Probabilistic Inference of Twitter Users' Age based on What They Follow

Benjamin Paul Chamberlain[1], Clive Humby[2], and Marc Peter Deisenroth[1]

[1] Department of Computing, Imperial College London, London, UK,
b.chamberlain14@ic.ac.uk
[2] Starcount Insights, 2 Riding House Street, London, UK

Abstract. Twitter provides an open and rich source of data for studying human behaviour at scale and is widely used in social and network sciences. However, a major criticism of Twitter data is that demographic information is largely absent. Enhancing Twitter data with user ages would advance our ability to study social network structures, information flows and the spread of contagions. Approaches toward age detection of Twitter users typically focus on specific properties of tweets, e.g., linguistic features, which are language dependent. In this paper, we devise a language-independent methodology for determining the age of Twitter users from data that is native to the Twitter ecosystem. The key idea is to use a Bayesian framework to generalise ground-truth age information from a few Twitter users to the entire network based on what/whom they follow. Our approach scales to inferring the age of 700 million Twitter accounts with high accuracy.

Introduction

Digital social networks (DSNs) produce data that is of great scientific value. They have allowed researchers to study the flow of information, the structure of society and major political events (e.g., the Arab Spring) quantitatively at scale.

Owing to its simplicity, size and openness, Twitter is the most popular DSN used for scientific research. Twitter allows users to generate data by *tweeting* a stream of 140 character (or less) messages. To consume content users *follow* each other. Following is a one-way interaction, and for this reason Twitter is regarded as an *interest network* (Gupta, 2013). By default, Twitter is entirely public, and there are no requirements for users to enter personal information.

The lack of reliable (or usually any) demographic data is a major criticism of the usefulness of Twitter data. Enriching Twitter accounts with demographic information (e.g., age) would be valuable for scientific, industrial and governmental applications. Explicit examples include opinion polling, product evaluations and market research.

We assume that people who are close in age have similar interests as a result of age-related life events (e.g., education, child birth, marriage, employment, retirement, wealth changes). This is an example of the well-known

Fig. 1. Twitter profile for @williamockam that we created to illustrate our method. The profile contains the name, Twitter handle, number of tweets, number of followers, number of people following and a free-text description field with age information.

homophily principle, which states that people with related attributes form similar ties (McPherson, 2001). For age inference in Twitter, we exploit that most Follows[3] are indicative of a user's interests. Putting things together, we arrive at our central hypothesis that (a) somebody follows what is interesting to them, (b) their interests are indicative of their age. Hence, we propose to infer somebody's age based on what/whom they Follow. We created the artificial @williamockam account shown in Fig. 1 to use as a running example of our method.

The contribution of this paper is a probabilistic model that is massively scalable and infers every Twitter user's age based on what/whom they Follow without being restricted by national / linguistic boundaries or requiring data that few users provide (e.g. photos or large numbers of tweets). Our model handles the high levels of noise in the data in a principled way. We infer the age of 700 million Twitter accounts with high accuracy. In addition we supply a new public dataset to the community.

Related Work

There is a large body of excellent research on enhancing social data with demographic attributes. This includes work on gender (Burger, 2011), political affiliation (Pennacchiotti, 2011), location (Cheng, 2010) and ethnicity (Mislove, 2011; Pennacchiotti, 2011). Also of note is the work of Fang (2015) who focus on modelling the correlations between various demographic attributes.

Following the seminal work of Schler (2006), the majority of research on age detection of Twitter users has focused on linguistic models of tweets (Nguyen, 2011; Rao, 2010; Al Zamal, 2012). Notably, Nguyen (2013) developed a linguistic model for Dutch tweets that allows them to predict the age category (using logistic regression) of Twitter users who have tweeted more than ten times in Dutch. They performed a lexical analysis of Dutch language tweets and obtained ground truth through a labour intensive manual tagging process. The principal features were unigrams, assuming that older people use more positive language, fewer pronouns and longer sentences. They concluded that age prediction works well for young people, but that above the age of 30, language tends to homogenise.

