A Machine Learning Approach to Twitter User Classification

Marco Pennacchiotti and **Ana-Maria Popescu**
Yahoo! Labs
Sunnyvale, USA
{pennac,amp}@yahoo-inc.com

Abstract

This paper addresses the task of user classification in social media, with an application to Twitter. We automatically infer the values of user attributes such as *political orientation* or *ethnicity* by leveraging observable information such as the user behavior, network structure and the linguistic content of the user's Twitter feed. We employ a machine learning approach which relies on a comprehensive set of features derived from such user information. We report encouraging experimental results on 3 tasks with different characteristics: political affiliation detection, ethnicity identification and detecting affinity for a particular business. Finally, our analysis shows that *rich linguistic features* prove consistently valuable across the 3 tasks and show great promise for additional user classification needs.

1 Introduction

Successful microblogging services such as Twitter have become an integral part of the daily life of millions of users. In addition to communicating with friends, family or acquaintances, microblogging services are used as recommendation services, real-time news sources and content sharing venues.

A user's experience with a microblogging service could be significantly improved if information about the demographic attributes or personal interests of the particular user, as well as the other users of the service, was available. Such information could allow for personalized recommendations of users to follow or user posts to read; additionally, events and topics of interest to particular communities could be highlighted.

Profile information including name, age, location and short summary of interests is available in most social network and micro-blog services, although it can be incomplete (a user may choose not to post bio details) or misleading (a user may choose to list an imaginary place - aka, "Wonderland" - as her location). Furthermore, other relevant attributes, such as explicit and implicit interests or political preferences are usually omitted.

In this work we address the task of *user classification*: we attempt to automatically infer the values of user attributes (e.g. political orientation, ethnicity) by leveraging observable information such as the user behavior, network structure and the linguistic content of the user's Twitter feed.

Our main contributions are the following:

- We describe a general machine learning framework for social media user classification which relies on four general feature classes: *user profile*, *user tweeting behavior*, *linguistic content of user messages* and *user social network* features.
- We show that the framework can be instantiated and used with good results for a popular microblogging service (Twitter) and three different tasks (political orientation, ethnicity and business fan detection).
- We provide an in-depth analysis of the relative value of feature classes both within specific tasks and across all tasks: we show experimentally that *content features* are in general highly valuable, and that *large-scale topic models* are consistently and specifically reliable and show promise for additional user classification tasks.

The paper is organized as follows. In Section 2 we introduce relevant previous work on user profiling for social media, Twitter user attribute detection and topic models for Twitter. In Section 3 we describe in detail our model and features for Twitter user classification, while in Section 4 we report an extensive experimental evaluation including a quantitative and qualitative discussion. Finally, in Section 5 we draw final conclusions and outline future work.

2 Related work

Detecting user attributes based on user communication streams. Previous work has explored the impact of people's profiles on the style, patterns and content of their communication streams. Researchers investigated the detection of *gender* from well-written, traditional text (Herring and Paolillo 2010; Singh 2001), blogs (Burger and Henderson 2010) reviews (Otterbacher 2010), e-mail (Garera and Yarovsky 2007), user search queries (Jones et al. 2007; Weber and Castillo 2010) and, for the first time, Twitter (Rao et al. 2010). Other previously explored attributes include the user's *location* (Jones et al. 2007; Fink et al. 2009; Cheng, Caverlee, and Lee 2010), *location of origin* (Rao et al. 2010), *age* (Jones et al. 2007; Rao et al. 2010), *political*

orientation (Thomas, Pang, and Lee 2006; Rao et al. 2010). While such previous work has addressed blogs and other informal texts, microblogs are just starting to be explored for user classification. Additionally, previous work uses a mixture of sociolinguistic features and n-gram models while we focus on richer features (e.g., features derived from large-scale-topic models) in order to better exploit the user-created content.

Twitter user attribute detection. (Rao et al. 2010) is the work most relevant to ours: authors present an exploratory study of Twitter user attribute detection which uses *simple* features such as n-gram models, simple sociolinguistic features (e.g., presence of emoticons), statistics about the user's immediate network (e.g., number of followers/friends) and communication behavior (e.g., retweet frequency). In comparison, our work confirms the value of *in-depth* features which reflect a deeper understanding of the Twitter user stream and the user network structure (e.g., features derived from large-scale topic models, tweet sentiment analysis and *explicit* follower-followed links).

Topic models for Twitter. (Ramage 2010) uses large-scale topic models to represent Twitter feeds and users, showing improved performance on tasks such as post and user recommendation. We confirm the value of large-scale topic models for a different set of tasks (user classification) and analyze their impact as part of a rich feature set.

3 A general model for user profiling

In this section we describe in detail four types of information which can help characterize a micro-blog user: *profile*, *messaging (tweeting) behavior*, *linguistic content of messages* and *social network information*. We use these four information types to derive a rich set of features for use in a general-purpose user classification model. Our goal is two-fold: first, to provide a general assessment of the relative value, robustness and generalization potential of features for user classification purposes and second, to explore the value of *linguistic information* for classifying users.

