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,
WWW home page: http://wp.doc.ic.ac.uk/sml
[2] Starcount Insights, 2 Riding House Street, London, UK

Appendix

Age Extraction Using REGEX Matching of Descriptions

We extracted user ages from the free text Twitter description using UNIX scripting REGEX matching tools. The exact REGEX strings are included in Listing 1.1. An initial run of the REGEX revealed some frequent false positives with terms like 'I feel like I am 80' or 'I am more than 10', which were manually corrected for in the final iteration.

Listing 1.1. Regex matching run against Twitter descriptions. The code detects age references in English, German, French and Portuguese. Terms including 'feel like', 'think I am' and 'more / less than' were a major source of error in early versions, which led us to write a REGEX that explicitly removes them.

```
awk '{for (i=2; i<=NF; i+=2) {gsub (/,/, "p1p2p3p4p", $i)} print $0 }' FS="\"" OFS="\""
awk '{print $2, $3, $6}' FS="," OFS="," temp |
sed 's/p1p2p3p4p/\,/g' | egrep -i "[\'a][mn] [][0-9][0-9][ \,\.\!\;y][ \/yea]|
[ua]is[ ][0-9][0-9][ \,\.\!\;a][an]|bin[ ][0-9][0-9][ ]| [hg]o[ ][0-9][0-9][ a][an][on
sed "s/.*[hghia\'][mnso][ ]\([0-9][0-9]\)[ \,\.\!\;ya]
[ \/yean].*/\1/I;s/.*bin[ ]\([0-9][0-9]\)[ ].*/\1/I" temp1.csv |
egrep -v -i "more than [0-9][0-9]| "think i am [0-9][0-9]|think i'm [0-9][0-9]|
i feel like [0-9][0-9] | depuis [0-9][0-9]| [a-ln-su-z] an [0-9][0-9]" > temp2.csv
awk '{getline a < "temp2.csv"; print $0","a}' temp1.csv > temp3.csv
```

The Most Popular Accounts Followed by Labelled Users

We split the Followers into ten age categories. Table 1 shows that general trends across features are that the age distribution is peaked towards "younger" ages and that not many older people reveal their age for the top features. The Followers column gives the total number of Followers of each feature across the Twitter network. There is a Pearson correlation of 0.86 between the support and the total Follower count for our data set.

Table 1. The accounts with the highest support within the labelled data set.

Twitter Handle	Support	<12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	≥65	Followers
justinbieber	**20,359**	1517	5179	5737	4202	3073	412	99	67	34	38	8.7×10^7
katyperry	**18,395**	1467	4180	4410	3604	3575	701	158	124	75	102	9.2×10^7
taylorswift13	**15,199**	1207	3417	3674	3045	2919	507	113	117	79	122	8.1×10^7
selenagomez	**14,264**	1270	3578	3691	2847	2339	367	76	43	26	27	4.6×10^7
ArianaGrande	**13,512**	1254	3404	3604	2631	2172	319	50	40	19	20	4.1×10^7
ddlovato	**13,259**	1099	3284	3562	2741	2135	301	53	37	19	28	3.8×10^7
onedirection	**12,834**	979	3472	3778	2767	1622	138	43	20	7	8	3.0×10^7
Harry_Styles	**12,830**	912	3468	3936	2751	1581	120	24	15	9	13	2.9×10^7
NiallOfficial	**12,498**	858	3431	3895	2702	1468	90	24	15	8	8	2.7×10^7
YouTube	**11,688**	926	2496	2687	2193	2287	495	183	154	99	169	6.4×10^7

The Most Discriminative Features in Each Category

For each feature we calculate the posterior probability of Following that feature given the user's age. We sort the posteriors within each age category and present the accounts with the five highest values in Table 2.

Table 2. In the model the features are popular Twitter accounts. This table contains the posterior distributions $p(X = 1|A = a)$ over age for the five most discriminative (useful) features in each age class.

