Religious Politicians and Creative Photographers: Automatic User Categorization in Twitter

Claudia Wagner*, Sitaram Asur† and Joshua Hailpern†
*Institute of Information and Communication Technologies, JOANNEUM RESEARCH, Graz, Austria
Email: claudia.wagner@joanneum.at
†HP labs, Palo Alto, California
Email: sitaram.asur@hp.com, joshua.hailpern@hp.com

***Abstract*—Finding the "right people" is a central aspect of social media systems. Twitter has millions of users who have varied interests, professions and personalities. For those in fields such as advertising and marketing, it is important to identify certain characteristics of users to target. However, Twitter users do not generally provide sufficient information about themselves on their profile which makes this task difficult. In response, this work sets out to automatically infer professions (e.g., musicians, health sector workers, technicians) and personality related attributes (e.g., creative, innovative, funny) for Twitter users based on features extracted from their content, their interaction networks, attributes of their friends and their activity patterns. We develop a comprehensive set of latent features that are then employed to perform efficient classification of users along these two dimensions (profession and personality). Our experiments on a large sample of Twitter users demonstrate both a high overall accuracy in detecting profession and personality related attributes as well as highlighting the benefits and pitfalls of various types of features for particular categories of users.**

I. Introduction

Manually built directory systems, such as *Wefollow*[1], have been created to allow end users to find new Twitter users to follow. In these ecosystems, users place themselves in explicitly defined categories (e.g. musician, doctor, or inspirational). While these categories are quite broad (e.g. geographical locations, brands, interests, personalities), their manual creation limits their scope. With an estimated 500 million users [1] on Twitter, only a small percentage of these handles are in systems like *Wefollow* – and therefore, the majority of Twitter users are not categorized. This information would be invaluable to both the end user, as well as advertisers.

This work directly addresses this need by focusing on automatic detection of attributes for users on Twitter. In particular, we focus on profession types/areas of Twitter users (e.g., musicians, entrepreneurs or politicians), and on online personality attributes of Twitter users (e.g., creative, innovative, funny). Because Twitter user metadata is highly limited, this work's analysis is based upon a comprehensive set of features that can be summarized into four key groups: *linguistic style*, *semantics*, *activity patterns* and *social-semantics*.

Our contributions in this paper are two-fold. First, we construct a comprehensive collection of features and examine their efficacy in classifying users on Twitter. Second, we consider two dimensions of user attributes, personality and profession, and show how efficient user classification can be performed on them. In order to create and explore these models we conducted extensive experiments using a corpus of more than 7k labeled Twitter users, crawled from *Wefollow*, and more than 3.5 million tweets. On this rich data set we trained Random Forest classifiers, and used correlation-based feature selection to prune correlated features.

Overall, the classifiers built on an optimal subset of features achieved an impressive accuracy ≥ 0.9 for most categories. When we examine the features independently, we observed that, not surprisingly, the feature groups differ in their accuracy across the various categories. For example, we found that social-semantic features are very efficient while classifying personality related attributes but achieve lower accuracy in the case of professions. Linguistic style features tend to work well with both personality and professions. However, not all features were universally useful for all categories. Specifically we found that *what a user says* (semantics) and *how a user behaves online* (activity patterns) tend to reveal less information about his professional areas and personality compared to *how a user says something* (linguistic style) and *what others say about/to another user* (social-semantic).

II. Related Work

Rao et al. [2] classify Twitter users according to a set of latent user attributes, including gender, age, regional origin, and political orientation. They show that message content is more valuable for inferring the gender, age, regional origin, and political orientation of a user than the structure or communication behavior of his/her social network. Rather than performing general user classification, [3] specifically models a user's political affiliation, ethnicity and affinity for one specific business, namely Starbucks. While their approach combines both user-centric features (profile, linguistic, behavioral, social), and social graph based features, their results suggest that user-centric features alone achieve good classification results, and social graph information has a negligible impact on the overall performance.

Our goal is to classify users along much broader planes of categories and we construct a comprehensive list of features for this purpose. Though the user attributes which we analyze in this work are substantially different from those analyzed in

[1]http://wefollow.com

[2] and [3], we also find content-based features more useful than activity features which capture amongst other structural similarities between users. However, unlike previous work we examine not only content which can directly be associated with a user (via an authorship relation) but also content which can indirectly be related with a user via other users.

Hong et al. [4] compare the quality and effectiveness of different standard topic models for analyzing social data. Their results suggest that topic features tend to be useful if the information to classify is sparse (message classification task), but if enough text is available (user classification task) simple TFIDF weighted words perform better. In our work we do not only compare the performance of TFIDF and topic models within the user classification task, but also explore the utility of explicit ontological concepts.

Unlike the above work which focused on individuals, [5] examine how networks emerging from user communication are closely replicated in the frequency of words used within these communities. In short, users who are strongly connected also talk about similar subjects and therefore use similar words. In addition, [5] also reveal that users who belong to one community tend to show similarities in the length of words they use or in their three letter word ending usage. This suggests that socio-linguistic features may help differentiate users in different communities. Therefore, we decided to incorporate linguistic style features which may have the potential to identify users who belong to the same community (e.g. group of users working in the same professional area or group of users sharing some personality characteristics).