ML framework for user classification. The set of features we explore below is used in conjunction with a supervised machine learning framework providing models for specific user classification tasks. As a learning algorithm, we use Gradient Boosted Decision Trees - GBDT (Friedman 2001) (any other algorithm could be adopted), which consists of an ensemble of decision trees, fitted in a forward step-wise manner to current residuals. Friedman (2001) shows that GDBT competes with state-of-the-art machine learning algorithms such as SVM (Friedman 2006) with much smaller resulting models and faster decoding time.

In the following, we describe our feature classes in more detail.

3.1 Profile features: "Who you are"

Most services (such as Twitter) publicly show by default profile information such as the user name, the location and a short bio. The Twitter API (2010) also provides access to other basic user information, such as the number of a user's friends, followers and tweets. In related work, Cheng and colleagues (2010) estimated that only 26% of users report a specific *location* such as a city, while the rest provide either general locations (states, countries) or imaginary places. We conducted a pilot study in the same vein to assess the direct use of such public profile information for basic user classification tasks, such as identifying a user's gender and ethnicity. Given a corpus of 14M users active in April 2010, we found that 48% of them provide a short bio and 80% a location. We then matched more than 30 regular expression patterns over the bio field to check if they are effective in extracting classification information. The following are 2 examples of such patterns for age and, respectively, ethnicity classification:

```
(I|i)(m|am|'m)[0-9]+(yo|year old)
white(man|woman|boy|girl)
```

We were able to determine the *ethnicity* of less than 0.1% users and to find the *gender* of 80%, but with very low accuracy. We then investigated the use of the profile avatar in determining the gender and ethnicity attribute values. We sampled 15,000 random users and asked a pool of editors to identify the ethnicity and gender of the user based on only the avatar picture: less than 50% of the pictures were correlated with a clear ethnicity while 57% were correlated with a specific gender. We found that pictures can often be misleading: in 20% of the cases, the editors verified that the picture was not of the account owner, but of a celebrity or of another person.

The above statistics show that the profile fields do not contain enough good-quality information to be directly used for user classification purposes, though they can be effectively used for bootstrapping training data. Yet, we implemented basic profile-based features (referred as PROF in the experiments): the length of the user name, number of numeric and alphanumeric characters in the user name, different capitalization forms in the user name, use of the avatar picture, number of followers, number of friends, friends/followers ratio, date of account creation, matching of various regular expression patters in the bio field as listed above, presence of the location field.

3.2 Tweeting behavior: "How you tweet"

Tweeting behavior is characterized by a set of statistics capturing the way the user interacts with the micro-blogging service: the average number of messages per day, number of replies, etc. Intuitively, such information is useful for constructing a model of the user; Java and colleagues (2007) suggest that users who rarely post tweets but have many followers tend to be information seekers, while users who often post URLs in their tweets are most likely information providers. Rao and colleagues (2010) instead suggest that tweeting behavior information is not useful for most classification tasks and that it is subsumed by linguistic features. In this paper we aim at verifying these claims, by experimenting with more than 20 tweeting behavior features (BEHAV), including: number of tweets posted by the user, number and fraction of tweets that are retweets, number and fraction of tweets that are replies, average number of hashtags

and URLs per tweet, fraction of tweets that are truncated, average time and std.dev. between tweets, average number and std.dev.of tweets per day, fraction of tweets posted in each of 24 hours.

3.3 Linguistic content: "What you tweet"

Linguistic content information encapsulates the main topics of interest to the user as well as the user's lexical usage. Simple linguistic information is helpful for classifying users in several media, such as formal texts, blogs, spoken conversational transcripts or search sessions.

We explore a wide variety of linguistic content features, as detailed below. [1]

Prototypical words (LING-WORD**).** In a classification task, classes can be described by prototypical words (hereafter 'proto words'), i.e. typical lexical expressions for people in a specific class as well as phrases denoting typical interests of people in that class. For example, younger people tend to use words such as 'dude' or 'lmao'; democrats tend to use the expression 'health care' more than republicans. Rao and colleagues (2010) explored this intuition by manually building a list of words which are likely to characterize sociolinguistic behaviors, e.g. emoticons and ellipses: however, their list is meant to be generic and it is not easy to translate into strong class-indicative features without manual effort. Instead, we employ a probabilistic model for automatically extracting proto words: it only needs a few seed users and it is easily portable to different tasks, similarly to what was proposed in (Pasca 2007).

Given n classes, each class c_i is represented by a set of seed users S_i. Each word w issued by at least one of the seed users is assigned a score for each of the classes. The score estimates the conditional probability of the class given the word as follows:

$$proto(w, c_i) = \frac{|w, S_i|}{\sum_{j=1}^{n} |w, S_j|} \quad (1)$$

where $|w, S_i|$ is the number of times the word w is issued by all users for class c_i. For each class, we retain as proto words the highest scoring k words [2].

