Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 57-64 recibido 18-04-2013 revisado 18-06-2013 aceptado 21-06-2013

Characterising social media users by gender and place of residence

Caracterización de los usuarios de medios sociales mediante lugar de residencia y género

Óscar Muñoz-García
Havas Media Group
Madrid - Spain
oscar.munoz@havasmg.com

Jesús Lanchas Sampablo, David Prieto Ruiz
Acceso
Madrid - Spain
jlanchas@acceso.com, dprieto@acceso.com

Resumen: La caracterización de los usuarios mediante atributos sociodemográficos es un paso necesario previo a la realización de estudios de opinión a partir de información publicada por dichos usuarios en los medios sociales. En este trabajo se presentan, comparan y evalúan diversas técnicas para la identificación de los atributos "género" y "lugar de residencia", a partir de los metadatos asociados a dichos usuarios, así como el contenido publicado y compartido por los mismos, y sus redes de amistad. Los resultados obtenidos demuestran que la información proporcionada por la red social es muy útil para identificar dichos atributos.
Palabras clave: demografía, género, lugar de residencia, usuarios, análisis de medios sociales

Abstract: Characterising users through demographic attributes is a necessary step before conducting opinion surveys from information published by such users in social media. In this paper, we describe, compare and evaluate different techniques for the identification of the attributes "gender" and "place of residence" by mining the metadata associated to the users, the content published and shared by themselves, and their friendship networks. The results obtained show that the social network is a valuable source of information for obtaining the sociodemographic attributes of single users.
Keywords: demographics, genre, place of residence, social media analysis, users

1. Introduction

Social media has revolutionized the way in which organizations and consumers interact. Users have adopted massively these channels to engage in conversations about content, products, and brands, while organizations are striving to adapt proactively to the threats and opportunities that this new dynamic environment poses. Social media is a knowledge mine about users, communities, preferences and opinions, which has the potential to impact positively marketing and product development activities (Weber, 2007).

Social media monitoring tools are being used successfully in a range of domains (including market research, online publishing, etc.). Most of these tools generate its reports from metrics based on volume of posts and on opinion polarity about the subject that is being studied. Although such metrics are good indicators of subject popularity and reputation, these metrics are often inadequate for capturing complex multi-modal dimensions of the subjects to be measured that are relevant to business, and must be complemented with ad-hoc studies such as opinion polls.

The validity of these social metrics depends to a large extent on the population over which they are applied. However, social media users cannot be considered a representative sample until the vast majority of people regularly use social media. Therefore, until then, it is necessary to identify the different strata of users in terms of socio-demographic attributes (e.g., gender, age or geographical precedence), in order to weight their opinions according to the proportion of each stratum in the population (Gayo-Avello, 2011). Author and content metadata is not enough for capturing such attributes. As an example, not all the social media channels qualify their

ISSN 1135-5948

users neither with gender nor with geographical location. Some channels, such as Twitter, allow their authors to specify their geographical location via a free text field. However, this text field is often left empty, or filled with ambiguous information (e.g., Paris - France vs. Paris - Texas), or with other data that is useless for obtaining real geographical information (e.g., "Neverland"). For these cases, the friendship networks and the content shared and produced by social media users can be used for estimating their socio-demographic attributes, applying techniques such as geographical entity recognition.

This paper explores different techniques for obtaining the place of residence and gender attributes. Such techniques exploit social users' metadata, the content published and shared by the users to be categorised, and their friendship networks.

The paper is structured as follows. Section 2 summarises related work. Section 3 describes techniques for the identification of the "place of residence" attribute. Section 4 describes techniques for gender recognition. Section 5 evaluates and compare the techniques. Finally, Section 6 presents the conclusions and future lines of work.

2. Related work

The identification of the geographical origin of social media users has been tackled in the past by several research works.

In (Mislove et al., 2011) geographical location is estimated for Twitter users by exploiting the self-reported location field in the user profile, which correspond to the technique described in Subsection 3.1.

