A Comparative Study of Demographic Attribute Inference in Twitter

Xin Chen
Emory University
xin.chen2015@gmail.com

Yu Wang
Emory University
yu.wang@emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Fusheng Wang
Stony Brook University
fusheng.wang@stonybrook.edu

Abstract

Social media platforms have become a major gateway to receive and analyze public opinions. Understanding users can provide invaluable context information of their social media posts and significantly improve traditional opinion analysis models. Demographic attributes, such as ethnicity, gender, age, among others, have been extensively applied to characterize social media users. While studies have shown that user groups formed by demographic attributes can have coherent opinions towards political issues, these attributes are often not explicitly coded by users through their profiles. Previous work has demonstrated the effectiveness of different user signals such as users' posts and names in determining demographic attributes. Yet, these efforts mostly evaluate linguistic signals from users' posts and train models from artificially balanced datasets. In this paper, we propose a comprehensive list of user signals: self-descriptions and posts aggregated from users' friends and followers, users' profile images, and users' names. We provide a comparative study of these signals side-by-side in the tasks on inferring three major demographic attributes, namely ethnicity, gender, and age. We utilize a realistic unbalanced datasets that share similar demographic makeups in Twitter for training models and evaluation experiments. Our experiments indicate that self-descriptions provide the strongest signal for ethnicity and age inference and clearly improve the overall performance when combined with tweets. Profile images for gender inference have the highest precision score with overall score close to the best result in our setting. This suggests that signals in self-descriptions and profile images have potentials to facilitate demographic attribute inferences in Twitter, and are promising for future investigation.

Introduction

Users' demographic attributes, such as ethnicity, gender, age, and education level, can provide invaluable information about users' characteristics and even their life experience. Thus, social science research often relies on demographic attributes to characterize and group users or survey participants. In Twitter, however, these attributes are often not explicitly coded by users through their profiles or metadata on social media web sites.

Previous work has investigated the efficacy of different user informations, for example, user names, tweets, friends, on inferring various demographic attributes (Pennacchiotti and Popescu 2011; Al Zamal, Liu, and Ruths 2012; Nguyen et al. 2013; Liu and Ruths 2013). These studies and experiments are conducted on different datasets and in different tasks in terms of which demographic attributes are inferred, making it difficult to have a comprehensive view of the state-of-the-art performance on demographic attribute inference.

We aim to extend and improve previous work on demographic inference from two perspectives: (1) we evaluate two important user signals, namely **profile self-descriptions** and **profile images**, which are not fully explored in prior work; (2) we compare 5 different user signals on inferring 3 types of demographic attributes (ethnicity, gender, age) side-by-side. The user signals we considered are users' names, self-descriptions, tweets, social networks, and profile images. Intuitively, these signals have very distinct nature and therefore require different models to meaningfully represent the user. For example, user names are very informative once jointed with Census data; profile images need image processing and computer vision techniques to extract features from. Thus, we carefully craft different models for each type of the user signals, and then evaluate their efficacy both individually and as feature groups.

We evaluate all features and feature groups on more than 2,000 Twitter users with human annotated demographic attributes. Results show that self-descriptions provide the strongest signal for inferring ethnicity and clearly improve the performance when combined with tweets. For gender, while tweets' n-gram feature has the best performance, the profile image achieves the best precision score and its overall performance is close to the best result in our setting. According to a crowdsourcing experiment (Nguyen et al. 2014), 10% of the Twitter users do not employ language that the crowd associates with their biological gender. While gender inference with linguistic features may be less effective for such users, profile images can be a good complementary signal. For both gender and ethnicity inference, users' names have shown a significant improvement for all evaluation metrics, especially for Asian and Hispanic users. For age, the Fleiss' kappa score of annotation is 0.239, implying that even human annotators find the task difficult. Nevertheless, all the best performing age inference systems have features from self-descriptions.

In summary, while linguistic features from tweets are more general and portable across different inference tasks, our work has shown the potential of self-descriptions and profile images for complementing or boosting the demographic attribute inference.

Data Collection and Human Annotation

To evaluate demographic inference with a more realistic datasets that share similar demographic makeups of tweet streaming, we collect datasets from Twitter streaming API which return a small random sample of all public statuses. As profile image is used as features and only English language is considered in this work, datasets are filtered to English speaking users with a profile image. As a result, 2,266 users, who use English in their posts and have profile images, are sampled.