The $n * k$ proto words collected across all classes serve as features for representing a given user: for each proto word wp the user u is assigned the score:

$$f\_proto\_wp(u) = \frac{|u, wp|}{\sum_{w \in W_u} |u, w|} \quad (2)$$

where $|u, wp|$ is the number of times the proto word w is issued by user u, and W_u is the set of all words issued by u. For each class, the user is also assigned an aggregated feature:

$$f\_proto\_c(u) = \frac{\sum_{wp \in WP} |u, wp|}{\sum_{w \in W_u} |u, w|} \quad (3)$$

where WP is the set of proto words for class c. Table 3 reports the highest scoring proto words for the user classes targeted in our paper.

[1] Note that as far as language models are concerned, we prefer the use of Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2002) and automatically bootstrapped prototyical words over a more simple bag-of-word model (various studies, e.g. (Rao et al. 2010), have showed that bag-of-words models are usually outperformed by more advanced linguistic ones).

[2] In our experiment we use $k = 200$, and discard all words occurring 5 or less times, and long less than 3 characters.

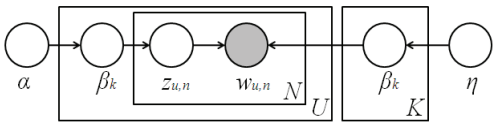

Figure 1: Plate representation of the user-level LDA model.

Prototypical hashtags (LING-HASH**).** Twitter users may use hashtags (sequences of characters prefixed by '#') to denote the topic(s) of their tweet; many times, the same or similar hashtags are used by Twitter users in order to facilitate the retrieval and surfacing of information on a particular topic. We hypothesize that if users from a class are interested in the same topics, the most popular such topics can be found by collecting statistics on used hashtags. The intuition is implemented similarly to LING-WORD. Given a seed user set S_i for a class c_i, we collect all the hashtags h contained in the tweets of each seed user. We then derive the set of prototypical hashtags, by applying Eq. 1 (where w is replaced by h). Finally, we retain the highest scoring 100 hashtags for each class, and compute feature values as in Eq. 2 and 3.

Generic LDA (LING-GLDA**).** Our generic LDA model is an adaptation of the original LDA proposed by Blei and colleagues (2002) where documents are replaced by users. Our hypothesis is that a user can be represented as a multinomial distribution over topics. This representation may help with classification: e.g., democrats may have, on average, a higher probability of talking about social reforms, while republicans may mention oil drilling more often. While Blei represents documents by their corresponding bag of words, we represent users by the words of their tweets.

Our generative model works as follows (see Figure 1). Given a number U of users and a number K of topics, each user u is represented by a multinomial distribution θ_u over topics, which is drawn from a Dirichlet prior with parameter α. Also a topic is represented by a multinomial distribution β_k drawn from another Dirichlet prior with parameter η. The generative model states that each word position n in a user vocabulary is assigned a topic $z_{u,n}$ drawn from θ_u, and that the word in that position $w_{u,n}$ is drawn from the distribution $\beta_{z_{u,n}}$.

The model is obtained by training a LDA parallel implementation (Smola and Narayanamurthy 2010) with 500 iterations over a set of 4M users, each represented by a maximum of 20,000 words collected from their tweets. As a result, we obtain 100 topics which will each be used to derive features for classification. The model is then applied to each

test user in order to obtain his topic distribution, i.e. the feature values for the classification task.

Domain-specific LDA (LING-DLDA**).** This LDA model differs from LING-GLDA in that it is not derived from a generic set of users, but from users drawn from the training set (e.g., the training set of Democrat and Republican users is used to build the model for the political affiliation task). The intuition is that while LING-GLDA returns coarse-grained topics such as soccer, music and politics, LING-DLDA should return fine-grained topics that are more discriminative for the classification task. The model is derived as for LING-GLDA, though the smaller training set allows us to run 1000 iterations. We again use 100 topics.

Sentiment words (LING-SENT**).** In some cases, it is possible to identify terms or entities about which a particular user class has an overall majority opinion which is not shared by a different class (e.g., "Ronald Reagan" is generally viewed positively by republicans and negatively by democrats). We manually collect a small set of such terms for our classes and implement sentiment analysis techniques to find the sentiment of a user with respect to the term.

Given user u, her set of tweets and each term t, we first identify the number of tweets in which a *positive*, *negative* or *neutral* sentiment is expressed with respect to t by relying on Opinion Finder 1.5 (Wiebe, Wilson, and Cardie 2005) term lexicon for positive, negative and neutral sentiment words. For each tweet and term t, we compute the dominant sentiment in the tweet with respect to t by inspecting the phrases in a window of $k = 4$ words to the left and right of t. If more than 50% of the words are neutral, or not in the Opinion-Finder lexicon, the tweet is classified as neutral with respect to t. Otherwise, we classify the tweet as *positive* if a majority of the terms are *positive* and *negative* otherwise. Given the set of tweets of user u annotated with u's sentiment towards t, we retain as features the percentage of positive tweets with respect to t , the percentage of negative tweets with respect to t and the percentage of neutral tweets with respect to t.

