PLOS ONE

RESEARCH ARTICLE

What demographic attributes do our digital footprints reveal? A systematic review

Joanne Hinds*, Adam N. Joinson

School of Management, University of Bath, Bath, United Kingdom

* J.Hinds@bath.ac.uk

Abstract

To what extent does our online activity reveal who we are? Recent research has demonstrated that the digital traces left by individuals as they browse and interact with others online may reveal who they are and what their interests may be. In the present paper we report a systematic review that synthesises current evidence on predicting demographic attributes from online digital traces. Studies were included if they met the following criteria: (i) they reported findings where at least one demographic attribute was predicted/inferred from at least one form of digital footprint, (ii) the method of prediction was automated, and (iii) the traces were either visible (e.g. tweets) or non-visible (e.g. clickstreams). We identified 327 studies published up until October 2018. Across these articles, 14 demographic attributes were successfully inferred from digital traces; the most studied included gender, age, location, and political orientation. For each of the demographic attributes identified, we provide a database containing the platforms and digital traces examined, sample sizes, accuracy measures and the classification methods applied. Finally, we discuss the main research trends/findings, methodological approaches and recommend directions for future research.

Citation: Hinds J, Joinson AN (2018) What demographic attributes do our digital footprints reveal? A systematic review. PLoS ONE 13(11): e0207112. https://doi.org/10.1371/journal.pone.0207112

Editor: David Garcia, ETH Zurich, SWITZERLAND

Received: August 9, 2017

Accepted: October 23, 2018

Published: November 28, 2018

Data Availability Statement: Data used in the review is found in the manuscript and supporting PRISMA checklist (S1 Table).

Funding: This work was funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1) to ANJ, www.crestresearch.ac.uk. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

We use the internet and digital devices in many aspects of our lives—to communicate, work, shop, bank, etc. Approximately 50% of the world's population now use the internet [1] and current estimates predict that around 30 billion devices will be connected to each other by 2020 [2]. With every click or online interaction, digital traces (also known as 'digital footprints') are created and captured (usually automatically), providing a detailed record of a person's online activity. This constant generation of digital data provides opportunities to harvest and analyse 'big data' at an unprecedented scale and gain insights to an individual's demographic attributes, personality, or behaviour. Such information can be incredibly valuable for organisations (e.g. marketers, researchers, governments) hoping to understand digital data and predict future outcomes. Computer and data scientists have used digital data to successfully predict events including: the spread of flu in the US [3], box office revenue for new films [4], election results [5] and reactions or opinions to events such as the Arab Spring [6].

Predicting individuals' demographic attributes has become a rapidly growing area of research in recent years. However, the innumerable attributes, traces and platforms available,

combined with diverse methodological approaches means that research is extremely disparate and published in a variety of journals and conference proceedings. In this article we systematically review existing research to address the questions: (i) what demographic attributes can be predicted from digital traces? (ii) what traces and platforms have been studied? and (iii) how effective are current methodologies and predictions? In synthesising this information, we review current findings and offer recommendations for future research.

Background

Inferring individuals' demographic attributes has a long history in fields such as computer forensics and marketing. For instance, computer forensic investigators seek to determine the legitimacy of communications and online activities in order to prevent crimes such as bullying, harassment, or the unauthorised conveyancing of information. Marketers seek to establish who people are in order to target products and services to their desired audiences. In some circumstances, inferring certain attributes such as gender, approximate age and ethnicity may be relatively easy if individuals disclose this information or if they are visible in photographs. Conversely, if such information is absent, or if individuals try to masquerade as someone else, inferring attributes accurately becomes much more difficult.

One way of addressing this challenge is to analyse digital traces that 'objectively reveal' a person's identity. For instance, personality researchers have suggested that individuals leave *behavioural residue* (unconscious traces of actions that may objectively depict their identity, e.g. web browsing histories) when they interact online (e.g. [7,8]). Thus, behavioural residue such as language patterns, smartphone metrics and meta-data (e.g. no. posts, no. followers), provide opportunities to infer demographic attributes with computational techniques (e.g. natural language processing, machine learning) that would be too complex for humans to process. To date, numerous studies have predicted demographic attributes accurately from digital traces including Facebook likes [9–11], smartphone logs [12–15], Flickr tags [16], and language-based features [17–20].

Network analysis is another approach that can be useful for attribute inference. Researchers studying social networks often examine if people who are similar in age, interests, location etc. tend to be closely located in their social networks. Homophily–the notion that *birds of a feather flock together* is incredibly useful within this context, because gathering data from a person's network may improve the predictive accuracy of individuals for whom we have little, or distorted, data. The downside is that highly sensitive, or private attributes may be identifiable from other people's data. Indeed, this possibility raises numerous ethical and privacy concerns about what true 'informed consent' is, and what can be considered 'personally identifiable information' when hidden traits can be discovered using a combination of seemingly innocuous unrelated digital traces. For instance, the data analytics company, Cambridge Analytica recently came under scrutiny in the news for using data collected from approximately 87 million individuals' Facebook accounts without their explicit consent [21]. The data was supposedly used to create targeted advertisements, which attempted to influence people's voting preferences in the 'Vote Leave' campaign in Britain's European Referendum, and Donald Trump's 2016 presidential election [21,22]. If we are going to be able to critique such efforts, and identify what information about a person should be considered 'protected', then it is important that we know what the current state-of-the-art is in terms of predicting attributes from digital traces. These joint concerns motivate the present systematic review.

Although demographic inference is almost entirely reported in computer science journals and conferences, there is extensive social psychology research that has explored how demographic attributes (particularly gender and age) relate to certain behaviours, such as language

[23], technology use [24–26] and social activities [27]. Unfortunately, the two fields tend to remain distinct, with each adopting different conventions in terms of focus, methods and publishing. Computer scientists typically focus on improving methods and prediction outcomes, whereas psychologists aim to understand people's behaviour. As such, the majority of research identified by our search was published within computer science outlets. However, we seek to bridge this gap, wherever possible by discussing related psychology research. In the following section we outline our methods and search criteria.

Method

Search strategy

We systematically searched for articles published up until October 2018 (i.e. our search had a cut-off date of 30th September 2018) using four strategies. First, we performed searches in the Web of Science, IEEE and ACM online libraries for all relevant articles by searching for keywords and topic-related terms. These included (predict* or identify or detect* or Facebook or Twitter or Instagram or YouTube) and (demographic* or age or gender) and (digital or internet or online or computer-mediated) and (social* or web* or mobile* or sms or big data). Second, we identified all first authors with 3 or more papers and individually searched for further relevant papers written by these authors (identified via Google Scholar, Research Gate and their personal university web pages). Third, we hand searched the references of the papers that met our inclusion criteria and retrieved all further references. We performed this step iteratively on each paper added to the set, until no further papers were retrieved. Fourth, experts in the field were contacted to request information about any other studies that we might not have located. The search generated no studies that were in non-English languages. Our search strategy and statistics are reported in accordance with the PRISMA (Preferred Reporting of Items for Systematic Reviews and Meta-Analysis, www.prisma-statement.org) guidelines. The supporting PRISMA checklist is available as supporting information (see the PRISMA checklist included as S1 Table).

Inclusion criteria

To be included in the review, studies had to: (i) report findings where at least one demographic attribute was predicted/inferred from at least one form of digital footprint, (ii) the method of prediction had to be automated—this could include supervised, semi-supervised or unsupervised machine learning and (ii) the digital footprints could either be public (e.g. tweets) or private (e.g. clickstreams). All studies meeting these criteria were included in the review. The search generated a total set of 327 papers. The PRISMA flow chart detailing the papers retrieved and refined according to our criteria is displayed in Fig 1.

Data collection

For each demographic attribute we extracted the following data from each article: platform and type of digital trace studied, classes used for classification (e.g. unemployed, employed for 'occupation'; divorced, married, single for 'family and relationships'), sample sizes, predictive features, accuracy measures (including accuracy (%), area under the ROC curve (AUC), F1-score, precision, and recall), types of classifier used, and publication data (i.e. year of publication, reference data, and the quality of the conference/journal). This data is available as a series of tables in the supplementary materials (S2–S16 for each demographic attribute, respectively).

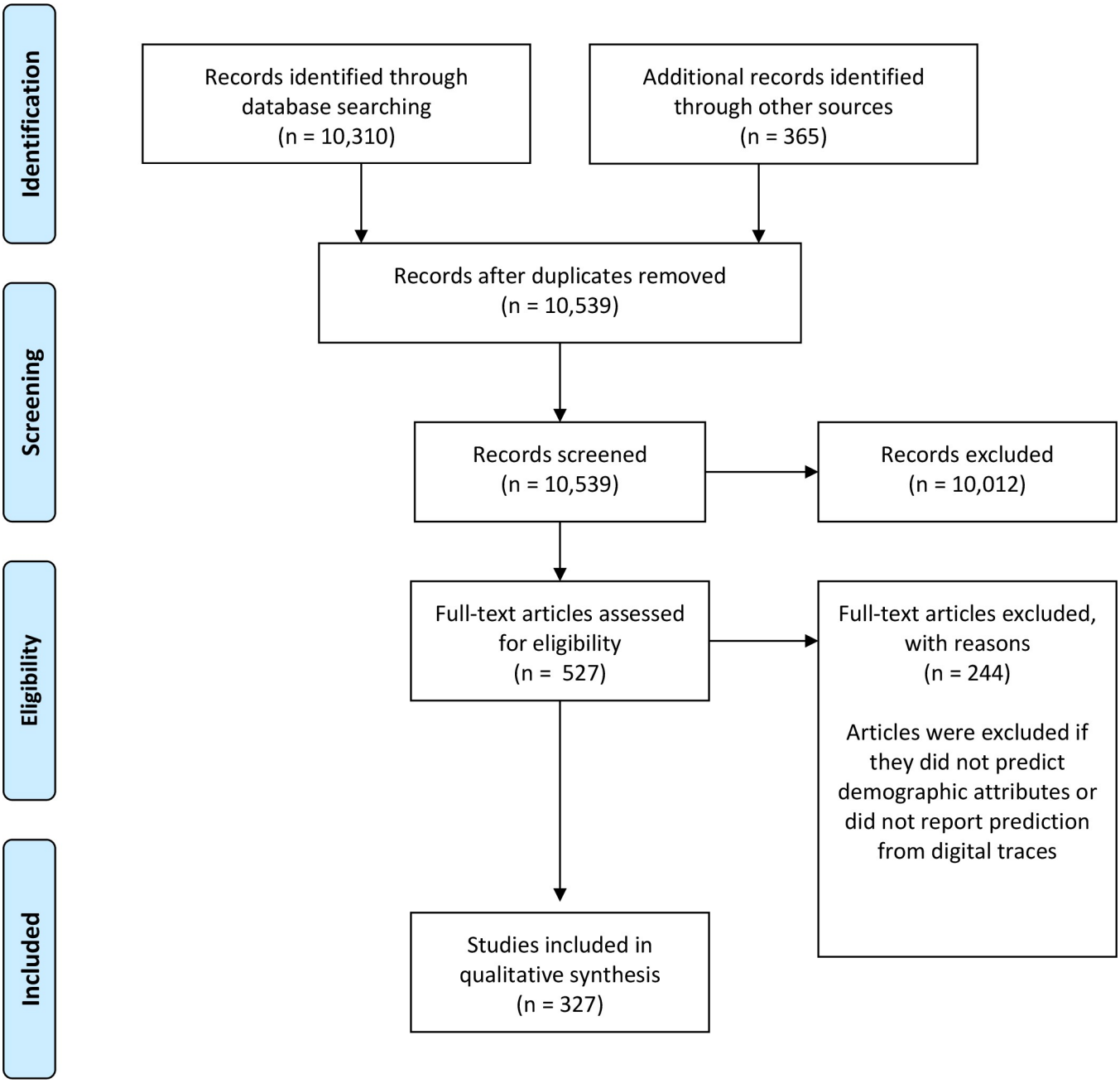

Fig 1. PRISMA Flowchart summarising study retrieval and selection.

https://doi.org/10.1371/journal.pone.0207112.g001

Study quality

To our knowledge, there are no existing protocols for assessing the quality of machine learning studies. As such, we assessed the quality of the articles by classifying them on the rank of their publication outlet (i.e. peer-reviewed conference proceedings and journals). We used highly regarded ranking systems of scientific value, specifically the SCImago Journal Rank (SJR) indicator (www.scimagojr.com) for journal articles, and the Excellence in Research in Australia (ERA), Qualis (2012), and Microsoft Academic's (MSAR 2014) field ratings for conferences

databases for conference proceedings. All values were taken from the rankings made in 2018. We scored articles across four categories as follows:

- High quality–journal articles in quartile 1 (Q1), and conference articles ranked as A, A1, or A2
- Medium quality–journal articles in quartile 2 (Q2), and conference articles ranked as B, B1, B2, B3, or B4
- Low quality–journal articles in quartile 3 (Q3) or quartile 4 (Q4), and conference articles ranked as B5, or C.
- Not reported (NR)–journal and conference articles that were not indexed in any of the ranking systems.

We assigned articles that were ranked in multiple categories or quartiles to the highest ranking, for example, articles ranked as B and B5 were classified as 'medium quality' (rather than 'low quality'). A similar approach was used by Azucar, Marengo and Settanni [28] in their review of personality prediction from digital footprints.

Results

Our search generated a total of 327 articles examining 14 demographic attributes including: gender (n = 241), age (n = 157), location (n = 32), political orientation (n = 33), sexual orientation (n = 7), family and relationships (n = 19), ethnicity and race (n = 20), education (n = 16), income (n = 13), language (n = 9), health (n = 9), religion (n = 8), occupation (n = 22), and social class (n = 1). Many of the articles studied multiple demographic attributes–Fig 2 displays the proportion of attributes studied across our entire dataset.

One of the reasons the number of articles retrieved for gender and age were markedly higher than the other attributes was because of a series of author profiling workshops (PAN) at the Conference and Labs of the Evaluation Forum (CLEF) (https://pan.webis.de). The workshops were held annually and involved teams reporting their solutions to gender and age profiling from a series of provided datasets. The results from the workshops resulted in 105 articles reporting gender, and 63 articles reporting age predictions.

Fig 3. displays the number of articles published per year (from 2000 up until Oct 2018) along with number published per quality quartile. The findings highlight that over the last few years, the majority of articles have been published in medium and high-quality journals and conference proceedings. Although a reasonable number of articles were published in journals/conferences that were not indexed in scientific databases, (i.e. we cannot assess the quality of those studies), the number of low-quality articles appears to be very low. In the remainder of this section we discuss the main research findings and trends for each demographic attribute.

Gender

Gender inference has a long history across numerous disciplines including computer forensics, linguistics and social psychology. In contrast to many other demographic attributes (with the exception of age), extensive research on inferring gender in offline contexts (e.g. conversations, texts, essays) existed prior to the digital-based studies that have proliferated in recent years. As such, it is perhaps unsurprising that gender is the most widely studied attribute within our set (241 articles in total, 136 independent articles, and 105 from the PAN workshops) and is often studied in tandem with age. Table 1 provides an overview of the articles published and associated references per platform. Table 2 provides an overview of the articles published and associated references per predictor. Because of the vast number of articles identified in the search,

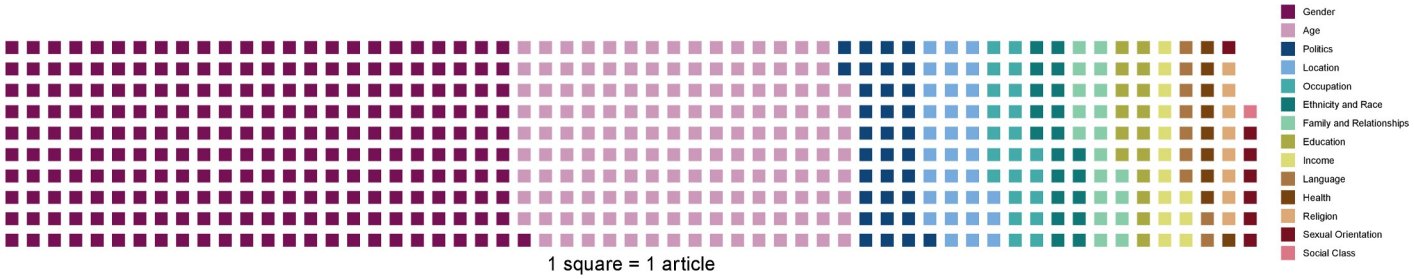

Fig 2. Waffle chart highlighting the proportion of demographic attributes comprising our dataset.

https://doi.org/10.1371/journal.pone.0207112.g002

we discuss the main trends and findings identified over a series of sub-sections, outlined below.

