Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter

Marco Pennacchiotti
Yahoo! Labs
pennac@yahoo-inc.com

Ana-Maria Popescu
Yahoo! Labs
amp@yahoo-inc.com

ABSTRACT

More and more technologies are taking advantage of the explosion of social media (Web search, content recommendation services, marketing, ad targeting, etc.). This paper focuses on the problem of automatically constructing *user profiles*, which can significantly benefit such technologies. We describe a general and robust machine learning framework for large-scale classification of social media users according to dimensions of interest. We report encouraging experimental results on 3 tasks with different characteristics: political affiliation detection, ethnicity identification and detecting affinity for a particular business.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Web-based services*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Design, Experimentation

Keywords

Microblogging, Social Media, User Profiling, Machine Learning

1. INTRODUCTION

The explosion of real-time microblogging services such as Twitter and Facebook has led to a wealth of efforts to make use of social media content as well as various social graphs. For example, major Web technology players such as Google, Bing and Yahoo! Search now incorporate microblog posts and trend analysis in their results; in addition to using social information in conjunction with existing search and retrieval models, significant efforts are dedicated to developing new applications (e.g., user and post recommendation services) for the new, real-time social realm.

In this context, a problem of significant interest is that of automatic *user classification and profiling*, i.e. mining values of various user attributes such as demographic characteristics (e.g., age, gender, ethnicity, origin), coarse- and fine-grained interests (e.g. politics, soccer, Starbucks, Glee TV series), stances on various issues (e.g. liberal, pro-choice), etc. These user models can help in multiple application scenarios, such as:

Authoritative users extraction : discovering expert users for a target topic. Bing Social, Klout [1] and other similar applications study the users' posted content and their interactions with others in order to suggest authoritative users to follow on a topic. A repository of user profiles may significantly help in estimating the authority of a user on a topic - for example, a user primarily interested in politics is more likely to be authoritative on political issues than a user whose attention is equally divided among many interests.

Post reranking in web search : reranking social media posts retrieved in response to search engine queries based on a particular quality function and the Web users' preferences. Improving and personalizing social media posts' retrieval and display would benefit from information about the authoritativeness and profile of the user writing the post, as well as from knowledge of the microblogging profile (if available) of the Web user issuing a query.

User recommendation : suggesting new interesting users to a target user. The Twitter's 'Who To Follow' and Google's 'Follow Finder' [2] applications suggest new accounts to follow to Twitter users by using exclusively the properties of a specific social network. Recently, several studies have shown the potential of user profiles to improve the quality and the coverage of those applications [16, 18]. Yet, much is still unexplored. User profile information can be used both to automatically match two users with similar profiles and to explicitly allow a user to specify the type of new "friends" he is looking for (e.g., people interested in sports).

The above applications, as well as many others, would significantly benefit from the existence of large-scale knowledge repositories of user profile attributes or classification information.

Most social network and microblog services already store profile information in the form of name, age, location and short summary of interests, but such information is often incomplete (e.g., a user may choose not to post bio details) or misleading (e.g., 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 mostly omitted. Also, Twitter *directories* such as WeFollow [3] are manually built examples of profile repositories. Such directories allow Twitter users to manually add themselves in specific categories such as 'Music' and 'Bloggers'; unfortunately they are limited in size, the categories are pre-defined and the manual addition process can be cumbersome.

In response to this shortage of user profile repositories, researchers

KDD'11, August 21–24, 2011, San Diego, California, USA.

[1] http://www.bing.com/social and http://klout.com/

[2] http://www.twitter.com and http://www.followfinder.googlelabs.com

[3] http://wefollow.com/

have experimented with automatic methods for populating values of choice attributes such a user's age and location of origin. However, a discussion of a general framework for attribute value extraction in the service of user profiling is still missing.

In this paper, we present an architecture which addresses the above problem by casting it as a **user classification** task and leveraging two types of information: *user-centric information* reflecting the linguistic content of the user's tweets, her social behaviors and likes, and; *social graph information* in the form of the distribution of the possible target class values for the people connected to the user by a social graph link. Our main contributions are:

(1) We introduce a novel architecture for social media user classification which is composed of: (i) a machine learning component that classifies users by leveraging user-centric features (profile, linguistic, behavioral, social), and; (ii) a graph-based updating component that integrates social graph information, by updating the classification result according to the class's values distribution across the social connections of the user.