We also derive aggregated features indicating the overall sentiment of the user u with respect to the target class, such as : the median and standard deviation of the above features across the entire term set; the number of terms t about which the user has overall, a mainly *positive*, *negative*, or *no opinion*.

3.4 Social network: "Who you tweet"

These features explore the social connections established by the user with others he follows, to whom he replies or whose messages he retweets.

"Friend" accounts (SOC-FRIE**).** Intuitively, Democrats are more likely to follow the accounts of Democratic politicians and Republicans those of Republican politicians. We hypothesize that users from other classes may also share specific "friend" accounts. We use the same basic mechanism employed to bootstrap proto words (Eq. 1) in order to bootstrap a set of class-specific prototypical "friend" accounts F, by exploring the social network of users in the training set. We then derive the following *aggregate* and *individual* social network-based features for a given user u: number of accounts in F which are *friends* of u (accounts which the user is following); percentage of F accounts which are *friends of* u; percentage of all Twitter accounts followed by u which are part of F.

For each prototypical "friend" account, a boolean feature is set to 1 if the user follows the account and 0 otherwise.

Prototypical replied (SOC-REP**) and retweeted (**SOC-RET**) users.** Similarly to SOC-FRIE, these two feature sets capture the idea that users from a particular class tend to reply to and retweet messages of specific accounts (e.g., young girls may tend to reply to Justin Bieber's account). These features are derived exactly as LING-WORD and LING-HASH, i.e. by first collecting accounts cited in tweets of users of a specific class, and prefixed by the reply and retweet tags ('@' and 'RT'); then discovering the 200 most significant replied/retweetd account applying Eq. 1; and, finally, deriving feature values as in Eq. 2, 3.

4 Experimental evaluation

We evaluate our classification system over three binary classification tasks: detecting political affiliation, detecting a particular ethnicity, and finally, identifying 'Starbucks fans'. Intuitively, these are very different use cases which allow for evaluating our feature families in different settings.

4.1 Experimental setup

Political affiliation. The task consists in classifying users as being either Democrats (positive set) or Republicans (negative set). Political affiliation detection is a very interesting task from many perspectives – e.g., from the perspective of tracking the concerns and interests of a party's base. We build the gold standard dataset by scraping lists of users that classified themselves as either Democrat or Republican in two major Twitter directories, namely WeFollow and Twellow[3]. We collect a total of 10,338 users, equally distributed in the two classes. [4]

Ethnicity. Our specific ethnicity identification task consists in classifying users as either African-Americans or not. This choice is motivated by Quantcast statistics indicating that African-Americans are the most represented ethnicity among Twitter users with respect to the average internet population (Quantcast 2010). The statistics mean that automatically identifying users of this ethnicity can have benefits from multiple perspectives: linguistic, sociological, as well as from the business perspective. We build the gold standard dataset by collecting users who explicitly mention their ethnicity in their profile, as described in Section 3.1. We then randomly sample 3000 African-American users (positive set) and 3000 users of other ethnicities (negative set).We performed a sanity check on the dataset and verified that the dataset is indeed a reliable gold standard.

[3] wefollow.com and www.twellow.com

[4] In this paper, the datasets are artificially balanced 50/50 in order to easily study feature behaviors. In future work we will experiment over realistic unbalanced data, by applying undersampling and skew insensitive measures. However, the real distribution for politican affiliation is close to that of our sample, as shown in recent Twitter demographic studies (Burson-Marsteller. 2010)

Starbucks fans. In addition to the more traditional user attribute identification tasks, we also consider the task of predicting whether a given user would likely follow a particular business. This task is particularly attractive from a business perspective, as it allows us to identify potential customers. For the purpose of this paper, we choose Starbucks, a business which attracts a large Twitter audience. The gold standard dataset is composed of 5,000 positive examples, represented by a random sample of users that already follow Starbucks on Twitter, and 5000 negative examples represented by a random sample of users who do not.

Evaluation metrics. For all tasks we report Precision, Recall and F-measure. In the case of the political affiliation task, we also report the overall accuracy, since both positive and negative examples are classes of interest. We experiment in a 10-folds cross validation setting, to compute statistical significance.

Comparisons and baselines. Our main system uses all features and is named FULL. We employ two baselines, B1 and B2, described below. B2 is a generic reference system represented by our machine learning system trained only on the profile and tweeting behavior features (basic information types readily available from Twitter). B1 denotes specific task-dependent baselines, as follows:
Political affiliation: B1 is a system which classifies as Democrats/Republicans all the users explicitly mentioning their political affiliations in the bio field (see Section 3.1). All other users are considered misses for the given class.
Ethnicity: B1 is an ideal system classifying users as African-Americans according to their profile picture. We simulate such a system by using the editorial annotations described in Section 3.1
Starbucks fans: B1 classifies as Starbucks fans all the users who explicitly mention Starbucks in their bio field.