Language. An individual's choice of language is largely related to their gender, a phenomenon that has been extensively studied by sociolinguists for decades, e.g. [263–265] in written texts, such as essays, poems, scientific articles or speech transcripts, e.g. [266,267]. In general, males and females have been found to differ in numerous ways; typically females tend to use more emotion words, negations and hedges, and males tend to use more assertion, swear words, and long words (over six letters in length), e.g. [268,269]. Lakoff [265] argued that these differences were caused by power differences in society, where women's lack of power would cause them to adopt more polite and uncertain forms of language. For comprehensive discussions on gender and language, see the work by Coates [270], Lakoff [271], or Holmes and Mayerhoff [272].

On the internet, people's interactions and communication patterns can change markedly for numerous reasons: a) non-verbal and prosodic cues are lost, b) the design of social media platforms, websites etc. influence the way people converse, and c) individuals may become more conscious of how they present themselves towards others. Digital language traces, combined with computational analytics or tools, such as natural language processing (NLP), and Linguistic Inquiry Word Count (LIWC) [273] enable researchers to study language and gender at mass scale, and in more naturalistic environments. In recent years, gender inference

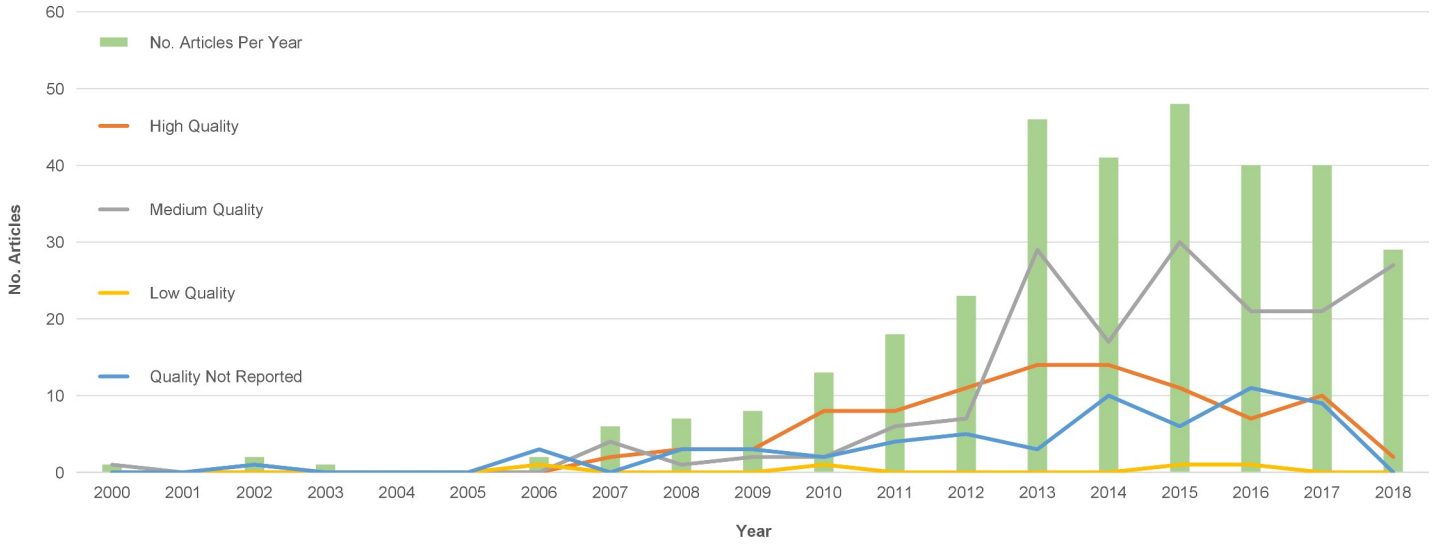

Fig 3. Number of articles published per year and by quality of publication.

https://doi.org/10.1371/journal.pone.0207112.g003

Table 1. Number of articles predicting gender, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (134)	Twitter (106)	[17,20,29–155]
	Facebook (7)	[10,156–161]
	YouTube (2)	[162,163]
	Netlog (2)	[164,165]
	Flickr (3)	[16,166,167]
	Pintrest (1)	[168]
	Instagram (1)	[169]
	Sina Weibo (1)	[170]
	Social Media (General) (25)	[93–100,102,103,127–133,135–142,171]
Digital Devices (22)	Smartphones (25)	[12–15,172–192]
	Tablets (1)	[193]
Websites (23)	News sites (3)	[194–196]
	Websites (6)	[179,197–201]
	IMDB (1)	[202]
	Hotel Reviews (25)	[93–100,102,103,127–133,135–142,171]
	Movielens (2)	[203,204]
	Crowdfunding Essays (1)	[88]
Blogs (58)	Blogger.com (4)	[18,205–207]
	Blogs (General) (51)	[88,93–100,102,103,127–133,135–142,171,208–232]
	Vietnamese Blogs (1)	[233]
	Tumblr (1)	[234]
Emails (9)	NR (9)	[196,235–242]
Radio (3)	Last.fm (3)	[243–245]
Search Engines (2)	Yahoo! (1)	[246]
	Bing (1)	[9]
Chat (20)	Chat Logs (General) (18)	[215,216,226–232,217–221,223–225]
	Heaven BBS (2)	[247,248]
Games (1)	World of Warcraft (1)	[249]
Other (18)	Wi-Fi (1)	[250]
	NA (1)	[251]
	Professional Writing (1)	[88]
	Essays (15)	[127–133,135–142]

https://doi.org/10.1371/journal.pone.0207112.t001

research has grown rapidly, with around 90 of the studies in our set performing some form of predictive analysis across a variety of platforms including Twitter [65,81,84,134], blogs [205,206,211,213], Facebook [157,159,160] and emails [235,236,240]. Researchers have also analysed how language differs by style, [134,205,235,248] sentiment, [74,157,171,255] structure [195,199,235] and content [18,67,84,199].

Overall, research has demonstrated that gender can be predicted from digital traces reasonably successfully, with accuracies often reaching 80% and above [66,72,81,112,193,195,207,209,274]. Studies have highlighted similar trends to offline studies of language, in that females are more likely to use pronouns, emotion words (e.g. happy, bored, love), interjections (e.g. urgh, hmm), while males tend to use more practical dictionary-based words, proper names (e.g. sports team names, numbers and technology words, e.g. [64,160,198,205]. Emoticons (e.g. <3, ☺ and abbreviations (e.g. lol, omg) (which are more often associated with online discourse) tend to be used more frequently by females, whereas males are more likely to post links to websites, videos etc. [67,160]. Gender prediction is also detectable at the level of

Table 2. Number of articles predicting gender, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (134)	Language (123)	[20,29–65,67,70,72–76,78,80,81,83–89,91–100,102–144,146–149,151–155,158–160,162–165,171,209,214,252–255]
	Network Data (8)	[51,61,62,66,69,78,162,252]
	Colours (4)	[79,90,101,163]
	Meta-data (17)	[61,63,66,69,72,74,78,134,159,171,209,210,252,256–259]
	Names (13)	[29,40,51,69,82,90,112,145,158,161,166,259,260]
	Images (30)	[37–39,41–60,76,77,82,166–169]
	Locations (2)	[29,209]
	Facebook Likes (2)	[10,156]
	Tags (3)	[16,167,169]
	Activity (1)	[169]
	Check-ins (1)	[170]
Digital Devices (22)	Application Data (9)	[12–14,172,178,182,188,189]
	Call Logs/SMS Data (11)	[14,15,174,178,181,182,187–189,191,192]
	Location Data (4)	[183–186]
Websites (23)	Language (35)	[62,93–100,102,103,127–133,135–142,171,194–196,198,199,202,204,208]
	Website Data (1)	[197]
	Network Traffic Traces (1)	[179]
	Background Colours (1)	[261]
	Video Tags/Titles (1)	[203]
	Web Usage Data (1)	[200]
Blogs (58)	Language (55)	[18,93–100,102,103,127–133,135–142,157,171,205–209,211–213,215–221,223–234,262]
	Behavioural Data (1)	[234]
	Meta-data (3)	[66,206,212]
Emails (9)	Language (9)	[196,235–242]
Radio (3)	Meta-data, Listening Habits (3)	[243–245]
Search Engines (2)	Query Log Data (1)	[246]
	Facebook Likes, Profile Data (1)	[9]
Chat (20)	Language (20)	[215,216,225–232,247,248,217–224]
Games (1)	Behavioural Data (1)	[249]
Other (17)	Wi-Fi Traffic (1)	[250]
	Academic Researcher Emails (1)	[251]
	Language (15)	[127–133,135–142]

https://doi.org/10.1371/journal.pone.0207112.t002

individual words, word-stems (parts of words) and ngrams (sequences of items or letters, e.g. a unigram = 1 letter, a bigram = 2 letters, a trigram = 3 letters and so forth) e.g. [64,70,72,252,275]. For instance, Mueller and Stuemme [72] found that females tended to use bab, feel and girl (word stems), aa, ah, ee (digrams), and aaa, aha, ee (trigrams), whereas males used scor, team, win (word stems), er, in, re (digrams) and ent, ing, ion (trigrams). S2 Table provides examples of the specific language markers that were particularly successful in predicting gender.

Although these studies have consistently demonstrated trends in gender inference, we should be careful not to generalise the extent to which gender manifests in digital-based language. Most research treats gender as a binary classification task, and attempts to find markers

that uniquely identify males and females. However, this disregards evidence and theoretical arguments that gender can be expressed in diverse ways [112], and that gender may manifest differently across social groups, cultures, and contexts. Another consideration is that research is heavily skewed toward inferring gender from English, meaning that there is little exploration of whether these trends extend to other languages. A small number of studies within our set examined other languages including Arabic [194,195], Japanese, Indonesian, Turkish, French [61], Vietnamese [233,276], Russian and Portuguese [63]. The construction of other languages presents numerous challenges–verbs and nouns are either masculine or feminine in French and Spanish for instance, and (to our knowledge), there is less theoretical/social psychology research that explores language-gender differences in other languages and cultures. However, there is evidence to suggest that gender prediction from other languages can be just as successful as English-based approaches. Ciot et al. [61] found that their classifiers which predicted gender from French, Indonesia, Turkish and Japanese tweets achieved similar accuracies to English datasets (with accuracies of 76%, 83%, 87% and 63% for each language respectively). Future research could therefore explore the nuances and effectiveness of gender prediction in other languages.

Network data and meta-data. Communications technologies such as social media, smartphones and other digital devices have provoked researchers to question whether an individual's gender can be predicted from their meta-data (e.g. number of posts, frequency of logins etc.) or through network data derived from their social connections. Researchers often combine such data with language in their classification models in attempt to improve predictive accuracy. In some circumstances, network data have helped to compensate for shortfalls in language-based predictions. For instance, Bamman et al. [260] found that misclassified males and females (i.e. males who were predicted to be female because of their predominant use of 'feminine' language and vice versa) were often connected to more members of the opposite gender within their networks. In other words, males who tended to use words commonly associated with females, often had more female followers/friends in their networks and vice versa. As such, males' different use of language in this context may result from individuals 'accommodating' their peers and strong ties by matching their language to maintain and build rapport [277,278].

Other research has used the homophily principle to infer gender directly. For instance, Al Zamal et al. [252] used data extracted from a person's network neighbours (rather than the individuals themselves) to predict gender on Twitter. Using features such as frequently-used words, stems, ngrams and hashtags, combined with popularity measures of an individual's network neighbours, Al Zamal et al. [252] inferred gender as accurately as when using the individual's own data (highest accuracy using network data = 80.02%, accuracy using individual's own data = 79.50%). Similarly, Jurgens et al. [67] predicted individuals' gender from their incoming communications (communications directed to an individual), achieving 80% accuracy. Jurgens et al. [67] suggested that because individuals tend to be similar to those in their networks (in terms of their demographic attributes), communication with others often focuses on common ground. This results in reciprocal self-disclosure, meaning that the content, sentiment etc. conveyed by an individual's friends, also becomes revealing of what an individual may be like.

Age

The study of age is a vast area of research, encompassing developmental, aging, and social psychology that examines how age is affected by various social processes and how people communicate over their lifespans. Age inference is commonly studied alongside gender and has

Table 3. Number of articles predicting age, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Website (32)	IMDB (1)	[293]
	Other (8)	[62,171,179,197–199,214,294]
	Hotel Reviews (24)	[93–100,102,103,127–133,135–142]
Search Engines (2)	Bing (1)	[9]
	Yahoo! (1)	[246]
Blogs (54)	Blogger.com (4)	[18,205,207,295]
	Blogs (General) (50)	[94–100,102,103,127–133,135–142,157,171,209,211,212,214–233,292,296]
	LiveJournal (1)	[297]
Smartphones (18)	NR (18)	[13–15,172–174,177–179,181–184,188,189,191,287,298]
Forums (2)	Vietnamese Forums (1)	[276]
	Breast Cancer Forum (1)	[292]
Social Media (84)	Twitter (75)	[20,51,62,65–67,71,73–75,82,84,87,89,91,93–100,102–111,113–142,171,209,214,252,285,286,290,293,299–303]
	Social Media (General) (25)	[93–100,102,103,127–133,135–142,171]
	Facebook (5)	[10,156,159,160,304]
	Flickr (1)	[166]
	Netlog (2)	[164,165]
	YouTube (2)	[162,163]
	Instagram (2)	[169,289]
	Pokec (1)	[305]
	Sina Weibo (3)	[170,288,306]
Emails (4)	NR (4)	[238–241]
Radio (3)	Last.fm (3)	[243–245]
Games (1)	World of Warcraft (1)	[249]
Chat (19)	NR (19)	[215,216,225–232,291,217–224]
Other (14)	Essays (14)	[127–133,135–142]

https://doi.org/10.1371/journal.pone.0207112.t003

received much attention from researchers trying to understand how online behaviour may signal how old a person is. Our search generated a set of 157 articles (94 independent articles and 63 articles from the PAN workshops) that reported some form of age inference from digital traces. Table 3 provides an overview of the articles published and associated references per platform. Table 4 provides an overview of the articles published and associated references per predictor. We discuss the main trends and findings for age inference over the following subsections.

Language. Similar to gender, extensive research has examined how language use is related to age, e.g. [23,279,280] and how a person's language is influenced by their emotional experiences, identity, social relationships and cognitive abilities over time e.g. [281–284]. Research on age and language has highlighted that individuals' use of positive emotion, future tense and cognitively complex words (causation words, insight words, long words) tends to increase with age, whereas negative emotion, first-person singular self-references and past tense words tends to decrease [23]. Around 60 articles in our set conducted some form of analysis related to age inferences and language across numerous platforms including Twitter [134,252,255,285,286], websites [197–200], smartphones [13,14,174,178,287], and emails [238,239,241]. Researchers have also analysed how language differs by style [18,74,205,288], content

[18,62,67,205,289,290], sociolinguistics [75,255], and ngrams [164,291,292]. S3 Table provides examples of the specific language markers that were particularly successful in predicting age.

Overall, research has demonstrated that age can be predicted from language reasonably successfully, with accuracies often reaching 70% and above [18,164,205,233,252]. Studies have highlighted numerous patterns related to language and age; in terms of content, younger people (in their teens and twenties) used words related to school, work, socialising, computer games and comedians, whereas older adults (over 30) tended to use more family related words and words associated with the news or society [18,62,211]. In terms of style, younger people tended to use more acronyms, slang, self-references, and varied forms of grammar, whereas older adults tended to use more mature and polite language, with less linguistic variation [67,207].

Although these findings seem to broadly align with research on language and age in offline contexts, current methods are quite limited. There is a tendency for researchers to treat age as categorical variables such as 13–17, 18–24, 25–35, and then simply using 'over 35' or 'over 40' when predicting older ages. This approach can severely undermine the accuracy of prediction, especially for adults over the age of 30 –surprisingly, only 15 studies treated age as a continuous variable, e.g. [160,177,189,287,305]. In some circumstances, obtaining a more approximate age may be acceptable, for instance it is highly unlikely that a person's choice of product will change vastly from the age of 23 to 24. Another factor that may have hindered research thus far is that younger people tend to use the internet more than older people, so it may have been more difficult to obtain decent ground truth/training data. For instance, a survey by the Pew Internet Centre, highlighted that as of 2018, 66% of US adults over 65 use the internet, compared to 98% of 18-29-year olds. These figures have increased from 14% and 70% respectively since 2000 [308]. Future research may therefore want to consider exploring more nuances in language use across specific ages.