(2) We show that the architecture can be instantiated and used with good results on Twitter and three different tasks (political orientation, ethnicity and business fans detection).

(3) We provide an in-depth analysis of the results, revealing that user-centric features achieve alone good classification results, while social graph information has a negligible impact on the overall performance. We also comment on the value of specific user-centric features: we show that linguistic content features are in general highly valuable, and that large-scale topic models are particularly robust, showing promise for additional user classification tasks.

The rest of the paper is organized as follows: Section 2 places our research in the context of previous work; Section 3 describes the machine learning component and its feature set; Section 4 presents the graph-based updating component; Section 5 outlines our experiments and reports an in-depth quantitative and qualitative analysis of the obtained results; Section 6 describes our conclusions and future work directions.

2. RELATED WORK

User attribute detection 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. For example, researchers investigated the detection of *gender* from well-written, traditional text [10], blogs [4] reviews [14], e-mail [9], user search queries [12, 24] and, for the first time, Twitter [19].

Another area of focus has been understanding how the political orientation of a user is reflected in their writings: for example, [23, 13, 22] investigate congressional transcripts as well as informal user posts while [19] focuses on microblogs (we discuss the latter in more detail below). Other previously explored attributes include the user's *location* [6, 12], *location of origin* [19] and *age* [12, 19].

While such previous work has addressed blogs and other informal texts, microblogs are just starting to be explored for user classification. In the case of microblogs, we are usually interested in mining a series of short updates over a period of time for a particular user in order to infer particular user attributes.

While previous work uses a mixture of sociolinguistic features, ngram models as well as deep sentence-level analysis, we focus on aggregate features (e.g., those from large-scale topic models) in order to better exploit the user-created microblog content.

Twitter-related research and applications. Social media in general and Twitter in particular are the subject of considerable ongoing research work: spam detection research [2], tweet sentiment analysis [1], conversation modeling [20] and more. While in the future we plan to integrate insights from a number of research directions, we focus herein on the work directly related to ours. The most relevant such paper is [19], 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).

A second relevant paper is [18] which 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.

In addition to the growing number of publications, Twitter users benefit from an entire ecosystem of applications which seek to make Twitter data easier to find and use. Of particular interest to us are *directories* such as WeFollow, whose goal is to make it easier to find users from particular categories (e.g., certain professions, political persuasions, etc.). As outlined in the previous section, our work can be used to automatically augment such already available user databases.

3. MACHINE LEARNING MODEL

Task. The goal of our work is to build a general, scalable and robust architecture for automatically computing the values of given user attributes for a large set of Twitter users.

Solution overview. We cast the problem as a classification task, and solve it using an architecture with two components: the first component is a machine learning algorithm that learns a classification model from labeled data and user-centric features. The model is then used to classify new incoming users (e.g., as being a Democrat or a Republican). The second component of the architecture is a graph-based label updating function which uses social graph information in order to revise the classification label assigned to each user by the initial machine learning component.

An important requirement for our architecture is *scalability*: as Twitter reports, the number of active Twitter users per month is in the tens of millions. We estimate that each month the number of active users increases by about 8%, and that every week 1.6M new active users join the network. [4] The post volume is also steadily increasing, amounting to about 102M tweets per day in January 2011. Our architecture adopts highly scalable algorithms and engineering solutions, in order to scale to terabytes of data, and potentially petabytes in the near future. Both our architectural components run over a Map/Reduce Hadoop system which is able to build and apply classification models on a large scale within a few hours.

The rest of this section describes in more depth the first machine learning component of the architecture, while Section 4 analyzes the updating function. The machine learning component takes as input a small training set of labeled examples for a given class (e.g. 'political affiliation') and learns a classification model which can then be used to label the large set of Twitter users. As a classification algorithm, we use the Gradient Boosted Decision Trees - GBDT framework [7], which consists of an ensemble of decision trees, fitted in a forward step-wise manner to current residuals.

[4] Statistics based on data collected between December 2010 and January 2011.