System and features setup. For all models, GBDT parameters were experimentally set as follows: number of trees=500, shrinkage= 0.01, max nodes per tree=10, sample rate=0.5. In the political affiliation task we use the full set of features. In the Starbucks and ethnicity tasks, we do not use SOC-FRIE, since these features would be intuitively difficult to apply. The set of controversial terms for LING-SENT is composed of 40 famous politicians (for the politican affiliation task) and 30 popular African Americans (for the ethnicity task), semi-automatically harvested from Wikipedia. As for LING-WORD, SOC-REPL, SOC-RETW, SOC-FRIE, the list of seed users is derived from the training set of each fold. All features and models used in the experiments are computed on a Twitter firehose corpus spanning the July 2009 - February 2010 time period. All gold standard datasets described above contain users who were active in the considered time period by posting at least 5 tweets, and who posted at least 50% of their tweets in English (this being verified via dictionary lookup).

4.2 Experimental results

This section describes our experimental results in detail: Table 1 summarizes our overall results, Tables 2, 5 and 6 analyze in-depth the performance of feature sets on each task.

System	PREC	REC	F-MEAS
democrats-B1	**0.989**	0.183	0.308
democrats-B2	0.735	0.896	0.808
democrats-FULL	0.894‡	**0.936**♭	**0.915**♭
republicans-B1	**0.920**	0.114	0.203
republicans-B2	0.702	0.430	0.533
republicans-FULL	0.878‡	**0.805**♭	**0.840**♭
ethnicity-B1	**0.878**	0.421	0.569
ethnicity-B2	0.579	0.633	0.604
ethnicity-FULL	0.646‡	**0.665**♭	**0.655**♭
starbucks-B1	**0.817**	0.019	0.038
starbucks-B2	0.747	0.723	0.735
starbucks-FULL	0.762	**0.756**♭	**0.759**♭

Table 1: Overall classification results. †, ‡ and ♭ respectively indicate statistical significance at the 0.95 level with respect to B1 alone, B2 alone, and both B1 and B2.

The set of semi-automatically fabricated features used is available in Table 3.

Overall results reported in Table 1 show that our system generally achieves good precision and recall. However, results vary across tasks: identifying political affiliation labels can be done with very high accuracy. Classifying a user as a Starbucks fan can also be achieved with good performance, while identifying users of African-American ethnicity proves to be the most challenging task.

Political Affiliation. Our models perform best on the task of classifying a user as Democrat vs. Republican - both overall accuracy and class-specific performance measures have values above 0.80 (see Table 2). As expected, the baseline B1 has high precision but very low recall which makes the method less useful. All our system configurations largely outperform B1 in F-measure and accuracy. Also, the FULL system, integrating all available features, outperforms B2 in F-measure by 11% for Democrats and 31% for Republicans. Since B2 is based only on profile and behavior features, this result shows the value of constructing sophisticated social and linguistic features for the target classification tasks.

Table 2 shows that social features overall (SOC-ALL) and follower features (SOC-FRIE) in particular perform best, followed by the linguistic and profile features. Results also show that combining the high quality social features with linguistic, behavior and profile information (FULL model) improves the accuracy of SOC-ALL alone by 2.6% , suggesting that these latter features do add value to the classification model. This conclusion is strengthened by the feature importance ranking returned by the GBDT algorithm: while the 3 most discriminative features are from the SOC-FRIE set, we find 9 linguistic and 5 behavioral and profile features among the top 20.

The high performance of social features is due to the typical characteristics of users interested in politics: such users tend to interact with media or party personalities with an established Twitter presence (see Table 3 for examples of

