An Interactive Method for Inferring Demographic Attributes in Twitter

Valentina Beretta
University of Milano-Bicocca
DISCo
Viale Sarca 336
Milano, Italy
v.beretta3@campus.unimib.it

Daniele Maccagnola
University of Milano-Bicocca
DISCo
Viale Sarca 336
Milano, Italy
daniele.maccagnola@disco.unimib.it

Timothy Cribbin
Brunel University
Department of Computer Science
Kingston Lane
Uxbridge, United Kingdom
timothy.cribbin@brunel.ac.uk

Enza Messina
University of Milano-Bicocca
DISCo
Viale Sarca 336
Milano, Italy
messina@disco.unimib.it

ABSTRACT

Twitter data offers an unprecedented opportunity to study demographic differences in public opinion across a virtually unlimited range of subjects. Whilst demographic attributes are often implied within user data, they are not always easily identified using computational methods. In this paper, we present a semi-automatic solution that combines automatic classification methods with a user interface designed to enable rapid resolution of ambiguous cases. TweetClass employs a two-step, interactive process to support the determination of gender and age attributes. At each step, the user is presented with feedback on the confidence levels of the automated analysis and can choose to refine ambiguous cases by examining key profile and content data. We describe how a user-centered design approach was used to optimise the interface and present the results of an evaluation which suggests that TweetClass can be used to rapidly boost demographic sample sizes in situations where high accuracy is required.

1. INTRODUCTION

Social scientists, policy makers and marketers are keen to find ways to mine social media (SM) data in order to gain insights into public attitudes and opinion. Traditional survey research methods (questionnaires, interviews etc.) are becoming less attractive, due to falling response rates and increasing costs [9]. At the same time, members of potential target populations are increasingly sharing their views, for free, on SM platforms such as Twitter and Facebook. For this reason, mining SM is seen by many as a key part of the next generation in survey research methods [23]. There are several features that make SM based research attractive. that are particularly attractive. First large datasets can be collected relatively cheaply and are already digitally encoded. Second, SM users tend to comment in a responsive, ad hoc manner, allowing a more timely polling of opinion on current events, in comparison to 'designed' research. Third, despite being public forums, the perceived anonymity of SM platforms means that views expressed online may often be more honest and expressive than those collected using designed instruments [21].

A key barrier to the use of SM data is the absence of explicit and/or reliable demographic attribute data. Such metadata is essential in survey research to make comparisons between population groups. Without ready demographic data, researchers tend to resort to making subjective judgments by explaining the qualitative characteristics of user's posted content and virtual profile.

However, this method is very time consuming. On the other hand, automatic techniques can be used for deriving the demographic attributes, but in some cases (for instance in age identification task) their results are not always reliable [22, 26].

Given this problem, we propose a semi-automatic framework to facilitate and accelerate the human judgment process. The framework relies as much as possible on automatic techniques, essential for handling the huge amount of data that originates from SM, only requiring human intervention for cases that cannot be classified with high confidence by the algorithms. We incorporate this approach into a proof-of-concept tool, called TweetClass, designed to support researchers in the identification of demographic attributes of a Twitter user sample. In order to evaluate the capabilities of our tool, our experiments include an extensive analysis of the interface design. Moreover, even if our focus is not on the classification method, we investigated the best approach among few popular techniques for facilitating the refinement process for the end-user.

The rest of the paper is structured as follows. In Section 2 we present previous work related to demographic attribute inference and semi-automatic approaches to classification. In Section 3 we explain the rationale behind combining automatic and interactive methods and how we combine them. Section 4 describes how we collected our dataset, followed by Section 5 which describes the experiments carried out to find the best approach for automatically identifying age and gender class. Finally, in Section 6, we focus our attention on an essential part of the work, the interface design and the method employed for evaluating it. In particular, here the description of the first prototype is followed by the description of

HT'15, September 1–4, 2015, Guzelyurt, TRNC, Cyprus.

DOI: http://dx.doi.org/10.1145/2700171.2791031.

his evaluation that highlighted several problem addressed with the development of a second prototype interface.

2. RELATED WORKS

Previous research has explored the problem of automatically inferring key demographic variables such as gender, age, ethnicity, political orientation, occupation and regional origin [20, 10, 18]. Here, we focus our attention on gender and age, since they are fundamental attributes required in social research. Existing approaches have exploited different feature types that can be used for deriving age and gender. The three main approaches can be distinguished as: profile-based, content-based and hybrid.

Profile-based approaches use metadata associated with the user's account or profile. In Twitter, such features include real name, description, location, followers and friends. For instance, the simplest profile-based method assigns gender class based on a dictionary look-up of the user's first-name, see [13, 22]. An alternative approach is to infer a user's gender based on profile colour preference [1]. When it comes to age inference, profile-based features tend not to be used alone, but combined with content-based features.

Content-based methods exploit the language expressed in the text of users' posts. One of the earliest content-based approaches focused on gender and age class inference is [20]. Their method processes unigrams and bigram features using a Support Vector Machine (SVM) algorithm. Similarly, in [17], n-grams are used with a combination of Chi-square based feature selection and SVM. Other work has also used n-gram features in combination with logistic and linear regression models [14, 15].

In addition to n-grams, stylistic features have also been studied. For instance, several approaches describe methods to derive gender and/or age based on the usage of smileys, abbreviations, punctuation, possessive bigrams, repeated letters, pronouns, hashtags and other grammatical features[20, 6, 4].