Friedman [7] shows that by drastically easing the problem of overfitting on training data (which is common in boosting algorithms), GDBT competes with state-of-the-art machine learning algorithms such as SVM with much smaller resulting models and faster decoding time [8], which is an advantage in our real-time setting. We use a distributed GBDT architecture running over Hadoop [26] which allows us to scale to the dimensionality of Twitter.

To learn the classification model, we use a large set of features, falling into four main categories (described below), according to the nature of information they aim to capture: *profile*, *messaging (tweeting) behavior*, *linguistic content of messages* and *social network information*.

3.1 Profile features

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 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 [6] 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 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, such as the following for respectively age and 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* for 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 material. 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 patterns in the bio field as listed above, presence of the location field. Such features are general in nature and portable across different classification tasks.

3.2 Tweeting behavior features

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 [11] suggest that users who rarely post tweets but have many friends tend to be information seekers, while users who often post URLs in their tweets are most likely information providers. Rao and colleagues [19] 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 standard deviation between tweets, average number and standard deviation of tweets per day, fraction of tweets posted in each of 24 hours.

Like profile features, tweeting behavior features are fully generalizable and portable across different classification tasks.

3.3 Linguistic content features

Linguistic content information encapsulates the user's lexical usage and the main topics of interest to the user. Simple linguistic information is helpful for classifying users in several media, such as formal texts, blogs, spoken conversational transcripts or search sessions (see Section 2 for a discussion). We explore a wide variety of linguistic content features, as detailed below. Note that as far as language models are concerned, we prefer the use of Latent Dirichlet Allocation (LDA) [3] and automatically bootstrapped prototypical words over a more simple bag-of-word model, since several studies, e.g. [19], have showed that bag-of-words models are usually outperformed by more advanced linguistic ones.

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 (e.g. younger people tend to use words such as 'dude' or 'lmao') and phrases denoting typical interests of people in that class (e.g., Democrats may tend to use the expression 'health care' more than Republicans.) Rao and colleagues [19] explored this intuition by manually building a list of words that 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 [15].

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:

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

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 [5].

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:

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

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:

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

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

where WP is the set of proto words for class c. Table 4 reports the highest scoring proto words for the classes targeted in our paper.

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 [3] 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. 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 an LDA parallel implementation [21] 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 wrt 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 wrt t by relying on Opinion Finder 1.5 [25] 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 in the window are neutral, or not in the OpinionFinder lexicon, the tweet is classified as neutral wrt t. Otherwise, we classify the tweet as *positive* if a majority of the terms are *positive* and *negative* otherwise (after accounting for the polarity flipping effect of simple negation modifiers such as "not" or "no"). Given the set of tweets of user u annotated with u's sentiment towards t, we retain as features the percentage of positive tweets wrt t , the percentage of negative tweets wrt t and the percentage of neutral tweets wrt t.

We also derive aggregated features indicating the overall sentiment of the user u wrt 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 features

These general 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. GRAPH-BASED LABEL UPDATE

In this section we describe a graph-based label updating algorithm that attempts to further improve the quality of the users' classification by taking into account social connections.

Twitter social connections are of two types: *friend* and *follower*. A user's friends are all users to whom she connected, in order to receive their messages and updates. Conversely, a user's followers are all users who have connected to her. Users with many followers and few friends are oftentimes called 'information sources' (e.g., news broadcasting organizations and celebrities), while users with many friends and few followers are called 'information seekers' (e.g., users who joined Twitter to keep up with the latest news).

Our intuition is that social connections can provide signals for correcting the errors of the machine learning component described in Section 3, by inverting the classification label. For example, a user mistakenly classified as a Democrat by the machine learning algorithm could be correctly classified as a Republican if her friends are predominantly classified as Republican.

In this paper we only experiment with users' friend connections, as they directly express the social likes of the user, and should thus intuitively provide better signal with respect to follower connections. In future work we will also explore the use of followers.

The graph-based updating algorithm is applied after the machine learning classification step, as follows. First, we collect the set of friends for each user in the original class dataset (e.g., Democrats and Republicans). We then apply the learned class-specific machine learning model on the friend set, assigning class-specific membership scores (and labels) to each friend. In a final *label updating* phase, we update the class labels for our original users based on the newly computed class-specific membership scores for their friends.

