# On Refining Twitter Lists as Ground Truth Data for Multi-Community User Classification

Ting Su[1], Anjie Fang[2], Richard McCreadie[3], Craig Macdonald[3], Iadh Ounis[3],

{[1]t.su.2, [2]a.fang.1}@research.gla.ac.uk, [3]{firstname.secondname}@glasgow.ac.uk
University of Glasgow, UK

**Abstract.** To help scholars and businesses understand and analyse Twitter users, it is useful to have classifiers that can identify the communities that a given user belongs to, e.g. business or politics. Obtaining high quality training data is an important step towards producing an effective multi-community classifier. An efficient approach for creating such ground truth data is to extract users from existing public Twitter lists, where those lists represent different communities, e.g. a list of journalists. However, ground truth datasets obtained using such lists can be noisy, since not all users that belong to a community are good training examples for that community. In this paper, we conduct a thorough failure analysis of a ground truth dataset generated using Twitter lists. We discuss how some categories of users collected from these Twitter public lists could negatively affect the classification performance and therefore should not be used for training. Through experiments with 3 classifiers and 5 communities, we show that removing ambiguous users based on their tweets and profile can indeed result in a 10% increase in F1 performance.

## 1 Introduction

Due to the popularity of social media platforms, such as Twitter, people with different backgrounds can express their views towards topics during events (e.g. elections). Indeed, such platforms have become a major channel to share ideas and opinions [1]. In previous studies, researchers have used text corpora (e.g. articles, books or speeches) to analyse how social groups influence one another (e.g. how journalists influence the public [2]). However, the emergence of social media as a popular communication medium and the relative ease of collecting large volumes of user and post data, provides new opportunities for researchers to better analyse how social groups/communities interact. On the other hand, users do not explicitly specify their social group/community affiliations. Hence, researchers need to resort to automatic approaches to infer this at scale. One popular means to classify Twitter users into different groups is to train learned classification models [3–5]. These approaches require a high quality training dataset to produce effective models, especially when it comes to difficult multi-class classification tasks, such as community classification. However, topical overlap between communities and ambiguous user affiliations makes training accurate and generalisable models challenging. Unsupervised clustering can also be used for community identification task [6], however the clusters obtained may not reflect predefined notions of communities, and hence supervised methods are of interest.

In this paper, we aim to produce a reliable dataset that can be used to effectively perform community classification of Twitter users into five classes:

Table 1: Lists used to extract users for each community.

| Community | Lists used | # users |
|---|---|---|
| ACA | Higher Ed Thought Leaders(@MSCollegeOpp), Edu-Scholars(@sesp_nu) Favourite academics(@AcademiaObscura), Northwestern(@sesp_nu), SESP Alumni(@sesp_nu), STEM Academic Tweeters(@LSEImpactBlog), The Academy(@AcademicsSay), Harvard(@hkslibrary) | 3592 |
| BSN | Social CEOs on Twitter(@debweinstein), Tech, Startups & Biz(@crblev) Tech Startup Founders(@realtimetouch), Top CEO's(@chrisgeorge187), Awesome Entrepreneurs(@vincentdignan) | 3013 |
| MDA | Mirror Political Journos(@MirrorPolitics), BBC News Official(@BBCNews) sundaymirror(@DailyMirror), Financial Tweets(@TIME), TIME Staff(@TIME), Sun accounts(@TheSun), Sun people(@TheSun), Mirror reporters/columist(@DailyMirror), BBC Asian Network(@BBC), BBC News(@BBC), Business staff(@guardian), Observer staff(@guardian), Money staff(@guardian),Technology staff (@guardian), Politics staff(@guardian), News staff (@guardian) | 1242 |
| PLT | UK MPs(@TwitterGov), US Governors(@TwitterGov), US Senate(@TwitterGov), US House(@TwitterGov), Senators(@CSPAN), New Members of Congress(@CSPAN) | 1899 |
| CTZ (celebrities) | celebrity(@mashable),the-celebrity-list(@buzzedition) celebrities(@GALUXSEE) | 774 |
| CTZ (normal users) | N/A | 800 |

- **Academic (ACA)**: users involved in research and/or teaching.
- **Business (BSN)**: company executives, managers and other white-collar workers.
- **Media (MDA)**: journalists or reporters working for news-media or as freelancers.
- **Politics (PLT)**: politically-active users, e.g. members of parliament or activists.
- **Citizen (CTZ)**: users who do not belongs to any of the 4 other classes.