Additionally, tweet-based methods struggle to make predictions for Twitter users with low tweet counts. In practice, this is a major problem since we calculated that the median number of tweets for the 700m Twitter users in our data

[3] we use capitalisation to indicate the Twitter specific usage of this word

set is only 4 (the *tweets* field shown in Fig. 1 is available as account metadata for all accounts).

The user name has also been considered as a source of demographic information. This was first done by Liu (2013) to detect gender and later by Oktay (2014) to estimate the age of Twitter users from the first name supplied in the free-text *account name* field (e.g. William in Fig. 1). In their research, they use US social security data to generate probability distributions of birth years given the name. They show that for some names, age distributions are sharply peaked. A potential issue with this approach is that methods based on the "user name" field rely on knowledge of the user's true first name and their country of birth (Oktay, 2014). In practice, this assumption is problematic since Twitter users often do not use their real names, and their country of birth is generally unknown.

Approaches to combine lexical and network features include Al Zamal (2012); Pennacchiotti (2011), who show that using the graph structure can improve performance at the expense of scalability. Kosinski (2013) used Facebook-Likes to predict a broad range of user attributes mined from 58,466 survey correspondents in the US. Their approach of solely using Facebook Likes as features for learning has the benefit of generalising readily to different locales. Culotta (2015) have applied a similar Follower based approach to Twitter to predict demographic attributes, however their approach of using aggregate distributions of website visitors as ground-truth is restricted to predicting the aggregate age of groups of users. Our work is inspired by the generality of the approaches of Kosinski (2013) and Culotta (2015), however our setting differs in two ways. We use data native to the Twitter ecosystem to generalise from a few examples to make individual predictions for the entire Twitter population. Secondly we do not make the assumption that our sample is an unbiased estimate of the Twitter population and we explicitly account for this bias to make good population predictions. For these reasons it is hard to get ground truth and careful probabilistic modelling is required to infer the age of arbitrary Twitter users.

Table 1. Ground-truth data set: Age categories and counts. "features" gives the average number of feature accounts followed.

idx	age	count	freq	features
0	under 12	7,753	5.9%	23.7
1	12–13	20,851	15.8%	27.9
2	14–15	30,570	23.1%	30.8
3	16–17	23,982	18.1%	28.7
4	18–24	33,331	25.2%	26.0
5	25–34	9,286	7.0%	23.1
6	35–44	3,046	2.3%	22.6
7	45–54	1,838	1.0%	16.0
8	55–64	962	0.7%	11.4
9	over 65	596	0.5%	11.2

Probabilistic Age Inference in Twitter

Our age inference method uses ground-truth labels (users who specify their age), which are then generalised to 700m accounts based on the shared interests, which we derive from Following patterns.

To extract ground-truth labels we crawl the Twitter graph and download user descriptions. To do this we implemented a distributed Web crawler using Twitter access tokens mined through several consumer apps. To maximize data throughput while remaining within Twitter's rate limits we built an asynchronous data mining system connected to an access token server using Python's Twisted library Wysocki (2011).

Our crawl downloaded 700m user descriptions. Fig. 1 shows the profile with associated metadata fields for the fictitious @williamockam account, which we use to illustrate our approach. We index the free-text description fields using Apache SOLR (Grainger, 2014) and search the index for REGular EXpression (REGEX) patterns that are indicative of age (e.g., the phrase: "I am a 22 year old" in Fig. 1) across Twitter's four major languages (English, Spanish, French, Portuguese). For repeatability we include our REGEX code in the git repository. Twitter is ten years old and contains many out-of-date descriptions. To tackle the stale data problem we restricted the ground-truth to active accounts, defined to be accounts that had tweeted or Followed in the last three months (we do not have access to Twitter's logs). This process discovered 133,000 active users who disclosed their age (i.e., 0.02% of the 700m indexed accounts), which we use as "ground-truth" labels. For each of these we download every account that they Followed. Fig. 1 shows that @williamockam Follows 73 accounts and we downloaded each of their user IDs. We use ten age categories with a higher resolution in younger ages where there is more labelled data. For our ground-truth data set, the age categories, number of accounts, relative frequency and average number of features per category are shown in Table 1.