In contrast, hybrid approaches leverage profile data for enhancing the accuracy of results obtained using content-based features. Notable examples of the hybrid approach include [26, 10, 11, 5, 3].

Comparing the efficacy of these and other methods is not straightforward because of the tendency to use different datasets for training and testing. Moreover, different studies tend to vary in the intervals used for age classes. Despite these problems, it is possible to draw some key conclusions:

- Regarding gender classification studies, it is evident that profile-based classification methods are faster than content-based ones, but the former achieve lower accuracy than the latter.
- Age inference tends to be the more challenging task, particularly with respect to older age groups, see [14]. This seems to be due to the fact that the way a person speaks is influenced by many factors, beside age: for instance, whilst adults tend to be more conservative in their language, factors such as their profession and culture can also impact on their content and style of expression.
- Generally, in studies where several machine learning techniques are compared, SVM performs better than other classification methods.
- Content-based methods have an high computational complexity due to the number of features generated from the text. This is particularly true for n-gram approaches.
- Gender and age classification are treated as independent tasks, although gender has been used as an additional binary feature of the age feature set ([16, 17]).

We hypothesised that features characterising age and gender might be co-dependent. Argamon et al. used factor analysis for identifying 20 coherent factors of words linking gender and age [2]. They show that male components of language increase with age, while female ones decrease. Therefore we decided to introduce a hierarchical approach, whereby the first step derives user's gender and in the second one derives user's age class conditioned to his gender class.

The results obtained with automatic methods, as described, tend to fall in the ranges of 70%-92% of accuracy for gender inference and 71%-88% for age class (levels of accuracy in excess 80% are only possible if the age classes are divided by a gap of several years). To become a credible alternative to designed survey research, it must be feasible and practicable to sample demographic groups to a much higher level of accuracy.To this end, we developed a semi-automatic approach in which a user interface presents the results of automatic classification and enables the user make refinements on the basis of additional profile and content information. This kind of method has been adopted successfully in other domains (e.g. [24, 27]).

3. APPROACH

TweetClass combines automatic classification with human interaction. The use of automatic methods is essential to manage the huge amount of data that originates from SM, however a reliable and accurate automatic classification may not be possible for all cases and human intervention may be required. Indeed in some cases determining a user's gender or age might be a simple task for a human, based on examination of a photograph or profile description, yet the same task is very difficult to reliably achieve using automatic methods. Moreover, for a given a Twitter user, humans are able to explore additional information. For instance, they can explore the user's digital footprint on the Web. If the name is not meaningful, they can see profile images from other platforms, explore SM relationships and so on. Just reading extracts from a user's timeline, can be sufficient to discover nuanced clues that might not be found by automatic methods amongst a much larger corpus of data.

Gender and age class inference is achieved through a process that is summarized in Figure 1. At each major step, the end-user is provided with the option to scrutinize and refine the results of automatic classification algorithms. The most critical part of the process regards the identification of age class demographic attribute. Our experimental data show that gender does influence the age class identification. As such, a two phase hierarchical procedure was used to build the classification model. Hence, gender is derived as first attribute to increase the classification accuracy of the user's age class.

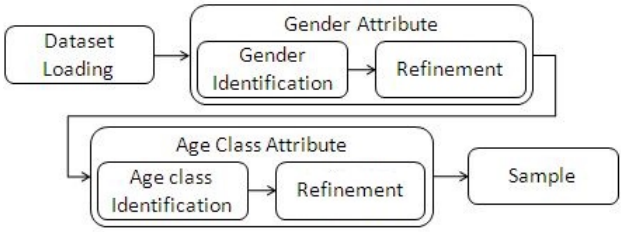

Figure 1: TweetClass process.

The gender of each Twitter user is determined using the user's first-name that appears in the profile. Identifying a person's gender

from their name is not always straightforward. For instance, users may use pseudonyms or transpose their surname and first-name. The latter case becomes more problematic if the user's surname is equivalent to a common first-name (e.g. Michael Stewart).

In our work, we decided to use a dictionary based approach, primarily because of its efficiency in terms of computational time. In particular, we used the 40N database [22] that contains a list a more than 44000 first names and related gender from 54 different states covering the vast majority of first names in all European Countries. Therefore, given a string u containing user's name, we have that $u \in \{F, mF, M, mM, U\}$, where F is female, mF is unisex but mostly female, M is male, mM is unisex but mostly male, and U is either unisex (with no prevalence of female or male) or not found. To increase the accuracy of this classification a refinement classification phase follows the automatic one in order to manually inspect the ambiguous classification $u \in \{mF, mM, U\}$.

Once the gender class is assigned, the process continues to the age inference step. During age inference phase the Twitter users are classified into two major demographic sets: users below 30 (younger) and users above 30 (older).

While this binary categorization may seem too simple, we must consider that age is a difficult attribute to learn. Not only does it change constantly, age-sensitive communication behavior differs based on numerous socioeconomic variables, and there is no well known indicator for age on Twitter [20].