In this work, ground-truth labels are collected through Amazon Mechanical Turk (AMT) that has been growingly used for obtaining annotations from human beings. Each of these Twitter users gets annotated by three different annotators based on their names, profile images, self-descriptions, and one sample tweet. User's gender is labeled as male, female, or *can't tell*. For ethnicity and age, we cast a multiple classification task rather than binary classification task which is a common practice in prior work. Table 1 shows the statistics of human annotations and corresponding Fleiss' kappa scores. Only the labels that receive more than 2 agreements out of the three human annotations are reserved as valid datasets. Table 2 shows the detailed composition for each demographic attribute.

Attribute		**Ethnicity**	**Gender**	**Age**
Agreement #	2	989	710	1340
	3	1020	1408	560
Total		2009	2118	1900
Fleiss' kappa		0.450	0.587	0.239

Table 1: Annotation statistics.

Ethnicity		**Gender**		**Age**	
Caucasian	967	**Male**	907	**Teenager**	427
Hispanic	154	**Female**	899	**Young Adult**	932
Black	316	-	-	**Adult**	172
Asian	58	-	-	-	-
Total	1495	**Total**	1806	**Total**	1501

Table 2: Statistics for labels with more than 2 annotations.

Framework for Demographic Inference

One common approach for demographic inference in social media is extracting a variety of features and feeding them into a supervised learning model. During the machine learning exercise, we can see which type of features is the most useful one for any particular classification exercise. This paper examines multiple types of signals: users' tweet documents and self-description documents aggregated from their neighors, profile images and user names.

Feature Extraction

Nbr-Tweet, Nbr-Des The interaction behaviors between Twitter users mainly include following and friending in addition to other more implicit ways like favoriting, retweeting and mentioning. A social circle is formed from such interaction behaviors. In this paper, we refer the social circle as a neighborhood. Users' neighborhood characterizes different aspects of users themselves. Results from prior work (Al Zamal, Liu, and Ruths 2012) have shown that inferences using only the features from a user's neighbors outperform those based on the user's features alone. In this paper, we do not use users' own posts or self-descriptions but only those generated from their neighbors. Such setup suffices our comparative study and will simulate the realworld inference scenario, for example when a user has a private or limited profile and only self-descriptions of the neighbors are available.

For each user in our datasets, the tweet document comprises a collection of tweets aggregated from neighbors including followers and friends with up to 200 tweets per neighbor. In a similar way, each user's self-description document comprises a collection of self-descriptions from neighbors including followers and friends.

To infer demographic attributes from users' contents and self-descriptions, the intuition is capturing their lexical usage and topics of interests. We enable this through two types of linguistic features, i.e., n-grams and hidden topic distributions derived from a Latent Dirichlet Allocation (LDA) model.

N-gram-Nbr-Tweet, N-gram-Nbr-Des The n-gram features include 10,000 most frequently appearing unigrams and digrams from each user's tweet document or self-description document in the training datasets.

LDA-Nbr-Tweet, LDA-Nbr-Des For the LDA feature, the assumption is that a user can be represented as a multinomial distribution over topics. The LDA model is trained over the collection of tweets or self-descriptions in training datasets and results in a number of hidden topics. Here we set the number of hidden topics as 100 and train the hidden topic model with 1,000 iterations to align with experimental settings in prior work (Pennacchiotti and Popescu 2011). For each user in the datasets, the trained LDA model and 100 generated topics are then applied to the tweet or self-description document and obtain a topic distribution as this user's LDA features.

Profile Image Profile images are quite common among all social media platform. In this study, we explore features extracted from profile images for demographic inference. The image features are based on the popular scale invariant feature transformation (SIFT) (Lowe 2004). A cookbook or dictionary of visual words is learned with k-means clustering over SIFT features. The SIFT descriptors from each profile image are then quantized with the visual word dictionary. The final image features for each user are the SIFT descriptor histogram from the quantization process.

Name Heuristic In our framework for demographic inference, we also enable a label updating module with name heuristic. We collect Census data with the frequency of gender categories by first name and racial categories by last name. If the tokens extracted from a user's profile name are matched to names in Census data, the gender or ethnicity category that most frequently occurs in the Census data would be assigned to this user. Table 3 shows name statistics and accuracy of name heuristic for ethnicity and gender inferences.

Demographic type	Ethnicity	Gender	Age
Rate of name identified	26.52%	55.07%	NA
Accuracy of name heuristic	73.14%	88.86%	NA

Table 3: Name Statistics and Accuracy of Name Heuristic.

Machine Learning Model

Prior work has explored multiple types of machine learining models for demographic inference that include support vector machines (SVM), gradient boosted decision trees, logistic regression, or customized models such as county regression (Mohammady and Culotta 2014). In this work, we focus on comparing the relative importance of different feature types for demographic inference, particularly n-grams and LDA features of tweets and self-descriptions, and profile image features. We use SVM with different types of features fed into the model and the model will assign higher weights to more effective features during the classification exercise. For gender and ethnicity inference, users' inferred labels from SVM models also get updated with name heuristic if their names are matched to those from Census data.