The label update function uses the classification confidence value (hereafter $score$) returned by GBDT for each user; the value is in the $[+1, -1]$ range, where positive values correspond to positively classified users, and negative values to negatively classified users. Higher absolute scores indicate a higher classification confidence.

For each user u_i in the original dataset, the function takes as input the $score(u)$ of the user and of each of her friends F_i, and returns a new updated score ($score'(u)$) for the user. If the score is positive, the user is classified as positive example, otherwise as a negative one. $score'$ is computed as follows:

$$score'(u_i) = \alpha \cdot score(u_i) + (1-\alpha) \cdot \frac{\sum_{j \in F_i} w_{ij} \cdot score(u_j)}{|F_i|} \quad (4)$$

where $w_{ij} \in [0, 1]$ is the connection weight between user u_i and a friend u_j; $|F_i|$ is the number of friends of user u_i, and; $\alpha \in [0, 1]$ is a parameter that indicates how relevant the original ML-based user score should be. Setting $\alpha = 1$ equates to using the original machine learning model described in Section 3 without any graph-based contribution. $\alpha = 0$ updates the score using only graph-based information.

The connection weight w_{ij} models the intuitions that not all friends are equally strongly connected to a user and that some friends are overall more influential than others. For example, a user may often engage in discussion with and refer to some friends more than others. We capture this intuition by giving a higher weight to friends which are often mentioned by the user. [6] Additionally, some friends are users with a larger Twitter-wide influence (e.g., "information sources"). We combine the two components (connection strength and network-wide influence) as follows [7]:

$$w_{ij} = \frac{1}{2} \cdot \frac{|mentions_{ij}|}{\sum_{k \in F_i} |mentions_{ik}|} + \frac{1}{2} \cdot \frac{|ratioFolFriends_j|}{\sum_{k \in F_i} |ratioFolFriends_k|} \quad (5)$$

where: $|mentions_{ij}|$ is the number of times that user u_i mentions user u_j in her tweets ; $|ratioFolFriends_j|$ is a simple measure of a user's influence in form of the ratio between the number of the user's followers to the number of the user's friends.

5. EXPERIMENTAL EVALUATION

We evaluate our architecture over three binary classification tasks: detecting political affiliation, detecting a particular ethnicity, and identifying 'Starbucks fans'.

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 for many applications – e.g., when employing user recommendation tools, a user may want to look for new friends with the same political ideas; social analytics applications may engage the audience by reporting opposing views on political issues or tracking the concerns and interests of a party's base.

We build the gold standard dataset by scraping lists of users who classified themselves as either Democrat or Republican in the Twitter directories WeFollow and Twellow[8]. We collect a total of 10,338 users, equally distributed in the two classes. In this paper, the datasets are 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 political affiliation is close to that of our sample, as shown in recent Twitter demographic studies [5]. For the graph-based component, we collect the list of friends for each user in the dataset. Overall, we found 2.7 million users, reduced to 1.2 millions by discarding those that posted less than 5 tweets in the considered Twitter corpus (see below).

Ethnicity. The 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 [17]. The statistics mean that automatically identifying users of this ethnicity can have benefits from multiple perspectives: linguistic and sociological (we can study the language, opinions or preoccupations of an important segment of the Twitter user population); for behavioral targeting applications (more focused marketing and ad targeting campaigns), if the previous analysis reveals that this segment of the user population has a set of unique interests.

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. We then collect about 909K friends for the users in the dataset, reduced to 508K after the 5 tweets cut.

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. The task of identifying potential business customers is particularly attractive for ad targeting and marketing campaign design applications. 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 who already follow Starbucks on Twitter, and 5000 negative examples represented by a sample of users following who do not. We finally collect about 1.9M friends for the users in the dataset, reduced to 981K after the 5 tweets cut.

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

[6] In Twitter, a user 'mentions' another user by including the tag '@' followed by the other user's name. These mentions usually correspond to replies and conversational exchanges between the two users.

[7] We experimented with a number of other combination functions, with comparable experimental results

[8] http://www.twellow.com

are classes of interest. We experiment in a 10-folds cross validation setting, to compute statistical significance.