We model the tweet contents, for each gender class, using a feature vector approach. Unigram features are selected after a preprocessing phase. Chi-Square feature selection is used to reduce the number of attributes and to take into account only the most predictive words. We derived the remaining stylistic features additionally using a POS-tagging procedure and partially using regular expressions, for instance the presence of stretched word (hellooo, SUNNY). Since an user can write more than one tweet, user's age class is identified taking into account a sample of their recent posts. Single tweets are then classified independently by using classification models such SVM or Naive Bayes. From this classification phase we obtain the label probability distribution and we assign to the tweet the label with the maximum likelihood. Once each single tweet of the user is classified the results are aggregated in order to obtain an overall age classification probability for each user. The probabilities for a user to belong to younger and older class are computed using the following formulas:

$$p\{u \in younger\} = \alpha \sum_{i=0}^{N} p\{tweet_i \in younger\}$$
$$p\{u \in older\} = \alpha \sum_{i=0}^{N} p\{tweet_i \in older\} \quad (1)$$

where α is a normalization factor and N is the number of tweets that belong to him/her. For each user u_j we define a confidence value $Conf_j$ given by the following formula:

$$Conf_j = max(p\{u_j \in younger\}, p\{u_j \in older\}). \quad (2)$$

The confidence level can be interpreted as the probability of how sure end-user can be regarding a certain classification. All user instances classified with a confidence level lower than a certain threshold are displayed to end-user.

4. DATASET

There is a lack of both gender and age labeled datasets in the public domain. Given this, we collected a new dataset using Twitter API. The absence of suitable datasets is a result of two key factors. First, to gather private information such as gender and age of a user is a resource intensive task. Second, issues relating to privacy and Twitter data user terms limit data diffusion. Indeed, datasets of Twitter Content or an API that returns Twitter Content can be downloaded only if they contain or return IDs (tweet IDs or user IDs).

To obtain our data collection we adopted a similar idea used in [26]. In order to identify age labeled users, they collected all tweets in which an individual announced his or her own birthday (e.g., "Happy ##th/st/nd/rd birthday to me").

As reported in Figure 2, we developed a crawler, based on the Streaming Search API provided by the Twitter site, able to filter only particular tweets from the stream.

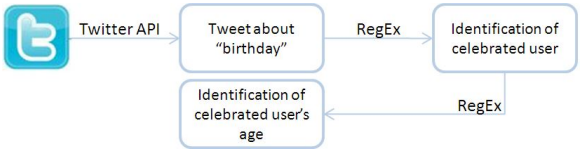

Figure 2: Dataset collection process.

We filtered all the tweets containing the word "birthday" and identifying the owner's birthday. At this point, we derived the age of each user using the regular expressions. We filtered all tweets including two consecutively digits that were not part of fractions, urls, usernames, hours, dates and three or more-digit numbers. In fact, with high probability, the remaining two digit numbers represent the age. We eliminated all users with follower number greater than 5000 because these users were likely to be celebrities or big companies that would not be representative of behaviour Twitter population.

In order to validate the dataset, a manual inspection was necessary and during this phase the gender labels were added.

Using this method we were able to collect a dataset of 386 users. It is composed of 62 younger male, 88 older male, 152 younger female and 84 older female. For each of these users we retrieved between 20 and 200 tweets from their timeline. In this way, we obtained a tweet dataset where 8368 tweets belong to younger male, 12868 to older male, 21288 to younger female and 12002 to older female authors. To create a balanced dataset for both gender and age class attributes, we randomly sampled the dataset retrieving for each gender-age class a number of tweets equal to 8368 (minimum number of tweet in gender-age class combination). Using the collected dataset we conducted the following experiments.

5. EXPERIMENTS

5.1 Gender identification

Since social scientists are interested in maximizing the accuracy of the assigned demographic attributes, the automatic gender classification should ideally either assign the correct gender if possible or leave the classification to the refinement phase. For this reason we decided to reduce from 5 to 3 the number of classes presented to the end-user, namely (female, male and unknown). During the experiments we used two configurations to identify these three classes:

1. $u^* \in \{F, M, U^*\}$ where $U^* = \{mF \cup mM \cup U\}$;
2. $u^+ \in \{F^+, M^+, U\}$ where $F^+ = \{F \cup mF\}$ and $M^+ = \{M \cup mM\}$;

Social scientists need a sample of users with specific demographic attribute values and it is very important that these values match the real ones. So, the first configuration seems to be more desirable than the second one. Indeed, in configuration 1 the users with names that are considered mostly male or mostly female are classified as "unknown". In this way, social scientists can be assured that if a user is classified as "male" or "female", he/she belongs to this class with high reliability. Nonetheless, we decided to analyse both the situations to obtain a much deeper investigation.

Moreover, we analysed performance changes considering either all words contained in the user name field or just the first word that appeared in it. Therefore we conducted experiments with the following resulting configurations:

- *method G1*, we derived gender of u^* using all user's name field;
- *method G2*, we derived gender of u^* using first word in user's name field;
- *method G3*, we derived gender of u^+ using first word in user's name field;

Note that, in all the experiments, we used the C library written by Micheal et al.[12] for exploiting the 40N database. The library is able to check first-names and determine their gender automatically.

Results

The results of these experiments are reported in Table 1. Note that we were interested in computing only the evaluation measures over the instances classified as "male" or "female". In fact, due to the huge amount of Twitter users, social scientists are likely to be less worried about loss of some users, if this results in a significant improvement in classification accuracy. So, the instances considered "unknown" were just discarded and we focused our attention only on male and female classified instances. We observed

	Class	Precision	Recall	F1	Acc
G1	F	97.7%	90.7%	94.07%	93%
	M	86.7%	96.6%	91.38%	
G2	F	98.4%	96.1%	97.2%	97%
	M	93.7%	97.4%	95.51%	
G3	F	97.4%	94.94%	96.15%	95%
	M	92.23%	95.96%	94.06%	

Table 1: Gender classification results

that the *method G2* achieves the best result. In this case, the error rate is only 3%. Instead, the worse performances were obtained with *method G1*. The difference between this method and the other ones was the part of user name field used to derive gender. As already explained, in *method G1* all name field was considered, while in *method G2* and in *method G3* only the first word of name field was used. Since typically Twitter users fill their name field writing first name followed by surname, taking into account the entire name field could be a problem. Indeed, several surnames are also first names. In *method G1*, this affects the gender classification task increasing the number of gender misclassified instances. For example, a possible name field could be "Rylee Ross" (first name + surname). In this situation, considering all the name field (*method G1*), user would be misclassified as male. Indeed, while "Rylee" is a mostly female name," Ross" can be also used as first name, and, in particular as a male name leading to an overall classification as male user.