System	Democrats			Republicans			All
	PREC	REC	F-MEAS	PREC	REC	F-MEAS	ACC
B1	$0.989_{\pm 0.006}$	$0.183_{\pm 0.016}$	$0.308_{\pm 0.023}$	$0.920_{\pm 0.011}$	$0.114_{\pm 0.002}$	$0.203_{\pm 0.011}$	$0.478_{\pm 0.013}$
BEHAV-ALL	$0.663_{\pm 0.011}$	$0.774_{\pm 0.011}$	$0.714_{\pm 0.009}$	$0.436_{\pm 0.011}$	$0.307_{\pm 0.011}$	$0.360_{\pm 0.009}$	$0.605_{\pm 0.009}$
PROF-ALL	$0.728_{\pm 0.009}$	$0.808_{\pm 0.016}$	$0.765_{\pm 0.006}$	$0.582_{\pm 0.024}$	$0.468_{\pm 0.016}$	$0.517_{\pm 0.011}$	$0.684_{\pm 0.007}$
SOC-REPL	$0.671_{\pm 0.008}$	$0.988_{\pm 0.002}$	$0.799_{\pm 0.006}$	$0.876_{\pm 0.023}$	$0.148_{\pm 0.010}$	$0.252_{\pm 0.015}$	$0.684_{\pm 0.008}$
SOC-RETW	$0.651_{\pm 0.009}$	$0.992_{\pm 0.003}$	$0.786_{\pm 0.007}$	$0.833_{\pm 0.056}$	$0.060_{\pm 0.009}$	$0.115_{\pm 0.016}$	$0.656_{\pm 0.010}$
SOC-FRIE	$0.857_{\pm 0.010}$	$0.933_{\pm 0.003}$	$0.893_{\pm 0.006}$	$0.860_{\pm 0.006}$	$0.726_{\pm 0.018}$	$0.787_{\pm 0.011}$	$0.858_{\pm 0.007}$
SOC-ALL	$0.863_{\pm 0.009}$	$0.932_{\pm 0.008}$	$0.896_{\pm 0.007}$	$0.862_{\pm 0.014}$	$0.741_{\pm 0.016}$	$0.796_{\pm 0.012}$	$0.863_{\pm 0.008}$
LING-HASH	$0.688_{\pm 0.010}$	$0.980_{\pm 0.003}$	$0.808_{\pm 0.007}$	$0.861_{\pm 0.016}$	$0.216_{\pm 0.018}$	$0.345_{\pm 0.023}$	$0.703_{\pm 0.010}$
LING-WORD	$0.745_{\pm 0.011}$	$0.885_{\pm 0.009}$	$0.808_{\pm 0.007}$	$0.697_{\pm 0.018}$	$0.466_{\pm 0.020}$	$0.558_{\pm 0.016}$	$0.733_{\pm 0.009}$
LING-GLDA	$0.723_{\pm 0.010}$	$0.790_{\pm 0.013}$	$0.755_{\pm 0.010}$	$0.559_{\pm 0.018}$	$0.468_{\pm 0.019}$	$0.509_{\pm 0.017}$	$0.674_{\pm 0.011}$
LING-DLDA	$0.798_{\pm 0.009}$	$0.838_{\pm 0.013}$	$0.817_{\pm 0.008}$	$0.688_{\pm 0.020}$	$0.627_{\pm 0.017}$	$0.656_{\pm 0.015}$	$0.761_{\pm 0.009}$
LING-SENT	$0.707_{\pm 0.011}$	$0.897_{\pm 0.012}$	$0.791_{\pm 0.010}$	$0.658_{\pm 0.033}$	$0.346_{\pm 0.020}$	$0.453_{\pm 0.023}$	$0.698_{\pm 0.013}$
LING-ALL	$0.804_{\pm 0.007}$	$0.847_{\pm 0.010}$	$0.825_{\pm 0.006}$	$0.702_{\pm 0.015}$	$0.636_{\pm 0.015}$	$0.668_{\pm 0.013}$	$0.770_{\pm 0.007}$
FULL	$0.894_{\pm 0.007}$	$0.936_{\pm 0.007}$	$0.915_{\pm 0.005}$	$0.878_{\pm 0.010}$	$0.805_{\pm 0.012}$	$0.840_{\pm 0.007}$	$0.889_{\pm 0.005}$

Table 2: Results for the political affiliation task.

such personalities). Linguistic features also have encouraging performance (especially, LING-DLDA, LING-WORD, LING-HASH) as different classes of users discuss either different topics or common topics in different ways: e.g., republicans are passionate about different issues ("liberty") than democrats ("inequality", "homophobia") and tend to use a specific vernacular ("obamacare") when discussing issues of interest to both sides (healthcare reform). Another reason for the good performance of linguistic features is the event of the Nov. 2010 elections, which precipitated party-specific, get-out-the-vote messages and voting-related discussions showcased by the hashtag features in Table 3. We notice that class-specific topic models (LING-DLDA) outperform generic topic models (LING-GLDA): generic topic models define coarse-grained topics shared by republicans and democrats, e.g. they inform us that users discuss the Nov. 2010 elections (e.g, *news, delaware, o'donnell, christine*), while domain specific topics reveal items of specific interest for republicans (*american, government, conservative, freedom..*) vs. democrats (*progressive, moveon, obama*), thus being more discriminative (see Figure 4 for a few examples.)

Starbucks Fans. As seen in Table 5, deciding whether a user is a potential follower of Starbucks can be done with reasonable precision (0.763) and recall (0.759). Results indicate that *profile* and *linguistic* information are the most helpful features. Profile features alone achieve performance close to the FULL system. A look at the most discriminative features for GBDT reveals that the ratio between followers and friends is the most relevant feature, suggesting that Starbucks afficionados are users who follow others more than they are followed: they are mostly *information seekers*, e.g. probably people looking for deals and coupons.

Both social and linguistic features do not perform as well as in the political affiliation task. We hypothesize that the potential of prototype-based features such as LING-WORD and SOC-FRIE is diluted by the heterogeneity of the large group of Starbucks fans. Within the set of linguistic features, LING-HASH and LING-DLDA perform best overall, while sentiment features LING-SENT have the highest precision but very low recall. This latter result is due to two facts: the fact that LING-SENT look at the sentiment attached by users to the word "Starbucks"; and the nature of Twitter accounts: on average, people mention the name of a particular business only sporadically, as the focus of the communication is mostly on personal developments, news tracking and sharing, etc. Under these circumstances, features which analyze in depth the totality of the user's account become even more important (hence the good relative performance of PROF-ALL).