Regarding content-analysis approaches, in (Cheng, Caverlee, and Lee, 2010) the authors propose to obtain user location based on content analysis. The authors use a generative probabilistic model that relates terms with geographic focuses on a map, placing 51 % of Twitter users within 100 miles of their actual location. This probabilistic model was previously described in (Backstrom et al., 2008). In (wen Chang et al., 2012) a similar approach is followed, consisting in estimating the city distribution on the use of each word.

In addition, in (Burger et al., 2011) the authors describe a method for obtaining user regional origin from content analysis, testing different models based on Support Vector Machines (Cortes and Vapnik, 1995), achieving a 71 % of accuracy when applying a model of socio-linguistic-features.

With respect to gender identification, in (Burger et al., 2011) the use of profile metadata to identify the gender of the authors is proposed. Using only the full name of the author, an accuracy of 0.89 is reached. Using the author description, the screen name and the tweet text the obtained accuracy is 0.92.

Another relevant related work regarding gender identification is described in (Rao et al., 2010). In this case the proposed method, based on SVM, tries to distinguish the author gender exclusively from the content and style of their writing. This solution needs an annotated seed corpus with authors classified as male or female, to create the model used by the SVM classifier. In this case the accuracy of the best model is 0.72, lower than considering the full name of the author.

3. Place of residence recognition

We have tested different techniques for identifying the place of residence of users, defining "place of residence of a user" as the geographical location where a user lives usually. Each technique is described next.

3.1. Technique based on metadata about locations of users

This technique makes use of the location metadata in the user profile, as for example, the *location* attribute returned by Twitter API when querying user details (Twitter, 2013). Users may express their location in different forms through this attribute, such as geographical coordinates, or the name of a location (e.g., a city, a country, a province, etc.). Therefore, a normalization stage is required in order to obtain a standard form for each location.

For normalising the location this technique makes use of a geocoding API. Our implementation uses Google Maps Web services. This technique invokes a method of the geocoding API that analyses a location and return a normalised tuple composed by a set of components that define the location, including *latitude*, *longitude*, *locality*, and *country*, among others. As for example, if the request "santiago" is sent to the Web service, the response will be a tuple containing "Chile" as the country and "Santiago" as the locality, among other location components. The complete list of components is listed in the API

```
1 function ResidenceFromLocationData(user)
2     return GeoCode(location(user))
```

Listing 1: Technique based on metadata about locations of users

documentation (Google, 2013). Please note that this query does not provide enough information for disambiguating locations, i.e., "santiago" may refer to many geographical locations, including *Santiago de Chile* and *Santiago de Compostela (Spain)*. Therefore the precision of this technique depends on how users describe their location when filling in their profiles. For example, geographical coordinates will define locations accurately, while combinations of city and country (e.g., "Villalba, Spain") will enhance disambiguation (although not completely). In addition, this technique does not return a place of residence when users have not filled in the location field contained in user's profile form of the social network. The technique described next deals with these precision and coverage issues.

Listing 1 summarises the steps executed by this technique.

3.2. Technique based on friendship networks

This technique exploits the homophily principle in social networks (McPherson, Smith-Lovin, and Cook, 2001) for obtaining the place of residence of users. Listing 2 summarises the steps executed by this technique, which are described next.

1. Firstly, we execute the previous technique for obtaining the place of residence of a given user. If a result is obtained, the process finishes. If not, the steps described next are executed (line 2).
2. Secondly, the friends of the user in her online community are collected. After that, the location of each friend is obtained by using the geocoding API. The normalised locations obtained are appended to a list (lines 5-6).
3. Finally, the list obtained in the previous step is filtered iteratively selecting on each iteration the locations that contain the value with the most frequency for a given location component, starting from the country and finishing in the street number, until there is only one location in the set. First the locations whose country are the most frequent are selected, then the locations whose first-order civil entity (e.g., a state in USA or an autonomous community in Spain) is the most frequent, and so forth. The location that remains in the list after completing the filtering iterations is selected as the place of residence of the user. This approach ensures that the most frequent regions in the friendship network of the user are selected(line 7).

```
1 function ResidenceFromFriends(u)
2     l ⇐ ResidenceFromLocationData(u)
3     if l = ∅ then
4         L ⇐ ∅
5         for each f in friends(u) then
6             L ⇐ L ∪ {GeoCode(location(f))}
7         l ⇐ MostFrequentLocation(L)
8     return l
```

Listing 2: Technique based on friendship networks

3.3. Technique based self-descriptions of users

This technique exploits the description published by users about themselves in their profiles for obtaining their place of residence, as for example, the *description* attribute returned by Twitter API when querying a user profile (Twitter, 2013). Listing 3 summarises the steps executed by this technique, which are described next.