Our goal is to cover these five particular roles in political events, based on users' jobs and/or social roles. However, in reality, users may have multiple roles, temporally/permanently switch roles, or act as if they have different roles. This can make it challenging to conduct accurate training and generalise models to categorise users based on their profile and past tweets. This also makes existing methods, which are often based on crowdsourced data (e.g. [7]) or automatic user behaviour analysis based on predefined rules (e.g. hashtag usage [8], words in profiles [9], location and name [10]) unsuitable, because human-labelled data is expensive, and user classification based on their behaviours can be vague when classes overlap with each other.

Hence, as a first step towards producing high quality training data for multi-community classification, we examine the effectiveness of a list-based approach [11], as well as investigate where it tends to fail. We first collect Twitter lists representing our target communities, and then crawl the posts and profiles for each user in those lists. We analyse these lists, with the aim of removing users that might make poor quality training examples, producing several (more refined) datasets. We then train several supervised community classification models based on the original and 5 refined datasets, and compare their performance when tested on a separate gold-standard human-labelled dataset. In this way, we establish the raw performance of models produced by the list-based approach, as well as show how removing potentially problematic users leads to better classification models that can increase performance by up to 10%. Finally, we discuss the main issues observed when relying on Twitter lists for use as training data.

Thus, the contributions of this paper are as follows: 1) We conduct a failure analysis of a training dataset for multi-community classification that is automatically generated from Twitter public lists. 2) We discuss four categories of problematic users collected from these Twitter public lists, and empirically show that they can negatively affect classification performance when used for training.

## 2 Analysis of Community Lists and Users

To evaluate the effectiveness of list-based approaches when used for training community classification models, we construct an evaluation dataset and analyse categories of users that may cause issues when training.

**Dataset Collection**: To create an initial dataset, we extract users from existing public lists on Twitter. For each community, the lists we use, and the number of users obtained from lists per class are shown in Table 1. These lists were selected based on their descriptions. For example, `Edu-Scholars` is described as "A selection of the nation's most influential academics in education" by its creator. Hence, we consider its users within our `ACA` category. The outlier is the 'citizen' (`CTZ`) class, for which we first extract celebrity users from lists as examples of citizens. Normal citizens are not likely to be collected in any public lists, hence we also randomly sample users with $< 200$ followers along with the celebrities.

**User Analysis**: Having collected users belonging to each community, we randomly select 1000 users per community to form an initial and balanced dataset for training classifiers, and manually analyse those users based on their Twitter profiles and previous posts. Based on this analysis, we identify four categories of users that might be problematic if used as examples to train a community classifier. The prevalence of each category within the communities is shown in Table 2.

1. **Users with ambiguous descriptions. (Category 1).** We observe that there are a subset of users that would clearly label themselves as a member of a community, but in practice mainly tweeted/shared content on other unrelated topics. These off-community tweets can confuse text-based classifiers as they include words that may be associated to other communities.
2. **Users retweeting/sharing links without adding comments. (Category 2).** We observe that there is an active group of users who only retweet the community-related popular topics (e.g. others tweets, links, links with title, etc), but without their own opinions. Since links and words used in article titles tend to exist among all communities, including users who only tweet about such topics when training may add noise to the resultant classification model.
3. **Users (re)tweet useless content. (Category 3).** We observe that some users only make tweets containing 'useless' content, such as motivational pictures or quotes, tweets generated by other platforms, and advertisements links. Such tweets can contain highly duplicated content and off-community words, which can potentially reinforce classifiers with false features, and may lead to weak classifier models.
4. **Non-active users. (Category 4).** Public lists can be quite old and unmaintained, and hence can include users that have been inactive for years. Using users that have been inactive for an extended period of time may be problematic for training purposes, as the types of discussion topics that help distinguish a community change over time. Hence, training on old users/tweets may hinder the development of accurate classification methods.

