On Classifying Twitter Users' Policy-Relevant Community Affiliations Using DBpedia

Anjie Fang[1,a], **Philip Habel**[2,b], **Iadh Ounis**[2,a], **Craig Macdonald**[2,a], **Xiao Yang**[2,a]
[1]a.fang.1@research.gla.ac.uk
[2]firstname.lastname@glasgow.ac.uk
[a]School of Computing Science
[b]School of Social and Political Sciences
University of Glasgow

Abstract

Scholars have aimed to understand the ways in which different audiences on social media communicate with one another, especially who influences whom among, for example, politicians, the media, the public, and policy experts. However, to examine the flows of information within and across such communities, it is necessary to design a classifier that accurately and reliably captures the community affiliation of any given social media user, with relevant communities including, for example, members of the mass media; academics, business elites; politically interested users, and citizens. To train a classifier to accomplish this task, we propose a rough labeling method using DBpedia entities. We show that this labeling method for Twitter users has a reasonable agreement with human judgments of users' community affiliations, which we obtain through a crowdsourced user study. To increase the confidence in our training data and to improve the accuracy of our classifier, we also propose a user pooling method, which improves the accuracy of our classifer by 2%.

1 Introduction

Today a diverse audience relies on social media networks such as Facebook and Twitter to express their views on topics including political issues and events. For decades social scientists have sought to understand a range of questions related to the political influence of individuals and groups, such as whether the views of citizens drive policymakers to take action; or whether the media set the agenda; or whether certain citizens and elites are able to persuade others. Today social media allows researchers to address these questions in new and innovative ways, as scholars can now observe communications in real-time—and across different audiences. However to model information flows and to understand information dynamics within and across given communities of users (e.g. members of the media or politicians or citizens), it is first necessary to be able to understand accurately to which community a given social media user belongs. For this we propose a machine learning approach.

To assess the quality of democratic representation, scholars have long been interested in understanding the views of various political actors. These actors and communities of actors include, for example, politicians who make public policy—with long-standing debates over whether they are constrained by the views of the public (Erikson, MacKuen and Stimson, 2002) or whether politicians lead the public toward their preferred policy positions (Jacobs and Shapiro, 2000). Researchers have also been interested in the role of the media in influencing policymakers and citizens (Habel, 2012), and whether policy elites such as the business community exert influence (Hillman, Keim and Schuler, 2004), or whether informed communities such as academics matter for policy outcomes (Mansfield, 1995).

To examine the information dynamics among and across such communities, we first investigate a means of classifying social media users into appropriate communities, here using Twitter data.[1] Twitter provides limited publicly accessible information on individual users; however, we argue below that the information individuals disseminate on Twitter in the way of their profile statements and/or Tweets offers researchers valuable insights regarding the community to which a user belongs. We assume that a member of academia, for example, is likely to list an academic affiliation in

[1]We focus our analysis on Twitter, as the data are straightforward to obtain and are in the public domain.

her profile statement, or talk about matters related to academia over social media, such as a pending conference presentation. Indeed scholars have shown that information spread over Twitter has widespread utility. For example, a user's network of followers contains useful information regarding one's own political ideology (Barbera, 2015). Of course one could classify social media users manually, but such an approach is fundamentally limited, as it does not allow one to speak to a massive number of Twitter users, and particularly, to new users. Some existing works leverage features such as friendship networks and communication behaviors (Rao et al., 2010). However this approach is not feasible to implement over millions of users.

To implement the classifier, we use word features, which are relatively easy and quick to obtain. To achieve a large number of training Twitter users for a classifier, we propose a rough labeling method using DBpedia entities. DBpedia is a comprehensive source listing affiliations for millions of people worldwide. This labeling method identifies whether a Twitter user's description and recent tweets imply certain community-related entities. We then turn to a crowdsourced user study to obtain the agreement between our rough labeling method and humans who categorized selected Twitter users into one of 5 different communities. The results show that our classifier has a mid-level agreement with the human assessments. The initial classification results indicate that our classifier has a moderate performance over Twitter users. Additionally, we propose a user pooling method to improve the performance of the classifier, which we show boosts accuracy by 2%. The performance of the classifier can likely be further improved by refining DBpedia entities using a deep nets model, which we save for future work.