1. Firstly, we execute the technique described in Subsection 3.1). If a result is obtained, the process finishes. If not, the steps described next are executed (line 2).
2. Secondly, we obtain the user self-description attribute. Such attribute usually consist on a sentence that have to be processed for extracting the geographical locations mentioned in the text (line 4).
3. After obtaining the description of the user, we identify the language in which user self-description is written. For doing so, we make use of the Freeling (Padró and Stanilovsky, 2012) language recognition feature (line 5).

```
1  function ResidenceFromDescription(u)
2      l ⇐ ResidenceFromLocationData(u)
3      if l = ∅ then
4          desc ⇐ description(u)
5          lang ⇐ IdentifyLanguage(desc)
6          E ⇐ NamedEntities(desc, lang)
7          L ⇐ ∅
8          for each entity in E do
9              if isLocation(entity) then
10                 L ⇐ L ∪ {GeoCode(entity)}
11         l ⇐ MostFrequentLocation(L)
12     return l
```

Listing 3: Technique based on self-descriptions of users

```
1  function ResidenceFromtTexts(u)
2      l ⇐ ResidenceFromDescription(u)
3      if l = ∅ then
4          L ⇐ ∅
5          for each text in publications(u) do
6              norm ⇐ Normalise(text)
7              lang ⇐ IdentifyLanguage(norm)
8              E ⇐ NamedEntities(norm, lang)
9              for each ent in E do
10                 if isLocation(ent) then
11                     L ⇐ L ∪ {GeoCode(ent)}
12         l ⇐ MostFrequentLocation(L)
13     return l
```

Listing 4: Technique based on content

4. Once the language of the user's description has been identified, we perform an entity detection and classification process. As we use Freeling for this purpose, this step is restricted to the languages for which Freeling is able to perform named entity classification (i.e., English, Spanish, Galician and Portuguese). We enable Freeling's multi-word detection (line 6).

5. After that, we filter the named entities obtained in the previous step taking only the entities that correspond to a location. Such entities are sent one by one to the geocoding API for obtaining a set of normalised locations (lines 8-10).

6. As several locations may be obtained in the previous step, once the normalised locations have been obtained, we select only one location by following the same selection approach described in step 3 of the technique explained previously, returning one location as the place of residence of the user (line 11).

3.4. Technique based on content

This technique mines the content published (e.g., tweets and posts) and shared (e.g., retweets and links) by the users to obtain their place of residence. Listing 4 summarises the steps executed by this technique, which are described next.

1. Firstly, we attempt to obtain the place of residence by using the location metadata, as explained in Subsection 3.1. If a result is obtained, the process finishes with a location. Otherwise, the process continues in the following step.

2. Secondly, we use the user self-description as explained in Subsection 3.3. Is a result is obtained, the process finishes, otherwise, the process continues.

3. If the previous steps do not return a location, we obtain the textual contents published and shared by the user. We process each document obtaining a list of normalized locations mentioned in user's generated content. The process followed for obtaining the locations from the content is explained in Subsection 3.4.1.

4. Finally, we select the place of residence of the user from the list of locations obtained in the previous step, by applying the same location selection criteria used for the techniques previously described.

3.4.1. Extracting locations from content

For obtaining the locations from the textual content, we firstly identify the language of the post by applying the method explained in Subsection 3.3.