Recently the prediction of personality related attributes of social media users gained interest in the research community [6] [7] [8], since characterizing users on this dimension would be useful for various applications such as recommender systems or online dating services. For example, [7] gathered personality data from 335 Twitter users by asking them to conduct a personality test and examined the relationship between personality and different types of Twitter users (popular users, influential users, listeners and highly read users). They identified those types of Twitter users by using publicly available counts of (what Twitter calls) "following," "followers," and "listed," and by using existing social media ranks. In our work we do not aim to predict users' personality based on the big five model of personality (since this requires users to complete a personality test first), but aim to predict self-reported personality related characteristics that form a user's distinctive character on Twitter. This allows us to study different aspects of user's online personality on a larger sample of Twitter users. Further, we explore a larger range of features which go far beyond the publicly available counts and ranks used in [7] and examine their utility for predicting self-reported personality characteristics.

III. Feature Extraction and Classification

In order to automatically categorize users on dimensions of interest and profession, a discriminative feature set must be extracted from Twitter for each user category. In this section, we describe various features that can capture user attributes and behavior on Twitter. We then show how the best features can be identified and used to build efficient classifiers.

A. Feature Engineering

1) Activity Features: Activity features capture various facets of user activities on Twitter including following, replying, favoriting, retweeting, tweeting, hashtagging and link sharing activities. The intuition behind this set of features is that *users who have similar online activity patterns are more likely to belong to the same category*. For example, people in advertising are more likely to reach out to other users. Celebrities such as musicians are likely to have many followers and follow fewer people.

a) Network-theoretic Features: Network-theoretic features describe user characteristics via their position in an activity network. Since we do not have access to the full social network of all users in our dataset, we construct three directed networks (*reply-*, *mention-*, and *retweet-network*) using information from the tweets of users. We use network-theoretic measures such as *in- and out-degree*, *clustering coefficient*, *hub and authority scores* obtained via HITS algorithm [9], *betweenness-, eigenvector-, and closeness-centrality*.

b) Following, Retweeting and Favoriting: Since we do not have access to the full following network of users, we compute simple ratios and counts (*Follower Count, Followee Count, Follower-Followee Ratio, Follower-Tweet Ratio, Favorite-Message Ratio*) which expose how popular a user is and/or how valuable and interesting his content might be for other users.

c) Diversity of Activities: The next set of features capture the diversity in a user's activity patterns. Our activity diversity features are based on Stirling's diversity measure [10] which captures three qualities of diversity - *variety*, *balance*, and *similarity*.

Social/Hashtag/Link/Temporal Variety: The social variety of a user is defined as the ratio between the number of different users a user interacted with (U_i) and the total number of messages published by this user (M). A high social variety indicates that a user mainly uses Twitter for a social purpose. The hashtag, link and temporal varieties are defined in the same way as the social variety.

Social/Hashtag/Link/Temporal Balance: To quantify the social balance of a stream, we define an entropy-based measure, which indicates how evenly balanced the social interactions of a user are. If a user's personal user stream has a high social balance, this indicates that the user interacts almost equally with a large set of users U_i. The hashtag, link and temporal balance are defined in the same way as the social balance as an entropy-based measure which quantifies how focused the hashtagging, the link sharing and the temporal tweeting activities of a user are.

Social/Hashtag/Link/Temporal Similarity: To measure the similarity between two users, we represent each user as a vector of users he interacted with, hashtags he used, links he used and time points he tweeted at. That means that we

use the interaction partners, hashtags, links and time points as features and count their frequency.

2) Semantic Features: Next we present a set of features that can be used to characterize users semantically via the content of their messages, or the content of their personal description (bio information). The intuition behind this set of features is that *users who talk about similar things are more likely to belong to the same category*.

Bag of Words: We represent each user by the union of all the published messages, excluding stopwords, and use term frequency-inverse document frequency (*TFIDF*) as the weighting schema. TFIDF allows us to emphasize the words which are most discriminative for a document (where a document in this case is a user).

Latent Topics: Topic modeling approaches discover topics in large collections of documents. The most basic topic modeling algorithm is Latent Dirichlet Allocation (LDA) [11]. In this work we fit an LDA model to a stratified sample of 10% of our training documents where each document consists of all messages authored by one user. We choose the default hyperparameters ($\alpha = 50/T$, $\beta = 0.01$) and optimize them during training by using Wallach's fixed point iteration method [12]. We choose the number of topics $T = 200$ empirically by estimating the log likelihood of a model with T= 50, 100, 150, 200, 250, 300 on held out data.