Ethnicity. Identifying African-American users proves to be a more challenging task (see Table 6), for which linguistic features (LING-ALL) prove to perform best. Within the set of linguistic features, LING-HASH and LING-WORD have the highest precision (albeit low-recall): Table 3 shows examples of the lexical usage (e.g., "betta", "brotha") and issues or entities (e.g. "jeezy", aka "Young Jeezy") in African-American user accounts which can help our automatic classification system. However, personalities and lexical usages which were once the province of the African-American community have long gained adoption by other groups, which leads to linguistic features being useful only up to a point for our task. LDA models are once again the most balanced in P/R, showing the highest F-measure. For this classification task, topics mostly capture lexical usage (one topic is (*gettin, watchin, tryna, finna*) and popular celebrities (*beyonce, smith, usher, kanyewest, atlanta*). We find that the task can also be helped by profile information (e.g. African Americans tend to have longer bio descriptions, as one of the most discriminative features reveals), but best classification performance is only achieved by combining the different classes of features.

Features	DEMOCRATS	REPUBLICANS	AFRICAN-AMERICANS	STARBUCKS FANS
LING-WORD	inequality, homophobia, woody, socialism	obamacare, liberty, taxpayer, patriots	betta, brotha, finna, jeezy	mocha, recipes, dining, espresso
LING-HASH	#itgetsbetter, #VOTE2010, #ProgCa, #voteDem	#cagop, #ConsNC, #ObamaTVShows, #RememberNovember	#sadtweet, #pissed, #PSA, #teamdroid	#Yelp!, #iPhone, #Starbucks, #Starbucks
SOC-REPL	txvoodoo, polipaca, liberalcrone, socratic	itsonlywords, glenasbury, RickSmall, astroterf	MonicaMyLife, serenawilliams, RayJ, MissyElliott	GoldenMiley,Heyitsmimila_, Aerocles, GoodCharlotte
SOC-RETW	ebertchicago, BarackObama, KeithOlbermann, GottaLaff	Drudge_Report, michellemalkin, fredthompson, mikepfs	WatchJ, DeRayDavis, TiaMowry, KDthunderup	TheBieberFun, Nordstrom, Starbucks, Orbitz, WholeFoods
SOC-FRIE	Barack Obama, Rachel Maddow, Al Gore, Keith Olbermann	Michelle Malkin, Heritage Foundation, Glenn Beck, Newt Gingrich		

Table 3: Examples of automatically induced features LING-WORD,LING-HASH,SOC-REPL,SOC-RETW and SOC-FRIE.

Dominant class	Topic id	Topic words
Democrats	2	anti, rights, justice, protest, reform
Republicans	7	america, country, conservative, constitution, tea
Democrats	72	tax, economy, spending, cuts, stimulus
Democrats	75	progressive, moveon, political, thinkprogress, corporations

Table 4: Examples of highly discriminative topics from LING-DLDA for the political affiliation task, together with the dominant class.

System	PREC	REC	F-MEAS
B1	$0.817_{\pm 0.190}$	$0.019_{\pm 0.006}$	$0.038_{\pm 0.012}$
BEHAV-ALL	$0.583_{\pm 0.023}$	$0.613_{\pm 0.009}$	$0.597_{\pm 0.010}$
PROF-ALL	$0.746_{\pm 0.018}$	$0.723_{\pm 0.023}$	$0.735_{\pm 0.020}$
SOC-REPL	$0.511_{\pm 0.020}$	$0.979_{\pm 0.007}$	$0.671_{\pm 0.018}$
SOC-RETW	$0.502_{\pm 0.016}$	$0.995_{\pm 0.003}$	$0.667_{\pm 0.014}$
SOC-ALL	$0.532_{\pm 0.048}$	$0.885_{\pm 0.180}$	$0.613_{\pm 0.112}$
LING-HASH	$0.528_{\pm 0.950}$	$0.950_{\pm 0.008}$	$0.678_{\pm 0.019}$
LING-WORD	$0.585_{\pm 0.024}$	$0.660_{\pm 0.023}$	$0.619_{\pm 0.017}$
LING-GLDA	$0.602_{\pm 0.026}$	$0.642_{\pm 0.033}$	$0.620_{\pm 0.021}$
LING-DLDA	$0.614_{\pm 0.016}$	$0.660_{\pm 0.024}$	$0.636_{\pm 0.016}$
LING-SENT	$0.700_{\pm 0.030}$	$0.125_{\pm 0.105}$	$0.211_{\pm 0.015}$
LING-ALL	$0.628_{\pm 0.026}$	$0.660_{\pm 0.021}$	$0.643_{\pm 0.015}$
FULL	$0.763_{\pm 0.021}$	$0.759_{\pm 0.004}$	$0.761_{\pm 0.010}$

Table 5: Results for the Starbucks fans task

5 Conclusions and future work

We presented a generic model for user classification in social media and provided extensive quantitative and qualitative analysis which shows that in the case of Twitter users, this is a feasible task, although results vary across classes. Linguistic features, especially topic-based, are found to be consistently reliable. Explicit social network features, though expensive to collect, are valuable and may especially help if the target class is rich in celebrities with an active Twitter presence.