Secondly, if the content processed is a micro-post (i.e., content posted on Twitter), we perform a syntactic normalisation. This step converts the text of the tweet, that often includes metalanguage elements, to a syntax more similar to the usual natural language. Previous results demonstrate that this normalisation step improves the accuracy of the part-of-speech tagger (Codina and Atserias, 2012), of which the named entity classification module depends. Specifically, we have implemented several rules for syntactic normalization of Twitter messages. Some of these rules haven described in (Kaufmann and

Jugal, 2010). The rules executed by the content normaliser are the following:

1. Transform to lower-case the text completely written with upper-case characters;
2. Delete the sequence of characters "RT" followed by a mention to a Twitter user (marked by the symbol "@") and, optionally, by a colon punctuation mark;
3. Delete mentions to users that are not preceded by a coordinating or subordinating conjunction, a preposition, or a verb;
4. Delete the word "via" followed by a mention to a user at the end of the tweet;
5. Delete the hash-tags found at the end of the tweet;
6. Delete the "#" symbol from the hash-tags that are maintained;
7. Delete the hyper-links contained within the tweet;
8. Delete ellipses points that are at the end of the tweet, followed by a hyper-link;
9. Delete characters that are repeated more than twice (e.g., "maaaadrid" is converted to "madrid");
10. Transform underscores to blank spaces;
11. Divide camel-cased words in multiple words (e.g., "FutbolClubBarcelona" is converted to "Futbol Club Barcelona").

After normalising the text, we use Freeling to extract the locations, as described in Subsection 3.3. We have evaluated this step by using the training set published by the Concept Extraction Challenge of the #MSM2013 Workshop (MSM, 2013). Such training set consist of a corpus of 2.815 micro-posts written in English. The precision obtained is 0.52, while the recall is 0.43 (F_1 =0.47).

Finally we invoke the geocoding API for obtaining the normalized list of locations.

3.5. Hybrid technique

This technique combines the ones described previously, executing one after the other, ordered by computational complexity. Listing 5 summarises the steps executed by this technique, which are described next.

1. Firstly, we execute the technique based on content, which has been described previously (line 2).

```
1 function ResidenceHybrid(u)
2     l ⇐ ResidenceFromTexts(u)
3     if l = ∅ then
4         L ⇐ ∅
5         for each f in friends(u) do
6             L ⇐ L ∪ {ResidenceFromTexts(f)}
7         l ⇐ MostFrequentLocation(L)
8     return l
```

Listing 5: Hybrid technique

2. Finally, if the previous step does not return a place of residence, we make use of the friendship network of the user, by applying this hybrid technique to the list of friends of the users, and selecting the location as described in Subsection 3.2 (lines 3-7).

4. Gender Recognition

We have tested two techniques for gender recognition which are described next.

4.1. Technique based on user name metadata

This technique exploits publicly available metadata associated with the user profile. Such metadata may include the user name, as for example, the *name*, and *screen_name* Twitter attributes (Twitter, 2013).

The technique makes use of two lists of first names that have been previously classified by gender (one list for male names, and one list for female names). The lists have been curated, so unisex names have been excluded for classification purposes, given the ambiguity that they introduce. Specifically, we have generated the lists of first names from the information published by the Spanish National Institute of Statistics (INE, 2013). The initial list contains 18,697 first names (single and composite) for males and 19,817 first names for females. After the curation process (removing the first names that appear in both lists) the male first names list is reduced to 18,391 entries and the female names list to 19,511. Some examples of removed first names are Pau, Loreto and Reyes, as they are valid for either males and females in Spain.

Given a user account, its name metadata is scanned within the lists and, if a match is found in one of the lists, we propose the gender associated to the list where the first name has been found as the gender of the user. Our

technique not only takes the current value for the name metadata, but also the previous values for each attribute, as our data collection system stores historical data.

The proposed method is mostly language independent, being the only language-dependent resource the lists of first names. Those lists could be manually created from scratch, but there are plenty resources readily available, such as population censuses that can be used to build them.

4.2. Technique based on mentions to users

This techniques exploits the information provided by mentions to users. As for example, if someone post in Twitter *"I'm going to visit to my uncle @Daureos to Florida"*, she is providing explicit information about the gender of the user mentioned. We know that *@Daureos* is male because of the word "uncle" written before the user identifier.