System Configurations. We experiment with different instantiations of our architecture, as follows. ML: the architecture using only user-centric features in the machine learning component, i.e. setting $\alpha = 1$ in Eq. 4. GRAPH: the architecture using only social network information (the graph-based updating component, i.e. setting $\alpha = 0$). HYBRID: using both the machine learning component and graph updating with equal weight, i.e. $\alpha = 0.5$. These three configurations allow us to check if including graph-based information in the classification task is helpful.

The performance of the various systems is compared against two baselines. B2 is a generic reference baseline represented by our machine learning component trained only on the profile and tweeting behavior features (the basic information types readily available from Twitter). For each of the three tasks, B1 denotes a different task-specific baseline:

Political affiliation: B1 is a system which classifies as Democrats all users explicitly mentioning their Democratic/liberal political affiliation in the bio field (see Section 3.1) and proceeds the same for Republicans.

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 is a system which classifies as Starbucks fans all the users who explicitly mention Starbucks in their bio field.

System and features setup. 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 political 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 September - October 2010 time period. All gold standard datasets described above contain users that were active in the considered time period by posting at least 5 tweets, and that posted at least 50% of their tweets in English.

5.1 Experimental results

Overall results for political affiliation, ethnicity and Starbucks fans are reported respectively in Tables 1 , 2 and 3. Table 4 reports the semi-automatically induced features, obtained by applying Eq. 1 as described in Section 3. Results 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 1). 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 HYBRID system, combining the graph update function and the machine learning component using all available features, outperforms B2 in F-measure of +11% for Democrats, and +31% for Republicans. Since B2 is based only on profile and behavior features, this result suggests it is worthwhile to explore and implement sophisticated social and linguistic features, in order to obtain good classification results.

As we can see by comparing the performance of the HYBRID and ML systems, revising the ML-derived score by taking into account the scores of a test user's neighbors has a positive, but small effect on the final results; the improvements are consistent across measures and political affiliations, but not statistically significant. We attribute the magnitude of the effect to a hard-to-beat baseline (in the form of the ML system). For Democrats, the profile of the neighborhood is highly predictable - Democrats tend to consistently have a large percentage of friends with the same affiliation (as evidenced by both Figure 1 and by the precision and recall of the GRAPH system). For Republicans, the political affiliation of the neighbors is more mixed (e.g., Republican Twitter users tend to have friends - and followers - with both probable Republican and Democrat affiliations). Using the updating function alone (GRAPH) gives good performance, but significantly worse than ML, confirming that social graph information are helpful but not necessary.

Table 1 also 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 (ML model) improves significantly over SOC-ALL alone of +2.6% accuracy, suggesting that these latter features add important value to the task. The feature importance values returned by the GBDT algorithm reveal that the 3 most discriminative features are from the SOC-FRIE set, but at the same time, among the first 20 features, 9 are linguistic and 5 behavioral/profile.

The high performance of social features is due to the typical characteristic of users interested in politics, of interacting with media or party personalities with an established Twitter presence, as those reported in the last three rows of Table 4. 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 November 2010 elections, which precipitated party-specific, get-out-the-vote messages and voting-related discussions showcased by the hashtag features in Table 4. We notice that class-specific topic models (LING-DLDA) outperform generic topic models (LING-GLDA); generic topic models define corse-grained topics shared by Republicans and Democrats, e.g. they inform us that users discuss the November 2010 elections (e.g, *news, delaware, o'donnell, christine*), while domain specific topics reveal items of specific interest for Republicans (*American, conservative, freedom*) vs. Democrats (*progressive, moveon, obama*), thus being more discriminative (see Table 5 for examples.)

Starbucks Fans. As seen in Table 2, deciding whether a user is a potential follower of Starbucks can be done with reasonable precision (0.764) and recall (0.756). The HYBRID system returns the best F-measure performance, though the improvement over the ML system is small and not statistically significant, this again indicating that revising the ML test user scores based on the scores is helpful to a certain extent. The correlation between the test user ML score and the scores of her neighbors for this task is not as strong as in the case of the political affiliation task, as seen both in Figure 2 and