Experimental Results

In our experimental setting, 25% of the total datasets are reserved for testing. The other 75% of datasets are used for training LDA topics, bag of visual words and SVM models. During the evaluation phase, five evaluation metrics are considered that include precision, recall, F1, accuracy and area under ROC curve (AUC).

Ethnicity Inference Table 4 shows the ethnicity inference results with different feature combinations. Self-descriptions provide the strongest signal for ethnicity inference and clearly improve the overall performance when combined with tweets. Name heuristic has shown a significant improvement for all evaluation metrics. For Asian and Hispanic users, neither linguistic nor image features appear to take effect for inferring their ethnic groups. The small portion of successful calls in testing datasets (13 out of 32 Hispanic users; 2 out of 11 Asian users) is all coming from the step of label updating with name heuristic.

Gender Inference For gender inference, table 5 shows the gender inference results with different feature combination. While tweets n-gram features have the best performance, the profile image features get the highest precision score and its overall performance is close to the best result. According to a crowdsourcing experiment (Nguyen et al. 2014), 10% of the Twitter users do not employ language that the crowd associates with their biological gender. While gender inference with linguistic features may be less effective for such users, profile images may provide a good complementary signal. Name heuristic has also shown a significant improvement for all evaluation metrics.

Age Inference Table 6 shows the age inference results with different feature combination. The LDA features from users' self-descriptions gain better performance than other features. While users in older age groups often exhibit less linguistic differences (in their tweets) with younger groups due to societal pressure in the workplace (Nguyen et al. 2014), users' self-descriptions from different age groups may contain certain distinguishable signals.

Discussion

In this work, we examine a comprehensive list of user signals, namely tweets and self-description from users' neighbors, users' profile images, and users' names, on inferring ethnicity, gender, and age in Twitter. A realistic unbalanced datasets are collected and utilized for such side-by-side comparison. The results have confirmed previous finding that linguistic features of tweets provide most robust performance across different classification tasks. However, our experiments also exhibit the value of self-descriptions and profile images for certain demographic attributes. For ethnicity, self-descriptions provide the strongest signal and improve the overall performance when combined with tweets. For age, LDA features from self-descriptions achieve better performance than all other features. Among the first to apply profile image features in inferring user demographics, we extract image features with SIFT and bag-of-visual-words model. For gender inference, such features have achieved the highest precision score with overall scores close to the best result among all feature types. In our future work, we will explore more sophisticate and customized image features with more careful parameter settings. Using more advance image learning models, such as multi-view learning and deep learning, is another research direction for profiling social user with images. Users' names have shown a significant improvement for all evaluation metrics in gender and ethnicity inference. As not all Twitter users provide valid names in their profiles, the usage of name heuristic based methods will be limited. In future work, we plan to combine the signal of profile names from users' neighbors.

This study feeds supervised learning models with features that are derived exclusively from training datasets rather than from a general set of users. It would be interesting to use cross-validation to further examine the domain-specific features from different combinations of training datasets or to compare them with general features derived from larger datasets. There are also several issues as regard to the training datasets and human annotations. To deal with unbalanced datasets for training models, possible solutions include undersampling for the datasets or enabling skew insensitive measures. For those Twitter accounts representing non-humans or with no agreement among human annotators, a hierarchical or separate classification task is worth experimenting. While

Feature Configuration	Precision	Recall	F1	Accuracy	AUC
N-gram-Nbr-Tweet	0.687 (0.735)	0.747 (0.753)	0.707 (0.739)	0.747 (0.753)	0.643 (0.692)
N-gram-Nbr-Des	0.725 (0.770)	0.774 (0.784)	0.740 (0.772)	0.774 (0.784)	0.668 (0.715)
N-gram-Nbr-Tweet + N-gram-Nbr-Des	0.677 (0.731)	0.747 (0.753)	0.707 (0.738)	0.747 (0.753)	0.641 (0.689)
LDA-Nbr-Tweet	0.685 (0.710)	0.675 (0.688)	0.678 (0.696)	0.675 (0.688)	0.641 (0.674)
LDA-Nbr-Des	0.695 (0.718)	0.565 (0.606)	0.608 (0.645)	0.565 (0.606)	0.630 (0.664)
LDA-Nbr-Tweet + LDA-Nbr-Des	0.691 (0.723)	0.671 (0.705)	0.677 (0.711)	0.671 (0.705)	0.654 (0.692)
N-gram-Nbr-Des + LDA-Nbr-Des	**0.740 (0.778)**	**0.788 (0.791)**	**0.753 (0.778)**	**0.788 (0.791)**	**0.733 (0.719)**
Profile Image	0.540 (0.603)	0.459 (0.555)	0.481 (0.572)	0.459 (0.555)	0.552 (0.609)
All features combined	0.677 (0.731)	0.747 (0.753)	0.707 (0.738)	0.747 (0.753)	0.641 (0.689)

Table 4: Ethnicity inference results with different feature combinations (brackets contain results combining name heuristic.)