Usually, in all the experiments, the number of incorrectly female classified instances was higher than the male one. Inspecting these misclassified instances we found that in several cases female users do not write their name, but acclaim related to some male celebrity, such as "I love you Ashton". After all these observations, since we want to assure to social scientist the best reliability, we decided to use the *method G2* to infer gender attribute of each user in TweetClass tool.

5.2 Age class identification

The feature set that represents each tweet was composed of 108 attributes where 80 are term-based and 28 are stylistic-based features. For the preprocessing phase we employed the StringToWordVector filter, present in Weka's libraries, to obtain the unigram features. This filter is able to transform a input char set into a output token set where each token has specific value. This value is also called weight and it is derived from frequency of the token. During the tokenization phase we applied the stopword list, but not the stemming. Then, we applied the Chi-square feature selection techniques for reducing the number of unigrams to 80. Then we used regular expressions and the "Twitter NLP and Part of Speech Tagging" for identifying the stylistic features, as detailed in [8].

During the experiments we investigated the best method to identify the age class, among a set of popular classification methods: SVM, Naive Bayes, Multinomial Naive Bayes and K Nearest Neighbours. Moreover, we were interested in understanding if gender attribute value influenced age classification. For this reason, we performed several experiments using different dataset compositions:

- *dataset A*, composed of all instances belonging to the dataset;
- *dataset B*, composed of only male instances of *dataset A*;
- *dataset C*, composed of only female instances of *dataset A*.

For each aspect that we explored, we conducted experiments using 10-fold cross validation technique.

Results

A set of experiments was conducted using several machine learning techniques, aimed at inferring the age class of each single tweet, over the three dataset. The SVM obtained the best results among the considered approaches as shown in Table 2. Furthermore, other

	Dataset A	Dataset B	Dataset C
SVM	64%	65%	66%
NB	59%	59%	60%
NBMultinomial	62%	62%	62%
kNN (k=7)	60%	60%	61%

Table 2: Tweet-level accuracy obtained using different machine learning techniques over the different dataset.

experiment outcomes reported in Table 3 show the different performances achieved, over *dataset A*, *dataset B* and *dataset C*, using SVM with different feature set: only unigram features, only stylistic features and both. Note that, in all the experiments, the highest performances were obtained considering all features, while the lowest were obtained considering only unigram features.

Once we discovered the best configuration to obtain the age class of each single tweet, we decided to study the performances from the user point of view. The experimental campaign was conducted using all the datasets (*dataset A*, *dataset B* and *dataset C*). Table 4 shows the outcomes. So, in order to obtain the best outcomes, we

Dataset	Features	Accuracy
Dataset A	Unigrams	61.88%
	Stylistic	63.12%
	All	64.45%
Dataset B	Unigrams	62.69%
	Stylistic	63.09%
	All	65.17%
Dataset C	Unigrams	62 .53%
	Stylistic	63.67%
	All	65.76%

Table 3: Accuracy performances using SVMs.

decided to learn two different model to infer age: one for male and one for female instances. Both the model are trained using SVM, using the entire feature set (unigrams + stylistic features).

Dataset	Accuracy
Dataset A	71.28%
Dataset B	72.82%
Dataset C	75.43%

Table 4: User-level accuracy performances achieved using SVM in combination with the entire feature set.

The effect of size variation in the tweet set available for each author to infer his/her age class was also analyzed. In particular, the experiments shown that the variation of accuracy for each Twitter user was reduced as the number of tweets available decreased (see Figure 3). To judge age class from just one tweet is a complex task. Therefore, to address this difficulty it is useful to increase the number of tweets examined for each Twitter user. In order to balance effectiveness and efficacy of age classification, in the tool a set of 55 tweets was considered for each user.

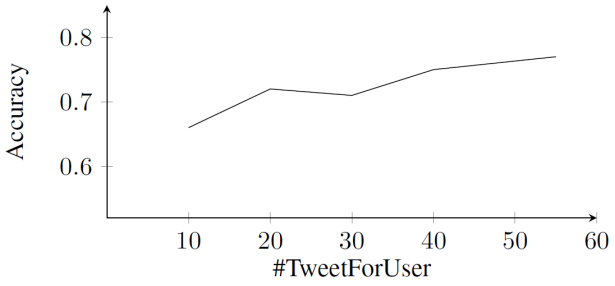

Figure 3: User-level accuracy variation respect to tweet test-set dimension variation.

6. INTERFACE DESIGN

The aim of the interface design is to support the users with all instances that are difficult to classify automatically. Any instance classified as "unknown" can be processed manually by the end-users through the refinement step. The interface is composed of two main areas: the process timeline viewer *(a)* and the data viewer *(b)* on the left and on the right part of Figure 4, respectively.

