Inferring Demographic data of Marginalized Users in Twitter with Computer Vision APIs

Panos Kostakos  
Center for Ubiquitous Computing  
University of Oulu  
Oulu, Finland  
panos.kostakos@oulu.fi

Abhinay Pandya  
Center for Ubiquitous Computing  
University of Oulu  
Oulu, Finland  
abhinay.pandya@oulu.fi

Olga Kyriakouli  
Dpt. of Informatics and Telematics  
Harokopio University  
Athens, Greece  
itp16401@hua.gr

Mourad Oussalah  
Center for Ubiquitous Computing  
University of Oulu  
Oulu, Finland  
mourad.oussalah@oulu.fi

Abstract—Inferring demographic intelligence from unlabeled social media data is an actively growing area of research, challenged by low availability of ground truth annotated training corpora. High-accuracy approaches for labeling demographic traits of social media users employ various heuristics that do not scale up and often discount non-English texts and marginalized users. First, we present a framework for inferring the demographic attributes of Twitter users from their profile pictures (avatars) using the Microsoft Azure Face API. Second, we measure the inter-rater agreement between annotations made using our framework against two pre-labeled samples of Twitter users (N1=1163; N2=659) whose age labels were manually annotated. Our results indicate that the strength of the inter-rater agreement (Gwet’s AC1=0.89; 0.90) between the gold standard and our approach is ‘very good’ for labelling the age group of users. The paper provides a use case of Computer Vision for enabling the development of large cross-sectional labeled datasets, and further advances novel solutions in the field of demographic inference from short social media texts.

Keywords— Social media mining; Twitter; Privacy; Demographic data.

I. INTRODUCTION

The popularity of social media enables public interest organizations to engage in cost-effective awareness raising and policy assessing campaigns; however, lack of demographic data in microblogging platforms (owing to the privacy issues), makes targeted outreach increasingly difficult [3]. As a result, various demographic groups, such as elderly, remain marginalized [3].

Increasingly, researchers in Computer Science and related disciplines are developing machine learning models that predict demographic traits like age, sex, race, and ethnicity using textual and non-textual features provided by user’s public profiles in Twitter [5,6,7,8,9,10,11]. Demographic inference is mostly approached as a supervised learning problem where researchers use a training set of annotated users to fit a model/function that best predicts users’ social attributes. A major challenge is developing valid ground truth data that can scale up to include different age groups and languages. A low-key approach in this domain exploits the re-use of public profile images that significantly improves the prediction scope of derived models [12,13,14].

The aim of the paper is threefold. First, we present a framework that leverages on the use of public profile images and off-the-shelf Computer Vision software (Face API) for estimating demographic features of public twitter profiles. Second, we benchmark the accuracy of our framework against two pre-existing datasets built using gold standard annotations methods [1,5]. Third, we briefly show how demographic data can be used to enhance decision making for marginalized users.

The remaining of the paper is organized as follows. First, we review the existing work to highlight the methods used for developing ground truth. Second, we provide an overview of our architecture and the resources we used. Third, we compare our prototype against two independent datasets containing Twitter users. Finally, we provide summary statistics of the corpus built.

II. RELATED WORK

In this section we summarize prior research related to the problem of age recognition of users in social media, and identify the main limitations in using images for developing annotated corpora. In the subsequent section, we focus on the data collection methods and corpus annotation.

The predominant way of developing accurate datasets of Twitter users with demographic labels chiefly follows these techniques: i) use of regular expressions to identify phrases that show the gender and the age of users [1,8] ii) matching first names from public lists with the screen names provided by users [15] iii) annotating using crowdsourcing tools like Mechanical Turk [4] iv) manually labeling by human annotators by searching across social media and other public sources [4,5] v) use of face recognition tools [12,13,14] vi) stacked approaches combining one or more of the above [11].

There are various shortcomings, as high accurate methods for example reported in the literature that involve the gender-name association or other anglophone linguistic features fail to capture non-English users. While all aforementioned approaches report high confidence results, the machine vision approach seems to hold greater promise for scalability, and has not been given proper attention.

However, exploiting social media avatar images for gender and age identification is challenging since many users prefer to keep the default image provided by Twitter, or users’ profile images may be non-human (pets, logos, objects, children), may contain more than one faces, may be of bad image quality (size, resolution, blur), or even could be fake images taken from google image search results (e.g., image of celebrities). The proposed framework overcomes all these limitations by adopting rigorous filtering of profile avatars which is explained in more detailed next.
III. METHODS

In this section we present our framework for building the age and gender annotated corpus of Twitter users. Subsequently, we describe the methods used for validating the performance of Microsoft Face API in correctly classifying the age group of Twitter users.