Table 2. Public dataset labels: age categories and counts.

idx	age range	count
1	10-19	4486
2	20-29	4485
3	30-39	4487
4	40-49	4485
5	50-59	4484
6	60-69	4481
7	70-79	4481

Applying REGEX matches to free-text fields inevitably leads to some false positives due to unanticipated character combinations when working with large data sets. In addition, many Twitter accounts, while correctly labelled, may not represent the interests of human beings. This can occur when accounts are controlled by machines (bots), accounts are set up to look authentic to distribute spam (spam accounts) or account passwords are hacked in order to sell authentic looking Followers. To reduce the impact of spurious accounts on the model we note that (1) incorrectly labelled accounts can have a large effect on the model as they are distant in feature space from other members of the class / label (2) incorrectly labelled accounts that have a small effect on the model (e.g. because they only follow one popular feature) do not matter much by definition. To measure the effect of each labelled account on the model we compute the Kullback-Leibler divergence $\mathrm{KL}(P||P_{\setminus i})$ between the full model and a model evaluated with one data point missing. Here, P is the likelihood of the full, labelled data set, and $P_{\setminus i}$ is the likelihood of the model using the labelled data set minus the i^{th} data point. This methodology identifies any accounts that have a particularly large impact on our predictive distribution. We flagged any

Table 3. Spurious data points identified by taking the Median Absolute Deviation of the leave-one-out KL-Divergence.

Handle	Twitter Description	REGEX age	Reason to Exclude
RIAMOpera	Opera at the Royal Irish ... Presenting: Ormindo Jan 11...	11	An Irish Opera
TiaKeough13	My name Tia I'm 13 years old.	13	Hacked account
39yearoldvirgin	I'm 39 years old... if you're a woman, I want to meet you.	39	Probably not 39
50Plushealths	Retired insurance Agent After 40 years of Services.	retired	Using reciprocation software
MrKRudd	Former PM of Australia... Proud granddad of Josie & McLean...	grandparent	Outlier. Former AUS PM

training examples that were more than three median absolute deviations from the median score for manual inspection. This process excluded 246 accounts from our training data and examples are shown in Table 3. We also randomly sampled 100 data points from across the full ground-truth set and manually verified them by inspecting the descriptions, tweets and who / what they Follow.

For reproducibility we make an anonymised sample of the data and our code publicly available [4]. The data is in two parts: (1) A sparse bipartite adjacency matrix; (2) a vector of age category labels. This dataset was collected and cleaned according to the methodology described above and then down-sampled to give approximately equal numbers of labels in each of seven classes detailed in Table 2. It includes only accounts that explicitly state an age (ie. no grandparents or retirees). The adjacency matrix is in the format of a standard (sparse) design matrix and includes only features that are Followed by at least 10 examples. The high level statistics of this network are described in Table 4.

Age Inference based on Follows

Given a set of 133,000 labelled data points (ground-truth, i.e., Twitter users who reveal their age) we wish to infer the age of the remaining 700m Twitter users. For this purpose, we define a set of features that can be extracted automatically. The features are based on the Following patterns of Twitter users. Once the features are defined, we propose a scalable probabilistic model for age inference.

Table 4. Public dataset adjacency matrix statistics. Subscript 1 describes labelled acounts and 2 describes features. V denotes vertices, E edges and D degree.

attribute	value
$\|V_1\|$	31,389
$\|V_2\|$	50,190
$\|E\|$	1,810,569
avg D_1	57.7
max D_1	2049
std D_1	95.2
avg D_2	36.1
max D_2	4405
std D_2	96.2

Our age inference exploits the hypothesis that someone's interests are indicative of their age, and uses Twitter Follows as a proxy for interests. Therefore, the features of our model are the 103,722 Twitter accounts that are Followed by more than ten labelled accounts, which can be found automatically. Of the 73 accounts Followed by @williamockam, 8 had sufficient support to be included in our model. These

[4] `https://github.com/melifluos/bayesian-age-detection`

Table 5. Follower counts for the eight @williamockam features. The support gives their total number of Followers in our labelled data set and Followers is their total number on Twitter. Fractional counts are from assigning a distribution to grandparents.