The process timeline viewer shows all the steps of the procedure, and highlights the one currently performed. The content of the data viewer varies, depending on the relevant information for the current process phase. For instance, during this refinement phase, additional information related to the Twitter profile of each user is

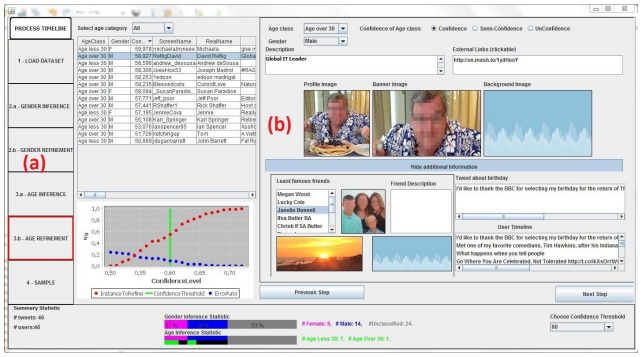

Figure 4: Tool interface, there are the process timeline bar (a) on the left and the data viewer (b) on the right.

shown to the end-user. For the gender inspection, screen name, description, profile image, banner image and background profile image are shown. All this information can be used to infer a user's gender. Indeed, frequently, the images could show a self-portrait and the description could give indications about their gender. In the same way, the age class identification step is followed by a user refinement step. Here, since during the automatic classification stage a confidence level (see Formula 2) is associated to each classified instance, all instances whose age class is identified with a low confidence level are refined by the end-user. Once again, during this phase, the end-user can process the data by using additional information. Beside all the information already described for the gender refinement step, the end-user can also see external links, a list of least famous friends, the user's timeline and the user's tweets about "birthday". All this kind of additional information is useful for different reasons. For instance, external links could be able to connect the unclassified users with their other web-pages. Here, suggestions about a user's age class could be found. In a user's timeline the end-user could read about a specific topic highly related to a particular age class. In user's tweets about birthday a user could celebrate his/her birthday indicating his/her age. Moreover, as reported in [26], an indicator of user's age class is the set of least famous friends. They are friends of user who have the fewest followers, which are more likely to be friends in real life (versus celebrities or organizations) and therefore belong to the same age group. For each friend of this kind, information like description, profile image, background image and banner image, is displayed. After the age class refinement phase, the final aim of the tool is reached: a set of users with their gender and age class is obtained.

6.1 Cognitive Walkthrough

To understand if the interface that we designed is intuitive and easy to use for typical end-users, we decided to conduct a formal evaluation using a method called cognitive walkthrough. The cognitive walkthrough entails an usability analyst stepping through the cognitive tasks that a user must carry out in interacting with technology. The aim of a walkthrough is to evaluate the design of a user interface, with special attention to how well the interface supports "exploratory learning", i.e., first-time use without formal training. In brief, users start with a goal and some sort of plan(task sequence) as how to achieve the goal. Users then look for apparently relevant actions, activate the most probable option, consider the system response (system feedback) and decide whether the right effect has been achieved.

For each individual step, performed by the end-users, in the interaction the analyst asks the following questions:

- Will the user try to achieve the right effect?
- Will the user notice that the correct action is available?
- Will the user associate the correct action with the effect to be achieved?
- If the correct action is performed, will the user see that progress is being made towards solution of the task? [25]

These questions permit to understand in which tool part end-user has difficult to do tasks. For instances the case in which a button produces unexpected effect, the order of displayed information is wrong, the end-user is not able to proceed in the process, etc. In all these cases improvements of the tool interface are required in order to maximize the easiness of it. In fact, the final aim of the cognitive walkthrough is to realize an interface that not required it explanation to end-users before it usage.

Procedure

For our cognitive walkthrough we recruited 2 domain experts (both male with age between 30 and 35). both are social scientists who work in a major market research organization in the UK that conducts surveys for a wide range of major clients, including commercial and government organisations. They are comfortable with computers and very familiar with Twitter Social Media.

Before beginning the cognitive walkthrough, the participants received a 10 minute of presentation about the tool, which presented the aim and all basic conceptual steps required to obtain it. The presentation described some background to the work and explained the limitations of automatic methods, but no reference to the interface was made at this stage. In this way the participants were not influenced in how to achieve the requested goals. We asked participants to attempt to reach the following two goals:

1. to obtain a sample composed of older users, only using automatic method procedure (no refinement phase for either gender and age class inference);
2. to obtain a sample composed of older users, using both automatic method procedure and refinement phase for just few users for either gender and age class inference. After that, to load another dataset.

The sequences of tasks needed for both first and second goal are reported in Table 5 and 6, respectively.

Step	Task
1	Load new tweet dataset
2	Start gender inference process
3	Do not do gender refinement phase
4	Start age class inference phase
5	Do not do age class refinement phase
6	Save author sample (save only older authors)

Table 5: Task sequence to achieve goal 1.

During the procedure we recorded our observations on paper data sheets. We used the “thinking aloud” methods, whereby participants were asked to verbalize their thoughts while performing the tasks. Comments made by the participants are often valuable complements to observed behaviors in the test, and “thinking aloud” can help participants communicate what they are feeling about a tool and problems they may encounter while using it.

