA pp. 226–231, August 1996
4. Gao, H., Tang, J., Liu, H.: Mobile location prediction in a spatio-temporal context. In: Proceedings of the Mobile Data Challenge by Nokia Workshop in conjunction with International Conference on Pervasive Computing, Newcastle, UK, June 2012
5. Kinsella, S., Murdock, V., O'Hare, N.: "I'm eating a sandwich in glasgow": modeling locations with tweets. In: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC 2011), Glasgow, UK, pp. 61–68, October 2011
6. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, pp. 230–237, April 2004
7. Lee, J., Ahn, J., Oh, J.S., Ryu, H.: Mysterious influential users in political communication on Twitter: user's occupation information and its impact on retweetability. In: Proceedings of the iConference 2015, Newport Beach, CA, USA, March 2015
8. Manning, C.D., Raghavan, P., Schuetze, H.: Introduction to Information Retrieval, pp. 272–275. Cambridge University Press, Cambridge (2008). Chap. 13.5.1
9. Narayanan, V., Arora, I., Bhatia, A.: Fast and accurate sentiment classification using an enhanced naive Bayes model. In: Yin, H., Tang, K., Gao, Y., Klawonn, F., Lee, M., Weise, T., Li, B., Yao, X. (eds.) IDEAL 2013. LNCS, vol. 8206, pp. 194–201. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41278-3_24
10. Otterbacher, J.: Inferring gender of movie reviewers: exploiting writing style, content and metadata. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010), Toronto, Canada, pp. 369–378, October 2010
11. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
12. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through twitter content. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), Beijing, China, pp. 1754–1764, July 2015

13. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC 2010), Toronto, ON, Canada, pp. 37–44, October 2010
14. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Proceedings of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs, Menlo Park, CA, USA, pp. 191–197, March 2006
15. Ye, M., Janowicz, K., Mülligann, C., Lee, W.C.: What you are is when you are: the temporal dimension of feature types in location-based social networks. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Chicago, IL, USA, pp. 102–111, November 2011
16. Zamal, F.A., Liu, W., Ruths, D.: Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors. In: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012), Palo Alto, CA, USA, pp. 387–390, June 2012