## 3 Investigating the Impact of User Filtering

Having produced a tweet dataset for community classification and identified some potential issues that might arise when using it for training, we now examine if the issues we have identified do indeed impact upon classification performance.

Table 2: Number of users in each categories.

| | ACA | BSN | CTZ | MDA | PLT |
|---|---|---|---|---|---|
| AllUsers | 1000 | 1000 | 1000 | 1000 | 1000 |
| Category 1 | 323 | 411 | 27 | 298 | 2 |
| Category 2 | 47 | 111 | 54 | 194 | 1 |
| Category 3 | 27 | 80 | 49 | 30 | 0 |
| Category 4 | 5 | 19 | 108 | 10 | 4 |
| Non-English | 38 | 33 | 60 | 5 | 0 |

Table 3: Number of users in training & test datasets.

| | ACA | BSN | CTZ | MDA | PLT |
|---|---|---|---|---|---|
| AllUsers | 1000 | 1000 | 1000 | 1000 | 1000 |
| AllUsers - Category 1 | 590 | 589 | 590 | 590 | 590 |
| AllUsers - Category 2 | 800 | 800 | 800 | 800 | 800 |
| AllUsers - Category 3 | 920 | 920 | 920 | 920 | 920 |
| AllUsers - Category 4 | 900 | 900 | 892 | 900 | 900 |
| All Filtered | 350 | 346 | 350 | 350 | 350 |
| Crowdsourced Test set | 80 | 163 | 337 | 159 | 57 |

### 3.1 Experimental Setup

**Methodology** To evaluate community classification, we train classification models based on the dataset discussed above. As discussed above, we randomly sample 1000 users from collected user lists of each of the five communities to form a balanced training dataset (denoted *AllUsers*). However, to determine what effect the four categories of potentially undesirable users have on classification performance, we produce alternative datasets (denoted *AllUsers - Category X*) that do not contain users from one of the identified categories, and adjust the number in each community to form balanced datasets. Finally, we create another dataset (denoted *AllFiltered*), by removing users of all the identified categories. Details about each dataset are provided in Table 3. We train classification models based on all 6 datasets using three types of learner, namely: Support Vector Machine (SVM), Multinomial Naive Bayes (NB) and Multilayer Perceptron (MLP).
**Gold Standard** Having defined the training datasets, we next need a gold standard that we can evaluate against. To create this, we randomly sample another 1000 Twitter users (who do not appear in any of the training datasets) and use crowdsourced workers to manually label each user's community affiliation by examining that user's profile and his/her 8 most recent tweets. Three workers labelled each user and a majority vote is used to produce the final label for a user. If a majority could not be reached for a user, then more workers labelled that user until a majority was obtained (7.3% of the users required such additional labels to reach a majority). [1] Details about the test set are provided in Table 3.
**Classifier Configuration** For the purposes of building the user classification models, we use the 20k most frequently occurring terms across all user's tweets and profile descriptions as features, after applying stopword removal and stemming. Each term is represented by its `TF-IDF` score. The configuration settings for the three learned models are: Multinomial NB $\alpha$=0.01; for SVM we use a Linear kernel, L2 penalty, $C$=1.0, $\gamma$=0.001, and *multi_class*=one-vs-all[2]; for MLP we use one hidden layer with 500 neurons. All of the above parameters are obtained using a 10-fold cross-validation on the training dataset.
**Baseline** To provide a basis for comparison, we also report the performance of a Random Classifier using uniform distribution (denoted as RDN) as a baseline.
**Metric** We report F1 for classes, and Micro F1 across all classifiers and datasets.