2 Approaches to Classification

Our work is not the first to aim to classify social media users based on the content of their communications. Rao et al. (2010), Al Zamal, Liu and Ruths (2012), Mislove et al. (2011) and Barbera (2016) have, for example, classified users' ages, political orientations, ethnicity and sex using features such as their last names, or the following networks, or communications (i.e. tweeting, re-tweeting and replying), or a combination thereof. Chen et al. (2015) take advantage of friends and follower networks, user profiles, and user images to derive politically relevant sociodemographic characteristics of users. Barbera (2016) goes further, training the classifier through a dataset comprised of geolocated Tweets to publicly available voter records and housing prices. Apart from these, Pennacchiotti and Popescu (2011) also showed that a superior performance can be achieved by leveraging a Twitter user's profile description. Indeed, most of users indicate their occupation, interests and organization affiliation, which may imply their political orientations or ethnicities. Fang et al. (2015) also reported that the text of Twitter user's communicated tweets (word features) can be used to effectively classify a user's political orientation. The effectiveness of word features has also be investigated by Lee et al. (2011) to improve the categorization of trending topics.

Although most of these existing works have demonstrated a relatively high accuracy, their approaches are not efficient for large scale social media user classification. It can be challenging to generate hand-crafted features for a massive number of users (i.e. features leveraging a user's following network, or communication behavior, for example.), as one quickly runs up against API rate limits. Barbera (2016) goes further than previous approaches, limiting draws from the API to 5. [2] Our proposed approach is able to effectively and efficiently classify a large number of Twitter users, as it does not draw from the Twitter API.

The training data—Twitter users with known community affiliations—can be rendered either from human annotations, e.g. (Rao et al., 2010), or it can be generated automatically. For the latter, Read (2005) and Go, Bhayani and Huang (2009) obtained their sentiment training data by identifying whether a tweet contained relevant emoticon symbols. Similarly, Fang et al. (2015) relied on the use of Scottish Independence Referendum hashtags to obtain Twitter users' preferences for or against independence. Bagdouri and Oard (2015) acquire their training data from a seed list of journalists so as to classify journalists on Twitter. In the next section, we explain how we use a rough labeling method to generate our training data. We focus on the community identification of the following set of users: *Media; Academics; Business Executives; Politically Interested Users*; and *Citizens* (the residual category). The messages dis-

[2] Twitter rest API has to be used to obtain such features.

Community	Predicate	Object	# of Entities Extracted
Academics	Subject:Category 22-rdf-syntax-ns #type 22-rdf-syntax-ns #type ...	Science_occupations University EducationalInstitution ...	167,495
Media	Subject:Category 22-rdf-syntax-ns #type 22-rdf-syntax-ns #type ...	Journalists Broadcaster Newspaper ...	59,601
Business Elites	Subject:Category 22-rdf-syntax-ns #type 22-rdf-syntax-ns #type ...	Business_and _financial_occupations Company108058098 BusinessPerson ...	18,048
Politically Interested Users	Subject:Category Subject:Category 22-rdf-syntax-ns #type ...	Legislators Political_occupations Political Party ...	191,330

Table 1: Combinations of DBpedia predicates and objects for 4 communities of users.

seminated within and across these communities, and the influence of these communications, should be of particular interest to social scientists.[3]

3 A Labeling Method Using DBpedia Entities

DBpedia has developed into an interlinking hub, which contains information regarding individuals, groups, and companies from millions of Wikipedia entries (Bizer et al., 2009). Roughly half of the resources are entities including persons, organizations and products (Mendes et al., 2011). Thus, we propose a labeling method using DBpedia entities to obtain our ground truth training data. This method allows us to get a very large training dataset from tweets, and then we can use the obtained training data to inform our classifier. In this section, we explain how we use DBpedia entities to generate our training data. There are two steps in our rough labeling method:

1. Extracting Community Related Entities. In DBpedia data, an entity is represented by several entries in n-triple format `<subject> <predicate> <object>`, where `<subject>` is the entity name. Usually, the combinations of `<predicate> <object>` describes the properties of an entity. For instance, `<Professor> <subject> <Academic_terminology>` is one entry

[3]Politicians are here considered members of the Political Interested Users set, but one can manually remove politicians by defining a set of politicians and locating their social media accounts manually.