Explicit Concepts: In [13] the authors evaluated several open APIs for extraction semantic concepts and entities from tweets. They found that the AlchemyAPI extracts the highest number of concepts, and has also the highest entity-concept mapping accuracy. We apply the concept extraction method provided by Alchemy[2] to the aggregation of a sample of users' messages and represent each user as a weighted vector of DBpedia concepts.

Possessives: Besides using topics, concepts and words as semantic features, we also employ words following personal pronouns as features (e.g. "my mac", "my wife", "my girlfriend") since previous work [2] has shown that self-reference pronouns are often useful for distinguishing certain properties of individuals.

3) Social Semantic Features: Beside textual content created by a given user, which can be explicitly attributed to the user, other users may also associate their textual content with the user. For example, Twitter allows users to create user lists. A user may use these lists to organize their contacts. Usually lists consist of a name and a short, optional description of the list. If a user adds another user to a list he/she directly associates the list name and description with this user (*List TFIDF*, *List Concepts*). Further, users can be indirectly associated with topics by extracting the topics the user's top friends are talking about (*Friend Concepts*). We determine the top friends of a user by analyzing how frequently he interacts with other users, since previous research has shown that communication intensity is second to communication intimacy in predicting tie strength [14]. This set of features examines what others are saying about/to particular users, and how that can aid in categorizing users.

4) Linguistic Style Features: The last set of features are designed to characterize users via their use of language. The motivation for this set of features is to consider not what the user is saying but *how he says it*. Given the short length of tweets, it would be interesting to observe if there are significant linguistic cues and how they vary over users of different categories.

We use LIWC [15] to classify words into 70 linguistic dimensions[3] which we used as features. Apart from LIWC we also use a Twitter-specific part-of-speech tagger [16] and compute how frequently a certain tag is used by a user on average. Tags include standard linguistic part of speech tags such as verbs, nouns, proper nouns, adjectives, but also include Twitter or social media specific tags such as emoticons, links, usernames or hashtags. Therefore we computed features such as the mean number of emoticons or the mean number of interjections (e.g., lol, haha, FTW, yea) a user is using. Finally, we assess how easily text (in our case the aggregation of all recent tweets authored by a user) can be read by using standard readability measures such as the *Flesch reading ease score*, the *Fog index*, and the *Flesch-Kincaid grade level score*.

B. Feature Selection

A number of feature selection metrics have been explored in text categorization, among which information gain (IG), chi-square (CHI), correlation coefficient (CC) and odds ratios (OR) are considered most effective. CC and OR are one-sided metrics while IG and CHI are two-sided. In this work we use the IG which measures the difference between the entropy of the class labels and the conditional entropy of the class labels given a feature and the CC which shows the worth of an attribute by measuring the Pearson correlation between it and the class.

C. Classification Models

Ensemble techniques such as Random Forests have an advantage in that they alleviate the small sample size problem and related overfitting problems by incorporating multiple classification models [17]. Random Forests grow a voting committee of decision trees by selecting a random set of $logM + 1$ features where M refers to the total number of features. Therefore, random forests are particularly useful for high-dimensional datasets because increased classification accuracy can be achieved by generating multiple prediction models, each with a different feature subset [18] [19].

However it is known that the performance of Random Forests depends on the correlation between trees as well as the prediction strength of each individual tree. Therefore, we decided to combine Random Forests with a greedy correlation based sub-feature-group selection method [20] which prefers subsets of features that are highly correlated with the class while having low intercorrelation. This ensures that the trees which are grown are strong and uncorrelated.

[2]http://www.alchemyapi.com

[3]http://www.liwc.net/descriptiontable1.php

To assess the performance of different classification models we first conduct a 5-fold cross validation and subsequently conduct a separate evaluation on a hold-out test dataset which consists of a random sample of users. We use the area under the ROC curve (AUC) as an evaluation measure. One advantage of ROC graphs is that they enable comparing classifiers performance without regard to class distributions which makes them very useful when working with unbalanced datasets. To have a realistic setup, we did not artificially balance our dataset and randomly chose three times more negative samples than positive ones for each class.

IV. Experimental Evaluation

In this section, we will first discuss the training and test datasets that we collected for this study. Then, we will describe our classification results in detail followed by a discussion of the implications of this study.

A. Datasets

1) Wefollow Top-500 Dataset: In order to construct our classification models, we need an established "ground truth" or gold standard. For this, we leveraged the manually curated *Wefollow* web directories. When a user wishes to place themselves within the *Wefollow* system, they either send a tweet with a maximum of 5 labels they wish to be associated with or register themselves via the *Wefollow* web application. It is important to note that users choose their own labels thus reflecting their opinion of themselves. While this therefore means that the labels are not guaranteed to reflect the consensus opinion of the user, it does mean that more hidden or subtle labels [21] are recorded. Each *Wefollow* directory corresponds to one label and users within each directory are ranked in a crowdsourced manner.