Network data. Network data has also been a reliable indicator of a person's age, with studies highlighting that people of similar ages tend to congregate in the same networks e.g. [177, 252,287] and communicate more with each other on social media e.g. [67,209]. Research has also identified patterns of homophily in smartphone records and applications [181,191,287] that varies across different age groups. For example, Park et al. [287] found that children (9 year olds), and teenagers (14-18-year olds) sent most of their SMS messages to others their own age and Dong et al. [191] found that 18–35 year-olds had more (same and opposite gender) contacts than people over 35, who had smaller, same-gender social circles. Similarly, children and teenagers were also identifiable from their communication patterns to people their parents age [287], which subsequently decreased as individuals became older. Although specific explanations from social psychology for these patterns of behaviour do not exist (to our knowledge), these types of findings highlight the potential to gain new understanding and extend existing explanations of how relationships and communication change over different age groups.

Location

Location-based services (LBS) are incredibly useful across many domains, including personalised services (e.g. local restaurants, hospitals, events), coordinating responses to disease or disasters, and detecting security intrusion. Using digital traces to infer location data enables researchers to examine the relationship between online behaviour and individuals' locations (e.g. regional nuances, countries etc.), rather than relying upon IP addresses. Because location or geo-location-based work is an area of research within itself, we were careful to restrict our inclusion criteria to studies that predicted location data relating specifically to individuals'

Table 4. Number of articles predicting age, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	Reference
Website (32)	Language (30)	[62,93,103,127–133,135,136,94,137–142,171,198,199,214,95–100,102]
	Website Data (3)	[197,199,294]
	Network Data (1)	[62]
	Network Traffic Data (1)	[179]
	Demographics, Names, Followers (1)	[293]
Search Engines (2)	Facebook Likes (1)	[9]
	Query Logs (1)	[246]
Blogs (54)	Language (54)	[18,93,103,127–133,135,136,94,137–142,171,205,207,209,95,211,212,214–221,96,222–231,97,232,233,292,295,297,98–100,102]
	Meta-data (6)	[157,207,209,211,212]
Smartphones (18)	Application Use (7)	[13,14,177,178,182,188,189]
	Call/SMS Data (15)	[13,14,183,188,189,191,298,15,172,174,177–179,181,182]
	Location Data (8)	[14,173,178,182–184,188,189]
	Accelerometer Data (5)	[14,178,182,188,189]
	Network Data (7)	[15,174,178,179,181,191,287]
Forums (2)	Language (2)	[276,292]
Social Media (84)	Language (81)	[20,51,84,87,91–98,62,99,100,102,103,105–110,65,111,114–122,66,123–132,67,133–142,71,159,160,162–165,171,209,252,255,73,285,286,288–290,299,300,302,306,307,74,82]
	Meta-data (7)	[66,169,285,289,300,306]
	Network Data (12)	[62,69,216,219,97,98,105,120,154,167,202,211]
	Facebook Likes (2)	[10,156]
	Names (4)	[51,82,166,301]
	Images (4)	[82,166,169,288]
	Check-ins (1)	[170]
Emails (4)	Language (4)	[238–241]
Radio (3)	Music Meta-data/Listening Habits (3)	[243–245]
	Profile Information (1)	[243]
Games (1)	Character Features/Behavioural Data (1)	[249]
Chat (19)	Language (19)	[215,216,225–232,291,217–224]
	Meta-data (1)	[291]
Other (14)	Language (14)	[127,128,138–142,129–133,135–137]

https://doi.org/10.1371/journal.pone.0207112.t004

home cities, countries etc. (as opposed to analyses of where individuals were at particular moments in time, e.g. [309]. For articles that cover geolocation prediction in more detail see the work by Jurgens et al. [310] and Stefanidis et al. [311]. 32 articles reported some form of location prediction, across a range of granularities (e.g. home, city, country), platforms (e.g. Twitter, Facebook, Flickr, Foursquare) and traces (e.g. language, network data, location fields in profiles) (see Table 5 and Table 6 for breakdowns of the platforms, predictors and references).

Inferring location accurately can be challenging due to the complexity of information available, individuals' personal circumstances and platform design. These challenges have been acknowledged in much of the research conducted to date. For instance, many applications enable individuals to self-report their location–Facebook provides the "Current City" and "Hometown" fields, and Twitter provides the profile "Location" field. Often these fields are

Table 5. Number of articles predicting location, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (24)	Facebook (2)	[156,312]
	Twitter (20)	[29,67,75,82,87,89,313–326]
	Flickr (3)	[16,321,322]
Location-based Applications (5)	Foursquare (3)	[67,317,319,327,328]
	Brightkite (1)	[328]
	Google+ (1)	[319]
	Gowalla (1)	[328]
Blogs (1)	NR (1)	[233]
Emails (3)	NR (1)	[238,239,241]
Smartphones (2)	NR (1)	[179,329]
Forums (1)	Webretho, Otofun, Tinhte (1)	[276]
Search Engines (1)	Yahoo! (1)	[246]
Websites (1)	NR (1)	[179]

https://doi.org/10.1371/journal.pone.0207112.t005

non-compulsory, and have no restrictions; as such, individuals can enter incorrect, non-existent or even fake information. For instance, Hecht et al. [313] found that 34% Twitter users did not provide location information in their profiles, and those that did rarely provided detail beyond their current city. Users who did provide data often replaced locations with false places (e.g. "outta space"), sarcastic comments (e.g. "redneck hell") or celebrities' names (e.g. "Justin Bieber's heart"). Despite the limited reliability of profile location fields, numerous studies have used them in their algorithms, but typically in combination with other digital traces such as network data [312] name data [29] and tweet contents [315,316]

Other approaches have involved inferring location solely from language without considering other geospatial cues ([315,326,330]. Language may reveal aspects of an individual's demographic location if they directly reference particular venues, places or use certain colloquialisms or slang. For instance, people from Texas may use "howdy" frequently, or people from London may reference Arsenal Football Club. Chang et al. [315] and Cheng et al., [325] predicted individuals' cities tweet location-related contents; their most accurate predictions were 50.93% (within a 100 mile radius) and 78.80% (within a 536 mile radius) respectively.

Table 6. Number of articles predicting location, with associated predictors and references.

Category (n = no. articles)	Predictor (n = no. articles)	Reference
Social Media (24)	Location Data (16)	[19,31,322,69,312,314–319]
	Network Data (7)	[67,82,312,315,316,318,323]
	Names (2)	[29,82]
	Facebook Likes (1)	[156]
	Language (16)	[67,75,82,87,89,255,313–318,323–326]
	Spatial, Visual, Temporal Features (1)	[321]
Location-based Applications (5)	Check-in Data (2)	[327,328]
	Location Data (3)	[67,317,319]
Blogs (1)	Language (1)	[233]
Emails (3)	Language (1)	[238,239,241]
Smartphones (2)	Applications (1)	[179,329]
Forums (1)	Language (1)	[276]
Search Engines (1)	Query Logs (1)	[246]
Websites (1)	Network Traffic Traces (1)	[179]

https://doi.org/10.1371/journal.pone.0207112.t006

Chang et al.'s method was particularly useful as it only required 250 local words, (selected by unsupervised methods) in contrast to Cheng et al.'s approach which relied on 3,183 local words (selected by supervised classification based on 11,004 hand-annotated ground truth data).

Although these studies have demonstrated that inference from tweet content alone is possible, the language contained within tweets can be very noisy, as people may discuss varied topics and may use language that does not readily link to specific locations (e.g. conjunctions, prepositions, adjectives, or generic terms like 'restaurant', 'city centre'). Network data may therefore provide a more objective measure for predicting location. Numerous studies incorporated various forms of network data in their models including 'friends" location data [312,320] or network data combined with tweet contents or other meta-data, e.g. [82,310,315]. Traditionally, one would predict that people would tend to know (or be 'friends' with) more people in close physical proximity to themselves, that is, they would be connected to people who live in the same town or city. Although the internet has the ability to change this drastically, by connecting people over vast distances, research has highlighted that homophily still holds within this context. Backstrom et al. [312] for instance found that the likelihood of friendship reduced as a function of distance, and their model based on network associations and address data was able to predict the locations of 69.10% of users within a 25-mile radius.

Finally, while the bulk of research has used Twitter data, other studies have examined other platforms and devices, including smartphone applications [329] web traffic data [244] Foursquare e.g. [310,317,328] and Google+ [319]. Foursquare in particular, is designed to provide users with personalised, location-based recommendations, based on their browsing histories, purchases and check-in behaviour. Findings to date have demonstrated accuracies of 67.41% for city [319,327], 80.92% for state, and 93.67% for country-level prediction [327].

Political orientation

In recent years, the internet has become a hotbed for publishing and promoting political activity. Social media in particular has become a forum where news stories are circulated, political parties disseminate their agendas, and where any individual can express political opinions and beliefs. As such, research exploring political related activity online has proliferated, with researchers attempting to use online data to understand people's political sentiments e.g. [331,332] and predict election outcomes, e.g. [333,334]. Thus, inferring an individual's political orientation from their digital traces is just one area among a rapidly growing field of research. Our search generated 33 articles that inferred political orientation from digital traces. Twitter is the most studied platform, with language and network-based features most commonly used for inference (see Table 7 and Table 8 for overviews).

Inferring an individual's political orientation accurately is particularly challenging because it can vary in strength and change over time. This is particularly pertinent when external

Table 7. Number of articles predicting political orientation, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (25)	Twitter (25)	[19,20,293,299,302,335–341,62,342–346,75,82,84,86,252,255,274]
	Facebook (2)	[10,11]
Websites (3)	IMDB (1)	[293]
Search Engines (1)	Bing (1)	[9]
Blogs (4)	Digg (1)	[347]
	Blogs (Other) (3)	[348–350]

https://doi.org/10.1371/journal.pone.0207112.t007

Table 8. Number of articles predicting political orientation, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (25)	Meta-data (4)	[252,335,337,342]
	Language (24)	[19,20,299,302,335–342,62,343–346,75,82,84,252,255,274,293]
	Network Data (10)	[62,82,86,252,293,299,336,339,340,344]
	Facebook Likes (2)	[10,11]
Websites (3)	Language (3)	[62,293,351]
	Network Data (2)	[62,293]
	Location Data (1)	[293]
	Name Data (1)	[293]
Search Engines (1)	Facebook Likes (1)	[9]
Blogs (4)	Language (4)	[347–350]

https://doi.org/10.1371/journal.pone.0207112.t008

factors, such as societal events or political campaigns directly attempt to sway peoples' ideologies. However, the subjective nature individuals' political preferences has generally not been reflected in existing research. The majority of studies in our set have treated prediction as a classification problem, where individuals are categorised into two [10,86,252,338], three [336,337], or four classes [341,342]. Given that most countries tend to be dominated by two political parties, these approaches may seem logical for gaining a simplistic overview of individuals' political preferences. However, the disadvantage is that such categorisations cannot capture the strength or idiosyncrasies of individuals' beliefs. Barberá [86] directly attempted to address this problem by developing a model that estimated ideology on a continuous scale. By using social ties (i.e. who individuals follow), Barberá [86] successfully inferred ideological alignment (strength in terms of right vs. left leaning) across European countries and the US, that correlated strongly with offline measures of voting records. As such, Barberá's method has since been widely adopted by other political scientists analysing political behaviour online, e.g. [352,353].

Another challenge for predicting political orientation is that gaining valid ground truth is often difficult. Many individuals do not explicitly state their political affiliation online, and those that do are likely to be more politically opinionated or active that the average person. For instance, Priante et al. (2016) claimed that fewer than 5% of Twitter members state their affiliation. Cohen and Ruths [338] suspected this may have caused studies that used explicit political preferences as ground truth to be biased in favour of political activists or those with strong political views. To examine this, Cohen and Ruths [338] constructed three separate Twitter datasets (comprising tweets and hashtags), each representing different strengths of political orientation: a) US politicians' accounts, b) users who self-reported their political orientation in their accounts, and c) 'modest' users who frequently mentioned politics in their tweets, (such that their orientation could be manually inferred), yet without any explicit declaration.

Cohen and Ruths' [338] findings demonstrated that classification accuracy decreased as visible political engagement decreased. In other words, US politicians' preferences were the easiest to predict, with 91% accuracy, followed by politically active users at 84% and modest users at 68%. Given that much of the previous research used self-reported political affiliation as ground truth, e.g. [252,255,340], these findings suggested that many of the reported accuracies were likely unrepresentative of the general population. Cohen and Ruths examined this further by testing the transferability of their classifiers and found that accuracy reduced significantly–to 11% when classifiers trained on political figures were tested on modest users.

Perhaps due to Cohen and Ruth's (somewhat concerning) findings, subsequent research has adopted more cautious approaches toward classification. Preotiuc-Pietro et al., [19] created

a language-based model using individuals' self-reported orientation, where individuals rated the strength of their political ideologies on a seven-point scale (ranging from 'Very Conservative' to 'Very Liberal'). This enabled them to account for varying strength of political preferences rather than limiting predictions to 2–3 classes. Similarly, obtaining self-reports in this instance enabled them to avoid the biased and unrealistic forms of data inherent in the previously used methods. Their accuracies ranged from 22–27%, highlighting that realistic, fine-grained political orientation is more nuanced and complex than that reported by previous research. Future research may therefore want to be mindful of selecting appropriate training data and examining degrees of political orientation to ensure that predictions are realistic.

Sexual orientation

To date, research on inferring sexual orientation has received little attention in comparison to other demographic attributes, with 7 studies generated from our search (see Table 9 and Table 10). Despite this, inferring an individual's sexuality has many important implications, especially with regards to individuals' privacy and how their data may be used. Across many types of social media, individuals have freedom over whether to disclose their sexual preferences, whereas in other platforms such as dating websites/applications, individuals may be required to provide such data in order to use the service.

The notion that individuals may unintentionally 'leak' clues to their sexuality in their digital traces may therefore be worrying to those who may want to keep such data private or hidden. In fact, all of the studies within our set examined inference from data that was unintentionally revealed by the individuals themselves or inferred through homophily [10,11,201,354–356]. For instance, Kosinski et al. [10] found that Facebook likes such as 'Ellen DeGeneres', 'Mac Makeup' and 'Wicked The Musical' were highly predictive of homosexual males, and 'Not Being Pregnant' and 'No H8 Campaign' were predictive of homosexual females. Further, 'Being Confused After Waking Up From Naps' and 'Nike Basketball' were highly predictive of heterosexual males, and 'Adidas Originals' and 'Yahoo' were predictive of heterosexual females.

Alternatively, research by Jernigan et al. [355], Sarigol et al. [356] and Garcia [354] used data derived from other people to infer individuals' sexuality—their findings highlighted accuracies of around 00.80 (AUC). In particular, Sarigol et al. [356] and Garcia [354] demonstrated how such techniques could be used to infer the sexuality of non-users, also referred to as the '*shadow profile hypothesis*'. By analysing data from profiles on the (discontinued) social networking site Friendster, Sarigol et al. [356] and Garcia [354] found that sexual orientation groups were affected by network size and disclosure parameters where, as size/disclosure increases, so does the likelihood of inferring a non-user's private data. Although there is limited work exploring shadow profiles, these findings highlight a concerning possibility that future research may want to consider when studying networks and individuals' privacy. That is, whether it is possible to infer sexuality (or indeed any other attributes) from other peoples' data, and in turn what can be done in order to protect peoples' privacy.

Table 9. Number of articles predicting sexual orientation, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (6)	Friendster (2)	[354,356]
	Facebook (3)	[10,11,355]
	Sina Weibo (1)	[170]
Dating Website (1)	NR (1)	[357]

https://doi.org/10.1371/journal.pone.0207112.t009

Table 10. Number of articles predicting sexual orientation, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (6)	Network Data (2)	[355,356]
	Gender, Relationship Status, Sexual Orientation (1)	[354]
	Facebook Likes (2)	[10,11]
	Check-ins (1)	[170]
Dating Website (1)	Images (1)	[357]

https://doi.org/10.1371/journal.pone.0207112.t010

Other demographic attributes

Numerous articles reported multiple demographics that were distinct from the main traits outlined thus far. In most cases, these attributes were not studied independently and, (to our knowledge) do not have extensive research histories or theoretical backgrounds from social psychology. Nevertheless, we believe inferring these attributes forms an important part in profiling individuals, and are likely to receive more research attention in the future. Because of the limited literature surrounding the remaining attributes, we display the main findings for each in the series of tables that follow and in the supplementary materials. The attributes identified include: family and relationships (Table 11, Table 12, S7 Table), ethnicity and race (Table 13, Table 14, S8 Table), education (Table 15, Table 16, S9 Table), income (Table 17, Tables 18 and S10 Income), language (Table 19, Table 20, S11 Table), religion (Table 21, Table 22, S12 Table), occupation (Table 23, Table 24, S13 Table), health (Table 25, Tables 26 and S14) and social class (Table 27, Table 28, S15 Table).