	Academics	Media	Business Elites	Politically Interested Users	Citizens
users	4982	5512	22K	4538	6M
Tweets	84K	97K	430K	106K	153M

Table 2: The number of the labeled user of 4 communities.

of the entity "Professor". By distinguishing combinations of `<predicate> <object>`, we can identify that entity "Professor" belongs to the community "Academics". Therefore, we pre-define the combinations of `<predicate> <object>` for 4 communities[4]. Some examples of these combinations are shown in Table 1. If entities contain the `<predicate> <object>` combination of community c, then these entities belong to community c. For example, users listing "Harvard University" as an affiliation belong to the academic community.[5] We extract the community-related entities from 4 DBpedia datasets[6]: `instance types`, `articles categories`, `YAGO types` and `UMBEL links`. For each community, we get a number of entities shown in Table 1. Note that all the used DBpedia datasets are in English, which means that the extracted entities are English only.

2. Filtering users. We check whether a Twitter user's profile description and recent tweets imply the extracted entities of the 4 communities listed above. If a Twitter user's description and more than 20% of their tweets imply the entities from the same community c, then this user is labeled as community c. In order to compile a large training dataset, we use an existing Twitter background data (Fang et al., 2016*a*,*b*), which contains about 10% random tweets posted from September 2015 to March 2016. Instead of checking whether the text of a tweet/profile contains the community-related entities, we use DBpedia Spotlight (Daiber et al., 2013) to extract the entities from tweets by taking into account the topical pertinence and contextual ambiguity. We obtain thousands of users (See Table 2) with community labels from the Twitter background dataset, which is larger than the training dataset in the aforementioned literature. For each labeled Twitter user in-

[4]We do not use the proposed labeling method to generate the training data for "Citizens". We consider the Twitter users who do not belong to the other 4 communities to be *Citizens*.

[5]The classifer has difficulties when encountering information pertaining to multiple communities. For example, an affiliation of Harvard Management Company has properties of both the academic and business communities.

[6]wiki.dbpedia.org/Downloads2015-10

stance (obtained by the rough labeling method), we also extract their tweets from the Twitter background dataset. As we mentioned previously, the extracted community-related entities are English. We only consider English tweets in the Twitter background datasets. Therefore, the labeled users could be from any English-speaking country. The labeled users along with their tweets are used to form our training data.

4 Crowd-sourced User Study

To check the quality of the generated training data, we conduct a crowdsourced assessment using the CrowdFlower[7] platform. We present the crowdsourcing worker with a given Twitter user's profile and 8 of his/her tweets. The worker is asked to choose one community label out of the 5 provided. The user interface of this assessment is shown in Figure 1. We instruct workers to label the given users by considering the rules below:[8]

Media: People who work in newspaper/broadcast companies are media people. They can be journalists, reporters, correspondents, etc. Most of them hold a neutral position when reporting some events/news. On Twitter, there is a type of user, who spreads/forwards the latest news or writes news stories by themselves. These users use a lot of hashtags and mentioned Twitter user names to get people's attentions. They may not work for any newspaper/broadcast companies, but they are also media people.

Academics: People who are doing research are academic people. They may focus on a specific field, e.g. Chemistry, Mathematics or Social Science. These people mostly come from universities or research institutes. A Twitter user who describes himself/herself as a researcher or who has an academic title (e.g. Professor, Associate Professor, Lecturer, etc.) are academic people. Their tweets may pertain to the latest techniques, research findings or projects.

Business Elites (BE): People from commercial company/industry/manufacture should be labeled as business users. They usually post tweets about the new products of their companies, how to manage a project/company/team, etc. Or, they interact with the other business people in similar field on Twitter. Their Twitter description may contain their work title (e.g. Manager, CEO). A business person may use a Twitter account to introduce his/her business on Twitter.