A. Corpus Annotation with Face API

When setting up a new account on Twitter, users are provided with a default profile picture called avatar. Users can replace the default avatar by uploading their own picture. Avatar images are publicly accessible, crawled regularly by search engines, and can also be seen by users who do not have a twitter account. Our framework exploits facial objects found in avatars to build a large corpus of Twitter users annotated with key demographic variables, such as age and gender. Fig. 1 depicts the pipeline for parsing, filtering and annotating the avatars. The following inclusion/exclusion criteria were used to address the shortcomings reported in the literature when using images for demographic annotation of social media profiles:

1. Create a wordlist with the 100 most frequently used Dutch words, and use the Twitter Streaming API to track tweets in Dutch. We used Dutch users to illustrate our methodology but our framework is generic to all languages/regions.

2. Retrieve the URLs of the avatar images for each user, but exclude gifs and default avatars (Twitter default images).

3. Parse the resulting list of URLs with Python’s OpenCV to detect avatars with one or more faces.

4. In order to remove fake images and/or celebrities, we pass the list of images to Google Cloud Vision API and removed images that appear more than 4 times in google image results.

5. Input the filtered list of image URLs to Google Cloud Vision API to detect avatars with one face and good exposure.

6. Input these image URLs to Microsoft Face API and retrieve demographic data. Subsequently, retrieve at least 70 tweets from each profile and remove unused accounts.

Below we elaborate on the previous steps in more detail. First, we collected a sample of 9.5 million twitter posts in July 2017 by filtering tweets that contained the 100 most frequent Dutch words. This enabled us to identify 1,043,116 unique users who shared public tweets in Dutch and English. Following the aforementioned exclusion criteria, we removed fake images, celebrities, pictures with more than one face and low exposure. Subsequently, for the remaining unique avatars (N= 82,781), we collected their most recent tweets along with other public data. Finally, we split the corpus into two sub-corpora based on the language of the tweets.

The English Twitter corpus contains a total 166,182 tweets while the Dutch corpus holds 297,544 observations. In addition to the stock features supplied by the Twitter API, Microsoft Face API, and Google Cloud Vision API, we calculated a number of social network metrics and linguistic features increasing the total number of features to 67. These features are available for re-use in future machine learning tasks. Next, we focus on how the framework was validated through a series of inter-rater agreement tests.

B. Inter-Annotator Agreement

Fig. 2 depicts our approach for validating the developed framework that predicts the age of twitter users based exclusively on their profile image. We deployed our existing pipeline on a known list of Twitter users whose age labels were assigned manually from two independent annotators (Sample 1 and Sample 2). We obtain the avatar pictures from the profiles of these users and retrieved the age estimates. These estimates were then compared against the pre-assigned labels. An index of agreement between pre-assigned labels and those estimated with our framework is found by computing inter-rater agreement.

In order to quantify the degree of agreement between different annotators, many methods have been proposed in the...
discipline of statistics. Out of various choices available for finding this agreement, such as Kappa statistics (Cohen’s kappa, Fleiss’s kappa), Correlation coefficients (Pearson’s r, Spearman’s rho, etc.), and probability-based methods (cross-entropy, KL divergence, etc.), we argue that Gwet’s AC1 and AC2 have been found to be more resilient to chance agreements [17]. Also, Kappa statistics suffer from another problem: Kappa can be low despite a high level of agreement [16,17,18]. Contrariwise, Gwet’s AC between two or multiple raters is defined as the conditional probability that two randomly selected raters will agree, given that no agreement will occur by chance.

C. Datasets

To assess the validity of our annotation framework, we utilize two existing data samples that were manually labeled with demographic information. Table I summarizes the two pre-labeled datasets, and contains the observations of three raters: Rater 1 refers to data sample 1 [5]; Rater 2 to data sample 2[1]; Rater 3 refers to results obtained through our framework for both samples 1 and 2.

The data collected in [5] is based on their work in 2013 and we found that many of the Twitter profiles are not active at the time of writing this paper. Similarly, for dataset in [1], we could find only about half of the profiles active at present time. However, for those profiles which are still available on Twitter, we processed their profile pictures through our framework and obtained the age and gender estimates after correcting for the difference in the number of years between the original annotation and our estimate. Also, since Face API produce age estimates as floating point numbers, we rounded up these numbers to the nearest integers. Some datasets also require human annotators to assign a confidence score to their annotations (e.g., dataset in [5]). For such datasets, we used age annotations with a confidence value of 0.5 or higher. Fig. 3 shows the distribution of age in these two datasets [1,5].