Twitter Handle	Support	<12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	≥65	Followers
Lord_Voldemort7	**273**	5	35	75	55	87	13	0	1	1	1	2.0×10^6
WaltDisneyWorld	**435**	61	100	89	80	65	20	4	7	4	4	2.5×10^6
Applebees	**191**	18	43	38	30	37	9	8	2.33	2.33	3.33	0.57×10^6
UniStudios	**60**	7	7	14	14	13	5	0	0	0	0	0.27×10^6
UniversalORL	**65**	5	13	10	15	14	4	0	1.66	1.66	0.66	0.40×10^6
HorrorNightsORL	**5**	0	0	0	1	3	1	0	0	0	0	0.04×10^6
HorrorNights	**18**	1	3	1	4	6	0	1	0.66	0.66	0.66	0.08×10^6
OlanRogers	**16**	0	2	0	7	7	0	0	0	0	0	0.11×10^6

Table 6. Posterior distributions (4) for the eight features Followed by @williamockam. Probabilities are $\times10^{-5}$

Twitter Handle	Support	<12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	≥65	Followers
Lord_Voldemort7	**273**	111.7	190.9	258.0	252.3	248.6	145.9	31.9	38.9	77.6	177.5	2.0×10^6
WaltDisneyWorld	**435**	725.0	538.2	441.2	377.6	267.3	233.2	194.2	270.7	254.5	224.4	2.5×10^6
Applebees	**191**	231.8	206.3	176.6	150.3	129.8	137.4	226.7	132.4	139.6	139.2	0.57×10^6
UniStudios	**60**	80.6	56.0	59.3	59.5	49.3	48.1	11.3	2.8	2.3	2.3	0.27×10^6
UniversalORL	**65**	67.4	63.0	56.6	60.5	50.7	42.0	21.1	62.7	86.4	40.6	0.40×10^6
HorrorNightsORL	**5**	0.3	0.7	1.5	4.0	8.3	9.4	2.0	0.3	0.1	0.1	0.04×10^6
HorrorNights	**18**	14.0	13.7	11.3	15.5	16.1	9.4	29.1	29.9	36.8	29.3	0.08×10^6
OlanRogers	**16**	4.3	9.1	10.6	21.9	19.8	5.0	1.6	1.3	1.3	1.3	0.11×10^6

were: Lord_Voldemort7, WaltDisneyWorld, Applebees, UniStudios, UniversalORL, HorrorNightsORL, HorrorNights and OlanRogers.

Table 5 shows the number of labelled accounts Following each feature for @williamockam. The support is the number of *labelled* Followers summed over all age categories, while Followers gives the total number of Followers (labelled and unlabelled). A general trend across all features (not only the ones relevant to @williamockam) is that the age distribution is peaked towards "younger" ages as not many older people reveal their age (we show this for the accounts with the highest support in our data set in the appendix on our git repo). To improve the predictive performance of the model in higher age categories we adapted our REGEX to search for grandparents and retirees. This augmented our training data with 176,748 people labelled as retired and 63,895 labelled as grandparents. In our ten-category model, retired people are added to the 65+ category. Grandparents are assigned a uniform distribution across the three oldest age categories, which roughly reflects the age distribution of grandparents in the US (UScensus, 2014)[5], such that we ended up with approximately 374,000 labelled accounts in our ground-truth data.

Probabilistic Model for Age Inference We adopt a Bayesian classification paradigm as this provides a consistent framework to model the many sources of uncertainty (noisy labels, noisy features, survey estimates) encountered in the problem of age inference.

[5] This value was used as the US is the largest *Twitter country*.

Our goal is to predict the age label of an arbitrary Twitter user with feature vector X given the set of feature vectors $\mathbf{X}$ and corresponding ground-truth age labels $\mathbf{A}$. Within a Bayesian framework, we are therefore interested in the posterior predictive distribution

$$P(A|X, \mathbf{X}, \mathbf{A}) \propto P(X|A, \mathbf{X}, \mathbf{A})P(A)\,, \tag{1}$$

where $P(A)$ is the prior age distribution and $P(X|A, \mathbf{X}, \mathbf{A})$ the likelihood.