The post-study survey was used to gather the test participants’ opinions about the tool, after the test. The participants were asked

Step	Task
1	Load new tweet dataset
2	Start gender inference process
3	Do gender refinement phase and set the gender for some authors
4	Start age class inference phase
5	Do age class refinement for some authors
6	Save author sample (save all authors)
7	Save author sample (save only older authors)
8	Load new dataset

Table 6: Task sequence to achieve goal 2.

to answer questions that cover all the different aspects of usability. For instance, the questions regard effectiveness, efficiency, information understanding, and easiness of use of the tool. All these aspects are very important to create a usable interface. Indeed, our aim was to create a tool that was usable by a non-technical user, therefore the easiness for learning and using it is essential. Also the easiness of information understanding is important: the end-user has to use the additional information to judge demographic attributes of Twitter users.

Results

The cognitive walkthrough highlighted several problems:

1. Both participants suggested that a continuous update about the age and gender composition of the current set of Twitter users should be available. In this way the end-users can decide with more confidence about the number of instances to refine.
2. Both complained about the absence of options for selecting the confidence threshold that split the entire set of Twitter users into instances to refine or not.
3. During the two refinement phases, we noticed that the attention of the experts was captured by the images, while the textual information was mainly ignored.
4. The pop-up messages that appeared between two phases was not clear. The participants suggested to simplify the messages in order to make more simple and fast end-user choice.
5. In the gender refinement screen, the experts clicked on the "Next" button for selecting other users to refine. Actually, they obtained an unexpected effect: the age class identification started. Moreover, for the same screen, one of the cognitive walkthrough participant complained about the absence of label indicating which kind of images were shown in the user panel.
6. In the age refinement screen they suggested we simplify the top part of the screen designed to set the new age class or gender of a Twitter user, but they appreciated the facility to modify gender class at this stage. Moreover, they suggested to explain more clearly the meaning of word "confidence". In this screen they also found another problem related to going back in the process using buttons.

On the basis of this feedback, a second interface prototype was designed. A key feature in the new interface is an additional visualization component: the summary panel (see Figure 5). It is the main improvement made on the previous version and displays a breakdown of labelled and unlabelled user cases. A new combo box was

Figure 5: Summary panel.

added which allows the end-user to set the confidence threshold for user refinement. Having chosen the threshold, the tool presents all instances that require examination (all instances with a confidence level lower than threshold level).

The graph presented in Figure 6 is particularly useful for the user to establish an optimal trade-off between accuracy and sample size. It can be used by the end-user for understanding which is the best confidence level threshold to choose based on his/her needs (a social scientist could prefer a highest level of error and refine a lowest number of instances or viceversa). For supporting the end user in the choice of the confidence level threshold, during the age class refinement step, we incorporated the graph reported in Figure 6 in the tool interface. This graph has been obtained over the dataset where 55 tweets are used for identifying the age class of each user. It shows the variation of the number of instances that required a refinement (classified with lower confidence level than a certain threshold) and the variation of the error rate with respect to the variation of confidence level. Using this graph the end-user can follow two possible strategies for choosing the best threshold: one based on error rate requirements and one based on the size of the user sample required.

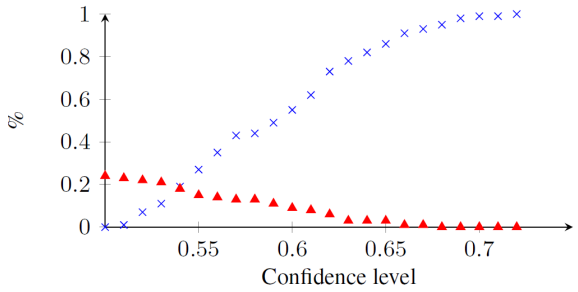

Figure 6: Percentage of instances to refine (x marks) and error rate (triangle marks) with respect to confidence level variation.

In the summary panel is also possible to see the sample composition in each step of the process. The information displayed is: number of initial users, number of female, male and unknown classified users during the gender inference phase, number of younger and older users classified with a certain confidence level. It is also possible to see the proportion of age classified users respect to gender. The rectangle below "Gender inference statistic" label represents percentage of Twitter users, belonging to the initial dataset, that are gender classified as female (fuchsia bar), male (blue bar) and unknown (gray bar). While the rectangle below "Age inference statistic" label represents also the proportion of females and males that are age classified with a confidence level higher than the threshold (green bar) and with a confidence level lower than the threshold (black bar).

The new gender refinement screen is presented in Figure 7. Here, for the gender inspection, screen name *(a)*, description *(b)*, profile image *(c)*, banner image *(e)* and background profile image *(d)* are shown. Essentially, we changed the order of information and we highlighted all the text boxes in order to attract user attention towards text areas.

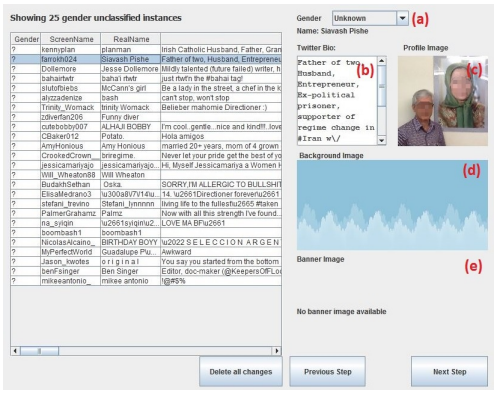

Figure 7: Gender refinement screen.