![Fig. 3. Age distribution of users in our two data samples [1,5]. Left: dataset from [1], Right: dataset from [5]](image)

| TABLE I. DESCRIPTION OF THE TWO VALIDATION DATASETS |
|----------------|----------------|----------------|----------------|
| Rater 1 | 3,086 | x | x | H | English |
| Rater 3 | 1,381 | x | x | CV | |
| Rater 2 | 1,470 | x | x | H | |
| Rater 3 | 709 | x | x | CV | |

N: Number of unique users in the original dataset. Some users have either deleted their Twitter profile or have removed public access. This is why the number of users in each sample is not the same. L: Labels included in the dataset. Gen: Gender of user. Age: Age of the users as continuous variable. AB: dataset contains age in bins. Lang.: Main language of the users. M: Method used for annotation. CV: Machine Vision, H: Heuristics.

Table II shows the results of the inter-rater agreement test.

| TABLE II. RESULTS OF RATER AGREEMENT USING GWET’S METHOD. |
|----------------|----------------|----------------|----------------|
| RATERS | SIZE | CAT | WEIGHTS | ESTIMATE | STDERR |
| 2 | 3 | 659 | 3 | linear | 0.9930670 | 0.0025695 |
| | | | | unweighted | 0.9877497 | 0.0043372 |
| | | | | squared | 0.9877497 | 0.0043372 |
| | 1 | 1163 | 3 | linear | 0.9696025 | 0.0039669 |
| | | | | unweighted | 0.9430740 | 0.0071778 |
| | | | | squared | 0.9430740 | 0.0071778 |
| | 2 | 659 | 6 | linear | 0.904401 | 0.0072728 |
| | | | | unweighted | 0.670151 | 0.020173 |
| | | | | squared | 0.670151 | 0.020173 |
| | 1 | 1163 | 6 | linear | 0.8635149 | 0.0072728 |
| | | | | unweighted | 0.629216 | 0.016152 |
| | | | | squared | 0.629216 | 0.016152 |

Each of the 12 experiments show the weighting scheme used for Gwet’s AC measurement, standard error, and the index of the inter-rater agreement obtained. CAT refers to the number of categorical bins used to create age-brackets. Our choices were limited by the available datasets [1,5] since the dataset in [5] does not have exact age information but only as a categorical variable. As can be seen from our experimental results, we find Gac values in the range indicating “very good” inter-rater agreement with three age groups, and “good” agreement when age is observed in 10-year volume increment.

Our results indicate that inter-rater agreement is high when the number of categorical bins is small, which is expected since the confidence of the raters is high when the number of categories is small. Also, it is observed that using “linear” weighting scheme for Gwet’s AC computation gives us better results than other weighting schemes. In order to test the results of Face API with the ones obtained manually, we treated age as an ordinal categorical variable and assigned age value to ordinal bins. Therefore, since age is treated as ordinal categorical value, weighted kappa computation is more appropriate here.

Also, it is noted that the number (size) of bins impact the agreement results. Increasing number of bins increases annotator’s uncertainty and therefore kappa score decreases. In theory, the results of agreement will depend largely on the choice of the category limits for bin creation. However, since the human annotators have already decided upon the bin sizes, we kept the same choices for computation of agreement.

IV. Results

A. Inter-annotator Agreement Tests

Various experiments were conducted measuring the inter-rater agreement between labels assigned by Microsoft Face API and those by humans and considered age both as continues variables with 10-year increments, and as categorical variables (<55, 55-67, <67). We used R’s ‘rel’ package to find GAC (Gwet’s coefficient), and used the AC2 (weighted agreement), specifically, “quadratic” weighing [17, 19].
B. The Obtained Twitter Corpus

Using the method outlined in Fig 1, we obtained age- and gender annotations for 82,781 unique Twitter users which is, to the best of our information, the biggest corpus for Twitter demographics. Fig. 4 summarizes our corpus. Most of the tweets are posted by young users below 40 years of age, but we also have a fair number of tweets for the elderly.

Fig. 4. (Above) age and gender distribution of the corpus. (Below) geographical location distribution of users.

V. CONCLUSIONS

In this paper, we proposed a scalable method to build a massive corpus of Twitter users with demographic data annotations. As a proof of validity of our method, we compared the results of our methods to the two publicly available datasets which show very good agreement with our predictions/estimates. Indeed, our method with carefully filtering fake images and harnessing off-the-shelf computer vision APIs holds promise to address the problem of dearth of data available for social research.

ACKNOWLEDGMENT

This work is (partially) funded by the European Commission grant 770469-CUTLER and 645706-GRAGE

REFERENCES


[19] https://cran.r-project.org/web/packages/rel/rel.pdf