We propose a technique for the Spanish language that performs a dependency parsing of the text with the aim of determining the gender of the terms related with the user mentioned. Therefore, for each tweet in which the user is mentioned, we attempt to estimate the gender of the user. Note that not all mentions to users provide information for estimating their genders (e.g., *"via @user"* and *"/cc @user"* at the end of the tweet). The dependency parser used is TXALA (Atserias, Comelles, and Mayor, 2005).

The steps executed by this technique are the following:

1. Firstly, we execute technique based on user name metadata described previously. If a gender is obtained, the process finishes.
2. If a gender is not identified in the previous step, we obtain all the tweets that mention the user.
3. For each tweet, we perform a dependency parsing. Once obtained the dependency tree, we assign a gender to the user for the tweet analysed according to the following heuristics: (1) if the gender of the term in the parent node, of the branch where the user is mentioned, is male or female, we consider that the user is male or female accordingly (e.g., *"Mi tío Daureos"*); (2) if some of the child nodes of the node corresponding to the user mention correspond to a term with a specific gender, we consider that the gender of the user correspond to the gender of such terms (e.g., *"Vio a Daureos enfermo y triste"*); (3) If there is a noun adjunct as the predicate of an attributive sentence where the user is the subject, we assign the gender of the noun adjunct as the gender of the user (e.g., *"Daureos es trabajador"*).
4. Finally, we select the gender that is associated the most to the tweets analysed for the user being analysed.

5. Evaluation

5.1. Place of residence recognition

We have evaluated the place of residence recognition techniques with an evaluation set of 1,080 users extracted from Twitter whose place of residence is known. Users in the evaluation set are distributed among 11 different countries (Argentina, Chile, Colombia, Spain, USA, Japan, Mexico, South Africa, Switzerland, Uruguay and Venezuela). Such users share and publish content in different languages (mainly in Spanish and English).

For evaluating the techniques that make use of the friendship networks, for practical reasons we have restricted the number of friends for each user to 20 (10 followers plus 10 persons followed by the user to be characterised), since Twitter limits the number of calls to its API. With respect to the techniques that make use of the content published and shared, we have restricted the number of tweets analysed to 20, for the same practical reasons, including tweets authored by the user and retweets.

All the techniques achieve a similar accuracy ($\approx 81\,\%$), with the exception of the technique based on friendship networks, which improves de accuracy to 86 %.

5.2. Gender recognition

To evaluate the techniques described we have considered an aleatory sample consisting on authors who have written a tweet in Spanish, as well as tweets that mention those authors between 29^{th} May 2012 and 27^{th} March 2013. The language of each tweet has been identified using LingPipe (Alias-i, 2008). The error of the language identification task causes the inclusion of authors in the evaluation corpus that might not be Spanish speakers, penalising the method recall.

	Predicted class		
Actual class	Male	Female	No gender
Male	530	42	49
Female	10	528	20
No gender	130	97	103

Table 1: Confusion matrix with the results of the technique based on mentions to users.

The evaluation set obtained for gender recognition contains 69,261 users. From these users, the technique based on profile metadata has been able to classify 46,030 users (9,284 female users and 36,746 male users), achieving a coverage of 66 % of the corpus. By contrast, the technique based on mentions to users has classified 46,396 users (9,386 female users and 37,010 male users), improving the coverage up to 67 %.

For evaluating the accuracy, we have annotated by hand the gender of 1,509 users (558 female users, 621 male users and 330 neutral users), and checked the automatic classification with respect to the manual annotation, obtaining an overall accuracy[1] of 0.9 for the technique based on user names, and of 0.84 for the technique based on mentions to users. By gender, for the technique based on user names, the precision obtained is 0.98 for male users and 0.97 for female users, while the recall is 0.8 and 0.87 respectively. For the technique based on mentions to users, the precision obtained is 0.8 for male users and 0.79 for female users, while the recall is 0.85 and 0.95 respectively. Therefore, the technique based on mentions to users achieves a smaller precision, but increases the recall with respect to the technique that only makes use of user names.