Discussion

The ability to predict individuals' demographic attributes from their online activity has many useful applications including marketing, criminal investigations, monitoring societal events and tracking health. Academic research attempting to use computational methods to infer attributes has proliferated in recent years and overall has demonstrated reasonable degrees of accuracy. This systematic review has highlighted the current state-osf-the art with regards to demographic prediction, in terms of the platforms, digital traces and methods currently employed. To date, age and gender are the most studied demographics—perhaps this is due to more established research histories within the social psychology literature, compared to other attributes.

A key factor in predicting such information is the type of digital footprint from which this information is derived. Many studies that perform linguistic analyses highlight trends in patterns of language use (in terms of style, content, slang etc.) that seem common across platforms and traits. For instance, females tend to use words such as *shopping*, *excited*, *sooo*, *yay* <*3*, e.g. [20,160,207], and males tend to use words such as *I've*, *fuck*, *league*, *youtube.com*, *system*, *software*, e.g. [18,20,160]. Younger adults tend to use shorter sentences and words such as *cuz*, *haha*, *school*, *don't*, *office*, *beer*, e.g. [20,160], and older adults (typically classified as over 30) tend to use words such as *kids*, *family*, *daughter*, *don't*, e.g. [160,207]. However, rarely are differences in either age or gender connected to theoretical perspectives on either life span development, or gender. For instance, there is considerable previous (earlier) work on the use of hedges and tag questions (e.g. it's a nice day, *isn't it*?) by female speakers, and how such language may reflect power differentials and inequalities in a patriarchal society, e.g. [265].

Similarly, differences in the challenges faced across life stages have been widely theorised, e.g. [376], as have the changing goals that people strive for as they age, e.g. [377,378]. However, it was rare to find consideration of *what* the predictive features might mean to a social scientist

Table 11. Number of articles predicting family and relationship status, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (8)	Facebook (3)	[10,11,156]
	Friendster (1)	[354]
	Twitter (4)	[20,84,87,358]
	Sina Weibo (1)	[170]
Smartphone (9)	NR (9)	[13,14,177,178,182,184,189,329,359]
Websites (1)	NR (1)	[200]

https://doi.org/10.1371/journal.pone.0207112.t011

within the papers reviewed, and often the predictive features were not even mentioned in the paper, making connection to social theory impossible. Instead, much of the time the approach taken was to compare classifiers, and to allow the machine learning program to identify the best features (or to include as many as possible in a training set, and then replicate with the 'best' features in a kept back sample for validation purposes. Although in many cases this likely results from conventions in different research fields–computer science approaches tend to focus more on successful methods and prediction, whereas psychology emphasises causes and explanations (for a detailed discussion of this, see the work by Yarkoni and Westfall [379]).

Network data, in the form of metrics derived from social network neighbours, structural features and popularity (e.g. mentions, follows) were also useful for predicting a range of attributes including age, gender, location and sexual orientation, e.g. [252,312,315,355]. The ability to use network data to infer attributes can be incredibly useful in identifying information that may not be disclosed directly by an individual. However, this has serious implications for privacy–individuals may want to keep their political beliefs, sexuality etc. private and may not realise they are inadvertently revealing them through their digital activity. Alternatively, the extent to which this is a concern is dependent on who the individual would want to conceal such information from–computer algorithms may be able to detect such information; however, it is unlikely that the average human or people within their network would be able to make such inferences accurately from looking at this type of data.

One aspect that was noticeable from the studies presented is that there was no focus on the more complex modes of interaction, such as deception or attempts by individuals to present themselves differently at different points in time/in different contexts. For instance, an individual's language is likely to differ when talking to friends in comparison to writing an online review. Would a computer be able to identify their demographic attributes as being the same across both contexts? Research on communication accommodation demonstrates that individuals co-ordinate their language use with those they are conversing with, e.g. [279,380], suggesting that the assessment of demographics from, say, the language used, should be more difficult

Table 12. Number of articles predicting family and relationship status, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (8)	Facebook Likes (2)	[10,156]
	Language (4)	[20,84,87,358]
	Relationship Status (1)	[354]
	Network Data (1)	[358]
	Check-ins (1)	[170]
Smartphone (9)	Application Data, Behavioural Data, Call Data (8)	[13,14,177,178,182,189,329,359]
	Location Data (1)	[184]
Websites (1)	Web Usage Data (1)	[200]

https://doi.org/10.1371/journal.pone.0207112.t012

Table 13. Number of articles predicting ethnicity or race, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (15)	Twitter (12)	[17,20,360,361,29,62,253,274,288,299,301,302]
	Facebook (3)	[10,158,330]
Websites (3)	News (1)	[362]
	Other (2)	[62,294]
Devices (2)	Smartphone (1)	[13]
	Tablet (2)	[193]
Radio (1)	Meta-data, Listening Habits (1)	[243]

https://doi.org/10.1371/journal.pone.0207112.t013

Table 14. Number of articles predicting ethnicity or race, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (15)	Names (6)	[29,51,158,301,330,360]
	Language (11)	[17,20,361,51,62,158,253,274,299,302,360]
	Network Data (2)	[51,62]
	Location Data (2)	[29,360]
	Meta-data (1)	[360]
	Facebook Likes (1)	[10]
	Profile Images (2)	[17,51]
Websites (3)	Names (1)	[362]
	Web Browsing Histories (1)	[294]
	Language, Network Data (1)	[62]
Devices (2)	Application Data (1)	[13]
	Actions, Keystrokes, Timestamps (1)	[193]
Radio (1)	Meta-data, Listening Habits (1)	[243]

https://doi.org/10.1371/journal.pone.0207112.t014

Table 15. Number of articles predicting education level, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (8)	Twitter (6)	[20,62,84,253,254,358]
	Facebook (1)	[363]
	Sina Weibo (1)	[170]
Websites (4)	NR (4)	[62,197,200,294]
Email (4)	NR (4)	[238–241]
Wi-Fi (1)	NA (1)	[250]

https://doi.org/10.1371/journal.pone.0207112.t015

in the context of interactions if one person's use of specific language is influenced by their conversational partners' use of the same linguistic features. Moreover, the degree to which people accommodate towards their conversational partner is influenced by a number of factors, including power differentials [381]. Indeed, there is evidence that deception in text-based communication can be identified by the language used by the person *being lied to* as well as via changes in the language of the deceiver [382] suggesting that analysing language from interactions as individual data points needs to be treated with particular caution. Future work could attempt to decipher whether computer models are able to use similar sociolinguistic

Table 16. Number of articles predicting education level, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (8)	Language (7)	[20,62,84,253,254,358,363]
	Network Data (2)	[62,358]
	Meta-data (2)	[358,363]
	Facebook Likes (1)	[363]
	Check-ins (1)	[170]
Websites (4)	Language (1)	[62]
	Network Data (1)	[62]
	Website Data (1)	[197]
	Meta-data (1)	[197]
	Web Browsing Histories (1)	[294]
	NR (1)	[200]
Email (4)	Language (4)	[238–241]
Wi-Fi (1)	Wi-Fi Traffic (1)	[250]

https://doi.org/10.1371/journal.pone.0207112.t016

Table 17. Number of articles predicting income, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (6)	Twitter (6)	[20,84,253,285,302,364]
Smartphone (5)	NR (5)	[13,187,365–367]
Websites (2)	NR (2)	[200,294]

https://doi.org/10.1371/journal.pone.0207112.t017

Table 18. Number of articles predicting income, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (6)	Language (6)	[20,84,253,285,302,364]
Smartphone (5)	Application Data (1)	[13]
	Call/SMS Data (4)	[187,365–367]
	Network Data (1)	[187]
Websites (2)	Web Usage Data (2)	[200,294]

https://doi.org/10.1371/journal.pone.0207112.t018

Table 19. Number of articles predicting language, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Devices (2)	Smartphone (1)	[329]
	Tablet (1)	[193]
Blogs (1)	Blogger.com (1)	[295]
Email (3)	NR (3)	[237–239]
Social Media (3)	Twitter (3)	[29,89,92]

https://doi.org/10.1371/journal.pone.0207112.t019

techniques to infer attributes from these types of interactions, and to unpick individual level characteristics from those dependent on the nature of the interaction or audience.

We also suspect that rather than simply comparing the effectiveness of classification algorithms, or mechanical turk workers vs. a classifier, in the future authors may wish to take a

Table 20. Number of articles predicting language, with associated predictors and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Devices (2)	Application Data (1)	[329]
	Actions, Keystrokes, Timestamps (1)	[193]
Blogs (1)	Language (1)	[295]
	Language (3)	[237–239]
	Names, Location (1)	[29]
	Language (2)	[89,92]

https://doi.org/10.1371/journal.pone.0207112.t020

Table 21. Number of articles predicting religion, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (6)	Twitter (4)	[20,67,84,368]
	Facebook (2)	[10,11]
Search Engines (1)	Bing (1)	[9]
Smartphones (1)	NR (1)	[329]

https://doi.org/10.1371/journal.pone.0207112.t021

Table 22. Number of articles predicting religion, with associated predictors and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (6)	Language (4)	[20,67,84,368]
	Facebook Likes (2)	[10,11]
Search Engines (1)	Facebook Likes, Profile Data (1)	[9]
Smartphones (1)	Application Data (1)	[329]

https://doi.org/10.1371/journal.pone.0207112.t022

Table 23. Number of articles predicting occupation, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (10)	Twitter (10)	[66,73,87,89,91,92,209,358,369,370]
Blogs (2)	NR (2)	[209,233]
Smartphones (8)	NR (8)	[14,177,178,182–184,188,189]
Websites (1)	NR (1)	[197]
Forums (1)	NR (1)	[276]

https://doi.org/10.1371/journal.pone.0207112.t023

more theoretically driven approach to feature selection. For instance, there is considerable evidence that pronoun use can be linked to a number of social and psychological theories–including ingroup ('we') and outgroup ('they') identification (e.g.[383]), leadership ([384]) and gender bias [385]. Given the existing body of work identifying differences between groups based on these features, one would expect that a classifier *should* be able to distinguish between categories based on existing theory. It would also further our understanding of an existing body of work if a theoretically derived model were compared against a 'best feature' model derived from a machine learning approach.

Finally, in reviewing the papers herein it became clear that summarising the results of studies across labs is particularly difficult. In many cases multiple, different algorithms are used, the most discriminating features aren't reported, or simple accuracy statistics are reported without the full confusion matrix or recall / sensitivity information provided. We would

Table 24. Number of articles predicting occupation, with associated predictors and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (10)	Language (9)	[66,73,87,91,92,209,358,369,370]
	Network Data (3)	[66,358,370]
	Meta-data (5)	[66,209,303,358,370]
Blogs (2)	Language (2)	[209,233]
	Meta-data (1)	[209]
Smartphones (8)	Application Data (6)	[14,177,178,182,188,189]
	Call Data (5)	[14,178,182,188,189]
	Location Data (2)	[183,184]
Websites (1)	Time/Day Data, Website Data (1)	[197]
Forums (1)	Language (1)	[276]

https://doi.org/10.1371/journal.pone.0207112.t024

Table 25. Number of articles predicting health, with associated predictors and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (7)	Twitter (5)	[20,67,84,371,372]
	Facebook (1)	[10]
	MyFitnessPal (1)	[373]
	Reddit (1)	[374]
Smartphones (2)	NR (2)	[176,201]

https://doi.org/10.1371/journal.pone.0207112.t025

Table 26. Number of articles predicting health, with associated predictors and references.

Category (n = no. articles)	Predictors (n = no. articles)	References
Social Media (7)	Language (5)	[20,67,84,371,372]
	Images (1)	[374]
	Facebook Likes (1)	[10]
	Behavioural Data (1)	[176]
Smartphones (2)	Application Data (2)	[176,201]
	Network Data (1)	[201]

https://doi.org/10.1371/journal.pone.0207112.t026

Table 27. Number of articles predicting social class, with associated platforms and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (1)	Twitter (1)	[375]
	Foursquare (1)	[375]

https://doi.org/10.1371/journal.pone.0207112.t027

Table 28. Number of articles predicting social class, with associated predictors and references.

Category (n = no. articles)	Platform (n = no. articles)	References
Social Media (1)	Language (1) [375]	[375]

https://doi.org/10.1371/journal.pone.0207112.t028

strongly advise that the field consider methods to standardise reporting across studies and labs, enabling replication and for future studies to build more ably from the basis of earlier work.

Supporting information

S1 Table. PRISMA checklist for systematic review.
(DOCX)

S2 Table. Gender articles and data. Note: All accuracy measures in this database (and those below) are summarised in ranges (lowest to highest) and are reported to 2 decimal places. This was performed in order to standardise the varied styles of reporting provided in the set of articles. Further, some articles reported their findings as graphs or other visualisations, meaning that we could not extract specific accuracies. In these instances, cells are left blank. For instances where data were simply not reported in an article, we denote this with NR (i.e. Not Reported).
(XLSX)

S3 Table. Age articles and data.
(XLSX)

S4 Table. Location articles and data.
(XLSX)

S5 Table. Political orientation articles and data.
(XLSX)

S6 Table. Sexual orientation articles and data.
(XLSX)

S7 Table. Family and relationships articles and data.
(XLSX)

S8 Table. Ethnicity and race articles and data.
(XLSX)

S9 Table. Education articles and data.
(XLSX)

S10 Table. Income articles and data.
(XLSX)

S11 Table. Language articles and data.
(XLSX)

S12 Table. Religion articles and data.
(XLSX)

S13 Table. Occupation articles and data.
(XLSX)

S14 Table. Health articles and data.
(XLSX)

S15 Table. Social class articles and data.
(XLSX)

S16 Table. Classifier codes.
(XLSX)

Acknowledgments

This work was funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/N009614/1). The authors would like to thank the anonymous reviewers, whose comments have significantly improved this manuscript.

Author Contributions

Conceptualization: Joanne Hinds, Adam N. Joinson.

Formal analysis: Joanne Hinds.

Funding acquisition: Adam N. Joinson.

Investigation: Joanne Hinds.

Project administration: Adam N. Joinson.

Supervision: Adam N. Joinson.

Validation: Adam N. Joinson.

Visualization: Joanne Hinds.

Writing – original draft: Joanne Hinds, Adam N. Joinson.

Writing – review & editing: Joanne Hinds, Adam N. Joinson.