Politically Interested Users (PIU): People who actively engage in discussion related to political topics, should be labeled as PIU. Even people do not have any political affiliations, they should be labeled as PIU if they post many politically relevant tweets. [Note that we classify whether a Twitter user is interested in politics or not. In this work, we combine politicians and people who are interested in politics into a community called *Politically Interested Users (PIU)*. In future work, we will differentiate between these communities by identifying the Twitter handles of those who have been elected or are running for office.]

Citizens: People who apparently do not belong to the other categories are labeled as *Citizens*. Most of their tweets are about life. These people may include normal citizens, workers and even famous people, as long as they post a lot of life related tweets.

In this study, we randomly sample 200 Twitter users from each community for classification. In total, 1000 Twitter users are evaluated by 124 different trusted Crowdflower workers.[9] Each Twitter user receives 3 assessments of his/her community membership from 3 different trusted Crowdflower workers. In our user study, 92.7% of Twitter users received at least 2 consistent categorizations from 3 human workers, which indicates that the workers have agreement for most of the given Twitter users. In other words, instances where there was no agreement across three workers were rare, occurring in less than 8 percent of the time.

Using human judgments as the ground-truth, the total accuracy of our labeling method is 52%. The `F1, Precision, Recall` scores are shown in Table 3. *Academics* and *PIU* get high recall but low precision, which indicates that the classifier better categorizes the user into these two communities than the others. *Media* and *Business Elites* have an average performance. Considering that the baseline probability for a user being classified into any one category is 0.2, our results are a marked improvement over chance. However, there is room for improvement, for example refining the `<predicate> <object>` for the communities.

[7] crowdflower.com

[8] The community labels we use in the paper are slightly different from those used in the CrowdFlower user interface.

[9] We set quality control strategies following (Fang et al., 2016*a*,*b*).

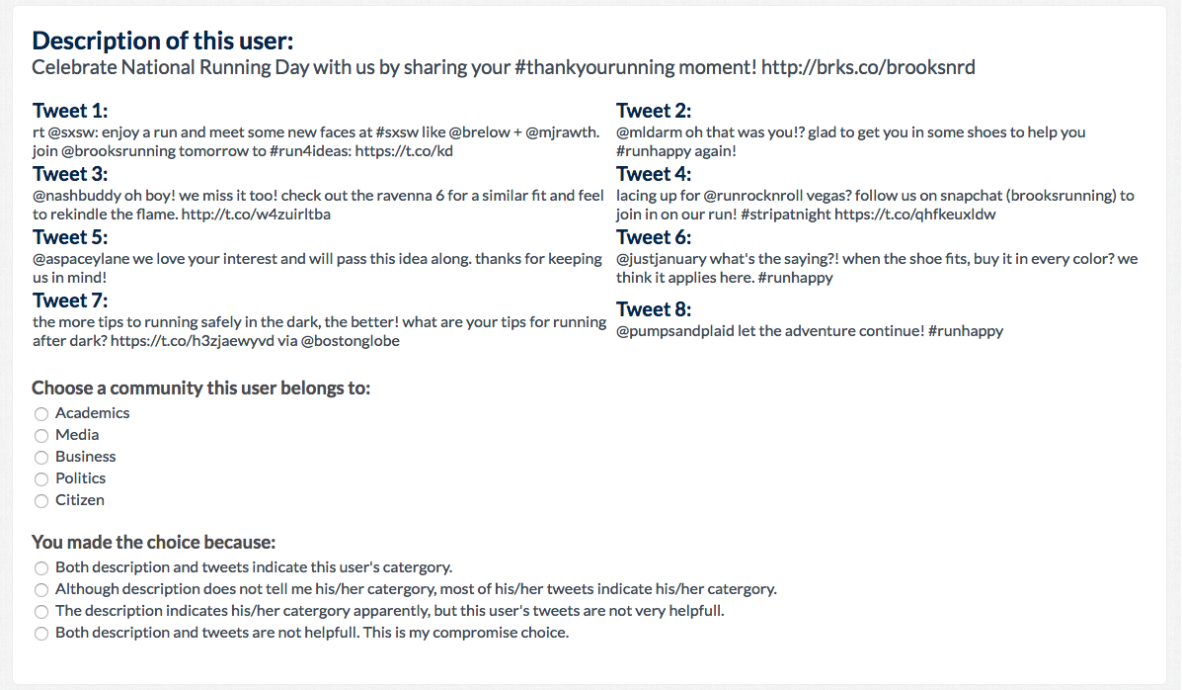

Figure 1: The CrowdFlower User Interface.