At the end of July 2012, we crawled the 100 largest *Wefollow* directories for Twitter handles. We then placed those directories into two broad dimensions: profession and personality[4]. For each relevant *Wefollow* directory, we extracted a list of users registered in this directory and their corresponding rank. *Wefollow* was using a proprietary algorithm to rank users at the time the data was crawled. According to their website, the algorithm took into account how many users in a directory follow another user in this directory. In order to ensure equal users for each class, we chose the top 500 users and mapped them to a dataset which was crawled between Sep 2009 and Apr 2011. To ensure that reasonable data was obtained, we excluded all users who had published less than 50 tweets in this time period, and for users with more than 3,000 tweets in our dataset we randomly sampled 3,000 of their tweets. After cleaning, the data amounted to 3,710,494 tweets from 7,121 users over these categories. It should be noted that 92% of the users provided a short bio description in their profile that we also extracted.

2) User Lists: An alternative to the self-assigned tags of *Wefollow* are user lists, which are categorizations users make of others on Twitter, and are public to view. Thus, for each of the 7,121 users in our dataset we crawled their 1,000 most recent list memberships. In our sample, which is obviously biased towards active and popular users, 96% of users were assigned to at least one list, with the median number of lists per user being 75, and the mean was 232.

Though the majority of user lists correspond to topical labels (e.g., "computer science" or "healthcare"), user lists may also describe how people feel about the list members (e.g., "great people", "geeks", "interesting twitterers") and how they relate with them (e.g., "my family", "colleagues", "close friends") [22]. Also, since user lists are created by the crowd they may be noisy, sparse or inappropriate.

3) Random Test Dataset: In addition to the *Wefollow* Top-500 dataset, which is biased towards users with high *Wefollow* rank, we crawled another random sample of *Wefollow* users which were not part of our original dataset. This, in theory, provides a broader base of users to sample from when testing our models (increasing generalizability), although it must be noted that, since these users were not highly ranked on *Wefollow* there is obviously a question regarding the reliability of their self-tags. This sample was collected by tracking new registrations made in April 2013 (to one of the above listed *Wefollow* directories). From this collection, we selected 100 random users from each directory.

B. Results

We trained multiple binary random forest classifiers for each category with different feature groups using the greedy correlation-based feature-group selection method [20]. The following section reports our results from the personality and profession classification tasks using cross fold validation and a separate test dataset of random users.

1) **Personality-related Categories:**

a) **WeFollow Top-500 Dataset:** Figure 1(a) shows that for all personality-related categories the best performance can be achieved when using a combination of all features. This provides an AUC score consistently ≥ 0.9 for 6 categories out of 8.

The highest performing individual feature group is the social-semantic group. These features achieve the highest AUC values for most categories (*advertising*, *creative*, *ecological* and *informational*). A separate performance comparison of the three different feature types of the social-semantic feature group (TFIDF based on user list memberships, concepts extracted from user list memberships and concepts extracted from the tweets of a users' top friends) shows that TFIDF based on user lists performs best ($AUC > 0.8$ for all categories except *inspirational* and *innovational*). This suggests that information about user list memberships is indeed useful for predicting personality-related user attributes. Also Table I which reveals the top five features for each category ranked via

[4]Not all the directories fit neatly into those dimensions. Directories that did not fit were excluded. The Wefollow directories *musician*, *dj*, *songwriter*, *singer* were merged into the category *musician*, the *developer*, *webdeveloper* and *computers* directory were merged into the category *IT*, the directories *business* and *entrepreneur* were merged into the category *business* and finally the directories *advertising* and *marketing* were merged into the category *marketing*. Each other *Wefollow* directory maps to exactly one category.

their Pearson correlation with the category label, shows that TFIDF weighted list names tend to be amongst the top features for all categories. For example *informational* users tend to be members of lists like *newspapers*, *outlets*, *newspapers*, *breaking* and *reporters*, while *ecological* users tend to be in lists called *eco*, *environmental* or *sustainable*.

Social-semantic features are closely followed by linguistic style features which achieve the highest AUC values for the category *funny* and *inspirational*. Ranking only features of the social-semantic feature group shows that *funny* users tend to use more swear words ($CC = 0.47$), body related words ($CC = 0.37$), negative emotions ($CC = 0.35$) and talk about themselves ($CC = 0.35$). On the other hand *inspirational* users have a high usage of the word "you" ($CC = 0.24$) and talk about positive emotions ($CC = 0.22$) and affective processes ($CC = 0.2$) – i.e., they use words which describe affection such as cried or abandon.

For the category *religious* semantic features achieve a slightly higher AUC value than linguistic style and social-semantic features. Ranking the features of the semantic features group reveals that religious (or more specific christian) Twitter users tend to talk about their religion and therefore use words such as worship ($CC = 0.44$), bible ($CC = 0.41$), church ($CC = 0.41$), praying ($CC = 0.39$) or god ($CC = 0.38$). Further Table I shows that *religious* users tend to use words which fall into the LIWC category of religious words ($CC = 0.48$) and tend to be mentioned in lists called *christian* or *church*. This indicates also that social-semantic and linguistic style features contribute to identifying religious users and Figure 1(a) shows that indeed *religious* users can be identified best when using an optimal combination of all features.