Future work directions include the integration of n-gram features as experimented in previous work, the use of link analysis algorithms to better incorporate the social dimension, experimenting with different user classes and finally, incorporating our methods into applications which benefit from user profiling.

References

Blei, D.; Ng, A.; and Jordan, M. 2002. Latent dirichlet allocation. *JMLR* (3):993–1022.

Burger, J., and Henderson, J. 2010. An exploration of observable features related to blogger age. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*, 710–718.

Burson-Marsteller. 2010. Press Releases Archives. In *Archive of Sept 10, 2010.*

System	PREC	REC	F-MEAS
B1	$0.878_{\pm 0.010}$	$0.421_{\pm 0.014}$	$0.569_{\pm 0.013}$
BEHAV-ALL	$0.534_{\pm 0.014}$	$0.496_{\pm 0.021}$	$0.514_{\pm 0.013}$
PROF-ALL	$0.578_{\pm 0.020}$	$0.643_{\pm 0.029}$	$0.609_{\pm 0.022}$
SOC-REPL	$0.813_{\pm 0.047}$	$0.090_{\pm 0.006}$	$0.161_{\pm 0.011}$
SOC-RETW	$0.709_{\pm 0.068}$	$0.061_{\pm 0.007}$	$0.112_{\pm 0.012}$
SOC-ALL	$0.671_{\pm 0.021}$	$0.367_{\pm 0.011}$	$0.474_{\pm 0.011}$
LING-HASH	$0.792_{\pm 0.033}$	$0.127_{\pm 0.007}$	$0.218_{\pm 0.011}$
LING-WORD	$0.671_{\pm 0.016}$	$0.333_{\pm 0.014}$	$0.445_{\pm 0.014}$
LING-SENT	$0.597_{\pm 0.029}$	$0.254_{\pm 0.015}$	$0.355_{\pm 0.015}$
LING-GLDA	$0.625_{\pm 0.020}$	$0.602_{\pm 0.018}$	$0.613_{\pm 0.015}$
LING-DLDA	$0.645_{\pm 0.017}$	$0.640_{\pm 0.013}$	$0.642_{\pm 0.021}$
LING-ALL	$0.655_{\pm 0.014}$	$0.641_{\pm 0.012}$	$0.647_{\pm 0.006}$
FULL	$0.646_{\pm 0.017}$	$0.665_{\pm 0.013}$	$0.655_{\pm 0.015}$

Table 6: Results for the ethnicity task.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet: A Content-based Approach to Geo-locating Twitter Users. In *Proceedings of CIKM*.

Fink, C.; Mayfield, J.; Piatko, C.; Finin, T.; and Martineau, J. 2009. Geolocating Blogs from Their Textual Content. In *Proceedings of ACL*, 710–718.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29(5):1189–1232.

Friedman, J. H. 2006. Recent advances in predictive (machine) learning. *Journal of Classification* 23(2):175–197.

Garera, N., and Yarovsky, D. 2007. Modeling latent biographic attributes in conversational genres. In *Proceedings of CIKM*.

Herring, S., and Paolillo, J. 2010. Gender and genre variation in weblogs. In *Journal of Sociolinguistics*, 710–718.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007*.

Jones, R.; Kumar, R.; Pang, B.; and Tomkins, A. 2007. I Know What you Did Last Summer - Query Logs and User Privacy. In *Proceedings of CIKM*.

Otterbacher, J. 2010. Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata. In *Proceedings of CIKM*.

Pasca, M. 2007. What you seek is what you get: Extraction of class attributes from query logs. In *Proceedings of IJCAI*.

Quantcast. 2010. Report May 2010. In *http://www.quantcast.com/twitter.com*.

Ramage, D. 2010. Characterizing Microblogs with Topic Models. In *Proceedings of ICWSM 2010*.

Rao, D.; D., Y.; Shreevats, A.; and Gupta, M. 2010. Classifying Latent User Attributes in Twitter. In *Proceedings of SMUC-10*, 710–718.

Singh, S. 2001. A pilot study on gender differences in conversational speech on lexical richness measures. In *Literary and Linguistic Computing*.

Smola, A., and Narayanamurthy, S. 2010. An architecture for parallel topic models. In *Proceedings of VLDB*.

Thomas, M.; Pang, B.; and Lee, L. 2006. Get out the vote: determining support or opposition from congressional floor-debate transcripts. In *Proceedings of EMNLP*.

Twitter. 2010. Twitter API documentation. In *http://dev.twitter.com/doc*.

Weber, I., and Castillo, C. 2010. The Demographics of Web Search. In *Proceedings of SIGIR*.

Wiebe, J.; Wilson, T.; and Cardie, C. 2005. Annotating expressions of opinions and emotions in language. In *Language Resources and Evaluation*, 165–210.