	Academics	Media	BE	PIU	Citizens
F1	0.549	0.451	0.570	0.465	0.547
Precision	0.385	0.445	0.580	0.306	0.890
Recall	0.905	0.458	0.560	0.968	0.395

Table 3: The agreement between human judgments and the rough labeling method.

5 Classification and User Pooling Method

As noted above in Section 2, it can be a challenge to generate hand-crafted features (features leveraging a user's following network, communication behavior, etc.) for a massive number of users. To quickly classify a large number of unknown Twitter users, we use word features in order to avoid rate limits from the Twitter rest API. In the following classification experiments, we randomly select roughly 5K users from each of the communities as the training data[10]. We use 1K verified users by our crowdsourced workers as the testing data. The training and test data are all transferred into `TF-IDF` vectors before the procedure. Several classification machine learning algorithms are applied including `Support Vector Machine` (SVM), `Navie Bayesian` (NB), etc. Since the Twitter users in our generated training data could be from any English-speaking country, this proposed classifier is applicable to English-speaking Twitter users.

We notice that our labeling method has a reasonable, but not especially high agreement with human judgments, 52% to be precise. This tells us that some aspect of training data is flawed. To deal with noise in the data, we propose a user pooling method. Such a pooling method has been used to improve the topic modeling for tweets (Mehrotra et al., 2013). More specifically, the probability of a training instance relevant to its given community label (confidence) is 52%. If we randomly combine n users from community c to form a combined user, confidence can be increased to $1-(52\%)^n$. We assume that every user instance is relevant to the given community labels, and thus the classifier should be able to learn the correct mutual patterns among the combined Twitter users from the same community. While a higher n results in a higher confidence, it is less likely that a combined user resembles a real one. Thus there is a tradeoff, and it is necessary to set up the correct parameter n. Here we set the parameter n in the pooling method to (2..10) for the SVM classifiers.In the following section, we show the initial results of our classification.

6 Classification Results

In this section, we show the initial results of the proposed classifiers and the performance of the

[10]We select balanced Twitter users for communities to avoid biases.

		Academics	Media	BE	PIU	Citizens	Avg. of 5 communities	Accuracy
Naive Bayes	F1	0.434	0.272	0.583	0.381	0.504	0.462	0.451
	Precision	0.333	0.272	0.530	0.247	0.893	0.609	
	Recall	0.624	0.273	0.647	0.841	0.351	0.451	
SVM	F1	0.418	0.381	0.471	0.427	0.512	0.465	0.452
	Precision	0.292	0.346	0.514	0.299	0.784	0.571	
	Recall	0.729	0.423	0.435	0.746	0.380	0.452	
SVM & Pooling	F1	0.458	0.372	0.560	0.458	0.503	**0.483**	**0.475**
	Precision	0.319	0.347	0.555	0.312	0.902	**0.636**	
	Recall	0.812	0.402	0.565	0.857	0.349	**0.475**	

Table 4: The classification results.

pooling method. The testing data is comprised of the 1000 verified users whose community labels are obtained from the crowdsourced workers. The training data consists of 25K randomly sampled users (excluding the 1000 testing users) spread across 5 communities. Due to the space limitations, here we only show the results of Naive Bayes (NB), Support Vector Machine (SVM) and SVM & User pooling method (SVM&Pooling) in Table 4.

The accuracy of NB and SVM are about 45%. However, their performances are distinct in different communities. For instance, NB works better on *Business Elites* (in terms of `F1, Precision, Recall`) while SVM works better on *Politically Interested Users*. The results show that *Media* is the most difficult community to classify, which is surprising given that this domain on its face seems relatively well defined compared with the others. In fact, even *Business Elites* and *Citizens* are relatively easier to classify than *Media*. To improve the performance of the classifier, we deploy the user pooling method mentioned above. The results in Table 4 demonstrate some improvement, although quite modest, indeed we see a 2% improvement in accuracy for SVM&Pooling with $n = 6$.