Activity features which describe users via their activity patterns tend to perform worse than social-semantic, semantic and linguistic style features for most categories except *advertising* and *informational*. Ranking features of the activity feature group by their information gain and correlation coefficient shows that users who are actively advertising tend to have a significantly higher temporal balance ($CC = 0.35$ and $IG = 0.09$) and temporal variance ($CC = 0.24$ and $IG = 0.08$) which indicates that they publish frequently and the same amount of messages. Informational users tend to have higher informational variety ($CC = 0.38$ and $IG = 0.09$) and informational balance ($CC = 0.2$ and $IG = 0.08$) which indicates that they tend to share a large variety of links but share each link only once – i.e. they provide more information.

Overall, the most difficult task was to classify *creative* users. One can see from Table I and Figure 1 that the best features for this category are social semantic features while all other feature groups perform pretty poor.

b) **Random Test Dataset:** Figure 1(b) shows that for the random test users, the overall accuracy is slightly lower than for the top-500 dataset. This can be attributed to the fact that these users are not the top-ranked users, and therefore there is question of reliability on their self-categorization. However, we observe that the AUC scores are still reasonably good over most of the categories (≥ 0.8) except for the category *creative*. The results on test users shows that social-semantic features and linguistic style features tend to be most useful for classifying random test users into personality categories. Again, activity features are only useful for the category *advertising* and *informational*. Interestingly, semantic features do not generalize well to random test users and are almost as useless as activity features. One possible explanation is that random test users may be less active on Twitter and may reveal less personal information when tweeting. Another possible explanation is that there might be a vocabulary mismatch between the train and test users which might become bigger if we reduce the feature space during training.

2) **Professional Areas:**

a) **WeFollow Top-500 Dataset:** One can see from Figure 2(a) that again using a optimal subset of all features provides excellent classification performance with AUC values ≥ 0.9 for all categories except *business*. The most useful feature groups for classifying users into professional areas are linguistic style and semantic features. It is interesting to note that social-semantic features which were most useful for identifying users' personality related attributes, are not as useful for identifying their professional areas. Particularly for identifying users interested in *business* or *health* and for identifying *politicians* and *writers* social-semantic features are pretty useless ($AUC < 0.6$).

Table II shows that indeed the features with the highest information gain tend to be semantic and linguistic style features. In the semantic feature group especially topical features and TFIDF weighted words were most useful which indicates that users working in different professional areas talk indeed about topics related to this area. For example, *photographers* tend to talk about photography and art and design in general, while *politicians* tend to talk about Obama, the republican party and health care.

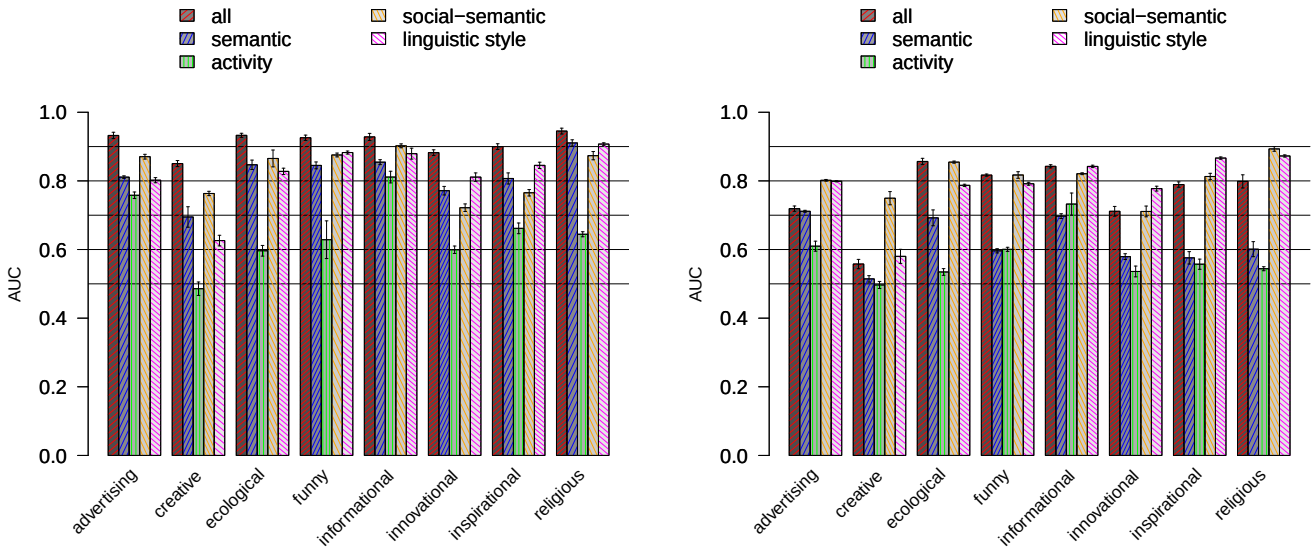
When comparing different types of semantic features we found for both tasks that concepts were pretty useless. One possible justification for this could be that the concept annotations tend to be too general. On the other hand, TFIDF weighted words work very well overall and TFIDF outperforms LDA on the random hold out dataset. This finding is in line with previous research [4] which shows as well that TF-IDF features perform almost twice as good as topic features within a user classification task.