References

1. Internet World Stats [Internet]. [cited 30 May 2018]. Available: https://www.internetworldstats.com/stats.htm
2. Nordrum A. Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated. IEEE Spectr. 2016;
3. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Predicting flu trends using twitter data. 2011 IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2011. 2011. https://doi.org/10.1109/INFCOMW.2011.5928903
4. Mestyán M, Yasseri T, Kertész J. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLoS One. 2013; https://doi.org/10.1371/journal.pone.0071226 PMID: 23990938
5. Bermingham A, Smeaton AF. On Using Twitter to Monitor Political Sentiment and Predict Election Results. Psychology. 2011;
6. Howard P, Hussain M. Digital media and the Arab Spring. Democr Fourth Wave. 2012; https://doi.org/10.1353/jod.2011.0041
7. Gosling SD, Ko S, Mannarelli T, Morris ME. A room with a cue: Personality judgments based on offices and bedrooms. J Pers Soc Psychol. 2002; https://doi.org/10.1037//0022-3514.82.3.379
8. Vazire S, Gosling SD. e-Perceptions: Personality impressions based on personal websites. Journal of Personality and Social Psychology. 2004. https://doi.org/10.1037/0022-3514.87.1.123 PMID: 15250797
9. Bi B, Shokouhi M, Kosinski M, Graepel T. Inferring the demographics of search users. Proceedings of the 22nd international conference on World Wide Web—WWW '13. 2013. https://doi.org/10.1145/2488388.2488401
10. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci. 2013; 110: 5802–5805. https://doi.org/10.1073/pnas.1218772110 PMID: 23479631
11. Theodoridis T, Papadopoulos S, Kompatsiaris Y. Assessing the Reliability of Facebook User Profiling. Proceedings of the 24th International Conference on World Wide Web—WWW '15 Companion. 2015. https://doi.org/10.1145/2740908.2742728
12. Hazan I, Shabtai A. Noise Reduction of Mobile Sensors Data in the Prediction of Demographic Attributes. Proceedings - 2nd ACM International Conference on Mobile Software Engineering and Systems, MOBILESoft 2015. 2015. https://doi.org/10.1109/MobileSoft.2015.25

13. Malmi, E., & Weber I. You Are What Apps You Use: Demographic Prediction Based on User's Apps. In ICWSM. 2016. pp. 635–638.
14. Mo K, Tan B, Zhong E, Yang Q. Report of Task 3: Your Phone Understands You. Mob Data Chall Work. 2012;
15. Sarraute C, Brea J, Burroni J, Blanc P. Inference of demographic attributes based on mobile phone usage patterns and social network topology. Soc Netw Anal Min. 2015; https://doi.org/10.1007/s13278-015-0277-x
16. Popescu A, Grefenstette G. Mining User Home Location and Gender from Flickr Tags. Fourth Int AAAI Conf Weblogs Soc Media. 2010; doi:papers3://publication/uuid/7DB41A8E-EE73-4B31-AD14-A9EE2D3C668A
17. Ardehaly EM, Culotta A. Co-training for demographic classification using deep learning from label proportions. IEEE International Conference on Data Mining Workshops, ICDMW. 2017. https://doi.org/10.1109/ICDMW.2017.144
18. Argamon S, Koppel M, Pennebaker JW, Schler J. Automatically profiling the author of an anonymous text. Commun ACM. 2009; https://doi.org/10.1145/1562764.1562781
19. Preoțiuc-Pietro D, Liu Y, Hopkins D, Ungar L. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017. https://doi.org/10.18653/v1/P17-1068
20. Volkova S, Bachrach Y. On Predicting Sociodemographic Traits and Emotions from Communications in Social Networks and Their Implications to Online Self-Disclosure. Cyberpsychology, Behav Soc Netw. 2015; https://doi.org/10.1089/cyber.2014.0609 PMID: 26652673
21. Cadwalladr C. Facebook suspends data firm hired by Vote Leave over alleged Cambridge Analytica ties. The Guardian. 2018. Available: https://www.theguardian.com/us-news/2018/apr/06/facebook-suspends-aggregate-iq-cambridge-analytica-vote-leave-brexit
22. Kitchgaessner S. Cambridge Analytica used data from Facebook and Politico to help Trump. The Guardian. 2018.
23. Pennebaker JW, Stone LD. Words of Wisdom: Language Use Over the Life Span. Journal of Personality and Social Psychology. 2003. https://doi.org/10.1037/0022-3514.85.2.291
24. Weiser EB. Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison. CyberPsychology Behav. 2000; https://doi.org/10.1089/109493100316012
25. Wang H-Y, Wang S-H. User acceptance of mobile internet based on the Unified Theory of Acceptance and Use of Technology: Investigating the determinants and gender differences. Soc Behav Personal an Int J. 2010; https://doi.org/10.2224/sbp.2010.38.3.415
26. Gefen D, Straub DW. Gender Differences in the Perception and Use of E-Mail: An Extension to the Technology Acceptance Model. MIS Q. 1997; https://doi.org/10.2307/249720
27. Oakley A. Sex and Social Role. Sex, Gender and Society. 2015.
28. Azucar D, Marengo D, Settanni M. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. Personality and Individual Differences. 2018. https://doi.org/10.1016/j.paid.2017.12.018
29. Bergsma S, Dredze M, Van Durme B, Wilson T, Yarowsky D. Broadly Improving User Classification via Communication-Based Name and Location Clustering on Twitter. Hlt-Naacl. 2013; https://doi.org/10.1007/s00256-005-0933-8
30. Miura, Y., Taniguchi, T., Taniguchi, M., Ohkuma T. Author profiling with word+ character neural attention network. In Cappellato et al [13]. 2017.
31. Ribeiro-Oliveira, R., Oliveira-Neto RF. Using character n-grams and style features for gender and language variety identification. In Cappellato et al [13]. 2017.
32. Schaetti N. UniNE at CLEF 2017: TF-IDF and Deep-Learning for author profiling: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.
33. Sierra S, Montes-Y-gómez M, Solorio T, González FA. Convolutional neural networks for author profiling: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.
34. Tellez ES, Miranda-Jiménez S, Graff M, Moctezuma D. Gender and language-variety identification with MicroTC: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.
35. Akhtyamova L, Cardiff J, Ignatov A. Twitter author profiling using word embeddings and logistic regression: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.
36. Ogaltsov, A., Romanov A. Language variety and gender classificationfor author profiling in pan 2017. In Cappellato et al [13]. 2017.
37. Aragon, M. E., Lopez-Monroy AP. A straightforward multimodal approach for author profiling. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.

38. Bayot, R. K., Goncalves T. Multilingual author profiling using lstms. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
39. Ciccone, G., Sultan, A., Laporte, L., Egyed-Zsigmond, E., Alhamzeh, A., Granitzer M. Stacked gender prediction from tweet texts and images. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
40. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating Gender on Twitter. Assoc Comput Linguist. 2011; https://doi.org/10.1007/s00256-005-0933-8
41. Daneshvar S. Gender identification in twitter using n-grams and lsa. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
42. Garibo-Orts O. A big data approach to gender classification in twitter. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
43. HaCohen-Kemer, Y., Yigal, Y., Shayovitz, E., Miller, D., Breckon T. Author profiling: Gender prediction from tweets and images. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
44. Karlgren, J., Esposito, L., Gratton, C., Kanerva P. Authorship profiling without topical information. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
45. Kosse, R., Schuur, Y., Cnossen G. Mixing traditional methods with neural networks for gender prediction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
46. Lopez-Santillan, R., Gonzalez-Gurrola, L., Ramfrez-Alonso G. Custom document embeddings via the centroids method: Gender classification in an author profiling task. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
47. Martinc, M., Skrlj, B., Pollak S. Multilingual gender classification with multi-view deep learning. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
48. Nieuwenhuis, M., Wilkens J. Twitter text and image gender classification with a logistic regression n-gram model. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
49. Patra, B. G., Das, K. G. DD. Multimodal author profiling for arabic, english, and spanish. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
50. Raiyani, K., Goncalves, P. Q. T., Beires-Nogueira V. Multi-language neural network model with advance preprocessor for gender classification over social media. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
51. Chen X, Wang Y, Agichtein E, Wang F. A Comparative Study of Demographic Attribute Inference in Twitter. Ninth Int AAAI Conf Web Soc Media. 2015;
52. Sandroni-Dias, R., Paraboni I. Author profiling using word embeddings with subword information. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
53. Schaetti N. Unine at clef 2018: Character-based convolutional neural network and resnet18 for twitter author profiling. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
54. Sezerer, E., Polatbilek, O., Sevgili, O., Tekir S. Gender prediction from tweets with convolutional neural networks. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
55. Sierra-Loaiza, S., Gonzalez FA. Combining textual and representations for multimodal author profiling. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
56. Stout, L., Musters, R., Pool C. Author profiling based on text and images. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
57. Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma T. text and image synergy with feature cross trechnique for gender identification. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
58. Tellez, E. S., Miranda-Jimenez, S., Moctezuma, D., Graff, M., Salgado, V., Ortiz-Bejar J. Gender identification through multi-modal tweet analysis using microtc and bag of visual words. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
59. Veenhoven, R., Snijders, S., van der Hall, D., van Noord R. Using translated data to improve deep learning author profiling models. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
60. von Daniken, P., Grubenmann, R., Cieliebak M. Word unigram weighing for author profiling at pan 2018. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). 2018.
61. Ciot M, Sonderegger M, Ruths D. Gender Inference of Twitter Users in Non-English Contexts. Proc 2013 Conf Empir Methods Nat Lang Process. 2013;

62. Culotta A, Ravi NK, Cutler J. Predicting twitter user demographics using distant supervision from website traffic data. J Artif Intell Res. 2016;
63. Filho JABL, Pasti R, De Castro LN. Gender classification of twitter data based on textual meta-attributes extraction. Advances in Intelligent Systems and Computing. 2016. https://doi.org/10.1007/978-3-319-31232-3_97
64. Fink C, Kopecky J, Morawski M. Inferring Gender from the Content of Tweets: A Region Specific Example. Int Conf Weblogs Soc Media. 2012;
65. Flekova L, Carpenter J, Giorgi S, Ungar L, Preoţiuc-Pietro D. Analyzing Biases in Human Perception of User Age and Gender from Text. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016. https://doi.org/10.18653/v1/P16-1080
66. Ito J, Hoshide T, Toda H, Uchiyama T, Nishida K. What is He/She Like?: Estimating Twitter User Attributes from Contents and Social Neighbors. Conf Adv Soc Networks Anal Min (ASONAM), 2013 IEEE/ACM Int. 2013; https://doi.org/10.1145/2492517.2492585
67. Jurgens D, Tsvetkov Y, Jurafsky D. Writer profiling without the writer's text. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017. https://doi.org/10.1007/978-3-319-67256-4_43
68. Zamal F Al, Liu W, Ruths D. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. Science (80-). 2011;
69. Liu W, Ruths D. What's in a Name? Using First Names as Features for Gender Inference in Twitter. Anal Microtext Pap from 2013 AAAI Spring Symp. 2013;
70. Miller Z, Dickinson B, Hu W. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. Int J Intell Sci. 2012; https://doi.org/10.4236/ijis.2012.224019
71. Moseley N, Alm CO, Rege M. Toward inferring the age of Twitter users with their use of nonstandard abbreviations and lexicon. Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI 2014. 2014. https://doi.org/10.1109/IRI.2014.7051893
72. Mueller J, Stumme G. Gender Inference using Statistical Name Characteristics in Twitter. 5th ASE Int Conf Soc Informatics (SocInfo 2016), Union, NJ, USA, August 15–17, 2016 Proc. 2016; https://doi.org/10.1145/2955129.2955182
73. Preoţiuc-Pietro, Daniel, Wei Xu and LU. Discovering User Attribute Stylistic Differences via Paraphrasing. Proc Thirtieth AAAI Conf. 2016;
74. Rangel F. Author Profile in Social Media: Identifying Information about Gender, Age, Emotions and beyond. Proceedings of the 5th BCS IRSG Symposium on Future Directions in Information Access. 2013.
75. Rao D, Yarowsky D. Detecting Latent User Properties in Social Media. Proc NIPS MLSN Work. 2010;
76. Sakaki S, Miura Y, Ma X, Hattori K, Ohkuma T. Twitter User Gender Inference Using Combined Analysis of Text and Image Processing. Proceedings of the 25th International Conference on Computational Linguistics. 2014.
77. Shigenaka, R., Tsuboshita, Y. & Kato N. Content-Aware Multi-task Neural Networks for User Gender Inference Based on Social Media Images. In Multimedia (ISM), 2016 IEEE International Symposium. 2016. p. 169–172.
78. Ugheoke, T. O., Saskatchewan R. Detecting the gender of a tweet sender. 2014.
79. Alowibdi JS, Buy U a., Yu P. Language independent gender classification on Twitter. Proc 2013 IEEE/ACM Int Conf Adv Soc Networks Anal Min—ASONAM '13. 2013; https://doi.org/10.1145/2492517.2492632
80. Verhoeven B, Daelemans W, Plank B. Twisty: a Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling. Proc 10th Lang Resour Eval Conf (LREC 2016). 2016;
81. Verhoeven, B., Škrjanec, I., & Pollak S. Gender Profiling for Slovene Twitter Communication: The Influence of Gender Marking, Content and Style. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing. 2017. pp. 119–125.
82. Vijayaraghavan P, Vosoughi S, Roy D. Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning. Proc 55th Annu Meet Assoc Comput Linguist (Volume 2 Short Pap. 2017; https://doi.org/10.18653/v1/P17-2076
83. Volkova S, Wilson T, Yarowsky D. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. Proc 2013 Conf Empir Methods Nat Lang Process. 2013;
84. Volkova S, Bachrach Y, Durme B Van. Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter. Proceedings—2016 IEEE 2nd International Conference on Big Data Computing Service and Applications, BigDataService 2016. 2016. https://doi.org/10.1109/BigDataService.2016.28

85. Volkova, S., & Yarowsky D. Improving gender prediction of social media users via weighted annotator rationales. In NIPS 2014 Workshop on Personalization. 2014.

86. Barberá P. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. Polit Anal. 2015; https://doi.org/10.1093/pan/mpu011

87. Ikeda K, Hattori G, Ono C, Asoh H, Higashino T. Twitter user profiling based on text and community mining for market analysis. Knowledge-Based Syst. 2013; https://doi.org/10.1016/j.knosys.2013.06.020

88. Radford J. Piloting a theory-based approach to inferring gender in big data. International Conference on Big Data. IEEE; pp. 4824–4826.

89. Sloan L, Morgan J, Housley W, Williams M, Edwards A, Burnap P, et al. Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter. Sociol Res Online. 2013; https://doi.org/10.5153/sro.3001

90. Alowibdi JS, Buy UA, Yu P. Empirical evaluation of profile characteristics for gender classification on twitter. Proceedings—2013 12th International Conference on Machine Learning and Applications, ICMLA 2013. 2013. https://doi.org/10.1109/ICMLA.2013.74

91. Sloan L. Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. Soc Media + Soc. 2017; https://doi.org/10.1177/2056305117733224

92. Sloan L, Morgan J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0142209 PMID: 26544601

93. Weren ERD, Moreira VP, De Oliveira JPM. Exploring Information Retrieval features for author profiling: Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings. 2014.

94. Baker CI. Proof of concept framework for author profiling: Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings. 2014.

95. Gressel, G., Hrudya, P., Surendran, K., Thara S., Aravind, A., Poomachandran P. Ensemble Learning Approach for Author Profiling. Notebook for PAN at CLEF 2014, In Cappellato et al. 2014.

96. López-Monroy AP, Montes-Y-gómez M, Escalante HJ, Villaseñor-Pineda L. Using intra-profile information for author profiling: Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings. 2014.

97. Maharjan S, Shrestha P, Solorio T. A simple approach to author profiling in MapReduce: Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings. 2014.

98. Marquardt J, Farnadi G, Vasudevan G, Moens M-F, Davalos S, Teredesai A, et al. Age and gender identification in social media. CEUR Workshop Proceedings. 2014. https://doi.org/10.1007/s00256-005-0933-8

99. Mechti S, Jaoua M, Belguith LH. Machine learning for classifying authors of anonymous tweets, blogs, reviews and social media: Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings. 2014.

100. Villena-Román J, González-Cristóbal JC. DAEDALUS at PAN 2014: Guessing tweet author's gender and age. CEUR Workshop Proceedings. 2014.

101. Alowibdi J. S., Buy U. A., & Philip SY. Say it with colors: Language-independent gender classification on twitter. In Online Social Media Analysis and Visualization. 2014. p. 47–62.

102. Castillo E, Cervantes O, Vilariño D, Pinto D, León S. Unsupervised method for the authorship identification task: Notebook for PAN at CLEF 2014. CEUR Workshop Proceedings. 2014.

103. Amigo E, Carrillo-De-albornoz J, Chugur I, Corujo A, Gonzalo J, Meij E, et al. Overview of RepLab 2014: Author profiling and reputation dimensions for Online Reputation Management. CEUR Workshop Proceedings. 2014. https://doi.org/10.1007/978-3-319-11382-1_24

104. Alvarez-Carmona, M. A., Lopez-Monroy, A. P., Montes-y-Gomez, M., Villasenor-Pineda, L., Jair-Escalante H. INAOE's participation at PAN'15: Author profiling task. Working Notes Papers of the CLEF. 2015.

105. Arroju, M., Hassan, A., Farnadi G. Age, gender and personality recognition using tweets in a multilingual setting. In 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction. pp. 22–31.

106. Bartoli A, Dagri A, Lorenzo A De, Medvet E, Tarlao F. An Author Verification Approach Based on Differential Features Notebook for PAN at CLEF 2015. Work Notes CLEF. 2015; https://doi.org/10.1007/s00256-005-0933-8

107. Giménez M, Hernández DI, Pla F. Segmenting target audiences: Automatic author profiling using tweets. CEUR Workshop Proceedings. 2015.

108. Palomino-Garibay A, Camacho-González AT, Fierro-Villaneda RA, Hernández-Farias I, Buscaldi D, Meza-Ruiz I V. A random forest approach for authorship profiling. CEUR Workshop Proceedings. 2015.

109. Gonzalez-Gallardo CE, Montes A, Sierra G, Nunez-Juarez JA, Salinas-Lopez AJ, Ek J. Tweets classification using corpus dependent tags, character and POS N-grams. CEUR Workshop Proceedings. 2015.

110. Grivas A, Krithara A, Giannakopoulos G. Author profiling using stylometric and structural feature groupings. CEUR Workshop Proceedings. 2015.