System	Democrats			Republicans			All
	PREC	REC	F-MEAS	PREC	REC	F-MEAS	ACC
B1	**0.989**	0.183	0.308	**0.920**	0.114	0.203	0.478
B2	0.735	0.896	0.808	0.702	0.430	0.533	0.727
BEHAV-ALL	0.663	0.774†	0.714†	0.436	0.307†	0.360†	0.605†
PROF-ALL	0.728	0.808†	0.765†	0.582	0.468$^\flat$	0.517†	0.684†
SOC-REPL	0.671	0.988$^\flat$	0.799†	0.876‡	0.148†	0.252†	0.684†
SOC-RETW	0.651	0.992$^\flat$	0.786†	0.833‡	0.060	0.115	0.656†
SOC-FRIE	0.857‡	0.933$^\flat$	0.893$^\flat$	0.860‡	0.726$^\flat$	0.787$^\flat$	0.858$^\flat$
SOC-ALL	0.863‡	0.932$^\flat$	0.896$^\flat$	0.862‡	0.741$^\flat$	0.796$^\flat$	0.863$^\flat$
LING-HASH	0.688	0.980$^\flat$	0.808†	0.861‡	0.216†	0.345†	0.703†
LING-WORD	0.745	0.885†	0.808†	0.697	0.466†	0.558†	0.733†
LING-GLDA	0.723	0.790†	0.755†	0.559	0.468$^\flat$	0.509†	0.674†
LING-DLDA	0.798‡	0.838†	0.817†	0.688	0.627$^\flat$	0.656$^\flat$	0.761$^\flat$
LING-SENT	0.707	0.897†	0.791†	0.658	0.346†	0.453†	0.698†
LING-ALL	0.804‡	0.847†	0.825$^\flat$	0.702	0.636$^\flat$	0.668$^\flat$	0.770$^\flat$
ML	0.893‡	0.927$^\flat$	0.910$^\flat$	0.863‡	0.805$^\flat$	0.833$^\flat$	0.883$^\flat$
GRAPH	0.844‡	**0.938**$^\flat$	0.888$^\flat$	0.865‡	0.695$^\flat$	0.770$^\flat$	0.850$^\flat$
HYBRID	0.895‡	0.936$^\flat$	**0.915**$^\flat$	0.878 ‡	**0.806**$^\flat$	**0.840**$^\flat$	**0.889**$^\flat$

Table 1: Overall classification results for the political affiliation task. †, ‡ and ♭ respectively indicate statistical significance at the 0.95 level wrt B1 alone, B2 alone, and both B1 and B2.

System	PREC	REC	F-MEAS
B1	**0.817**	0.019	0.038
B2	0.747	0.723	0.735
BEHAV-ALL	0.583	0.613†	0.597†
PROF-ALL	0.746	0.723†	0.735†
SOC-REPL	0.511	0.979$^\flat$	0.671†
SOC-RETW	0.502	**0.995**$^\flat$	0.667†
SOC-ALL	0.532	0.885$^\flat$	0.613†
LING-HASH	0.528	0.950$^\flat$	0.678†
LING-WORD	0.585	0.660†	0.619†
LING-GLDA	0.602	0.642†	0.620†
LING-DLDA	0.614	0.660†	0.636†
LING-SENT	0.700	0.125	0.211†
LING-ALL	0.628	0.660†	0.643†
ML	0.760	0.752$^\flat$	0.755$^\flat$
GRAPH	0.706	0.702†	0.695†
HYBRID	0.764	0.756†	**0.758**$^\flat$

Table 2: Overall classification results for the Starbucks fan task. †, ‡ and ♭ respectively indicate statistical significance at the 0.95 level wrt B1 alone, B2 alone, and both B1 and B2.

System	PREC	REC	F-MEAS
B1	**0.878**	0.421	0.569
B2	0.579	0.633	0.604
BEHAV-ALL	0.534	0.496†	0.514
PROF-ALL	0.578	0.643†	0.609†
SOC-REPL	0.813‡	0.090	0.161
SOC-RETW	0.709‡	0.061	0.112
SOC-ALL	0.671‡	0.367	0.474
LING-HASH	0.792‡	0.127	0.218
LING-WORD	0.671‡	0.333	0.445
LING-SENT	0.597	0.254	0.355
LING-GLDA	0.625‡	0.602†	0.613†
LING-DLDA	0.645‡	0.640†	0.642$^\flat$
LING-ALL	0.655‡	0.641†	0.647$^\flat$
ML	0.629‡	**0.799**$^\flat$	**0.703**$^\flat$
GRAPH	0.604‡	0.621†	0.611†
HYBRID	0.630‡	0.753$^\flat$	0.686$^\flat$