In the linguistic style feature group LIWC features were most useful. One can see from Table II that for example *photographers* tend to focus on the perceptual process of seeing (i.e., they use words like *seeing* or *viewing*) while *musicians* focus on the perceptual process of hearing (i.e., they use words like *hearing* or *listening*). Not surprising, people working in the *health* sector tend to use words related with biological processes such as health-, body and ingestion-related words.

Finally, our results suggest that for some professional areas such as the movie industry, social activity features add value, since users who interact with key-players in their domain (such

TABLE I: Top features ranked by their Pearson correlation coefficient with each category. Topics are represented via their three most likely words. Feature Group: social-semantic, semantic, linguistic-style

advertising	creative	ecological	funny	informational	innovational	inspirational	religious
tfidf_list: advertising (0.51)	tfidf_list: twibes-creative (0.3)	green, energy, climate (0.58)	liwc: swear (0.47)	tfidf_list: newspapers (0.59)	tfidf_list: twibes-innovation (0.46)	love, I, we (0.47)	tfidf_list: christian(0.53)
tfidf_list: ad (0.44)	tfidf_list: com/creative/ twitter-list (0.3)	tfidf_list: eco (0.51)	was, he, is (0.47)	tfidf_list: outlets	tfidf_list: com/innovation/ twitter-list (0.46)	tfidf_list: twibes-inspiration (0.45)	liwc: relig (0.48)
tfidf_list:marketers (0.44)	tfidf_list: twibes-creative (0.23)	tfidf_list: environmental (0.5)	shit, fuck, ass (0.44)	tfidf_list: newsnews (0.58)	business, research, model (0.41)	tfidf_list: com/inspiration/ twitter-list (0.45)	tfidf_list:church (0.47)
tfidf_list: marketing (0.42)	tfidf_list: outofbox (0.22)	tfidf_list: sustainable (0.5)	I, me, you(0.42)	tfidf_list: breaking (0.56)	tfidf_tweetbio: innovation (0.37)	you, is, it (0.44)	god, lord, jesus (0.47)
tfidf_tweetbio: marketing (0.4)	tfidf_list: ly/oaomgp (0.22)	tfidf_list: environment (0.5)	liwc: body (0.37)	tfidf_list: reporters (0.55)	tfidf_list: twibes-innovation (0.32)	tfidf_list: spirituality (0.28)	tfidf_list: christians (0.45)



(a) *Wefollow* Top-500 Dataset

(b) Random Test Dataset

Fig. 1: Binary classifiers for different personality related user attributes.

as the film critic writer Scott Weinberg and Peter Sciretta) are more likely to work in this industry.

b) **Random Test Dataset:** Once again the test dataset reduces accuracy but the values are still reasonably good considering the unreliability of the ground truth in this case (see Figure 2(b)).

Our results from the random test dataset show that linguistic style features work best, which means that they generalize very well compared to other feature groups. An exception of this general pattern is the category *photographer*. For this category linguistic style features are not very useful and social-semantic features work best.

Again we observe that semantic features perform well within the cross fold validation but do not generalize well to the random test users. Overall, *writers* were most difficult to classify and the best performance could be achieved for the category *health-related professions* and *musicians* ($AUC > 0.9$) by using linguistic style features.