The prior $P(A)$ is based on a survey of American internet users conducted by Duggan (2013). They sampled 1,802 over-18-year olds using random cold calling and recorded their demographic information and social media use. 288 of their respondents were Twitter users, yielding a small data set that we use for the prior distributions of over 18s. For under 18s we inferred the corresponding values of the prior using US census data (UScensus, 2010), which leads to the categorical prior

$$P(A) = \mathrm{Cat}(\pi) = [1, 2, 2, 3, 14, 23, 23, 22, 6, 4] \times 10^{-2}\,. \tag{2}$$

The likelihood $P(X|A, \mathbf{X}, \mathbf{A})$ is obtained as follows: For scalability we make the Naive Bayes assumption that the decision to Follow an account is independent given the age of the user. This yields the likelihood

$$P(X|A, \mathbf{X}, \mathbf{A}) = \prod_{i=1}^{M} P(X_i|A, \mathbf{A}, \mathbf{X})^{X_i}\,, \tag{3}$$

where $X_i \in \{0, 1\}$ and i indexes the features. $X_i = 1$ means "user χ Follows feature account i".[6]

We model the likelihood factors $P(X_i|A, \mathbf{A}, \mathbf{X})$ as Bernoulli distributions

$$P(X_i|A = a) = \mathrm{Ber}(\mu_{ia}), \tag{4}$$

$i = 1, \ldots, M$, where M is the number of features and there are 10 age categories indexed by $a = 1, \ldots, 10$. Since our labelled data is severely biased towards "younger" age categories we cannot simply learn multinomial distributions $P(A|X_i)$ for each feature based on the relative frequencies of their followers (see Table 1). To smooth out noisy observations of less popular accounts we use a hierarchical Bayesian model. Inference is simplified by using the Bernoulli's conjugate distribution, the beta distribution

$$\mathrm{Beta}(\mu_{ia}|b_{ia}, c_a) \tag{5}$$

on the Bernoulli parameters μ_{ia}. We seek hyper-parameters b_{ia}, c_{ia} of the prior $\mathrm{Beta}(\mu_{ia}|\mathbf{X}, \mathbf{A})$, which do not have a large effect when ample data is available, but produce sensible distributions when it is not. To achieve this we set c_a to be

[6] We only consider cases where $X_i = 1$ since the Twitter graph is sparse: In the full Twitter graph there are 7×10^8 nodes with 5×10^{10} edges, which implies a density of 1.6×10^{-7}, i.e., the default is to follow nobody. Hence, not following an account does not contain enough information to justify the additional computational cost.

constant across all features X_i (hence dropping the i subscript) and proportional to the total number of observations n_a in each age category (the count column in Table 1). We then set $b_{ia} \propto \frac{n_a n_i}{K}$, where $K = 7 \times 10^8$ is the total number of Twitter users and n_i is the number of Followers of feature i (the Followers column of table 5 for @williamockam's features). Then, the expected prior probability that user χ Follows account i is $\mathbb{E}[\mu_{ia}|A = a] = \frac{b_{ia}}{b_{ia}+c_a} = \frac{n_i}{K+n_i}$, i.e., it is constant across age classes and varies in proportion to the number of Followers across features. The effect of this procedure is to reduce the model confidence for features where data is limited. Due to conjugacy, the posterior distribution on μ_{ia} is also Beta distributed. Integrating out μ_{ia} we obtain

$$P(X_i = 1|A{=}a, \mathbf{X}, \mathbf{A}) = \int_0^1 P(X_i{=}1|\mu_i, A)P(\mu_i|\mathbf{X}, \mathbf{A}, A)d\mu_i \tag{6}$$

$$= \int_0^1 \mu_{ia} P(\mu_{ia}|\mathbf{X}, \mathbf{A})d\mu_{ia} = \mathbb{E}[\mu_{ia}|\mathbf{X}, \mathbf{A}] = \frac{n_{ia}+b_{ia}}{n_a+b_{ia}+c_a}\,, \tag{7}$$