Table 1 shows the confusion matrix for the technique based on mentions to users. Users manually annotated as "no gender" correspond to non-personal Twitter accounts (e.g., a brand or a corporation), while those automatically classified as "no gender" are the users for which the algorithm was not able to identify a gender. Mainly, the confusions are produced between the male and female classes and the residual class.

In (Mislove et al., 2011) the authors propose techniques to compare Twitter population to the US population along three axes (geography, gender and race). Regarding the gender identification task, the method proposes a gender for 64.2 % of the authors (in our experiment this percentage is 66.45 %). In addition the 71.8 % of the users identified are males, while our experiment identifies the 79.8 %, obtaining similar distributions by gender.

[1] Accuracy $= \frac{tp+tn}{tp+tn+fp+fn}$

6. Conclusions

In this paper, we have described different techniques for obtaining the demographic attributes "place of residence" and "gender".

The evaluation results obtained for the techniques for identifying the place of residence of Twitter users show that the techniques that make use of the user's community achieve better performance than the techniques based on the analysis of the content published and shared by the user. While the major part of the community of a user uses to share the place of residence (because of the homophily principle in social networks), the mentions to locations included in the content published by the users are not related necessarily with their place of residence. Therefore, the hybrid technique does not perform better than the other techniques based on content.

We have achieved very satisfactory results for gender identification by just making use of user profile metadata, since the precision obtained is high and the technique used is very simple with respect to computational complexity, which leads to a straightforward set up in a production environment. The technique based on mentions to users increases the recall in the cases where the previous technique is not able to identify the gender of a given user, because for the Spanish language there exists grammatical agreement with respect to gender between nouns and other part-of-speech categories (e.g., adjectives and pronouns). However, such technique requires a language-depending dependency parser.

Future lines of work include experimenting with the detection of more demographic and psycho-graphic user characteristics which are relevant to the marketing and communication domains, including: age, political orientation and interests, among others.

7. Acknowledgement

This research is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (`http://www.cenitsocialmedia.es`).

References

Alias-i. 2008. LingPipe 4.1.0. `http://alias-i.com/lingpipe`. [Online; accessed 8-April-2013].

Atserias, Jordi, Elisabet Comelles, and Aingeru Mayor. 2005. Txala: un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural*, (35):455–456, September.

Backstrom, Lars, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. 2008. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 357–366, New York, NY, USA. ACM.

Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geolocating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, New York, NY, USA. ACM.

Codina, Joan and Jordi Atserias. 2012. What is the text of a tweet? In *Proceedings of @NLP can u tag #user_generated_content?! via lrec-conf.org*, Istanbul, Turkey, May. ELRA.

Cortes, Corinna and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September.

Gayo-Avello, Daniel. 2011. Don't turn social media into another 'literary digest' poll. *Communications of the ACM*, 54(10):121–128, October.

Google. 2013. The Google Geocoding API. `https://developers.google.com/maps/documentation/geocoding/`. [Online; accessed 8-April-2013].

INE. 2013. INEbase: Operaciones estadísticas: clasificación por temas. `http://www.ine.es/inebmenu/indice.htm`. [Online; accessed 8-April-2013].

Kaufmann, Max and Kalita Jugal. 2010. Syntactic normalization of twitter messages. In *Proceedings of the International Conference on Natural Language Processing (ICON-2010)*.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain, July.

MSM. 2013. Making Sense of Microposts (#MSM2013) – Concept Extraction Challenge. `http://oak.dcs.shef.ac.uk/msm2013/challenge.html`. [Online; accessed 8-April-2013].

Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 37–44, New York, NY, USA. ACM.

Twitter. 2013. REST API v1.1 (GET users/show). `https://dev.twitter.com/docs/api/1.1/get/users/show`. [Online; accessed 8-April-2013].

Weber, Larry. 2007. *Marketing to the Social Web: How Digital Customer Communities Build Your Business*. Wiley, June.

wen Chang, Hau, Dongwon Lee, M. Eltaher, and Jeongkyu Lee. 2012. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 111–118.