Table 3: Overall classification results for the ethnicity task. †, ‡ and ♭ respectively indicate statistical significance at the 0.95 level wrt B1 alone, B2 alone, and both B1 and B2.

in the performance of the GRAPH system. This result is to be expected, as the preference for a particular business in itself is a very specific, low-level attribute for a user while a political affiliation label is a more encompassing, broader attribute which is a more probable basis for community building. Additionally, our graph-based update is itself affected by the quality of the scores assigned to the neighbors of a test user - this quality is not as high as in the case of the political affiliation task, introducing additional noise.

Regarding the different features of the machine learning system, results show that *profile* and *linguistic* information are the most helpful features. Profile features alone achieve performance close to the ML 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 that follow others more than they are followed: they are mostly *information seekers*, e.g. probably people looking for deals. Both social and linguistic features do not offer performance as good as in the political affiliation task. This is probably due to the fact Starbucks fans are a very heterogeneous demographic group (as also the lists in Table 4 suggest), thus diluting the potential of prototype-based feature (e.g. LING-WORD and SOC-FRIE). Within the set of linguistic features, LING-HASH and LING-DLDA perform best overall, while sentiment features LING-SENT have the highest precision but a very low recall. This result is due to two facts: the

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

Table 5: Examples of highly discriminative topics from LING-DLDA for the political affiliation task, with the dominant class .

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 user's account become even more important (hence the good performance of PROF-ALL).

Ethnicity. The classification of African-American proves to be the most challenging task, as shown in Table 3. The ML has the best F-measure, significantly better than the HYBRID system; while the precision remains basically the same, the recall drops significantly. As we can see from the low-recall and F-measure numbers for the GRAPH system, using neighborhood information to predict the class membership of a test user is not very promising. Part of the problem is the imbalance in the real Twitter data for the target class; an additional aspect is the fact that African-American Twitter users are not a closed community, but rather connect to users of other ethnicities as well. We are again in the presence of an attribute which is not necessarily, by itself, the basis for community forming.

Regarding the different machine learning features, linguistic ones (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 4 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, e.g. one topic is (*gettin, watchin, tryna, finna*) and popular celebrities, e.g. (*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), but best classification performance is only achieved by combining the different classes of features.

Final observations. As a general comment, we note that the machine learning component alone achieves good performance, without the need of the social graph information embodied in the graph update function. Indeed, graph-based information seems helpful (to a small extent) only in the case of attributes such as political affiliation, for which a user may seek to connect with others which share the same conviction. Attributes such as the preference for a specific business or product and, finally, ethnicity are either too specific or too broad to alone be the basis for a connection. In such cases, graph-based information is not particularly helpful and can even hurt. Our experiments with neighborhood-based updates underscore the importance of a comprehensive ML model which can address the case of attributes not correlated with community building or the case of users with a small or inactive neighborhood. Conversely, we see that for some attributes (e.g., political affiliation), graph-based information alone can be used to confidently predict the class label for users with a shorter/sparser message history as long as they have some connections to other more active users.

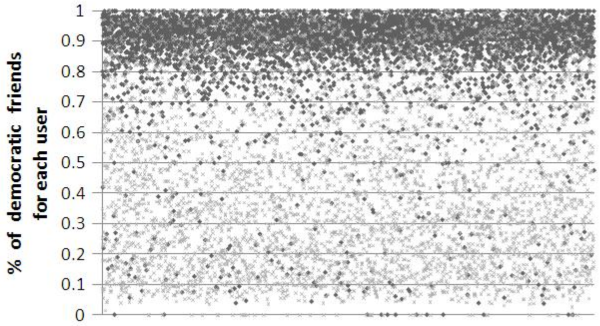

Figure 1: Percentage of friends that are Democrats for each user in the 'political affiliation' task. Black dots are Democratic users, gray dots are Republican users.