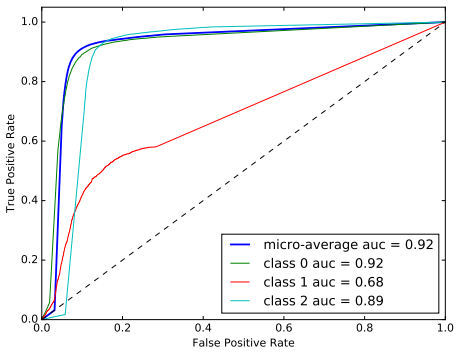

Fig. 2. Receiver operator characteristics for three class age detection (0 = under 18, 1 = 18–45, 2 = 45+). The dashed line indicates random performance.

where n_{ia} is the number of labelled Twitter users in age category a who Follow feature X_i, which are given in Table 5 for the @williamockam features and n_a is the number of Twitter users in category a in the ground-truth (See Table 1). Performing this calculation yields the likelihoods for the @williamockam features shown in Table 6. We are now able to compute the predictive distribution in (1) to infer the age of an arbitrary Twitter user. The predictive distribution for @williamockam is shown in Fig. 4 and is calculated by taking the product of the likelihoods from Table 6 with the prior in (2) and normalising.

The generative process in our model for the likelihood term in (1) is as follows.

1. Draw an age category $A \sim \text{Cat}(\pi)$
2. For each feature i draw $\mu_{ia} \sim \text{Beta}(\mu_{ia}|b_{ia}, c_a)$
3. For each account draw the Follows: $X_i \sim \text{Ber}(\mu_{ia})$

In Table 7, we report the five features with the highest posterior age values of $P(A|X_i = 1)$ for each age category. The account descriptions are taken from the first line of the relevant Wikipedia page. The youngest Twitter users are characterised by an interest in internet celebrities and computer games players. Music genres are important in differentiating all age groups from 12–45. 25–34 year olds are in part marked by entities that saw greater prominence in the

Table 7. The most discriminative features based on the posterior distribution over age in (6). Descriptions are taken from the 1^{st} line of their Wikipedia pages. See the git repo for a full table with probabilities and handles.

<12	**12–13**	**14–15**	**16–17**	**18–24**
vlogger	child presenter	child singer	singer	metalcore band
minecraft gamer	YouTuber	child singer	metalcore band	rock band
internet personality	child actress	child singer	deathcore singer	rapper
vlogger	child actress	child singer		
gaming commentator	girl band	child singer	electronic band	rock band
25–34	**35–44**	**45–64**[7]	**65+**	
hip hop duo	hip hop artist	evangelist	political journalist	
boy band	rapper	evangelist	retired cyclist	
boy band	history channel	evangelist	golf channel	
comedian	record label	faith group	retired rugby player	
adult actress	boxer	faith magazine	boxer	

past. This group is also distinguished by an interest in pornographic actors. Age categories 45–54 and 55–64 have the same top five and are differentiated by their interest in religious topics. Users older than 65 are identifiable through an interest in certain sports and politics.

Experimental Evaluation

We demonstrate the viability of our model for age inference in huge social networks by applying it to 700m Twitter accounts. We conducted three experiments: (1) We compare our approach with the language-based model by Nguyen (2013), which can be considered the state of the art for age inference. (2) We compare our age inference results with the survey by Duggan (2013). (3) We assess the quality of our age inference on a 10% hold-out set of ground-truth labels and compare it with results obtained from inference based solely on the prior derived from census and survey data in (2) for age prediction.

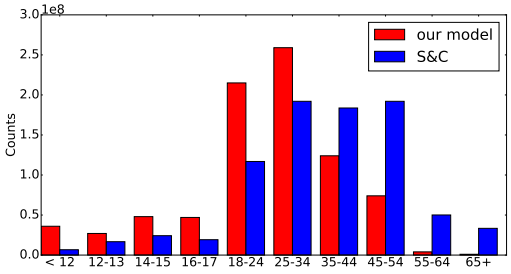

Fig. 3. Red bars show #accounts that our model allocated to each age class using the mode of the predictive posterior. Blue bars show #accounts that would have been allocated to each age class if ages were drawn from the Survey and Census (S&C) prior.