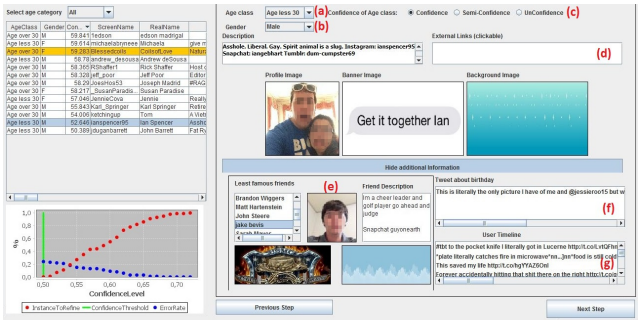

Figure 8: Age refinement screen.

The second age refinement screen is shown in Figure 8. Here, the user can see all the information already described for the gender refinement step, as well as external links *(d)*, list of least famous friends *(e)*, user's timeline *(f)* and user's tweets about "birthday" *(g)*. Another improvement on the age refinement screen was to move the text boxes above the images and highlight their texts. Moreover, we introduced combo boxes *(a)* and *(b)* for simplify how to set the new age/gender class. Both in age class and gender combo box the actual class value of an instance is shown. When the user selects another class, this class is automatically assign to the selected user. In this way, the process of setting age or gender class becomes more fast and simple. The user can specify their confidence level about refined instances, using a three radio buttons *(c)* that replace the earlier slider control. Now, the user only needs to specify if he/she is confident, semi-confident or completely unconfident about their classification.

6.2 Summative Evaluation

We also conduced a summative evaluation of the second interface prototype. This evaluation was quantitative and about testing if the objective of enabling users to make quick and confident judgments was met.

Through a traditional "time and errors" usability test, three dependent variables were collected for both gender and age class refinement tasks: completion time, inter-rater agreement and success rate. Task completion time was measured by recording the time when users clicked on a row related to a user until they made a final decision by selecting one of the possible classes. Inter-rater agreement is the degree of agreement among two or more evaluators. It describes how frequently they assign the exact same rating (if all give the same rating, they are in agreement) and it gives a score of

how much homogeneity, or consensus, there is in the ratings given by judges. It is useful in refining the tools given to human judges, for example by determining if a particular scale is appropriate for measuring a particular variable. Success rate is the percentage of users' correct classification decision.

Procedure

We recruited 22 participants (15 males and 7 females), of which 12 PhD students, 7 researchers and 3 master students in the Department of Computer Science both of Brunel University and Milano-Bicocca University. All of them were comfortable with computers and familiar with Twitter. A dataset for trial was collected using an experimental version of Chorus Tweetcatcher (TCD)[1] that is able to collect a table of Twitter users where each of them has all the attributes listed in Table 7.

UserId	ScreenName	RealName
Description	URL	Language
Location	GeoEnabled	TimeZone
UTCOffset	Followers	Friends
StatusCount	Favourites	CreatedAt
Verified	Protected	ProfileImage
BannerImage	BackImage	TweetSample
BirthdayTweets		

Table 7: Meta-data available for each user in the table obtained using Chorus TCD.

The dataset obtained through Chorus TCD was composed of 50 Twitter users. The test participants had to inspect 25 gender unclassified users and 21 age unclassified users belonging to this dataset.

Before beginning the study, participants received 10 minute training on final interface. The training session consisted of a brief explanation of the tool's purpose and basic concepts, and a short demonstration of the interface and detailed instruction on the usage of the interface. The tutorial was administered by the same person following a basic script (explanations and demonstrations). In addition to demonstrating the features of the interfaces, the administrator explained basic strategies to complete the tasks (for example, comparing the additional information of the considered user to assign gender class).

Tasks did not have a time limit. Once the participants completed the refinement of gender, they repeated the same procedure for assigning age to those user falling beneath the specified threshold.

After the participants finished all the tasks, they were asked to complete post-study questionnaire about the interface. Subjective measures including satisfaction, usability, and learnability were collected along with participants' comments and suggestions during the post-study survey session.

Results

Regards the summative evaluation the results that we collected show that the gender refinement phase required less time than the age one. The assignment of a gender to an instances takes around 8.3 sec, while the assignment of an age class takes around 16 sec. This confirms the idea that age classification is also more difficult for human judges than the gender one. A deep investigation of the decision time shows that the users do not tend to slow down in their judgment task as the trials proceeded, with decision time depending on how clear the additional information is for them. Figure 9 shows bar charts that represent, respectively, average task completion time to refine user gender and age class.

[1]This software is available at http://chorusanalytics.co.uk

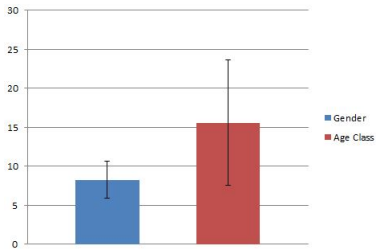

Figure 9: Average time in sec for gender and age refinement.

Furthermore, we find that in age refinement phase more time is required to load all the information (4 sec) than in gender refinement phase (0.5 sec). Indeed, least famous friend information, for user's age class inspection, was collected real-time, while all the other information required for gender refinement was already present as attributes of the user table. Indeed, this data was part of the input-data collected, a priori, using Chorus TCD.

Then, we studied the inter-agreement rate between participants for both gender and age refinements. We used an adaptation of Cohens' Kappa test for multiple raters proposed by Fleiss [7]. Fleiss' kappa is a variant of Cohen's kappa, a statistical measure of inter-rater reliability. Whereas Cohen's kappa works for only two raters, Fleiss' kappa works for any constant number of raters giving categorical ratings, to a fixed number of items. It is a measure of the degree of agreement that can be expected above chance. The Fleiss' Kappa statistic for gender was equal to 77.34%, and for age was equal to 70.45%. We find again that age inference seems to be more difficult than gender. Although 70.45% is a good level of agreement is less than the one obtained in gender refinement.