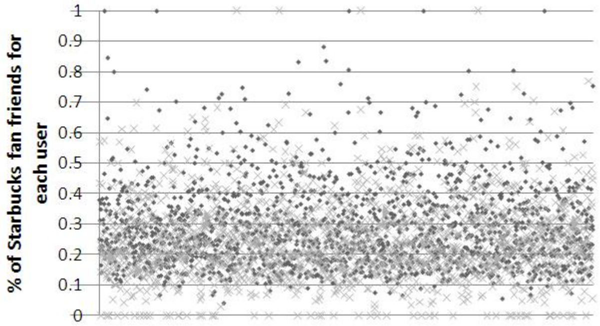

Figure 2: Percentage of friends that are Starbucks fan for each user in the 'Starbucks fan' task. Black dots are Starbucks fan users, gray dots are non Starbucks fan users.

A second important general conclusion is that linguistic features, especially topic-based LDA, are consistently reliable across all tasks, thus indicating an encouraging level of robustness and fostering further research in this area. Profile features are also quite stable across tasks, though their performance is not comparable to that of linguistic ones. Explicit social network features may be valuable in some tasks (though expensive to collect), especially when the user class of interest is rich in celebrities with which a user may connect.

6. CONCLUSIONS AND FUTURE WORK

This paper described a large-scale architecture for *user classification* in social media and extensive quantitative and qualitative results for three Twitter user classification tasks. We showed that: user classification is in general a feasible task, though results vary across different classes; a machine learning algorithm using user-centric features achieves alone good performance that is hard to improve by enriching the classification with social graph information; most machine learning features are general enough to be used in different classification and linguistic features show especially robust performance across tasks.

We are currently working on deploying and testing our models in

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
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	Michelle Malkin, Heritage Foundation, Glenn Beck, Newt Gingrich	Barack Obama, Rachel Maddow, Al Gore, Keith Olbermann		

Table 4: Example of automatically induced features LING-WORD,LING-HASH,SOC-REPL,SOC-RETW **and** SOC-FRIE**.**

a real-time, Twitter-based content aggregation and display application. Our user classification architecture will help in improving the user engagement with the application. An example use case involves the application being given a query (e.g., "Cairo protests") and retrieving high quality recent content from both Twitter and the Web at large about the anti-government protests in Egypt. Our architecture will help highlight content shared by users of particular political persuasions (e.g., Democrats vs. Republicans), as well as highlight authoritative users with opposing political views. Our architecture will also support the automatic analysis of the overall mood of a particular user class in conjunction with the topic at hand. In the longer term, we plan to integrate the user classification models into systems for content display personalization, which would benefit from knowing the profiles of users who create or share the displayed information.

7. REFERENCES

[1] L. Barbosa and F. J. Robust Sentiment Detection on Twitter from Biased and Noisy Data. In *Proceedings of COLING*, 2010.

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Proceedings of CEAS*, 2010.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *JMLR*, (3):993–1022, 2002.

[4] J. Burger and J. Henderson. An exploration of observable features related to blogger age. In *Computational Approaches to Analyzing Weblogs*, pages 710–718, 2010.

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

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

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

[8] J. H. Friedman. Recent Advances in Predictive (Machine) Learning. *Journal of Classification*, 23(2):175–197, 2006.

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

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

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

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

[13] S. Kim and E. Hovy. CRYSTAL: Analyzing Predictive Opinions on the Web. In *Proceedings of EMNLP*, 2007.

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

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

[16] M. Pennacchiotti and S. Gurumurthy. Investigating Topic Models for Social Media User Recommendation. In *Proceedings of WWW*, 2011.

[17] Quantcast. Report May 2010. In *http://www.quantcast.com/twitter.com*, 2010.

[18] D. Ramage, S. Dumais, and D. Liebling. Characterizing Microblogs with Topic Models. In *Proceedings of ICWSM*, 2010.

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

[20] A. Ritter, C. Cherry, and B. Dolan. Unsupervised Modeling of Twitter Conversations. In *Proceedings of HLT-NAACL*, 2010.

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

[22] S. Somasundaran and J. Wiebe. Recognizing Stances in Ideological On-Line Debates. In *Proceedings of NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, 2010.

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

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

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

[26] J. Ye, C. Jyh-Herng, C. Jang, and Z. Zhaohui. Stochastic gradient boosted distributed decision trees. In *Proceedings of CIKM*, 2009.