[1] ∼20% of accounts have been removed from Twitter, and are excluded from our test dataset. [2] One-vs-all is the recommended setup for multi-class classification using SVM [12].

Table 4: The F1 scores with different training data.

| Classifier | Training Dataset | ACA | BSN | CTZ | MDA | PLT | Micro |
|---|---|---|---|---|---|---|---|
| RDN | AllUsers | 0.10 | 0.15 | 0.24 | 0.15 | 0.08 | 0.18 |
| NB | AllUsers | 0.45 | 0.47 | 0.59 | 0.35 | 0.46 | 0.49 |
| | AllUsers - Category 1 | **0.47▲** | 0.47 | **0.61** | **0.41▲** | 0.37 | **0.51** |
| | AllUsers - Category 2 | 0.43 | 0.44 | 0.59 | **0.40** | 0.38 | 0.49 |
| | AllUsers - Category 3 | **0.46▲** | 0.45 | 0.59 | **0.42** | 0.36 | **0.50** |
| | AllUsers - Category 4 | 0.44 | 0.47 | 0.59 | **0.37** | **0.47** | 0.49 |
| | AllFiltered | **0.49** | 0.43 | **0.62** | **0.39** | 0.34 | **0.50** |
| SVM | AllUsers | 0.45 | 0.42 | 0.61 | 0.34 | 0.45 | 0.49 |
| | AllUsers - Category 1 | **0.48** | **0.53** | **0.64** | **0.42▲** | 0.34 | **0.54▲** |
| | AllUsers - Category 2 | 0.42 | **0.52** | **0.63** | **0.35** | **0.46▲** | **0.52** |
| | AllUsers - Category 3 | 0.45 | **0.49▲** | **0.63** | **0.41** | **0.46▲** | **0.53** |
| | AllUsers - Category 4 | 0.45 | 0.41 | 0.60 | **0.36** | **0.47** | 0.49 |
| | AllFiltered | 0.40 | 0.39 | **0.62** | **0.38** | 0.37 | 0.49 |
| MLP | AllUsers | 0.44 | 0.43 | 0.60 | 0.33 | 0.43 | 0.48 |
| | AllUsers - Category 1 | **0.46▲** | **0.49▲** | **0.63** | **0.41▲** | 0.28 | **0.51** |
| | AllUsers - Category 2 | 0.44 | **0.47** | 0.59 | **0.34** | **0.48▲** | **0.50** |
| | AllUsers - Category 3 | **0.45** | **0.45▲** | **0.62▲** | **0.36** | **0.45** | **0.50** |
| | AllUsers - Category 4 | 0.44 | 0.43 | **0.61** | **0.34** | **0.44** | **0.49** |
| | AllFiltered | 0.44 | 0.41 | **0.62** | 0.31 | **0.34▲** | 0.48 |

### 3.2 Results

In this section, we report the outcome of our comparison between models trained on the AllUsers dataset and all other datasets. Table 4 reports classification performances for three learned models across 6 datasets. Scores highlighted in bold indicate increased performance over AllUsers. "▲" denotes statistically significant increases in performance (McNemar's test, $p < 0.05$) over AllUsers.