111. Kiprov Y, Hardalov M, Nakov P, Koychev I. SU @ PAN 2015: Experiments in Author Profiling. CLEF 2015 Labs Work Noteb Pap. 2015;

112. Bamman D, Eisenstein J, Schnoebelen T. Gender identity and lexical variation in social media. J Socioling. 2014; https://doi.org/10.1111/josl.12080

113. Kocher M. Unine at CLEF 2015: Author profiling—notebook for PAN at CLEF 2015. CLEF 2015 Labs and Workshops, Notebook Papers CEUR Workshop Proceedings In:Cappellato et al [8]. 2015.

114. Maharjan S, Solorio T. Using wide range of features for author profiling. CEUR Workshop Proceedings. 2015.

115. McCollister C, Huang S, Luo B. Building topic models to predict author attributes from Twitter messages. CEUR Workshop Proceedings. 2015.

116. Miculicich Werlen L. Statistical Learning Methods for Profiling Analysis Notebook for PAN at CLEF 2015. CLEF 2015 Labs Work Noteb Pap. 2015;

117. Najib F, Cheema WA, Muhammad R, Nawab A. Author's Traits Prediction on Twitter Data using Content Based Approach. CLEF 2015 Labs Work Noteb Pap. 2015;

118. Nowson S, Perez J, Brun C, Mirkin S, Roux C. XRCE personal language analytics engine for multilingual author profiling. CEUR Workshop Proceedings. 2015.

119. Pervaz I, Ameer I, Sittar A, Nawab RMA. Identification of author personality traits using stylistic features. CEUR Workshop Proceedings. 2015.

120. Posadas-Durán JP, Markov I, Gómez-Adorno H, Sidorov G, Batyrshin I, Gelbukh A, et al. Syntactic N-grams as features for the author profiling task. CEUR Workshop Proceedings. 2015.

121. Poulston, A., Waseem, A., Stevenson M. Using tf-idf n-gram and word embedding cluster ensembles for suthor profiling. In Cappellato et al [13]. 2017.

122. Przybyła P, Teisseyre P. What do your look-alikes say about you? Exploiting strong and weak similarities for author profiling. CEUR Workshop Proceedings. 2015.

123. Bayot RK, Gon T. Age and Gender Classification of Tweets Using Convolutional Neural Networks. Machine Learning, Optimization, and Big Data (MOD 2017). 2018. https://doi.org/10.1007/978-3-319-72926-8

124. Iqbal HR, Ashraf MA, Muhammad R, Nawab A. Predicting an author's demographics from text using Topic Modeling approach Notebook for PAN at CLEF 2015. CLEF 2015 Labs Work Noteb Pap. 2015;

125. Şulea OM, Dichiu D. Automatic profiling of Twitter users based on their tweets. CEUR Workshop Proceedings. 2015.

126. Weren ERD. Information retrieval features for personality traits. CEUR Workshop Proceedings. 2015.

127. Agrawal M, Gonçalves T. Age and gender identification using stacking for classification? CEUR Workshop Proc. 2016;

128. Ashraf S, Iqbal HR, Nawab RMA. Cross-Genre author profile prediction using stylometry-based approach. CEUR Workshop Proceedings. 2016.

129. Bayot R, Gonçalves T. Author profiling using SVMs and Word embedding averages. CEUR Workshop Proceedings. 2016.

130. Bilan I, Zhekova D. CAPS: A cross-genre author profiling system. CEUR Workshop Proceedings. 2016.

131. Bougiatiotis K, Krithara A. Author profiling using complementary second order attributes and stylometric features. CEUR Workshop Proceedings. 2016.

132. Deyab, R. B., Duarte, J., Goncalves T. Author Profiling Using Support Vector Machines. In CLEF (Working Notes). 2016. pp. 805–814.

133. Dichiu D, Rancea I. Using machine learning algorithms for author profiling in social media. CEUR Workshop Proceedings. 2016.

134. Beretta V, Maccagnola D, Cribbin T, Messina E. An Interactive Method for Inferring Demographic Attributes in Twitter. Proc 26th ACM Conf Hypertext Soc Media—HT '15. 2015; https://doi.org/10.1145/2700171.2791031

135. Gencheva P, Boyanov M, Deneva E, Nakov P, Kiprov Y, Koychev I, et al. PANcakes team: A composite system of genre-Agnostic features for author profiling. CEUR Workshop Proceedings. 2016.

136. Markov, I., Gomez-Adorno, H., Sidorov, G., Gelbukh AF. Adapting Cross-Genre Author Profiling to Language and Corpus. In CLEF (Working Notes). 2016. pp. 947–955.

137. Modaresi P, Liebeck M, Conrad S. Exploring the effects of cross-genre machine learning for author profiling in PAN 2016. CEUR Workshop Proceedings. 2016.

138. Pimas O, Rexha A, Kröll M, Kern R. Profiling microblog authors using concreteness and sentiment. CEUR Workshop Proceedings. 2016.

139. Ucelay, M. J. G., Villegas, M. P., Funez, D. G., Cagina, L. C., Errecalde, M. L., Ramirez-de-la-Rosa, G., Villatoro-Tello E. Profile-based Approach for Age and Gender Identification. In CLEF (Working Notes). 2016.

140. op Vollenbroek, M. B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Nissim M. Gronup: Groningen user profiling. Notebook Papers of CLEF. 2016.

141. Zahid, A., Sampath, A., Dey, A., Farnadi G. Cross-Genre Age and Gender Identification in Social Media. In CLEF (Working Notes). 2016. pp. 1014–1017.

142. Kocher M. UniNE at CLEF 2016: Author Clustering. CEUR Workshop Proceedings. 2016.

143. Adame-Arcia Y., Castro-Castro D., Bueno R. O., Munoz R. Author Profiling, Instance-based Similarity Classification. Notebook for PAN at CLEF2. 2017.

144. Alrifai K, Rebdawi G, Ghneim N. Arabic tweeps gender and dialect prediction: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

145. Bergsma S, Durme B Van. Using Conceptual Class Attributes to Characterize Social Media Users. Acl. 2013;

146. Basile A, Dwyer G, Medvedeva M, Rawee J, Haagsma H, Nissim M. N-GRAM: New groningen author-profiling model: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

147. Ciobanu AM, Zampieri M, Malmasi S, Dinu LP. Including dialects and language varieties in author profiling: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

148. Franco-Salvador M, Plotnikova N, Pawar N, Benajiba Y. Subword-based deep averaging networks for author profiling in social media: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

149. Khan JA. Author profile prediction using trend and word frequency based analysis in text. In Cappellato et al [13]. 2017.

150. Kheng G, Laporte L, Granitzer M. INSA Lyon and UNI passau's participation at PAN@CLEF'17: Author Profiling task: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

151. Kocher, M., Savoy J. Unine at ALEF 2017: Author profiling reasoning. Cappellato et al [13]. 2017.

152. Kodiyan D, Hardegger F, Neuhaus S, Cieliebak M. Author Profiling with bidirectional rnns using Attention with grus: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

153. Lopez-Monroy, A. P., Gomez, M. M., Jair-Escalante, H., Pineda, L. V., Solorio T. Uh-inaoe participation at PAN17: Author profiling. In Cappellato et al [13]2. 2017.

154. Markov I, Gómez-Adorno H, Sidorov G. Language- and subtask-dependent feature selection and classifier parameter tuning for author Profiling: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017.

155. Martinc M, Škrjanec I, Zupan K, Pollak S. PAN 2017: Author profiling—Gender and language variety prediction: Notebook for PAN at CLEF 2017. CEUR Workshop Proceedings. 2017. https://doi.org/10.1016/j.paid.2017.02.037

156. Chaabane A, Acs G, Kaafar MA. You Are What You Like! Information Leakage Through Users' Interests. Netw Distrib Syst Secur Symp. 2012;

157. Rangel F, Rosso P. On the identification of emotions and authors' gender in Facebook comments on the basis of their writing style. CEUR Workshop Proceedings. 2013.

158. Rao D, Paul MJ, Fink C, Yarowsky D, Oates T, Coppersmith G. Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. Int Conf Weblogs Soc Media. 2011;

159. Sap M, Park G, Eichstaedt J, Kern M, Stillwell D, Kosinski M, et al. Developing Age and Gender Predictive Lexica over Social Media. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014. https://doi.org/10.3115/v1/D14-1121

160. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS One. 2013; https://doi.org/10.1371/journal.pone.0073791 PMID: 24086296

161. Tang C, Ross K, Saxena N, Chen R What's in a name: A study of names, gender inference, and gender behavior in facebook. Database Syst Adv Appl. 2011; https://doi.org/10.1007/978-3-642-20244-5_33

162. Filippova K. User Demographics and Language in an Implicit Social Network. Conf Empir Methods Nat Lang Process Comput Nat Lang Learn. 2012;

163. Ulges A, Koch M, Borth D. Linking visual concept detection with viewer demographics. Proc 2nd ACM Int Conf Multimed Retr—ICMR '12. 2012; https://doi.org/10.1145/2324796.2324827

164. Peersman C, Daelemans W, Van Vaerenbergh L. Predicting age and gender in online social networks. Int Conf Inf Knowl Manag Proc. 2011; https://doi.org/10.1145/2065023.2065035

165. van de Loo J, De Pauw G, Daelemans W. Text-Based Age and Gender Prediction for Online Safety Monitoring. Int J Cyber-Security Digit Forensics. 2016;

166. Gallagher AC, Chen T. Estimating age, gender, and identity using first name priors. 2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008. https://doi.org/10.1109/CVPR.2008.4587609

167. Elather M., Lee J. User profiling of Flickr: Integrating multiple types of features for gender classification. J Adv Inf Technol. 6.

168. You Q, Bhatia S, Sun T, Luo J. The eyes of the beholder: Gender prediction using images posted in online social networks. IEEE International Conference on Data Mining Workshops, ICDMW. 2015. https://doi.org/10.1109/ICDMW.2014.93

169. Han, K., Jo, Y., Jeon, Y., Kim, B., Song, J., Kim S. Photos Don't Have Me, But How Do You Know Me? Analyzing and Predicting Users on Instagram. Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization. 2018. pp. 251–256.

170. Zhong Y, Yuan NJ, Zhong W, Zhang F, Xie X. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins. WSDM. 2015. https://doi.org/10.1145/2684822.2685287

171. Rangel F, Rosso P. On the multilingual and genre robustness of EmoGraphs for author profiling in social media. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. https://doi.org/10.1007/978-3-319-24027-5_28

172. Qin Z, Wang Y, Xia Y, Cheng H, Zhou Y, Sheng Z, et al. Demographic information prediction based on smartphone application usage. Proceedings of 2014 International Conference on Smart Computing, SMARTCOMP 2014. 2014. https://doi.org/10.1109/SMARTCOMP.2014.7043857

173. Roy A, Pebesma E. A Machine Learning Approach to Demographic Prediction using Geohashes. Proceedings of the 2nd International Workshop on Social Sensing—SocialSens'17. 2017. https://doi.org/10.1145/3055601.3055603

174. Sarraute C, Blanc P, Burroni J. A study of age and gender seen through mobile phone usage patterns in Mexico. Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. 2014. https://doi.org/10.1109/ASONAM.2014.6921683

175. Seneviratne S, Seneviratne A, Mohapatra P, Mahanti A. Your Installed Apps Reveal Your Gender and More! ACM SIGMOBILE Mob Comput Commun Rev. 2015; https://doi.org/10.1145/2721896.2721908

176. Weiss GM, Lockhart JW. Identifying user traits by mining smart phone accelerometer data. Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data—SensorKDD '11. 2011. https://doi.org/10.1145/2003653.2003660

177. Ying JJ, Chang Y, Huang C, Tseng VS. Demographic Prediction Based on User ' s Mobile Behaviors. Nokia. 2012;

178. Zhong E, Tan B, Mo K, Yang Q. User demographics prediction based on mobile data. Pervasive and Mobile Computing. 2013. https://doi.org/10.1016/j.pmcj.2013.07.009

179. Liu Y, Bermudez I, Mislove A, Baldi M, Systems C, Torino P. Identifying Personal Information in Internet Traffic. COSN. 2015;

180. Wang Y, Tang Y, Ma J, Qin Z. Gender prediction based on data streams of smartphone applications. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2015; https://doi.org/10.1007/978-3-319-22047-5_10

181. Dong Y, Yang Y, Tang J, Yang Y, Chawla N V. Inferring user demographics and social strategies in mobile social networks. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '14. 2014. https://doi.org/10.1145/2623330.2623703

182. Nadeem S WMC. Demographic prediction of mobile user from phone usage. Age (Omaha). 2012;

183. Kelly D., Smyth B., Caulfield B. Uncovering measurements of social and demographic behaviour from smartphone location data. IEEE Trans Human-Machine Syst. 2013; 43: 188–198.

184. Solomon A, Bar A, Yanai C, Shapira B, Rokach L. Predict Demographic Information Using Word2vec on Spatial Trajectories. Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization—UMAP '18. 2018. https://doi.org/10.1145/3209219.3209224

185. Wang P, Feiyang S, Di W, Tao J, Guan X, Bifet A. Inferring Demographics and Social Networks of Mobile Device Users on Campus From AP-Trajectories. World Wide Web. 2017; https://doi.org/10.1145/1235

186. Wang P., Sun F., Wang D., Tao J., Guan X., Bifet A. Predicting attributes and friends of mobile users from AP-Trajectories. Inf Sci (Ny). 2018;

187. Aarthi S, Bharanidharan S, Saravanan M, Anand V. Predicting customer demographics in a Mobile Social Network. Proceedings—2011 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2011. 2011. https://doi.org/10.1109/ASONAM.2011.13

188. Akter S, Holder L. Using Graphical Features To Improve Demographic Prediction From Smart Phone Data. Proceedings of the 2nd International Workshop on Network Data Analytics—NDA'17. 2017. https://doi.org/10.1145/3068943.3068948

189. Brdar S, Culibrk D, Crnojevic V. Demographic attributes prediction on the real-world mobile data. Proc Mob Data Chall by Nokia 2012;

190. Choi Y, Kim Y, Kim S, Park K, Park J. An on-device gender prediction method for mobile users using representative wordsets. Expert Syst Appl. 2016; https://doi.org/10.1016/j.eswa.2016.08.002

191. Dong Y, Chawla N V., Tang J, Yang Y, Yang Y. User Modeling on Demographic Attributes in Big Mobile Social Networks. ACM Trans Inf Syst. 2017; https://doi.org/10.1145/3086701

192. Frias-Martinez V, Frias-Martinez E, Oliver N. A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. . . . Intell Dev. 2010;

193. Alharbi AR, Thornton MA. Demographic group classification of smart device users. Proceedings—2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015. 2016. https://doi.org/10.1109/ICMLA.2015.16

194. Alsmearat K, Shehab M, Al-Ayyoub M, Al-Shalabi R, Kanaan G. Emotion analysis of Arabic articles and its impact on identifying the author's gender. Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA. 2016. https://doi.org/10.1109/AICCSA.2015.7507196

195. Alsmearat K, Al-Ayyoub M, Al-Shalabi R, Kanaan G. Author gender identification from Arabic text. J Inf Secur Appl. 2017; https://doi.org/10.1016/j.jisa.2017.06.003

196. Cheng N, Chandramouli R, Subbalakshmi KP. Author gender identification from text. Digit Investig. 2011; https://doi.org/10.1016/j.diin.2011.04.002

197. De Bock K, Van Den Poel D. Predicting website audience demographics forweb advertising targeting using multi-website clickstream data. Fundam Informaticae. 2010; https://doi.org/10.3233/FI-2010-216

198. Hu J, Zeng H-J, Li H, Niu C, Chen Z. Demographic prediction based on user's browsing behavior. Proceedings of the 16th international conference on World Wide Web—WWW '07. 2007. https://doi.org/10.1145/1242572.1242594

199. Kabbur S, Han EH, Karypis G. Content-based methods for predicting web-site demographic attributes. Proceedings—IEEE International Conference on Data Mining, ICDM. 2010. https://doi.org/10.1109/ICDM.2010.97

200. Murray D, Durrell K. Inferring demographic attributes of anonymous internet users. Web Usage Anal User Profiling. 2000; https://doi.org/10.1007/3-540-44934-5_1

201. Wang Z, Derr T, Yin D, Tang J. Understanding and Predicting Weight Loss with Mobile Social Networking Data. Proceedings of the 2017 ACM Conference on Information and Knowledge Management—CIKM '17. 2017. https://doi.org/10.1145/3132847.3133019

202. Otterbacher J. Inferring gender of movie reviewers. Proceedings of the 19th ACM international conference on Information and knowledge management—CIKM '10. 2010. https://doi.org/10.1145/1871437.1871487

203. Feng T, Guo Y, Chen Y, Tan X, Xu T, Shen B, et al. Tags and titles of videos you watched tell your gender. 2014 IEEE International Conference on Communications, ICC 2014. 2014. https://doi.org/10.1109/ICC.2014.6883590

204. Weinsberg U, Bhagat S, Ioannidis S, Taft N. BlurMe: Inferring and Obfuscating User Gender Based on Ratings. Proc 6th ACM Conf Recomm Syst—RecSys '12. 2012; https://doi.org/10.1145/2365952.2365989

205. Argamon S, Konnel M, Pennebaker JW, Schier J. Mining the Blogosphere: Age, gender and the varieties of self-expression. First Monday. 2007; https://doi.org/10.5210/fm.v12i9.2003

206. Mukherjee A, Liu B. Improving Gender Classification of Blog Authors. Proc 2010 Conf Empir Methods Nat Lang Process. 2010;

207. Rustagi M, Prasath RR, Goswami S, Sarkar S. Learning age and gender of blogger from stylistic variation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009. https://doi.org/10.1007/978-3-642-11164-8_33

208. Sarawgi R, Gajulapalli K, Choi Y. Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre. Fifteenth Conf Comput Nat Lang Learn. 2011;

209. Ito J, Nishida K, Hoshide T, Toda H, Uchiyama T. Demographic and psychographic estimation of twitter users using social structures. Online Social Media Analysis and Visualization. 2014. https://doi.org/10.1007/978-3-319-13590-8

210. Rangel F, Rosso P. Use of Language and Author Profiling: Identification of Gender and Age. Proc 10th Work Nat Lang Process Cogn Sci. 2013;

211. Rangel F, Rosso P. On the impact of emotions on author profiling. Inf Process Manag. 2016; https://doi.org/10.1016/j.ipm.2015.02.003

212. Weren ERD, Kauer AU, Mizusaki L, Moreira VP, Oliveira JPM De, Wives LK. Examining Multiple Features for Author Profiling. J Inf Data Manag. 2014;

213. Zhang, C., & Zhang P. Predicting gender from blog posts. 2010.

214. Claude F, Konow R, Ladra S. Fast compressed-based strategies for author profiling of social media texts. Proceedings of the 4th Spanish Conference on Information Retrieval—CERI '16. 2016. https://doi.org/10.1145/2934732.2934744

215. Aleman Y, Loya N, Vilariño D, Pinto D. Two methodologies applied to the author profiling task. CEUR Workshop Proceedings. 2013.