We also attempted to understand the accuracy level obtained by participants. We performed a 3-fold cross-validation separating the evaluation given by participants into two groups. Each time with the evaluation of 66% of the raters we created a gold standard and with the evaluation of other 33% of the raters, we created the test set from which we computed their accuracy. We found that for gender refinement the level of accuracy reached is equal to 92%, while for age refinement the level of performances achieved is 91%. Also from this point of view, we obtained worse results in the age identification, the most difficult task, than in gender derivation.

Moreover, we studied how effective is the confidence value in finding the users misclassified by the automatic method, and whether the manual labeling had improved the results. We analyzed the agreement between the participants and the automatic method, and we found that the 80% of the misclassifications happen when the confidence value is in the range 50-60%.

In the post-study survey, the participants were asked to answer questions about easiness of use, easiness of learning, easiness of navigation and easiness of information understanding. Their overall satisfaction and confidence toward the interface was high. The main qualitative feedback was to increase the size of tweet windows in order to facilitate their reading.

As the summative evaluation showed, the visualization of additional information helped users make decisions faster. Users who have participated in our trials have been very positive about the interactive approach supported by the TweetClass tool.

7. CONCLUSION AND FUTURE WORKS

In this paper, we presented a semi-automatic approach to boost the accuracy of demographic attribution of users contributing to a Twitter corpus. We presented TweetClass, as a proof-of-concept

tool, that supports social scientist researchers in the identification of demographic attributes of a Twitter users sample by combining both automatic and interactive class inference methods. This is a difficult problem because these attributes (e.g. age class, gender) are not directly obtainable from tweets or user account meta-data. As first step we built the hierarchical model to automatically identify demographic attributes, then we developed an interface that enables the user to intervene in the classification process. The interface is necessary for inspecting and refining the Twitter users of the initial set for which an automatic classification of their gender or age class is very hard. In this way, the end-users can increase the quality or/and the dimension of Twitter user samples.

Future work of this study will concern how to exploit the knowledge provided by the end-users' refinement. For instance, we could use the inspected Twitter users as new instances in the training set obtaining an active learning model able to improve itself each time that new Twitter users are refined by TweetClass end-users. Moreover, another future work could relate to incorporate in the tool other automatic techniques able to increase the performance of the existing demographic attribute classification or new automatic techniques to identify other demographic attribute such as profession, marital status and so on. Since, after the end of the study, a new dataset, with both gender and age label, became available [19], we want to expand our experiments over this large dataset for a better evaluation of our approach.

8. REFERENCES

[1] J. S. Alowibdi, U. A. Buy, and P. Yu. Language independent gender classification on twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 739–743. ACM, 2013.

[2] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

[3] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, 2011.

[4] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[5] M. Ciot, M. Sonderegger, and D. Ruths. Gender inference of twitter users in non-english contexts. In *EMNLP*, pages 1136–1145, 2013.

[6] C. Fink, J. Kopecky, and M. Morawski. Inferring gender from the content of tweets: A region specific example. In *ICWSM*, 2012.

[7] J. L. Fleiss, B. Levin, and M. C. Paik. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2:212–236, 1981.

[8] 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*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] R. M. Groves. Three eras of survey research. *Public Opinion Quarterly*, 75(5):861–871, 2011.

[10] J. Ito, T. Hoshide, H. Toda, T. Uchiyama, and K. Nishida. What is he/she like?: Estimating twitter user attributes from contents and social neighbors. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1448–1450, New York, NY, USA, 2013. ACM.

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

[12] J. Michael. 40000 namen, anredebestimmung anhand des vornamens, 2007.

[13] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11:5th, 2011.

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

[15] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. Tweetgenie: automatic age prediction from tweets. *ACM SIGWEB Newsletter*, 4(Autumn):4, 2013.

[16] D. Nguyen, N. A. Smith, and C. P. Rosé. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics, 2011.

[17] C. Peersman, W. Daelemans, and L. Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

[18] 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*, KDD '11, pages 430–438, New York, NY, USA, 2011. ACM.

[19] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the author profiling task at pan 2013. *Notebook Papers of CLEF*, pages 23–26, 2013.

[20] 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*, pages 37–44. ACM, 2010.

[21] C. Seale, S. Ziebland, and J. Charteris-Black. Gender, cancer experience and internet use: a comparative keyword analysis of interviews and online cancer support groups. *Social science & medicine*, 62(10):2577–2590, 2006.

[22] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online*, 18(3):7, 2013.

[23] T. W. Smith. Survey-research paradigms old and new. *International Journal of Public Opinion Research*, page eds040, 2012.

[24] R. O. Tachibana, N. Oosugi, and K. Okanoya. Semi-automatic classification of birdsong elements using a linear support vector machine. *PloS one*, 9(3):e92584, 2014.

[25] C. Wharton, J. Rieman, C. Lewis, and P. Polson. Usability inspection methods. chapter The Cognitive Walkthrough Method: A Practitioner's Guide, pages 105–140. John Wiley & Sons, Inc., New York, NY, USA, 1994.

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

[27] Y. Zhang, D. Wang, and T. Li. idvs: an interactive multi-document visual summarization system. In *Machine Learning and Knowledge Discovery in Databases*, pages 569–584. Springer, 2011.