First, in Table 4, we observe that all the classifiers across all tested datasets achieve Micro F1 scores higher than 0.48, which is markedly higher than the RDN classifier (0.18). This indicates that the models produced are able to distinguish between the user classes. Next, comparing the classification models produced on each dataset, we observe that, for all three classifiers tested using all 5 filtered datasets, Micro F1 performance is greater than or equal to (by up to 10%) than that of AllUsers. Hence, the user categories identified in Section 2 do have negative impacts on community classification when used as training examples,

Among the four categories proposed in Section 2, we see that for all three classifiers, using dataset AllUsers - Category 1 as training set provides the highest Micro F1 scores, and obtains a significantly benefited SVM classifier compared to using AllUsers (McNemar's test, $p < 0.05$). As described above, AllUsers - Category 1 is the list that excludes ambiguous users, who mostly tweet about other communities. Indeed, by excluding users that appear to overlap with other communities, the classifiers perform better, as the difference between classes is clearer. Surprisingly, using the most sanitised dataset, namely AllFiltered, does not improve the result significantly. One reason can be that, as the size of dataset AllFiltered is only a third of the AllUsers dataset, the variety of text for the classifier to learn from is reduced, resulting in lower performance.

For the most difficult community observed, namely `MDA`, excluding ambiguous users (i.e. Category 1) results in an up-to 24.2% increase in the F1 score across almost all datasets and models. However, excluding the other 3 categories does not demonstrate a consistent benefit to F1 across the classifiers. Indeed, it is clear that ambiguous users are the most harmful for classifying `MDA` users.

## 4 Conclusions

In this paper, we examined how to construct a robust community classification dataset for Twitter and investigated challenges associated with selecting users as training examples. In particular, we first collected Twitter lists representing target communities and collected associated posts and profiles from each user. We analysed these lists, with the aim of performing a failure analysis, thereby identifying four categories of user that might be problematic and make poor training examples for classification. Therefore, we produced various datasets, by excluding users from each of the identified categories. We then trained several supervised community classification models based on the original and filtered datasets, and compared their performance when tested on a separate human-labelled gold-standard. We showed that public Twitter lists can be used as training data when analysing Twitter users, as all classifiers using AllUsers dataset achieved at least 0.48 Micro F1. On the other hand, the 4 categories of users we identified can be problematic, as Micro F1 scores increased by up to 10% when excluding each category in turn from the training dataset. Removing Category 1 (ambiguous users) in particular cause the largest increase in performance. Future studies are needed to develop automatic methods to identify and exclude such users for more effective community classification training datasets.

## References


1. Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T.: Understanding the participatory news consumer. Pew Internet and American Life Project **1** (2010) 19–21
2. Erikson, R., MacKuen, M., Stimson, J.: The macro polity. Cambridge University Press (2002)
3. Culotta, A., Kumar, N., Cutler, J.: Predicting the demographics of Twitter users from website traffic data. In: Proc. of AAAI. (2015)
4. Pennacchiotti, M., Popescu, A.: A machine learning approach to Twitter user classification. In: Proc. of ICWSM. (2011)
5. De Choudhury, M., Diakopoulos, N., Naaman, M.: Unfolding the event landscape on Twitter: classification and exploration of user categories. In: Proc. of CSCW. (2012)
6. Sachan, M., Dubey, A., Srivastava, S., Xing, E.P., Hovy, E.: Spatial compactness meets topical consistency: jointly modeling links and content for community detection. In: Proc. of the ICWSDM. (2014)
7. Chen, X., Wang, Y., Agichtein, E., Wang, F.: A comparative study of demographic attribute inference in Twitter. In: Proc. of ICWSM. (2015)
8. Fang, A., Ounis, I., Habel, P., Macdonald, C., Limsopatham, N.: Topic-centric classification of Twitter user's political orientation. In: Proc. of SIGIR. (2015)
9. Feng, V., Hirst, G.: Detecting deceptive opinions with profile compatibility. In: Proc. of IJCNLP. (2013)
10. Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., Yarowsky, D.: Broadly improving user classification via communication-based name and location clustering on Twitter. In: Proc. of HLT-NAACL. (2013)
11. Bagdouri, M., Oard, D.: Profession-based person search in microblogs: Using seed sets to find journalists. In: Proc. of CIKM. (2015)
12. Rifkin, R., Klautau, A.: In defense of one-vs-all classification. Journal of machine learning research **5** (2004)