216. Cruz FL, Haro RR, Ortega FJ. ITALICA at PAN 2013: An ensemble learning approach to author profiling: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

217. De-Arteaga M, Jimenez S, Dueñas G, Mancera S, Baquero J. Author profiling using corpus statistics, lexicons and stylistic features: Notebook for PAN at CLEF-2013. CEUR Workshop Proceedings. 2013.

218. Hernández DI, Guzmán-Cabrera R, Reyes A, Rocha MA. Semantic-based features for author profiling identification: First insights: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

219. Flekova L, Gurevych I. Can we hide in the web? Large scale simultaneous age and gender author profiling in social media: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

220. Jankowska M, Kešelj V, Milios E. CNG text classification for authorship profiling task: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

221. Lim WY, Goh J, Thing VLL. Content-centric age and gender profiling: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

222. Lopez-Monroy, A. P., Montes-Y-Gomez, M., Escalante, H. J., Villasenor-Pineda, L. Villatoro-Tello E. INAOE's participation at PAN'13: Author profiling task. In CLEF 2013 Evaluation Labs and Workshop. 2013.

223. Meina M, Brodzińska K, Celmer B, Czoków M, Patera M, Pezacki J, et al. Ensemble-based classification for author profiling using various features: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

224. Moreau E, Vogel C. Style-based distance features for author profiling: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

225. Patra BG, Banerjee S, Das D, Saikh T, Bandyopadhyay S. Automatic author profiling based on linguistic and stylistic features: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

226. Pavan A, Mogadala A, Varma V. Author profiling using LDA and maximum entropy: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

227. Santosh K, Bansal R, Shekhar M, Varma V. Author Profiling: Predicting Age and Gender from Blogs Notebook for PAN at CLEF 2013. PAN—Uncovering Plagiarism, Authorship, Soc Softw Misuse a benchmarking Act uncovering plagiarism, Authorsh Soc Softw misuse. 2013;

228. Sapkota U, Solorio T, Montes-Y-Gómez M, Ramírez-De-La-Rosa G. Author profiling for English and Spanish text: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

229. Mechti, S., Jaoua, M., Belguith, L. H., Faiz R. Author Profiling Using Style-based Features. Notebook Papers of CLEF2. 2013.

230. Weren, E. R., Moreira, V. P., Oliveira J. Using simple content features for the author profiling task. Notebook Papers of CLEF. 2013.

231. Gillam L. Readability for author profiling? Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN). In proceesings of the Working Notes of CLEF 2013. 2013.

232. Diaz, A. A. C., Hidalgo JMG. Experiments with SMS translation and stochastic gradient descent in Spanish text author profiling. In proceesings of the Working Notes of CLEF 2013. 2013.

233. Pham DD, Tran GB, Pham SB. Author profiling for Vietnamese blogs. 2009 International Conference on Asian Language Processing: Recent Advances in Asian Language Processing, IALP 2009. 2009. https://doi.org/10.1109/IALP.2009.47

234. Grbovic M, Radosavljevic V, Djuric N, Bhamidipati N, Nagarajan A. Leveraging blogging activity on tumblr to infer demographics and interests of users for advertising purposes. CEUR Workshop Proceedings. 2016.

235. Cheng N, Chen X, Chandramouli R, Subbalakshmi KP. Gender identification from e-mails. 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009—Proceedings. 2009. https://doi.org/10.1109/CIDM.2009.4938643

236. Corney M, De Vel O, Anderson A, Mohay G. Gender-preferential text mining of e-mail discourse. Proceedings—Annual Computer Security Applications Conference, ACSAC. 2002. https://doi.org/10.1109/CSAC.2002.1176299

237. Vel O De, Corney M, Anderson A. Language and gender author cohort analysis of e-mail for computer forensics. Digit Forensics Res Work. 2002;

238. Estival D. Author attribution with email messages. J Sci. 2008; 1–9.

239. Estival D, Gaustad T, Pham SB, Radford W, Hutchinson B. Author profiling for English emails. 10th Conference of the Pacific Association for Computational Linguistics. 2007.

240. Estival D, Gaustad T, Pham SB, Radford W, Hutchinson B. TAT: An Author Profiling Tool with Application to Arabic Emails. Proceedings of the Australasian Language Technology Workshop 2007. 2007.

241. Estival, D., Gaustad, T., Hutchinson, B., Pham, S., Radford W. Author Profiling for English and Arabic Emails. 2008.

242. Deitrick W, Miller Z, Valyou B, Dickinson B, Munson T, Hu W. Gender Identification on Twitter Using the Modified Balanced Winnow. Commun Netw. 2012; https://doi.org/10.4236/cn.2012.43023

243. Krismayer T, Schedl M, Knees P, Rabiser R. Prediction of User Demographics from Music Listening Habits. Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing—CBMI '17. 2017. https://doi.org/10.1145/3095713.3095722

244. Liu, J. Y., & Yang YH. Inferring personal traits from music listening history. In Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. ACM.; 2012. pp. 31–36.

245. Wu M, Jang J, Lu C. Gender Identification and Age Estimation of Users Based on Music Metadata. 15th Int Soc Music 2014;

246. Jones R, Kumar R, Pang B, Tomkins A. I know what you did last summer: query logs and user privacy. CIKM. 2007. https://doi.org/10.1145/1321440.1321573

247. Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C., & Can F. Chat mining for gender prediction. In International Conference on Advances in Information Systems. Springer, Berlin, Heidelberg.; 2006. pp. 274–283.

248. Kucukyilmaz T, Cambazoglu BB, Aykanat C, Can F. Chat mining: Predicting user and message attributes in computer-mediated communication. Inf Process Manag. 2008; https://doi.org/10.1016/j.ipm.2007.12.009

249. Likarish P, Brdiczka O, Yee N, Ducheneaut N, Nelson L. Demographic Profiling from MMOG Gameplay. HotPETs 2011 Hot Top Priv Enhancing Technol. 2011;

250. Li H., Zhu H., Ma D. Demographic Information Inference through Meta-Data Analysis of Wi-Fi Traffic. IEEE Trans Mob Comput. 2018; 5: 1033–1047.

251. Gu X., Yang H., Tang J., Zhang J., Zhang F., Liu D., Fu X. Profiling Web users using big data. Soc Netw Anal Min. 2018; 8.

252. Al Zamal, F., Liu, Q., Ruths D. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. International Conference on Weblogs and Social Media. 2012.

253. Ardehaly EM, Culotta A. Mining the demographics of political sentiment from twitter using learning from label proportions. Proceedings—IEEE International Conference on Data Mining, ICDM. 2017. https://doi.org/10.1109/ICDM.2017.84

254. Moseley N, Alm CO, Rege M. User-annotated microtext data for modeling and analyzing users' sociolinguistic characteristics and age grading. Proceedings—International Conference on Research Challenges in Information Science. 2014. https://doi.org/10.1109/RCIS.2014.6861046

255. Rao D, Yarowsky D, Shreevats A, Gupta M. Classifying latent user attributes in twitter. Proceedings of the 2nd international workshop on Search and mining user-generated contents—SMUC '10. 2010. https://doi.org/10.1145/1871985.1871993

256. Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, et al. Overview of the 2nd Author Profiling task at PAN 2014. CEUR Workshop Proceedings. 2014. doi: 1613–0073

257. Rangel Pardo, F. M., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans W. Overview of the 3rd Author Profiling Task at PAN 2015. In CLEF 2015 Evaluation Labs and Workshop Working Notes Papers. 2015. pp. 1–8.

258. Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. CEUR Workshop Proceedings. 2016.

259. Rangel F Rosso P, Potthast M, Stein B. Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. CEUR Workshop Proceedings. 2017.

260. Bamman D, Eisenstein J, Schnoebelen T. Gender in Twitter: Styles, Stances, and Social Networks. arXiv. 2012; https://doi.org/10.1371/journal.pone.0087041

261. Yan X, Yan L. Gender Classification of Weblog Authors. Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

262. López-Monroy AP, Montes-Y-Gómez M, Escalante HJ, Villaseñor-Pineda L, Villatoro-Tello E. INAOE's participation at PAN'13: Author profiling task: Notebook for PAN at CLEF 2013. CEUR Workshop Proceedings. 2013.

263. Coates J., Pichler P. Language and gender: A reader. Blackwell Oxford; 1998.

264. Eckert P., McConnell-Ginet S. Language and gender. Cambridge University Press; 2013.

265. Lakoff R. Language and woman's place. New York: Harper & Row; 1975.

266. Pennebaker JW, King LA. Linguistic styles: Language use as an individual difference. J Pers Soc Psychol. 1999; https://doi.org/10.1037/0022-3514.77.6.1296

267. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology. 2010. https://doi.org/10.1177/0261927X09359521

268. Mehl MR, Vazire S, Ramírez-Esparza N, Slatcher RB, Pennebaker JW. Are women really more talkative than men? Science. 2007. https://doi.org/10.1126/science.1139940 PMID: 17615349

269. Mulac A, Bradac JJ, Gibbons P. Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. Hum Commun Res. 2001; https://doi.org/10.1093/hcr/27.1.121

270. Coates J. Women, men and language: A sociolinguistic account of gender differences in language: Third edition. Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language: Third Edition. 2015. https://doi.org/10.4324/9781315645612

271. Sharma D. Language and woman's place: Text and commentaries. Gender and Language. 2007. https://doi.org/10.1558/genl.v1i2.319

272. Holmes J, Meyerhoff M. The handbook of language and gender. The Handbook of Language and Gender. 2008. https://doi.org/10.1002/9780470756942

273. Pennebaker J. W., Francis M. E., & Booth R. J. Linguistic inquiry and word count: LIWC 2001. Mahw Lawrence Erlbaum Assoc. 2001;71.

274. Pennacchiotti M, Popescu A. to Twitter User Classification. Proc Fifth Int AAAI Conf Weblogs Soc Media A. 2011; https://doi.org/10.1145/2542214.2542215

275. Nowson S, Oberlander J. The Identity of Bloggers: Openness and gender in personal weblogs. AAAI Spring Symp Comput Approaches to Anal Weblogs. 2005;

276. Duong DT, Pham SB, Tan H. Using content-based features for author profiling of Vietnamese forum posts. Studies in Computational Intelligence. 2016. https://doi.org/10.1007/978-3-319-31277-4_25

277. Giles H, Ogay T. Communication accommodation theory. Explaining Communication: Contemporary Theories and Exemplars. 2006. https://doi.org/10.4324/9781410614308

278. Gallois C, Ogay T, Giles H. Communication Accommodation Theory: A Look Back and a Look Ahead. WB Gudykunst Theor About Commun Cult. 2005; https://doi.org/10.1002/9781118611463.wbielsi066

279. Coupland N, Coupland J, Giles H. Language, Society and the Elderly: Discourse, Identity, and Ageing. Lang Soc. 1993;

280. Sankoff G. Language Change Across the Lifespan. Annu Rev Linguist. 2018; https://doi.org/10.1146/annurev-linguistics-011516-034253

281. Baltes PB. Theoretical propositions of life-span developmental psychology: On the dynamics between growth and decline. Dev Psychol. 1987; https://doi.org/10.1037/0012-1649.23.5.611

282. Schaie KW. The course of adult intellectual development. Am Psychol. 1994; https://doi.org/10.1037/0003-066X.49.4.304

283. Carstensen LL. Evidence for a Life-Span Theory of Socioemotional Selectivity. Curr Dir Psychol Sci. 1995; https://doi.org/10.1111/1467-8721.ep11512261

284. Carstensen LL, Pasupathi M, Mayr U, Nesselroade JR. Emotional experience in everyday life across the adult life span. J Pers Soc Psychol. 2000; https://doi.org/10.1037//0022-3514.79.4.644

285. Flekova L, Preo\ctiuc-Pietro D, Ungar L. Exploring Stylistic Variation with Age and Income on Twitter. Proc 54th Annu Meet Assoc Comput Linguist (Volume 2 Short Pap. 2016;

286. Nguyen D, Gravel R, Trieschnigg D, Meder T. "How old do you think I am?": A study of language and age in Twitter. Proc seventh Int AAAI Conf weblogs Soc media, 8–11 July 2013, Cambridge, Massachusetts, USA. 2013; https://doi.org/10.1007/s00256-005-0933-8

287. Park SH, Lee HJ, Han SP, Lee DH. User age profile assessment using SMS network neighbors' age profiles. Proceedings—International Conference on Advanced Information Networking and Applications, AINA. 2009. https://doi.org/10.1109/WAINA.2009.136

288. Chen, L., Qian, T., Wang, F., You, Z., Peng, Q., & Zhong M. Age Detection for Chinese Users in Weibo. In International Conference on Web-Age Information Management. Springer International Publishing.; 2015. pp. 83–95.

289. Han K, Lee S, Jang JY, Jung Y, Lee D. "Teens are from Mars, Adults are f rom Venus": Analyzing and Predicting Age Groups with Behavioral Characteristics in Instagram. ACM Web Sci 2016. 2016; https://doi.org/10.1145/2908131.2908160

290. Nguyen D, Gravel R, Trieschnigg D, Meder T. TweetGenie: automatic age prediction from tweets. ACM SIGWEB Newsl. 2013; https://doi.org/10.1145/2528272.2528276

291. Tam J, Martell CH. Age detection in chat. ICSC 2009–2009 IEEE International Conference on Semantic Computing. 2009. https://doi.org/10.1109/ICSC.2009.37

292. Nguyen D, Smith N, Rosé C. Author Age Prediction from Text using Linear Regression. LaTeCH '11 Proc 5th ACL-HLT Work Lang Technol Cult Heritage, Soc Sci Humanit. 2011;

293. Ardehaly EM, Culotta A. Learning from noisy label proportions for classifying online social data. Soc Netw Anal Min. 2018; https://doi.org/10.1007/s13278-017-0478-6

294. Goel S, Hofman JM, Sirer MI. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media. 2012.

295. Schler J, Koppel M, Argamon S, Pennebaker J. Effects of Age and Gender on Blogging. Artif Intell. 2006; https://doi.org/10.1155/2015/862427

296. Moon Y. Personalization and personality: Some effects of customizing message style based on consumer personality. J Consum Psychol. 2002; https://doi.org/10.1207/15327660260382351

297. Rosenthal S, McKeown K. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. Proc 49th Annu Meet Assoc Comput Linguist Hum Lang Technol 1. 2011;

298. Brea, J., Burroni, J., Minnoni, M., & Sarraute C. Harnessing mobile phone social network topology to infer users demographic attributes. In Proceedings of the 8th Workshop on Social Network Mining and Analysis ACM. 2014. p. 1.