Comparison with Dutch Language Model

For comparison with the state-of-the-art work of Nguyen (2013) based on linguistic features (Dutch tweets) we consider the performance of our model as a three-class classifier using age bands: under 18, 18–44 and 45+.

Table 8. Statistics for age prediction on a held-out test set.

		<12	**12–13**	**14–15**	**16–17**	**18–24**	**25–34**	**35–44**	**45–54**	**55–64**	**≥65**
	Test Cases	651	1,731	2,678	2,036	2,670	776	230	5,058	5,145	20,487
Ours	Recall	**0.19**	**0.20**	**0.38**	**0.23**	**0.33**	**0.25**	0.18	**0.32**	**0.41**	**0.30**
	Precision	**0.22**	**0.33**	**0.36**	**0.24**	**0.31**	**0.15**	**0.07**	**0.14**	**0.19**	**0.79**
	Micro F1	**0.31**									
S&C	Recall	0.01	0.02	0.02	0.03	0.14	0.23	**0.23**	0.22	0.06	0.04
	Precision	0.02	0.04	0.06	0.05	0.06	0.02	0.01	0.12	0.12	0.49
	Micro F1	0.07									

Fig. 9 lists the performance of our age inference algorithm on a 10% hold-out test set and the Dutch Language Model (DLM) proposed by Nguyen (2013). The corresponding performance statistics are shown in Table 9.

Both methods perform equally well with a Micro F1 score of 0.86. The precision and recall show that the DLM approach is efficient, extracting information from only a small training set (support). This is because significant engineering work went into labelling and feature design. In contrast, our feature generation process is automatic and scalable. While we do not achieve the same performance for the lower age categories, for the oldest age category, our approach performs substantially better than the method by Nguyen (2013), suggesting that a hybrid method could perform well. We leave this for future work.

The major advantages of our model to the state-of-the-art approach are twofold: First, we have applied our age inference to 700m Twitter users, as opposed to being limited to a sample of Dutch Twitter users with a relatively high number of Tweets. Second, generating our training set is fully automatic and relies only on Twitter data[8], i.e., no manual labelling or verification is required.

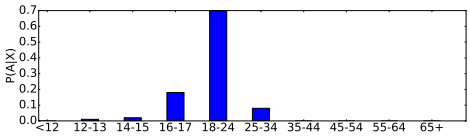

Fig. 4. Posterior age distribution for @williamockam.

Fig. 2 shows the areas under the receiver-operator characteristics (ROC) curves for our three-class model. The curves are generated by measuring the true positive and false positive rates for each class over a range of classification thresholds. A perfect classifier has an area under the curve (AUC) equal to one, while a completely random classifier follows the dashed line with an $AUC = 0.5$. Performance is excellent for classes under 18 and over 45, but weaker for 18–45 where training data was limited, which we note as an area for improvement in future work.

Comparison with Survey and Census Data

We report results on inferring the age of arbitrary Twitter users with the ten category model. Fig. 3 shows aggregate classification results for 700m Twitter

[8] Nguyen (2013) used additional LinkedIn data for labelling

Table 9. Performance for three-class age model.

	Our Approach			DLM (Nguyen, 2013)		
	<18	**18–44**	**≥45**	**≤18**	**18–44**	**≥45**
Support	7,096	3,676	30,690	1,576	608	310
Precision	0.76	0.39	**0.96**	**0.93**	**0.67**	0.82
Recall	0.68	0.50	**0.95**	**0.98**	**0.75**	0.45
Micro F1	**0.86**			**0.86**		

accounts compared with expected counts based on survey data (S&C) Duggan (2013). Our model predicts that over 50% of Twitter users are between 18 and 35, i.e., the bias of the original training set has been removed due to the Bayesian treatment. It is likely that S&C under-represents young people as we did not factor in the increased rates of technology uptake amongst the younger people when converting census data.