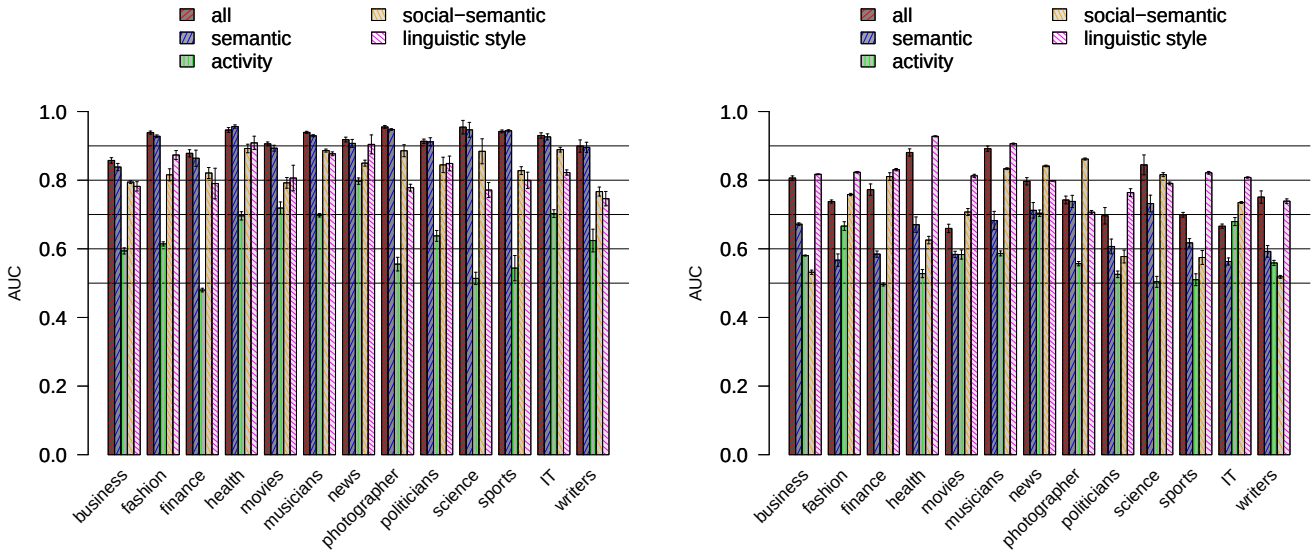
C. Discussion of Results

Our results show that random forests built on an optimal subset of the features demonstrate an impressive accuracy of above 0.8 (random test users) and 0.9 (top-500 *Wefollow* users), for most categories. For both tasks (personality and professional area classification) our results suggest that linguistic style features are most useful since they tend to generalize well on random test users. Further, the feature analysis reveals that LIWC based personal concern, linguistic and psychological process features, as well as our Twitter specific style and readability features, are useful for identifying users' professional areas and personality related attributes. This suggests that *not only what a user says but also how he expresses himself on Twitter* may reveal useful information about his professional areas and personality related attributes.

Further, we found that social-semantic features are very useful for predicting personality related attributes but less useful for predicting professional areas (especially for business, health, politics and writer, where the $AUC < 0.6$ when trained with social-semantic features). Since the best social-semantic

TABLE II: Top features ranked by their Information gain for each professional area. Topics are represented via their three most likely words. The cell color indicates to which group (semantic, linguistic-style, activity) a feature belongs.

business	fashion	finance	health	movies	music	news	photogr.	politician	science	sports	IT	writers
startup, startups, entrepreneurs (0.1)	dress, ebay, date (0.12)	dollar, #forex, u.s. (0.08)	liwc: health (0.1)	movie, trailer, review (0.11)	music, new, album (0.21)	u.s., obama, news (0.1)	photo, photography, photos (0.22)	obama, gop, palin (0.14)	science, research, data (0.12)	game, team, win (0.1)	code, web, project (0.24)	book, books, writing (0.08)
great, what, looking (0.09)	fashion, internship, intern (0.08)	liwc: money (0.06)	health, may, healthy (0.09)	video, movie, film (0.07)	liwc: hear (0.14)	death, two, men (0.09)	art, design, work (0.03)	obama, health, care (0.12)	space, nasa, science (0.06)	jets, jacksonville, nfl (0.06)	iphone, ipad, app (0.15)	book, ever, idea (0.05)
social, facebook, app (0.08)	so, love, oh (0.05)	$$, long, short (0.05)	liwc: bio (0.08)	mention slashfilm (0.03)	mix, remix, dj (0.12)	new, more, has (0.07)	liwc: see (0.03)	u.s., obama, news(0.07)	green, energy, climate (0.05)	bulls, lakers, nba (0.05)	new, just, site (0.11)	not, this, why (0.04)
twitter, social, media (0.06)	free, win, sale (0.05)	today, nice, $aapl (0.03)	health, patients, medical (0.07)	RT slashfilm (0.03)	liwc: work (0.1)	liwc: I (0.07)	new, more, has (0.02)	#tcot, #teaparty, #gop (0.06)	book, ever, idea (0.03)	yankees, baseball, mets (0.04)	google, twitter, apple (0.11)	also, actually, thing (0.03)
business, research, model (0.05)	show, girl, says (0.04)	business, apple, stock (0.03)	media, health, today (0.02)	RT scottEweinberg (0.03)	got, is, me (0.1)	liwc: assent (0.06)	rain, weather, snow (0.02)	new, more, has (0.05)	health, patients, medical (0.03)	world, cup, 2010 (0.03)	john, david, review (0.06)	#amwriting, las vegas, writing (0.03)



(a) *Wefollow* Top-500 Dataset

(b) Random Test Dataset

Fig. 2: Binary classifiers for different professional areas.

feature are TFIDF weighted list names, we can conclude that users' list memberships may indeed reveal information about users' personality related attributes (at least for those which were explored within this work). However, for professional areas the utility of social-semantic features may depend on the professional area.

Interestingly, semantic features perform well within the cross fold validation but do not generalize well to the random test users. One possible explanation is that there might be a vocabulary mismatch between the train and test users which is likely to become bigger if we reduce the feature space during training. Another potential explanation is that test users tend to be less active on Twitter and may reveal less personal information when tweeting. When comparing different types of semantic features we found that concepts did not provide much value. One possible justification could be that concept annotations tend to be too general. However, TFIDF weighted words work very well overall and TFIDF outperforms LDA on the random hold out dataset. This finding is in line with previous research [4] which shows as well that TFIDF features perform almost twice as good as topic features within a user classification task.