7 Conclusion & Future Work

We investigated a way for classifying Twitter users into communities of interest to social science, as a first step in examining the dynamics of information influence across groups such as members of the mass media, politicians, elites, and citizens. We proposed a labeling method from DBpedia to generate a large-scale training dataset. This method labels a Twitter user to 5 communities. i.e. *Media, Academics, Business Elites, Politically Interested Users*, and *Citizens*. By using this method, we obtained a larger training dataset, and we were able to apply the community labels to Twitter users. We then conducted a crowdsourced user study and showed that the rough labeling method has middle-level agreement with human judgements. We used the generated training data to implement our classifier using the word features. To improve the accuracy, we proposed a user pooling method to increase the confidence in the training process. Our initial results showed that our classifier have a moderate performance and that the new pooling method improves the accuracy of the classifier by 2%.

Our ultimate research aim will be to examine the information dynamics among communities on social media networks. As we move forward, we can apply topic modelling approaches to model information flows within and across communities, and dynamically. That is, we plan to integrate the features of both time and community in topic modelling. The proposed classifier in this paper is one important component in our larger work on information influence, being the starting place to identify which subsets of the population various users belong to.

References

Al Zamal, Faiyaz, Wendy Liu and Derek Ruths. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *Proceedings of ICWSM*.

Bagdouri, Mossaab and Douglas W Oard. 2015. Profession-Based Person Search in Microblogs:

Using Seed Sets to Find Journalists. In *Proceedings of CIKM*.

Barbera, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91.

Barbera, Pablo. 2016. "Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data." Working Paper.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak and Sebastian Hellmann. 2009. "DBpedia-A crystallization point for the Web of Data." *Web Semantics: science, services and agents on the world wide web* 7(3):154–165.

Chen, Xin, Yu Wang, Eugene Agichtein and Fusheng Wang. 2015. "A Comparative Study of Demographic Attribute Inference in Twitter." *International AAAI Conference on Web and Social Media* (9th):590–592.

Daiber, Joachim, Max Jakob, Chris Hokamp and Pablo N. Mendes. 2013. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of SEMANTiCS*.

Erikson, Robert S, Michael B MacKuen and James A Stimson. 2002. *The macro polity*. Cambridge University Press.

Fang, Anjie, Craig Macdonald, Iadh Ounis and Philip Habel. 2016*a*. Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data. In *Proceedings of ECIR*.

Fang, Anjie, Craig Macdonald, Iadh Ounis and Philip Habel. 2016*b*. Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data. In *Proceedings of SIGIR*.

Fang, Anjie, Iadh Ounis, Philip Habel, Craig Macdonald and Nut Limsopatham. 2015. Topic-centric Classification of Twitter User's Political Orientation. In *Proceedings of SIGIR*.

Go, Alec, Richa Bhayani and Lei Huang. 2009. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford* 1:12.

Habel, Philip D. 2012. "Following the opinion leaders? The dynamics of influence among media opinion, the public, and politicians." *Political Communication* 29(3):257–277.

Hillman, Amy J., Gerald D. Keim and Douglas Schuler. 2004. "Corporate Political Activity: A Review and Research Agenda." *Journal of Management* 30(6):837–857.

Jacobs, Lawrence R. and Robert Y. Shapiro. 2000. *Politicians Don't Pander: Political Manipulation and the Loss of Democratic Responsiveness*. Chicago: University ofChicago Press.

Lee, Kathy, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal and Alok Choudhary. 2011. Twitter trending topic classification. In *Proceedings of ICDMW*.

Mansfield, Edwin. 1995. "Academic research underlying industrial innovations: sources, characteristics, and financing." *The review of Economics and Statistics* pp. 55–65.

Mehrotra, Rishabh, Scott Sanner, Wray Buntine and Lexing Xie. 2013. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of SIGIR*.

Mendes, Pablo N, Max Jakob, Andrés García-Silva and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings of SEMANTiCS*.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J. Niels Rosenqueist. 2011. "Understanding the Demographics of Twitter Users." *ICWSM* .

Pennacchiotti, Marco and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of ICWSM*.

Rao, Delip, David Yarowsky, Abhishek Shreevats and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of SMUC*.

Read, Jonathon. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of SRW*.