Feature Configuration	Precision	Recall	F1	Accuracy	AUC
N-gram-Nbr-Tweet	0.831 (0.866)	0.835 (0.89)	0.833 (**0.878**)	0.831 (**0.875**)	0.831 (**0.875**)
N-gram-Nbr-Des	0.741 (0.799)	0.676 (0.764)	0.707 (0.781)	0.717 (0.783)	0.717 (0.784)
N-gram of Tweets + N-gram-Nbr-Des	0.851 (0.87)	0.819 (0.846)	**0.835** (0.858)	**0.836** (0.858)	**0.836** (0.858)
LDA-Nbr-Tweet	0.787 (0.836)	**0.852 (0.896)**	0.818 (0.865)	0.808 (0.858)	0.808 (0.858)
LDA-Nbr-Des	0.673 (0.776)	0.769 (0.835)	0.718 (0.804)	0.694 (0.794)	0.694 (0.794)
LDA-Nbr-Tweet + LDA-Nbr-Des	0.798 (0.834)	0.802 (0.857)	0.8 (.846)	0.797 (0.842)	0.797 (0.841)
N-gram-Nbr-Des + LDA-Nbr-Des	0.738 (0.802)	0.665 (0.758)	0.699 (0.780)	0.711 (0.783)	0.712 (0.784)
Profile image	**0.861 (0.879)**	0.714 (0.835)	0.781 (0.856)	0.797 (0.858)	0.798 (0.859)
All features combined	0.851 (0.87)	0.819 (0.846)	**0.835** (0.858)	**0.836** (0.858)	**0.836** (0.858)

Table 5: Gender inference results with different feature combinations (brackets contain results combining name heuristic.)

Feature Configuration	Precision	Recall	F1	Accuracy	AUC
N-gram-Nbr-Tweet	0.568	0.584	0.569	0.584	0.598
N-gram-Nbr-Des	0.559	0.571	0.561	0.571	0.615
N-gram-Nbr-Tweet + N-gram-Nbr-Des	0.589	0.595	0.588	0.595	0.629
LDA-Nbr-Tweet	0.557	0.547	0.551	0.547	0.622
LDA-Nbr-Des	**0.605**	0.584	0.588	0.584	**0.669**
LDA-Nbr-Tweet + LDA-Nbr-Des	0.561	0.561	0.561	0.561	0.621
N-gram-Nbr-Des + LDA-Nbr-Des	0.554	0.564	0.557	0.564	0.613
Profile Image	0.494	0.463	0.468	0.463	0.580
All features combined	0.595	**0.601**	**0.595**	**0.601**	0.634

Table 6: Age inference results with different feature combinations

the resulting demographic attributes can be used as the inputs for many analytical applications such as psychological, personality-based and behavior studies, how to avoid the issue of inferential circularity remains a challenge.

In conclusion, our work investigates different types of user signals, including text (tweets and self-descriptions) from their neighorhood context (friends and followers), users' profile images, and users' profile names, for inferring social users' ethnicity, gender, and age. The results of profile images and self-descriptions are promising. The side-by-side comparison also generates a more comprehensive picture for automatically inferring demographic attributes in Twitter, which provides useful references for applications in social science research.

Acknowledgments

This work is supported in part by the DARPA grant D11AP00269 and by the National Science Foundation under Grant Nos. IIS1218488, CCF0937094, CISE-RI 1302668, NSF IIS 1350885, and NSF ACI 1443054.

References

Al Zamal, F.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Liu, W., and Ruths, D. 2013. What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium: Analyzing Microtext*.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60(2):91–110.

Mohammady, E., and Culotta, A. 2014. Using county demographics to infer attributes of twitter users. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*.

Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. How old do you think i am? a study of language and age in twitter. In *ICWSM*.

Nguyen, D.; Trieschnigg, D.; Dogruöz, A. S.; Gravel, R.; Theune, M.; Meder, T.; and de Jong, F. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING*.

Pennacchiotti, M., and Popescu, A.-M. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*.