Consistently, we found that activity features are rather useless for most categories except those where users show very

specific activity patterns (e.g., the category of informational users who tend to post much more links than others). This finding is inline with previous research [2] [3] which found that user-centric content features are more useful than features which capture structural relations and similarities between users.

One finding that was not inline with existing work was the utility of self-referring possessives (i.e., my followed by any word). Unlike [2], performance did not improve when self-referring possessives were added. It is important to note that the classification dimensions described in [2] are very different from those which we use in our work. For example, it is intuitive that self-referring possessives are useful for predicting the gender of a user since a user who talks e.g. about his wife (i.e., uses the bigram "my wife") is almost certainly male. For professional areas and personality related attributes we could not find self-referring possessives with similar predictive power.

One limitation of our work is that both datasets used consist of users who registered themselves at *Wefollow* and those users may not be representative for the Twitter population as a whole. Thus, as future work, we propose an in-depth investigation into the relationship and model performance between those users that explicitly promote themselves via services like *Wefollow* and those that do not use such services

V. Conclusions

In this work we have constructed a comprehensive collection of features (around 20k features) and examined their efficacy in classifying Twitter users according to two broad different dimensions: professions and personality. We showed that the large set of features can be pruned to around 100 features per category using a greedy correlation-based subset feature selection. Further, random forests built on the selected subset of features obtained an impressive accuracy of ≥ 0.9 for most categories using our top-500 *Wefollow* dataset, and an accuracy of around ≥ 0.8 for most categories using our random test user dataset. Based on the varying utility of the features across categories, we believe that in order to create new classifications, a large initial set of features is required that can be pruned based on the characteristics of each category. This ensures that the idiosyncrasies of different categories are captured well by the features. Overall, we observed in both tasks that using only linguistic style features lead to consistently good results.

References

[1] R. Holt. (2013) Half a billion people sign up for twitter. [Online]. Available: http://www.telegraph.co.uk/technology/9837525/Half-a-billion-people-sign-up-for-Twitter.html

[2] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 37–44. [Online]. Available: http://doi.acm.org/10.1145/1871985.1871993

[3] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks afficionados: user classification in twitter," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 430–438. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020477

[4] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the IGKDD Workshop on Social Media Analytics (SOMA),*, 2010.

[5] J. Bryden, S. Funk, and V. A. A. Jansen, "Word usage mirrors community structure in the online social network twitter," *EPJ Data Science*, vol. 2, no. 1, pp. 3+, 2013. [Online]. Available: http://dx.doi.org/10.1140/epjds15

[6] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *SocialCom*. IEEE, 2011, pp. 149–156. [Online]. Available: http://dblp.uni-trier.de/db/conf/socialcom/socialcom2011.html#GolbeckRET11

[7] D. S. Daniele Quercia, Michal Kosinski and J. Crowcroft, "Our twitter profiles, our selves: predicting personality with twitter," in *SocialCom*, 2011.

[8] D. J. Hughes, M. Rowe, M. Batey, and A. Lee, "A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage," *Comput. Hum. Behav.*, vol. 28, no. 2, pp. 561–569, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2011.11.001

[9] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment." in *SODA*, H. J. Karloff, Ed. ACM/SIAM, 1998, pp. 668–677. [Online]. Available: http://dblp.uni-trier.de/db/conf/soda/soda98.html#Kleinberg98

[10] A. Stirling, "A general framework for analysing diversity in science, technology and society." *Journal of the Royal Society*, vol. 4, no. 15, pp. 707–19, Aug 2007. [Online]. Available: http://rsif.royalsocietypublishing.org/cgi/content/abstract/4/15/707

[11] D. M. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[12] H. M. Wallach, "Structured topic models for language," Ph.D. dissertation, University of Cambridge, 2008.

[13] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter." Boston, UA: Springer, 2012, pp. 508–524. [Online]. Available: http://ceur-ws.org/Vol-838/paper_01.pdf

[14] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 211–220. [Online]. Available: http://doi.acm.org/10.1145/1518701.1518736

[15] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," 2010. [Online]. Available: http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html

[16] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 42–47. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002736.2002747

[17] T. G. Dietterich and D. Fisher, "An experimental comparison of three methods for constructing ensembles of decision trees," in *Bagging, boosting, and randomization. Machine Learning*, 2000, pp. 139–157.

[18] B. B. Z. Pengyi Yang, Yee Hwa Yang and A. Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, pp. 296–308, 2010.

[19] L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.

[20] M. A. Hall, "Correlation-based feature selection for machine learning," Tech. Rep., 1998.

[21] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," 2013. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1218772110

[22] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: Modeling topical expertise of twitter users," in *Proceedings ASE/IEEE International Conference on Social Computing (SocialCom2012)*, 2012.