Quality Assessment

In the following, we assess the quality of our age inference model (10 categories) on a 10% hold-out test data set.

Table 8 shows the performance statistics for this experiment. The majority of the test cases are in the younger age categories (due to the bias of young people revealing their age) and in older age categories (due to the inclusion of grandparents and retirees). Table 8 shows that the precision depends on the size of the data (e.g., predicting 25–44 year categories is hard) whereas the recall is fairly stable across all age categories.[9] Our model significantly outperforms an approach based only on the survey and census data (S&C), which we use as a prior. This highlights the ability of our model to adapt to the data.

Conclusion

We proposed a probabilistic model for age inference in Twitter. The model exploits generic properties of Twitter users, e.g., whom/what they follow, which is indicative of their interests and, therefore, their age. Our model performs as well as the current state of the art for inferring the age of Twitter users without being limited to specific linguistic or engineered features. We have successfully applied our model to infer the age of 700 million Twitter users demonstrating the scalability of our approach. The method can be applied to any attributes that can be extracted from user profiles.

Acknowledgements

This work was partly funded by an Industrial Fellowship from the Royal Commission for the Exhibition of 1851. The authors thank the anonymous reviewers for providing many improvements to the original manuscript.

[9] Without the inclusion of grandparents and retirees in the training set, the predictive performance would rapidly drop off for ages greater than 35.

Bibliography

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

J. D. Burger, J. Henderson, G. Kim and G. Zarrella. Discriminating gender on Twitter. In *EMNLP*, 2011.

Z. Cheng, J. Caverlee and K. Lee. You are where you tweet: a content-based approach to geo-locating Twitter users. In *CIKM*, 2010.

A. Culotta, R. K. Nirmal and J. Cutler. Predicting the Demographics of Twitter Users from Website Traffic Data. In *AAAI*, 2015.

M. Duggan and J. Brenner. *The Demographics of Social Media Users—2012*. Retrieved Sep 12 2015 from `http://tinyurl.com/jk3v9tu`

Q. Fang, J. Sang, C. Xu, and M. S. Hossain. Relational user attribute inference in social media. In *IEEE Transactions on Multimedia*, 17(7), 1031-1044. 2015.

T. Grainger and T. Potter. Solr in action. Manning Publications Co. 2014

P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang and R. Zadeh. WTF: The Who to Follow Service at Twitter. In *WWW*, 2013.

M. Kosinski, D. Stillwell, and T. Graepel. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *PNAS*, 110(15), 2013.

W. Liu and D. Ruths Whats in a name? using first names as features for gender inference in twitter. *AAAI Spring Symposium on Analyzing Microtext*, 2013.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. In *Annual Review of Sociology*, 27(1):415–444, 2001.

A. Mislove, S. Lehmann, and Y. Y. Ahn. Understanding the Demographics of Twitter Users. In *ICWSM*, 2011.

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.

D. Nguyen, A. Noah, A. Smith, and Carolyn P. Rose. Author age prediction from text using linear regression. In *LaTeCH*, 2011.

H. Oktay, A. Firat, and Z. Ertem. Demographic Breakdown of Twitter Users: An Analysis based on Names. In *BIGDATA*, 2014.

M. Pennacchiotti and A. M. Popescu A machine learning approach to twitter user classification. In *ICWSM*, 2011.

D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in Twitter. In *SMUC*, 2010.

J. Schler, M. Koppel, S. Argamon and J. W. Pennebaker Effects of age and gender on blogging. In *AAAI-CAAW*, 2006.

R. Wysocki and W. Zabierowski. Twisted Framework on Game Server Example. In *CADSM*, 2011.

U.S. Census Bureau, 2010 Census. Profile of General Population and Housing Characteristics: 2010 Retrieved Sep 12, 2015 from `https://goo.gl/VAGMNl`

U.S Census Bureau, American Community Survey, 2014 Grandparent Statistics Retrieved Nov 15, 2015 from `https://goo.gl/CqGXWI`