twitter_handle	description	<12	12–13	14–15	16–17	18–24	25–34	35–44	45–54	55–64	65+
Under 12-year olds											
RosannaPansino	vlogger	**0.40**	0.22	0.15	0.09	0.07	0.02	0.01	0.01	0.01	0.02
AntVenom	minecraft gamer	**0.40**	0.25	0.15	0.09	0.06	0.02	0.01	0.01	0.01	0.01
Bajan_Canadian	internet personality	**0.37**	0.25	0.17	0.10	0.06	0.02	0.00	0.01	0.01	0.01
shaycarl	vlogger	**0.36**	0.20	0.14	0.10	0.07	0.04	0.02	0.02	0.02	0.02
InTheLittleWood	gaming commentator	**0.34**	0.23	0.16	0.11	0.08	0.02	0.01	0.01	0.01	0.02
12–13 year olds											
ivandorschner	child TV presenter	0.18	**0.27**	0.20	0.11	0.09	0.03	0.03	0.02	0.03	0.04
Vikkstar123	youtuber	0.29	**0.26**	0.20	0.14	0.07	0.02	0.01	0.01	0.01	0.02
PeytonList	child actress	0.29	**0.25**	0.20	0.14	0.07	0.02	0.01	0.01	0.01	0.01
G_Hannelius	child actress	0.31	**0.25**	0.18	0.13	0.07	0.02	0.02	0.01	0.01	0.01
Cimorelliband	girlband	0.20	**0.25**	0.23	0.17	0.09	0.02	0.01	0.01	0.01	0.01
14–15 year olds											
therealsavannah	child pop singer	0.10	0.18	**0.27**	0.21	0.12	0.02	0.01	0.03	0.03	0.03
jessicajarrell	child pop singer	0.12	0.21	**0.26**	0.24	0.10	0.02	0.01	0.01	0.01	0.01
TheDylanHolland	child R&B singer	0.12	0.22	**0.26**	0.24	0.11	0.02	0.01	0.01	0.01	0.01
OfficialBirdy	child singer	0.10	0.17	**0.26**	0.24	0.13	0.04	0.01	0.02	0.02	0.02
officialjman	child singer	0.10	0.18	**0.26**	0.28	0.13	0.02	0.01	0.01	0.01	0.01
16–17 year olds											
TannerPatrick	singer	0.05	0.13	0.25	**0.30**	0.18	0.03	0.01	0.01	0.01	0.01
TheWordAlive	metalcore band	0.04	0.11	0.19	**0.29**	0.22	0.09	0.02	0.01	0.01	0.01
MitchLuckerSS	deathcore singer	0.05	0.14	0.23	**0.29**	0.20	0.04	0.01	0.01	0.01	0.02
metrostation	electronic band	0.03	0.07	0.15	**0.29**	0.18	0.10	0.04	0.06	0.06	0.03
BreatheCarolina	electronic band	0.06	0.15	0.22	**0.29**	0.19	0.06	0.01	0.01	0.01	0.01
18–24 year olds											
wecameasromans	metalcore band	0.05	0.13	0.22	0.28	**0.21**	0.06	0.01	0.01	0.01	0.01
Sum41	rock band	0.07	0.11	0.18	0.24	**0.21**	0.09	0.02	0.02	0.03	0.03
hopsin	rapper	0.04	0.09	0.13	0.19	**0.20**	0.09	0.09	0.06	0.05	0.07
Diablo	computer game	0.03	0.06	0.09	0.13	**0.20**	0.17	0.09	0.05	0.06	0.12
paparoach	rock band	0.04	0.09	0.14	0.19	**0.20**	0.12	0.07	0.06	0.06	0.04
25–34 year olds											
icp	hip hop duo	0.02	0.04	0.05	0.09	0.19	**0.37**	0.09	0.04	0.05	0.05
kevinrichardson	boyband member	0.02	0.03	0.05	0.09	0.16	**0.35**	0.12	0.07	0.06	0.04
skulleeroz	boyband member	0.02	0.04	0.06	0.09	0.16	**0.33**	0.12	0.07	0.06	0.05
LeeEvansNews	comedien	0.02	0.03	0.06	0.07	0.17	**0.32**	0.09	0.08	0.09	0.09
miko_lee	adult actress	0.04	0.03	0.03	0.05	0.17	**0.31**	0.08	0.07	0.08	0.14
35–44 year olds											
djspooky	hip hop artist	0.01	0.02	0.03	0.02	0.04	0.15	**0.45**	0.14	0.06	0.08
Mr_Mike_Jones	rapper	0.01	0.01	0.01	0.01	0.03	0.14	**0.44**	0.16	0.09	0.10
HISTORYTV18	history TV channel	0.02	0.03	0.03	0.05	0.09	0.14	**0.36**	0.10	0.06	0.13
TopDawgEnt	record label	0.03	0.07	0.05	0.07	0.11	0.11	**0.36**	0.09	0.03	0.07
DannySwift	boxer	0.02	0.03	0.04	0.07	0.06	0.09	**0.33**	0.12	0.08	0.16
45–54 and 55–64-year olds (identical most-discriminant features)											
JohnBevere	evangelist	0.00	0.00	0.00	0.00	0.01	0.01	0.07	**0.36**	**0.39**	0.15
edstetzer	evangelist	0.00	0.00	0.00	0.00	0.00	0.01	0.07	**0.36**	**0.39**	0.16
ChristineCaine	evangelist	0.00	0.00	0.01	0.00	0.01	0.01	0.07	**0.36**	**0.38**	0.15
womenoffaith	faith group	0.00	0.00	0.00	0.00	0.00	0.02	0.08	**0.36**	**0.38**	0.16
RELEVANT	faith magazine	0.00	0.01	0.00	0.01	0.01	0.01	0.07	**0.35**	**0.38**	0.17
People over 65											
afneil	political journalist	0.00	0.00	0.01	0.01	0.02	0.02	0.04	0.17	0.25	**0.48**
Chris_Boardman	retired cyclist	0.01	0.01	0.01	0.02	0.01	0.01	0.04	0.17	0.25	**0.47**
SkySportsGolf	golf TV channel	0.01	0.02	0.02	0.02	0.03	0.01	0.04	0.16	0.22	**0.46**
IamAustinHealey	retired rugby player	0.04	0.02	0.01	0.01	0.01	0.01	0.04	0.17	0.25	**0.45**
anthonyfjoshua	boxer	0.02	0.03	0.03	0.04	0.09	0.06	0.03	0.08	0.15	**0.45**