299. Ardehaly EM, Culotta A. Inferring latent attributes of Twitter users with label regularization. HLT-NAACL 2015—Hum Lang Technol Conf North Am Chapter Assoc Comput Linguist Proc Main Conf. 2015;

300. Guimarães RG, Rosa RL, Gaetano D De, Rodríguez DZ, Bressan G. Age Groups Classification in Social Network Using Deep Learning. IEEE Access. 2017; https://doi.org/10.1109/ACCESS.2017.2706674

301. Oktay H, Firat A, Ertem Z. Demographic breakdown of Twitter users: An analysis based on names. ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conf. 2014;

302. Barberá P. Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. Work Pap NYU. 2016;

303. Sloan L, Morgan J, Burnap P, Williams M. Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0115545 PMID: 25729900

304. Dey R, Tang C, Ross K, Saxena N. Estimating age privacy leakage in online social networks. Proceedings—IEEE INFOCOM. 2012. https://doi.org/10.1109/INFCOM.2012.6195711

305. Perozzi, B., & Skiena S. Exact age prediction in social networks. In Proceedings of the 24th International Conference on World Wide Web. 2015. pp. 91–92.

306. Zheng L, Yang K, Yu Y, Jin P. Predicting Age Range of Users over Microblog Dataset. Int J Database Theory Appl. 2013; https://doi.org/10.14257/ijdta.2013.6.6.08

307. Kocher M, Savoy J. UniNE at CLEF 2015: Author Identification. CEUR Workshop Proceedings. 2015.

308. Pew Internet Research. Internet use by age [Internet]. 2017. Available: http://www.pewinternet.org/chart/internet-use-by-age/

309. Vieweg S, Hughes AL, Starbird K, Palen L. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. Proceedings of the 28th international conference on Human factors in computing systems—CHI '10. 2010. https://doi.org/10.1145/1753326.1753486

310. Jurgens D, Finnethy T, McCorriston J, Xu YT, Ruths D. Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. 9th Int Conf Weblogs Soc Media. 2015; https://doi.org/10.1002/0471264385.wei0223

311. Stefanidis A, Crooks A, Radzikowski J. Harvesting ambient geospatial information from social media feeds. GeoJournal. 2013; https://doi.org/10.1007/s10708-011-9438-2

312. Backstrom L, Sun E, Marlow C. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. Proc 19th Int Conf World Wide Web. 2010; https://doi.org/10.1145/1772690.1772698

313. Hecht B, Hong L, Suh B, Chi EH. Tweets from Justin Bieber's heart. Proceedings of the 2011 annual conference on Human factors in computing systems—CHI '11. 2011. https://doi.org/10.1145/1978942.1978976

314. Kinsella S, Murdock V, O'Hare N. "I'm eating a sandwich in Glasgow." Proceedings of the 3rd international workshop on Search and mining user-generated contents—SMUC '11. 2011. https://doi.org/10.1145/2065023.2065039

315. Li, R., Wang, S., & Chang KCC. Multiple location profiling for users and relationships from social network and content. Proceedings of the VLDB Endowment. 2012. pp. 1603–1614.

316. Li R, Wang S, Deng H, Wang R, Chang KC-C. Towards social user profiling. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '12. 2012. https://doi.org/10.1145/2339530.2339692

317. Mahmud J, Nichols J, Drews C. Where Is this tweet from? Inferring home locations of Twitter users. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12). 2012. doi:papers3://publication/uuid/8AAE166A-DE81-42AB-83BC-5E014B7B0039

318. McGee J, Caverlee J, Cheng Z. Location Prediction in Social Media Based on Tie Strength. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. 2013. https://doi.org/10.1145/2505515.2505544

319. Pontes T, Magno G, Vasconcelos M, Gupta A, Almeida J, Kumaraguru P, et al. Beware of what you share: Inferring home location in social networks. Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012. 2012. https://doi.org/10.1109/ICDMW.2012.106

320. Rout D, Bontcheva K, Preoţiuc-Pietro D, Cohn T. Where's @wally?: A Classification Approach to Geolocating Users Based on their Social Ties. Proceedings of the 24th ACM Conference on Hypertext and Social Media. 2013. https://doi.org/10.1145/2481492.2481494

321. Zheng D, Hu T, You Q, Kautz H, Luo J. Inferring Home Location from User's Photo Collections based on Visual Content and Mobility Patterns. Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia—GeoMM '14. 2014. https://doi.org/10.1145/2661118.2661123

322. Zheng D, Hu T, You Q, Kautz H, Luo J. Towards Lifestyle Understanding: Predicting Home and Vacation Locations from User ' s Online Photo Collections. Aaai. 2015;

323. Chandra S, Khan L, Muhaya F Bin. Estimating Twitter User Location Using Social Interactions—A Content Based Approach. 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing. 2011. https://doi.org/10.1109/PASSAT/SocialCom.2011.120

324. Chang HW, Lee D, Eltaher M, Lee J. Phillies tweeting from philly? Predicting twitter user locations with spatial word usage. Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012. 2012. https://doi.org/10.1109/ASONAM.2012.29

325. Cheng Z, Caverlee J, Lee K. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. Proc 19th ACM Int Conf Inf Knowl Manag. 2010; https://doi.org/10.1145/1871437.1871535

326. Eisenstein J, O'Connor B, Smith N a., Xing EP. A latent variable model for geographic lexical variation. Proc 2010 Conf Empir Methods Nat Lang Process. 2010; https://doi.org/10.1038/nrm2900

327. Pontes T, Vasconcelos M, Almeida J, Kumaraguru P, Almeida V. We know where you live: privacy characterization of foursquare behavior. UbiComp. 2012. https://doi.org/10.1145/2370216.2370419

328. Rossi L, Musolesi M. It's the Way you Check-in: Identifying Users in Location-Based Social Networks. Second ACM Conf ONLINE Soc NETWORKS. 2014; https://doi.org/10.1145/2660460.2660485

329. Seneviratne S, Seneviratne A, Mohapatra P, Mahanti A. Predicting user traits from a snapshot of apps installed on a smartphone. ACM SIGMOBILE Mob Comput Commun Rev. 2014; https://doi.org/10.1145/2636242.2636244

330. Chang J, Rosenn I, Backstrom L, Marlow C. epluribus: Ethnicity on social networks. . . . Weblogs Soc Media 2010;

331. Bollen J, Mao H, Pepe A. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011.

332. Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. Proc 50th Annu Meet Assoc Comput Linguist. 2012; https://doi.org/10.1145/1935826.1935854

333. Tumasjan A, Sprenger T, Sandner P, Welpe I. Predicting elections with Twitter: What 140 characters reveal about political sentiment. Proc Fourth Int AAAI Conf Weblogs Soc Media. 2010; https://doi.org/10.1074/jbc.M501708200

334. Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. Election forecasts with Twitter: How 140 characters reflect the political landscape. Social Science Computer Review. 2011. https://doi.org/10.1177/0894439310386557

335. Boutet, A., & Yoneki E. Member classification and party characteristics in twitter during uk election. In DYNAM: Proceedings of the 1st International Workshop on Dynamicity. 2011. p. 18.

336. Boutet A, Kim H, Yoneki E. What ' s in Your Tweets? I Know Who You Supported in the UK 2010 General Election. Labour. 2012;

337. Boutet A, Kim H, Yoneki E. What's in Twitter, I know what parties are popular and who you are supporting now! Soc Netw Anal Min. 2013; https://doi.org/10.1007/s13278-013-0120-1

338. Cohen R, Ruths D. Classifying Political Orientation on Twitter: It's Not Easy! Seventh Int AAAI Conf Weblogs 2013;

339. Colleoni E, Rozza A, Arvidsson A. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. J Commun. 2014; https://doi.org/10.1111/jcom.12083

340. Conover MD, Gonçalves B, Ratkiewicz J, Flammini A, Menczer F. Predicting the political alignment of twitter users. Proceedings—2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011. 2011. https://doi.org/10.1109/PASSAT/SocialCom.2011.34

341. Lampos V, Preotiuc-Pietro D, Cohn T. A user-centric model of voting intention from Social Media. Proc 51st Annu Meet Assoc Comput Linguist. 2013;

342. Makazhanov A, Rafiei D, Waqar M. Predicting political preference of Twitter users. Soc Netw Anal Min. 2014; https://doi.org/10.1007/s13278-014-0193-5

343. Sylwester K, Purver M. Twitter language use reflects psychological differences between Democrats and Republicans. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0137422 PMID: 26375581

344. Volkova S, Coppersmith G, Durme B Van. Inferring User Political Preferences from Streaming Communications. Acl. 2014;

345. Wong FMF, Tan CW, Sen S, Chiang M. Quantifying political leaning from tweets, retweets, and retweeters. IEEE Trans Knowl Data Eng. 2016; https://doi.org/10.1109/TKDE.2016.2553667

346. Wong, F. M. F., Tan, C. W., Sen, S., & Chiang M. Quantifying Political Leaning from Tweets and Retweets. ICWSM. 2013. pp. 640–64.

347. Zhou DX, Resnick P, Mei Q. Classifying the Political Leaning of News Articles and Users from User Votes Semi-Supervised Learning Algorithms. Icwsm. 2011;

348. Jiang, M., & Argamon S. Finding political blogs and their political leanings. In Proceedings of SIAM Text Mining Workshop. 2008.

349. Jiang M, Argamon S. Exploiting subjectivity analysis in blogs to improve political leaning categorization. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval—SIGIR '08. 2008. https://doi.org/10.1145/1390334.1390472

350. Jiang, M., & Argamon S. Political Leaning Categorization by Exploring Subjectivities in Political Blogs. In DMIN. 2008. pp. 647–653.

351. Durant KT, Smith MD. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2007; https://doi.org/10.1007/978-3-540-77485-3_11

352. Vaccari C, Valeriani A, Barberá P, Bonneau R, Jost JT, Nagler J, et al. Political expression and action on social media: Exploring the relationship between lower- and higher-threshold political activities among twitter users in Italy. J Comput Commun. 2015; https://doi.org/10.1111/jcc4.12108

353. Imai K, Lo J, Olmsted J. Fast estimation of ideal points with massive data. Am Polit Sci Rev. 2016; https://doi.org/10.1017/S000305541600037X

354. Garcia D. Leaking privacy and shadow profiles in online social networks. Sci Adv. 2017; https://doi.org/10.1126/sciadv.1701172 PMID: 28798961

355. Jernigan C, Mistree BFT. Gaydar: Facebook friendships expose sexual orientation. First Monday. 2009; https://doi.org/10.5210/fm.v14i10.2611

356. Sarigol, E., Garcia, D., & Schweitzer F. Online privacy as a collective phenomenon. In Proceedings of the second ACM conference on Online social networks. 2014. pp. 95–106.

357. Wang Y, Kosinski M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. J Pers Soc Psychol. 2018; https://doi.org/10.1037/pspa0000098 PMID: 29389215

358. Li J, Ritter A, Hovy E. Weakly Supervised User Profile Extraction from Twitter. Acl. 2014;

359. Reinhardt D, Engelmann F, Moerov A, Hollick M. Show me your phone, i will tell you who your friends are: Analyzing smartphone data to identify social relationships. 14th Int Conf Mob Ubiquitous Multimedia, MUM 2015. 2015; https://doi.org/10.1145/2836041.2836048

360. Huang W, Weber I, Vieweg S. Inferring nationalities of Twitter users and studying inter-national linking. Proc 25th ACM Conf Hypertext Soc media—HT '14. 2014; https://doi.org/10.1145/2631775.2631825

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

362. Ambekar A, Ward C, Mohammed J, Male S, Skiena S. Name-ethnicity classification from open sources. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD '09. 2009. https://doi.org/10.1145/1557019.1557032

363. Mislove A, Viswanath B, Gummadi K, Druschel P. You are who you know: inferring user profiles in online social networks. Third ACM Int Conf Web Search Data Min. 2010; https://doi.org/10.1145/1718487.1718519

364. Preoţiuc-Pietro D, Volkeva S, Lampos V, Bachrach Y, Aletras N. Studying user income through language, behaviour and affect in social media. PLoS One. 2015; https://doi.org/10.1371/journal.pone.0138717 PMID: 26394145

365. Fixman, M., Berenstein, A., Brea, J., Minnoni, M., & Sarraute C. Inference of Socioeconomic Status in a Communication Graph. In Simposio Argentino de GRANdes DAtos (AGRANDA 2016)-JAIIO 45 (Tres de Febrero, 2016). 2016.

366. Fixman M, Berenstein A, Brea J, Minnoni M, Travizano M, Sarraute C. A Bayesian approach to income inference in a communication network. Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016. 2016. https://doi.org/10.1109/ASONAM.2016.7752294

367. Fixman, M., Minnoni, M., & Sarraute C. Comparison of Feature Extraction Methods and Predictors for Income Inference. In Simposio Argentino de GRANdes DAtos (AGRANDA)-JAIIO 46 (Córdoba, 2017). 2017.

368. Nguyen M-T, Lim E-P. On predicting religion labels in microblogging networks. Proc 37th Int ACM SIGIR Conf Res Dev Inf Retr—SIGIR '14. 2014; https://doi.org/10.1145/2600428.2609547

369. Proserpio D, Counts S, Jain A, Acm. The Psychology of Job Loss: Using Social Media Data to Characterize and Predict Unemployment. Proceedings of the 2016 Acm Web Science Conference. 2016. https://doi.org/10.1145/2908131.2913008

370. Preoţiuc-Pietro D, Lampos V, Aletras N. An analysis of the user occupational class through Twitter content. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015. https://doi.org/10.3115/v1/P15-1169

371. Abbar, S., Mejova, Y., & Weber I. You tweet what you eat: Studying food consumption through twitter. n Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015. pp. 3197–3206.

372. Wang Y, Weber I, Mitra P. Quantified Self Meets Social Media. Proceedings of the 6th International Conference on Digital Health Conference—DH '16. 2016. https://doi.org/10.1145/2896338.2896363

373. Weber, I., & Achananuparp P. Insights from machine-learned diet success prediction. In Biocomputing 2016: Proceedings of the Pacific Symposium. 2016. pp. 540–551.

374. Kocabey, E., Camurcu, M., Ofli, F., Aytar, Y., Marin, J., Torralba, A., & Weber I. Face-to-bmi: Using computer vision to infer body mass index on social media. 2017.

375. Filho RM, Borges GR, Almeida JM, Pappa GL. Inferring User Social Class in Online Social Networks. Proceedings of the 8th Workshop on Social Network Mining and Analysis—SNAKDD'14. 2014. https://doi.org/10.1145/2659480.2659502

376. Pinquart M, Silbereisen RK. Human development in times of social change: Theoretical considerations and research needs. Int J Behav Dev. 2004; https://doi.org/10.1080/01650250344000226

377. Erikson EH. Childhood and society. New York: Norton; 1950.

378. McAdams DP. The Psycology of Life Stories. Review of General Psychology. 2001. https://doi.org/10.1037/1089-2680.5.2.100

379. Yarkoni T, Westfall J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspect Psychol Sci. 2017; https://doi.org/10.1177/1745691617693393 PMID: 28841086

380. Danescu-Niculescu-Mizil C, Gamon M, Dumais S. Mark my words! Proc 20th Int Conf World wide web —WWW '11. 2011; https://doi.org/10.1145/1963405.1963509

381. Muir K, Joinson A, Cotterill R, Dewdney N. Characterizing the Linguistic Chameleon: Personal and Social Correlates of Linguistic Style Accommodation. Hum Commun Res. 2016; https://doi.org/10.1111/hcre.12083

382. Hancock JT, Curry LE, Goorha S, Woodworth M. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. Discourse Process. 2008; https://doi.org/10.1080/01638530701739181

383. Smith LGE, Gavin J, Sharp E. Social identity formation during the emergence of the occupy movement. Eur J Soc Psychol. 2015; https://doi.org/10.1002/ejsp.2150

384. Kacewicz E, Pennebaker JW, Davis M, Jeon M, Graesser AC. Pronoun Use Reflects Standings in Social Hierarchies. J Lang Soc Psychol. 2014; https://doi.org/10.1177/0261927X14555549

385. Sendén MG, Sikström S, Lindholm T. "She" and "He" in news media messages: Pronoun use reflects gender biases in semantic contexts. Sex Roles. 2015; https://doi.org/10.1007/s11199-014-0437-x